Normalized, HOS-Based, Blind Speech Separation Algorithms

Phillip De Leon and Yunsheng Ma
New Mexico State University
Klipsch School of Electrical and Computer Engineering
Las Cruces, New Mexico 88003-8001
{pdeleon,yuma}@nmsu.edu

Abstract

Techniques for blind separation of mixed speech signals (co-channel speech) have been recently reported in the literature. One computationally simple method for linear mixtures (suitable for real-time separation), employs a gradient search algorithm to maximize kurtosis of the outputs (hopefully separated speech signals). In this paper, we report the results of an enhancement to the algorithm which involves a normalization to the correction matrix used in the update of the separation matrix. Simulation results (using the TIMIT speech corpus) generally indicate improved (sometimes significantly) separation quality, a higher probability in producing distinct source outputs, and robustness in noisy cases.

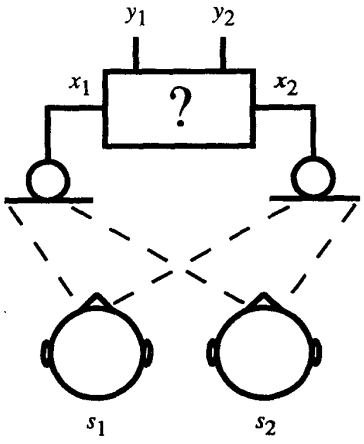

Figure 1. Speech signal separation problem.

1. Introduction

In many audio-interface, multimedia, and speech recognition applications, mixtures of speech signals from various competing speakers must be separated out before processing [1]. Given the complicated nature of speech signals this is a difficult problem compounded by environmental effects such as noise and reverberation and a strong desire for a simple algorithm suitable for real-time operation. Several methods have been proposed some of which have shown moderate success but often at the expense of high computational complexity [2],[3].

The basic problem is illustrated in Fig. 1. As a first step, we assume two unknown speech source signals, s_1 and s_2 are mixed in a linear fashion to produce two mixture signals x_1 and x_2. (Certainly the more realistic mixing model is convolutional but this more difficult to separate and to our knowledge, still an open problem. It is hoped good algorithms for source separation of linear mixtures can lead to methods for separation of convolutional mixtures.) Thus given x_1 and x_2 and no further information, we wish to produce y_1 and y_2 which approximate s_1 and s_2. Such a problem formulation is referred as the "blind source separation" problem. The problem is illustrated in a more convenient form in Fig. 2 where $\mathbf{A}(n)$ is the unknown, possibly time-varying, 2×2 mixing matrix composed of scalar elements. In this case we have

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

where

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

are the vectors of source, mixture signals respectively. The objective is to determine a separation matrix, $\mathbf{W}(n)$ such that

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

where

$$\mathbf{y}(n) = [\ y_1(n) \quad y_2(n)\ ]^T \tag{4}$$

0-7803-6514-3/00/$10.00©2000 IEEE

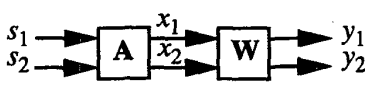

Figure 2. Basic signal separation setting.

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 not known.

2. Kurtosis-Based Speech Separation Algorithm

One previously reported method for separation of linear mixtures which is suitable for real-time applications is based on the fundamental assumption that linear mixtures of speech signals have a kurtosis, defined as

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

less than that for either source [4]. 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

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

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. The Kurtosis Maximization Algorithm (KMA) which implements (6) is listed in Fig. 3. Finally, we note that the fundamental assumption has been found to be generally true for long speech signals and a recent study of the validity of this assumption has also found it to be generally true over short windows ($\approx$ 0.5s in duration) of speech [5].

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

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

Figure 3. Kurtosis maximization algorithm (KMA) for speech separation.

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.

In this paper, we report the results of a normalized version of KMA for speech separation. Specifically, we examine two normalizers of the correction matrix $\mathbf{C}(n)$: 1) ℓ_2 norm and 2) Frobenius norm. Our results generally indicate both normalized algorithms improved (sometimes significantly) separation quality, probability of producing distinct and separate source outputs, and robustness in noisy cases. In addition, convergence speed appears to improve as well.

3. Normalizing the HOS-Based Speech Separation Algorithm

In practice, KMA often produces large, unnatural amplitude variations in the separated speech output signals. Analysis of these variations leads to the observation that the correction matrix, $\mathbf{C}(n)$ often changes by "large" amounts. We therefore consider two normalizations of the correction matrix. The first employs an ℓ_2 normalization

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

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

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

The second normalization employs a Frobenius norm

$$\mathbf{W}(n+1) = \mathbf{W}(n) + \frac{\tilde{\mu}}{||\mathbf{C}(n)||_F^2}\mathbf{C}(n) \tag{9}$$

where

$$||\mathbf{C}(n)||_F^2 = \sum_{i=1}^{L}\sum_{j=1}^{L}|\mathbf{C}_{ij}(n)|^2. \quad (10)$$

The latter norm is investigated due to its simpler implementation.

Simulations with both unnormalized and normalized KMA were conducted using 50 pairs of speech signals (each 10s in duration) from the TIMIT speech corpus [6]. A single mixing matrix, $\mathbf{A}$ was chosen at random and used for synthesizing mixture signals. The separation matrix was initialized to

$$\mathbf{W}(0) = \begin{bmatrix} 1.0 & 0.1 \\ 0.1 & -1.0 \end{bmatrix}, \quad (11)$$

$\mu = 2.0e-6$, $\tilde{\mu} = 0.0001$, and $\lambda = 0.99995$. The following four observations were made: 1) large, unnatural amplitude variations were reduced (informal listening tests) in the normalized cases, 2) separation ratios of the strongest separated source were comparable in all algorithms, 3) separation ratios of the second separated source were much higher in the normalized cases, i.e. two distinct and fully separated source outputs, and 4) at a 16kHz sampling rate, convergence was informally observed to be faster in the normalized cases (4 seconds, typical) than in the unnormalized case (8 seconds, typical).

Simulation results are summarized in Table 1. We note that the Mean SIR is computed by ensemble-averaging the separation ratios (ratio of row elements in product matrix, $\mathbf{WA}$) for the experiments. The result is then time-averaged over the duration of the simulation. Clearly if the algorithm converges quickly and has a high SIR, the Mean SIR will be high. Conversely, if the algorithm converges slowly and/or has a low SIR, the Mean SIR will be low. In addition, metrics regarding distinct sources are computed by examining the percentage of experiments where the Mean SIR for both sources exceeds 20dB. Figs. 4-6 illustrate learning curves for each algorithm for a particular experiment (same speech signals).

Table 1. Speech Separation Measures.

	Mean SIR (dB)	% Experiments w/ Distinct Sources
Unnormalized	35, 5	2%
Normalized-ℓ_2	38, 26	78%
Normalized-Fro	37, 25	74%

4. Normalized KMA: A Possible Solution to the Duplicate Source Output Problem

As noted in the original work involving unnormalized KMA for speech separation, duplicate source outputs were observed quite frequently [4]. It was assumed that some method of residual signal processing using the mixture signals and the separated source resulting from KMA could separate out the other remaining source. From the previous section, we find that experiments which use the normalized KMA usually produce distinct sources thus eliminating the need for further residual processing. In this section, we explore how the normalization of the correction matrix and initialization of the separation matrix can more frequently lead to distinct source outputs.

In the separation equation given in (3), for each output signal we linearly combine two mixture signals so that one of the sources is eliminated. This results in a scaled version of a speech source described mathematically as

$$y_i(n) = W_{i1}x_1(n) + W_{i2}x_2(n). \quad (12)$$

Again the goal of the algorithm is to maximize the kurtosis of y_i. Fig. 7 contains a plot of the output kurtosis, κ_{y_i} versus W_{i1} and W_{i2} (we choose a small range of W_{ij} for convenience) for a sample speech mixture. From the plot, we note the following:

- There are two "kurtosis ridges" each representing maximum kurtosis for a particular source output, i.e. for one set of (W_{i1}, W_{i2}) pairs we obtain one source, for another set of (W_{i1}, W_{i2}) pairs we obtain the other source.
- Formation of the ridges is due to the fact that the kurtosis measure is "scale invariant" thus if any one pair (W_{i1}, W_{i2}) maximizes kurtosis, then all pairs of the form (cW_{i1}, cW_{i2}) where c is a constant will also maximize kurtosis.
- At the $(0, 0)$ pair the output signal, according to (12), is zero and thus no kurtosis.
- Kurtosis ridges always intersect at the $(0, 0)$ pair and are never co-linear since we assume the mixing matrix is nonsingular, i.e. each mixture has different proportions of source speech signals.

The goal in the gradient search of (6) is to find two pairs (W_{11}, W_{12}) and (W_{21}, W_{22}) which lead to the top of each ridge (distinct source outputs) but not to the top of the same ridge (duplicate source outputs). In order to achieve this goal we must:

1. Initialize each pair so that they "point" to different kurtosis ridges and search in the correct direction.

2. Constrain corrections to the pair so that a "jump" from one ridge to another is not possible.

Since the kurtosis ridges intersect at $(0,0)$ and are assumed to not be co-linear, we initialize the two pairs to be *orthogonal* as in (11). Simulations using a variety of pairs (including non-orthogonal pairs), supports the use of this initialization strategy. The normalizations in (7) and (9) help us achieve our search goal by preventing jumps from one ridge to another.

5. Robustness of Normalized Speech Separation Algorithm to Additive Noise

In many applications, additive noise in the mixture signals can reduce the effectiveness of the speech separation algorithms. In this section, we evaluate speech separation using KMA under additive zero-mean, white, Gaussian noise, $\mathbf{v}(n) = \left[\begin{array}{cc} v_1(n) & v_2(n) \end{array} \right]^T$ with

$$\mathbf{x}(n) = \mathbf{A}(n)\mathbf{s}(n) + \mathbf{v}(n). \quad (13)$$

Simulations were conducted as in Section 3 with noise added to the mixture to achieve desired SNRs. Mean SIRs are computed as described in Section 3 and results are listed in Table 2. We note all algorithms can separate out at least one source under noisy conditions (except at 0dB SNR). In addition, the normalized algorithms can also separate out the other sources at SNRs above 40dB.

Table 2. Speech Separation (Noisy Case) Measures.

	Mixture-to-Noise Ratio (dB)	Mean SIR (dB)
Unnormalized	60	36, 5
	40	35, 5
	20	22, 4
	0	12, -9
Normalized-ℓ_2	60	38, 25
	40	29, 14
	20	19, 5
	0	7, 0
Normalized-Frobenius	60	37, 25
	40	31, 15
	20	19, 5
	0	5, 0

6. Conclusions

In this paper, we have presented a new normalized HOS-based speech separation algorithm. Both ℓ_2- and Frobenius-normalization of the correction term provide distinct advantages over their unnormalized predecessor including 1) reduction in unnatural amplitude variations in the separated output signals, 2) higher probability of distinct, separated source outputs, 3) better separation in the presence of noise, and 4) apparently faster convergence.

Acknowledgment

The authors wish to acknowledge the support of this research by the U.S. Air Force Research Laboratories, Grant #F41624-99-0001.

References

[1] D. Morgan, E. George, L. Lee, and S. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 407–424, Sep. 1997.

[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, pp. 320–327, May 2000.

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

[5] P. De Leon, "Short-time kurtosis of speech signals with application to co-channel speech separation," *Proc. Int. Conf. Multimedia and Expo*, 2000.

[6] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The darpa speech recognition research database: specifications and status," *Proc. DARPA Workshop on Speech Recognition*, pp. 93–99, 1986.