A COMPARISON OF DECORRELATION CRITERIA FOR THE BLIND SOURCE SEPARATION OF NONSTATIONARY SIGNALS

Craig Fancourt & Lucas Parra
Sarnoff Corp.
201 Washington Road
Princeton, NJ 08543
fancourt@sarnoff.com, lparra@sarnoff.com

ABSTRACT

We compare and contrast two frequency-domain decorrelation criteria for the blind source separation of convolutive mixtures of nonstationary signals. Both criteria operate on the coherence function matrix measured on the outputs obtained by filtering and combining the sensor signals. One criteria involves the Frobenius norm of the coherence function matrix, which is related to the sums of squared coherence functions between all possible pairs of outputs. The other criteria is the determinant of the coherence function matrix, which is related to the volume of state space occupied by the outputs. These criteria are identical for two outputs, but differ for three or more outputs. Full gradient descent of either criteria involves simultaneously decorrelating the outputs while maximizing the output energy. Thus, we will also compare and contrast the resulting algorithms to decorrelation alone. Quantitative comparison will be done using data recorded in a real acoustic environment.

1. INTRODUCTION

Informally, the problem we seek to solve is the following: a set of sources are convolutively filtered, mixed, and measured at an array of sensors, where each sensor measures a different mixture of the sources. We then seek to filter and recombine the sensor signals so as to best recover the sources.

If the sources are statistically independent, then this can be used as a criteria on the outputs for adapting the filter weights. Unfortunately, full independence requires an infinite amount of data and time to measure, and therefore much of the research in blind source separation involves relaxing the full criteria of independence by examining the effect of various priors on the sources. In this paper, we consider the separation problem when the source signals are known to be nonstationary. As was suggested by [3], and shown more rigorously in [6], the independence criteria can then be relaxed to one of decorrelation of the outputs.

However, even within these constraints, there are many ways to measure decorrelation, and even less is known about how to properly exploit non-stationarity. We have previously shown that the coherence function and its power independent normalization is an excellent criteria for measuring decorrelation between a pair of outputs [4]. However, the most common extension of this criteria to more outputs, namely the sum of the squared coherence functions between all pairs of outputs, does not perform as well at separating more than two sources. We therefore seek a stronger decorrelation criteria still based on the coherence function. In this paper, we examine the determinant of the coherence function matrix as a separation criteria, and compare it to our previous criteria.

2. PROBLEM STATEMENT

Formally, the problem we seek to solve is the following: N unknown nonstationary source signals are convolutively mixed and measured by M sensors

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

where **s** is an unknown (Nx1) vector of source signals, **A** is an unknown (MxN) mixing matrix of channel impulse responses, and **x** is a measured (Mx1) vector. The convolution operator * here implies both matrix multiplication and convolution. We then seek a matrix of filters operating on the sensor measurements

$$\mathbf{y}(t) = \mathbf{H}*\mathbf{x}(t) \tag{2}$$

such that the components of the (Nx1) output **y** are statistically independent, where **H** is a (NxM) matrix of filter impulse responses.

In the time domain, independence must be tested not only at the same instant of time, but for all possible combinations of delays of the components of **y**. This problem can be ameliorated by performing the separation in the frequency domain. In the frequency domain, convolution becomes multiplication and (2) becomes

$$\mathbf{Y}(\omega) = \mathbf{W}(\omega) \cdot \mathbf{X}(\omega) \tag{3}$$

where $\mathbf{W}(\omega)$ is a (NxM) matrix of filter frequency responses.

Equations (2) and (3) describe *any* linear system. Ultimately, we must implement them in a specific architecture. In this paper, we use finite impulse response (FIR) filters because this allows the actual filtering operation to be carried out in the frequency domain.

3. COHERENCE FUNCTION MATRIX

Both criteria we will introduce are based on the matrix of coherence functions, $\mathbf{C}_{\mathrm{YY}}$, between the outputs. This matrix has components

$$C_{Y_iY_j}(\omega) = \frac{S_{Y_iY_j}(\omega)}{\sqrt{S_{Y_iY_i}(\omega) \cdot S_{Y_jY_j}(\omega)}} \quad (4)$$

where $S_{Y_iY_j}(\omega)$ is the *cross-power spectral density* between outputs i and j at digital frequency ω. The coherence function matrix can also be expressed directly in terms of a matrix of cross-power spectral densities, $\mathbf{S}_{\mathrm{YY}}$, as

$$\mathbf{C}_{\mathrm{YY}} = \mathbf{P}_{\mathrm{YY}}^{-1/2} \cdot \mathbf{S}_{\mathrm{YY}} \cdot \mathbf{P}_{\mathrm{YY}}^{-1/2} \quad (5)$$

where

$$\mathbf{P}_{\mathrm{YY}} = \mathrm{diag}[\mathbf{S}_{\mathrm{YY}}] \quad (6)$$

is a diagonal matrix of output powers in the frequency domain.

The coherence function matrix is the frequency domain analog of the time-domain correlation matrix. The coherence function matrix at each frequency: (1) is unity along the diagonal; (2) is Hermitian; (3) is positive semi-definite; and (4) has a determinant that is bounded by 0 and 1.

3.1 Estimating the cross-power spectral density

In the frequency domain, the cross-power spectral density matrix between the outputs is formally given by

$$\mathbf{S}_{\mathrm{YY}}(\omega) = E[\mathbf{Y}(\omega)\mathbf{Y}^H(\omega)] \quad (7)$$

In order to efficiently estimate this online and capture nonstationarity, we use a recursive estimator

$$\mathbf{S}_{\mathrm{YY}}(\omega, t) = \gamma\ \mathbf{S}_{\mathrm{YY}}(\omega, t-T) + (1-\gamma)\ \mathbf{Y}(\omega, t) \cdot \mathbf{Y}^H(\omega, t) \quad (8)$$

where γ is a forgetting factor, constrained to $0 < \gamma < 1$ for stability, and T is a block processing time (frame rate) that represents the time it takes to estimate $\mathbf{Y}$. Note that because the signals are *assumed* nonstationary, we have written their frequency response as a function of time. The forgetting factor and block processing time combine to make the effective memory of the estimator to be $T/(1\text{-}\gamma)$. Taking the expected value of both sides of (8) readily shows that it is an *unbiased* estimator for *stationary* signals.

Later, we will also need to make use of the cross-power spectral density between the outputs and the *inputs*

$$\mathbf{S}_{\mathrm{YX}}(\omega, t) = \gamma\ \mathbf{S}_{\mathrm{YX}}(\omega, t-T) + (1-\gamma)\ \mathbf{Y}(\omega, t) \cdot \mathbf{X}^H(\omega, t) \quad (9)$$

3.2 Simultaneous decorrelation and power maximization

The big advantage of any criteria we might form based on the coherence function matrix is that it will be independent of the power of the sources as measured at the microphones.

However, note that since the cross-power terms appear in the numerator of (4), and the power terms appear in the denominator of (4), the full derivative of any criteria based on the coherence matrix will inevitably involve simultaneously decorrelating the outputs while maximizing their power. Maximizing the output power is a common operation in geometric beamforming.

From a beamforming perspective [8], using FIR filters in (3) means that the system can place infinite nulls but only finite maximums in a given direction. For this reason, separating the outputs through decorrelation alone usually means that the beamformer associated with each output places nulls in the directions of all the sources *except one*, with each output ignoring a different source. Our goal has been to find a separation algorithm that, in addition to placing nulls, also places maximums in the directions of the sources. Our hope is that the coherence functions will aid in this by allowing simultaneous decorrelation and power maximization.

4. CRITERIA I: FROBENIUS NORM

The first criteria we examine is the Frobenius norm of the coherence function matrix. This criteria first appeared in [4], but is closely related to the one introduced in [5]-[7]. The Frobenius norm is the sum of the squares of all the matrix elements, and thus the criteria is proportional to the sum of the squared coherence functions between all possible pairs of outputs. It can be concisely written as

$$J_1 = \sum_t \left\| \mathbf{C}_{\mathrm{YY}}(\omega, t) \right\|^2 = \sum_t \mathrm{trace}[\mathbf{C}_{\mathrm{YY}}^H(\omega, t) \cdot \mathbf{C}_{\mathrm{YY}}(\omega, t)] \quad (10)$$

This criteria is bounded by N and N^2.

4.1 Weight update

To obtain the weight update, we take the derivative of the criteria (10) with respect to the complex weights and employ the stochastic gradient approximation by dropping the summation over time. We have previously shown [4] that taking the derivative with respect to the weights only through the cross terms of $\mathbf{S}$ results in

$$\Delta\mathbf{W}_{\mathrm{cross}} = -\eta\mathbf{P}_{\mathrm{YY}}^{-1}\cdot(\mathbf{S}_{\mathrm{YY}}-\mathbf{P}_{\mathrm{YY}})\cdot\mathbf{P}_{\mathrm{YY}}^{-1}\cdot\mathbf{S}_{\mathrm{YX}} \tag{11}$$

where η is a learning rate. The full derivative is obtained by taking the derivative with respect to the weights through both the cross *and* power terms of **S**:

$$\Delta\mathbf{W}_{\mathrm{full}} = -\eta\mathbf{P}_{\mathrm{YY}}^{-1/2}\cdot(\mathbf{C}_{\mathrm{YY}}-\mathrm{diag}[\mathbf{C}_{\mathrm{YY}}^{H}\cdot\mathbf{C}_{\mathrm{YY}}])\cdot\mathbf{P}_{\mathrm{YY}}^{-1/2}\cdot\mathbf{S}_{\mathrm{YX}} \tag{12}$$

Note that both the cross and full derivatives involve the recursive cross-power spectral density estimate, $\mathbf{S}_{\mathrm{YX}}$, between the outputs and inputs.

5. CRITERIA II: DETERMINANT

The second criteria we consider is based on the determinant of the coherence function matrix, which is proportional to the volume of state space occupied by the normalized output vector. Because the outputs are maximally decorrelated when this volume is maximal, the criteria we adopt is one minus the determinant

$$J_2 = \sum_t\{1-|\mathbf{C}_{\mathrm{YY}}(\omega,t)|\} = \sum_t\left\{1-\frac{|\mathbf{S}_{\mathrm{YY}}(\omega,t)|}{|\mathbf{P}_{\mathrm{YY}}(\omega,t)|}\right\} \tag{13}$$

This criteria is bounded by 0 and 1. To the best of our knowledge, this criteria appears here for the first time in the context of source separation.

5.1 Weight update

Once again, we obtain the weight update equation by finding the stochastic gradient of the criteria (13). We haven't yet isolated the weight update equation obtained through the cross terms of **S** alone: $\Delta\mathrm{W}_{\mathrm{cross}}$. However, for simulations where the number of inputs and outputs are known apriori, we obtain the gradient analytically using a symbolic math program.

Nevertheless, we have found the full gradient, where the derivative is taken through both the cross and power terms of **S**, resulting in:

$$\Delta\mathbf{W}_{\mathrm{full}} = -\eta\cdot|\mathbf{C}_{\mathrm{YY}}|\cdot(\mathbf{S}_{\mathrm{YY}}^{-1}-\mathbf{P}_{\mathrm{YY}}^{-1})\cdot\mathbf{S}_{\mathrm{YX}} \tag{14}$$

6. EXAMPLES

6.1 Two outputs

For the case of two outputs, both criteria I and II are identical to within a constant:

$$J_1 = 2+2\sum_t\frac{|S_{Y_1Y_2}|^2}{S_{Y_1Y_1}S_{Y_2Y_2}} \tag{15}$$

$$J_2 = \sum_t\frac{|S_{Y_1Y_2}|^2}{S_{Y_1Y_1}S_{Y_2Y_2}} \tag{16}$$

Thus, the weight updates are also identical. Note that the two criteria are identical independent of the number of inputs.

6.2 Three outputs

For the case of three outputs, criteria I is

$$J_1 = 3+2\sum_t\left\{\frac{|S_{Y_1Y_2}|^2}{S_{Y_1Y_1}S_{Y_2Y_2}}+\frac{|S_{Y_1Y_3}|^2}{S_{Y_1Y_1}S_{Y_3Y_3}}+\frac{|S_{Y_2Y_3}|^2}{S_{Y_2Y_2}S_{Y_3Y_3}}\right\} \tag{17}$$

while criteria II is

$$J_2 = \sum_t\left\{\begin{array}{c}\dfrac{|S_{Y_1Y_2}|^2}{S_{Y_1Y_1}S_{Y_2Y_2}}+\dfrac{|S_{Y_1Y_3}|^2}{S_{Y_1Y_1}S_{Y_3Y_3}}+\dfrac{|S_{Y_2Y_3}|^2}{S_{Y_2Y_2}S_{Y_3Y_3}}\\ -\dfrac{Re[S_{Y_1Y_2}S_{Y_3Y_1}S_{Y_2Y_3}]}{S_{Y_1Y_1}S_{Y_2Y_2}S_{Y_3Y_3}}\end{array}\right\} \tag{18}$$

Note that the first three terms of criteria II (18), to within a constant, embody the entire criteria I (17). However, the fourth term represents interaction between all three outputs. In fact, it is not difficult to show that the second criteria always includes the first criteria plus extra higher-order products between the 2nd-order moments. We would thus expect the criteria based on the determinant to be stronger than the one based on the Frobenius norm. However, it is also more computationally demanding, particularly if the full derivative is used, due to the presence of the inverse of the cross-power spectral density matrix. Of course, the inverse can always be computed analytically in advance for a given problem with a known number of outputs.

7. EXPERIMENT

7.1 Data

We now present the results of applying the algorithms to audio data generated by live speakers in a real room. Because we have previously presented extensive results for the case of two sources [4], for which both criteria are identical, here we concentrate on mixtures of three sources.

Three live male speakers were recorded in a real room of dimensions 3 m x 3.6 m x 2.3 m using 7 unidirectional microphones, 20 cm apart and 2 m from the speakers. The three speakers were located in line with the 1st, 4th, and 7th microphones in the array in order to ensure that at least some of the room transfer functions from the speakers to the microphones were minimum phase. The audio was sampled at 44.1 kHz and downsampled to 11.025 kHz for input to the algorithms.

The training data is a ~12 second recording where all speakers are continuously and simultaneously talking. The test data consists of another ~6 second recording where the three speakers alternately say the digits such that only one speaker is active at a time. The training and testing data sets were recorded consecutively to ensure that the speakers maintained their position and thus that the room responses would not change.

7.2 Results

The four algorithms analyzed were the Frobenius norm with cross terms only, the Frobenius norm with full derivative, the determinant with cross terms only, and the determinant with full derivative.

With 7 inputs and 3 outputs, each algorithm used a 3x7 matrix of FIR filters, each of length 128 taps. They were all run with two different initial conditions on the weights. The first condition was on the filters associated with the direct path microphone-output combinations of 1-1, 4-2, and 7-3. These direct path filters were set to a unity delay of 64 samples. The other filters were set to zeros. The second initial condition was that a delay-sum beamformer associated with each output was pointed at one of the sources.

The filters weights were also updated using two different procedures. The first one didn't adapt the direct path filters, while the second one renormalized all the filter weights (as a single large vector) to have unity norm.

Each algorithm variation was trained for 20 passes through the training data using the same learning rate. For the test data, the active periods of each speaker were hand segmented in order to obtain an accurate measurement of signal separation. To obtain a performance measure after training, the training weights were used to filter the entire test data. The resulting output was then analyzed using the aforementioned segmentation such that whenever a speaker was talking, the power in both the enhanced and rejection channels were measured and accumulated. The *signal to interference ratio* (SIR) was then calculated as

$$SIR = 10 \cdot \log_{10}\left[\frac{P_{enhanced}}{P_{rejection}}\right] \qquad (19)$$

For comparison purposes, the SIR of the microphones alone was measured to be -4.6 dB. Likewise, the SIR of the delay-sum beams pointed at the sources was 0.2 dB.

The best results for all algorithms, initializations, and update variations, was obtained using the Frobenius norm cross power rule using the beamforming initialization and the renormalization update. It achieved an SIR of 9.8 dB, for a total SIR improvement of 14.4 dB. However, an examination of the resulting filter weights interpreted as beamformers showed that this was accomplished by eliminating the low frequencies where resolution is poor, as was confirmed by the “tinny” sound of the resulting output waveforms.

For the direct path delta function weight initialization, the full derivative of ther determinant criteria was the best, but only resulted in an SIR of -2.1, or an SIR improvement of only 2.5 dB. However, unlike the Frobeius norm cross power update, the output suffered no discernable speech distortion.

8. CONCLUSIONS

We have shown and contrasted two criteria that can be formed from the coherence function matrix: namely, the Frobenius norm and the determinant. The two criteria are identical for the case of two outputs, but differ for three or more outputs. We also showed that the full gradient of any criteria based on the coherence function inevitably leads to simultaneously decorrelating the outputs while maximizing their output power.

Experimental results showed that decorrelation alone is superior for signal to interference ratio (SIR) improvement, but only if the filter weights are initialized to beams that point in the direction of the sources.

Adding power maximization to decorrelation achieved an SIR similar to that of designed delay-sum beams but without requiring any special initial conditions on the weights. However, the SIR improvement was minimal.

Nevertheless, the results show some promise that an algorithm can be found that simultaneously nulls interferers while enhancing sources.

9. REFERENCES

[1] S. Roberts and R. Everson (eds.), *Independent component analysis: principles and practice*, Cambridge Univ. Press, Cambridge, U.K., 2001.

[2] S. Haykin (ed.), *Unsupervised adaptive filtering, vol. 1: blind source separation*, Wiley & Sons, New York, 2000.

[3] E. Weinstein, M. Feder, and A.V. Oppenheim, Multichannel signal separation by decorrelation, *IEEE Trans. Speech Audio Processing*, vol. 1, no. 4, pp. 405-413, 1993.

[4] C. Fancourt and L. Parra, The coherence function in blind separation of convolutive mixtures of nonstationary signals, *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 303-312, 2001

[5] L. Parra, C. Spence, and B. De Vries, Convolutive blind source separation based on multiple decorrelation, in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 23-31, 1998.

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

[7] L. Parra and C. Spence, On-line convolutive blind source separation of nonstationary signals, *Journal of VLSI Signal Processing*, vol. 26, pp. 39-46, 2000.

[8] L. Parra and C. Fancourt, An adaptive beamforming perspective on convolutive blind source separation, chapter in *Noise Reduction in Speech Applications*, CRC Press, Boca Raton, 2002.