An Adaptive Beamforming Perspective on Convolutive Blind Source Separation

Lucas Parra, Craig Fancourt
Sarnoff Corporation, CN-5300, Princeton, NJ 08543

Abstract

The purpose of this chapter is to review efficient methods for convolutive blind source separation (BSS) and their combination with geometric beamforming. Ambiguities inherent to convolutive blind separation can be resolved by introducing geometric constraints similar to those used in adaptive beamforming, resulting in more robust algorithms. We concentrate here on a criteria of cross-power spectral minimization, which is sufficient for separation of convolutive mixtures of non-stationary signals. A class of algorithms is presented that minimizes cross-power while linearly constraining the filter structure. Two of these algorithms, which we have termed Geometric Source Separation and the Generalized Sidelobe Decorrelator, will be presented in detail and validated on real room recordings.

1 Introduction

Microphones in an acoustic environment typically capture a mixture of several sources. The goal of convolutive blind source separation (BSS) is to filter the signals from a microphones array to extract the original sources while reducing interfering signals. Due to the spatial variability of a room response, different microphones receive different convolved versions of each source. Separation therefore also requires a convolutive filtering of the sensor signals, effectively resulting for each source in a spatially selective filter or 'beam'. As opposed to conventional geometric or adaptive beamforming, in BSS no assumptions on array geometry or source location are made. Instead, it is only assumed that the desired sources are statistically independent. Convolutive blind source separation can therefore be understood as multiple adaptive beamformers that generate statistically independent outputs, or more simply, outputs with minimal cross-talk.

Considerable progress has been made in formulating sufficient conditions on the source signals and deriving corresponding optimization criteria. Strict independence criteria involve higher order statistics (HOS) of the signals. Unfortunately, HOS are difficult to estimate and lead to complex and computationally demanding algorithms. An alternative to HOS is to constrain the cross-talk minimization to a second-order criteria and instead exploit the non-stationarity of the signals [1, 2]. Estimating second-order statistics is numerically more robust and the criteria lead to simpler algorithms. Most results reported in the literature on real room recordings are based on second order methods while higher-order separation algorithms are often only demonstrated on simulated data.

Aside from this, the independence criteria itself has a number of ambiguities: (1) the recovered sources are only determined up to an arbitrary convolution, (2) more microphones than sources results in under-constrained filter coefficients, and (3) frequency bins may not be assigned consistently to the correct channels. We propose to reduce the inherent ambiguities of convolutive BSS by introducing geometric constraints similar to those used in the Linearly Constraint Minimum Variance (LCMV) algorithm and Generalized Sidelobe Canceler (GSC) [3]. We have termed the resulting algorithms Geometric Source Separation (GSS) [4] and the Generalized Sidelobe Decorrelator (GSD) [5], respectively. Efficient frequency domain on-line and off-line implementations will be outlined. Results on noise reduction for speech recognition in different real room environments and applications will be given.

2 Convolutive blind source separation

Consider M uncorrelated sources, $\mathbf{s}(t) \in \mathbb{R}^M$, originating from different spatial locations and $N > M$ sensors detecting signals $\mathbf{x}(t) \in \mathbb{R}^N$. In a multi-path environment each source j couples with sensor i through a linear transfer function $A_{ij}(\tau)$, such that $x_i(t) = \sum_{j=1}^{M} \sum_{\tau=0}^{P-1} A_{ij}(\tau) s_j(t-\tau)$. Using matrix notation and denoting the convolutions by $*$ we can write this briefly as $\mathbf{x}(t) = A(t) * \mathbf{s}(t)$, or after applying the discrete-time Fourier transform (DTFT),

$$\mathbf{x}(\omega) = A(\omega)\mathbf{s}(\omega)\,. \tag{1}$$

The task of convolutive source separation is to find filters $W_{ij}(\tau)$ that invert the effect of the convolutive mixing $A(\tau)$. One generates model sources

$\mathbf{y}(\omega)$

$$\mathbf{y}(\omega) = W(\omega)\mathbf{x}(\omega) \tag{2}$$

that correspond to the original sources $\mathbf{s}(t)$. Although any linear system is compatible with (2), we restrict ourselves to finite impulse response (FIR) filters since this allows the algorithms to be efficiently implemented in the frequency domain.

2.1 Higher order methods vs. second order non-stationarity

Different criteria for convolutive separation have been proposed [1, 6, 7, 8, 9, 10, 2]. All criteria can be derived from the assumption of statistical independence of the unknown signals. However, typically only pairwise independence of the model sources is used. Pairwise independence implies that all cross-moments factor, yielding a set of necessary conditions for the model sources

$$\forall\, t,n,m,\tau, i \neq j: \quad E\left[y_i^n(t) y_j^m(t+\tau)\right] = E\left[y_i^n(t)\right] E\left[y_j^m(t+\tau)\right] . \tag{3}$$

$E[\,.\,]$ represents the ensemble average and will in practice be replaced with a sample average over a given time window surrounding time t. Convolutive separation requires these conditions to be satisfied for multiple delays τ, corresponding to the delays of the filter taps of $W(\tau)$. For stationary signals, multiple n, m, i.e., higher order criteria, are required. For non-stationary signals multiple t with $n = m = 1$ are sufficient [1, 11, 2]. In this case, conditions (3) state that cross-correlation matrices $R_{\mathbf{yy}}(\tau, t) = E\left[\mathbf{y}(t)\mathbf{y}^T(t+\tau)\right]$ have to be diagonal at all times.

2.2 Separation based on second-order non-stationarity

Joint diagonalization of $R_{\mathbf{yy}}(\tau, t)$ has to find filters $W(\tau)$ that decorrelate model sources $\mathbf{y}(t)$ at multiple t. This can be efficiently implemented in the frequency domain [2] using the Fourier transform of the cross-correlations – the cross-power spectra. Currently we obtain the best results with a diagonalization criteria based on the coherence function [12], defined as

$$C_{y_i y_j}(\omega, t) = \frac{R_{y_i y_i}(\omega, t)}{\sqrt{R_{y_i y_i}(\omega, t) R_{y_j y_j}(\omega, t)}} \tag{4}$$

where $R_{y_i y_j}(\omega, t)$ is the cross-power spectra between outputs y_i and y_j at frequency ω and time t. In matrix notation this can be written as,

$$C_{\mathbf{yy}}(\omega, t) = \Lambda_{\mathbf{yy}}^{-1/2}(\omega, t) R_{\mathbf{yy}}(\omega, t) \Lambda_{\mathbf{yy}}^{-1/2}(\omega, t) \tag{5}$$

with $\Lambda_{\mathbf{yy}}(\omega, t) = \operatorname{diag} R_{\mathbf{yy}}(\omega, t)$. The squared coherence function is real and constrained to lie between 0 and 1 for all frequencies. The coherence function matrix $C_{\mathbf{yy}}(\omega, t)$ is identically equal to one on the diagonal. Its off-diagonal elements vanish only if $R_{\mathbf{yy}}(\omega, t)$ is diagonal, and so we can use the following diagonalization criteria

$$J(W) = \sum_t \sum_\omega \| C_{\mathbf{yy}}(\omega, t) \|^2 \tag{6}$$

with the Frobenius norm, $\|C\|^2 = \operatorname{Tr}\left[C^H C\right]$, representing the square sum of all the elements in the matrix C. The minmization of (6) can be solved using gradient descent methods. The advantage of the coherence function criteria is that the normalization guarantees uniform convergence speed irrespective of the power present in any given frequency bin. The optimization of (6) requires multiple estimates of the cross-power spectra estimated at different times t. In [2] this is done using an off-line algorithm that first estimates the cross-power spectra of the microphones over different time windows, $R_{\mathbf{xx}}(\omega, t)$, and in a second step computes the simultaneously diagonalizing filters $W(\omega)$. The approximation of linear and circular convolution is used there, $R_{\mathbf{yy}}(\omega, t) \approx W(\omega) R_{\mathbf{xx}}(\omega, t) W(\omega)^H$, which is valid if the filters are short in comparison to the length of the discrete Fourier transform (DFT).

2.3 Online decorrelation

In attempting to convert the off-line algorithm into an on-line algorithm, we are faced with the problem of designing an algorithm that *requires* non-stationary signals for convergence. The reason for this is that what we do with each new measurement depends on whether it is part of the previous stationary regime, or represents a transition to a new stationary regime. In the first case, the new data should be used to improve the estimate of the current covariance, implying the use of a long effective memory. In the second case, the data represents the beginning of new covariance matrix for simultaneous diagonalization with previous covariance matrices, implying a short memory is appropriate. Therefore, in addition to the conventional trade-off between convergence speed and misadjustment, we now have a trade-off between estimation accuracy and novel information when measuring correlation.

Note that there are actually two sums over time in (6). First, there is the explicit summation over multiple coherence matrices estimated at different times. There is also an implicit summation over the block of time necessary to estimate each coherence matrix. The key insight is that these two sums are not interchangable because the criteria is non-linear in the power estimates. Often on-line second order decorrelation has proposed a stochastic optimization method whereby the sums over time are entirely removed. In doing so, however, non-stationarity is not properly captured and the algorithms reduce to simple decorrelation which is not sufficient for separation. Therefore, we propose to preserve the time averaging process by recursively estimating the cross-power spectra to capture short-term non-stationarity [12]

$$R_{\mathbf{yy}}(\omega,t) = \gamma R_{\mathbf{yy}}(\omega,t-T) + (1-\gamma)\mathbf{y}(\omega,t)\mathbf{y}^H(\omega,t) \tag{7}$$

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 $y(\omega,t)$. The forgetting factor and block processing time combine to make the effective memory of the estimator to be $T/(1-\gamma)$.

We consider the sum in (6) as an estimator of the instantaneous cost, $\sum_\omega \|C_{\mathbf{yy}}(\omega,t)\|^2$. Stochastic gradient descent uses the instantaneous cost for the weight updates. We take the derivative with respect to the complex weights in the frequency domain, and update the weights at the end of each time block

$$\Delta W = -\mu \left(\Lambda_{\mathbf{yy}}^{-1} R_{yy} \Lambda_{\mathbf{yy}}^{-1} - \mathrm{diag}[\Lambda_{\mathbf{yy}}^{-2} R_{\mathbf{yy}} \Lambda_{\mathbf{yy}}^{-1} R_{\mathbf{yy}}]\right) R_{\mathbf{yx}} \tag{8}$$

where μ is the learning rate and $R_{\mathbf{yx}}$ is a matrix of cross-power spectra between the outputs and the inputs:

$$R_{\mathbf{yx}}(\omega,t) = \gamma R_{\mathbf{yx}}(\omega,t-T) + (1-\gamma)\mathbf{y}(\omega,t)\mathbf{x}^H(\omega,t)\,. \tag{9}$$

The on-line blind source separation algorithm consists of equations (2) and (7)-(9) and is entirely compatible with the overlap-save method of frequency domain adaptive filtering [13]. The overlap-save method implements linear convolution in the frequency domain with the discrete Fourier transform (DFT), or its efficient counterpart, the fast Fourier transform (FFT). However, since the DFT corresponds to circular convolution in the time domain, the filters must be padded with zeros, in turn requiring the use of a larger input buffer. As a result, only the latter part of the output in the time domain is valid. In the context of the present algorithm, it is thus

incorrect to directly use the complex output (2) in updating the cross-power spectral densities in (7) and (9). Rather, they must first be transformed into the time domain (also required to obtain the system output), and the invalid parts zeroed prior to transforming back into the frequency domain for use in (7) and (9). Note that this is not required for $\mathbf{x}$, since the input buffer is always filled with valid input samples prior to transforming into the frequency domain.

The computational complexity of the algorithm scales linearly in the number of inputs and quadratically in the number of outputs. Although other frame rates relative to the filter size can be used, a 50% overlap is the most computationally efficient. For a two input - two output problem at a sampling rate of 8 kHz with 512 taps, the algorithm runs in approximately 1/10 real-time on a 866 MHz Pentium III. It is thus entirely suitable for real-time operation in many-input, many-output problems.

3 Combining source separation with beamforming

In this section the ambiguities of convolutive blind source separation will be discussed. We will review how geometric information is utilized in conventional adaptive beamforming and suggest that second-order BSS can readily be combined with adaptive beamforming methods, as they both operate on the power spectra of the signals.

3.1 Ambiguities of independence criteria

Regardless of the independence criteria, there remains an ambiguity of permutation and scaling in the separating filters. In the convolutive case the scaling ambiguity applies to each frequency bin, resulting in a convolutive ambiguity for each source signal. This expresses the fact that filtered versions of independent signals remain independent. Furthermore, when defining a frequency domain independence criteria such as

$$\forall\, n, m, \omega, i \neq j: \quad E\left[y_i^n(\omega) y_j^m(\omega)\right] = E\left[y_i^n(\omega)\right] E\left[y_j^m(\omega)\right] \tag{10}$$

there is a permutation ambiguity for each frequency. The criteria (10) is equally well satisfied with arbitrary scaling and assignment of indices i, j to the model sources, i.e.

$$W(\omega) A(\omega) = P(\omega) S(\omega) \tag{11}$$

where $P(\omega)$ represents an arbitrary permutation matrix and $S(\omega)$ an arbitrary diagonal scaling matrix per frequency. The most immediate problem

with this is that contributions of a given source may not be consistently assigned to a single model source across different frequency bins [14, 8, 2, 15]. In [4] it is argued that the permutation problem (10) also exists in the time domain criteria (3).

In practice one may want to use a larger number of microphones to improve spatial resolution or reduce aliasing. Aside from the permutation and scaling ambiguity, equation (11) suggests that for a given $A(\omega)$ there is a $N - M$ dimensional linear space of solutions $W(\omega)$. In effect, this indicates that there are additional degrees of freedom in terms of shaping a beam pattern represented by the separating filters $W(\omega)$.

3.2 Linear constraints in geometric beamforming

To disambiguate the permutation, convolution, and under-determined filter coefficients one can use geometric information. In conventional geometric and adaptive beamforming, information such as microphone position and source location are often utilized. A good review of these methods is given in [3]. We want to emphasize that geometric assumptions can be incorporated and implemented as linear constraints to the filter coefficients.

If the source location, array geometry, and microphone response characteristics are known, then we can specify an *array response vector*, $\mathbf{d}(\omega, \mathbf{q}) \in \mathbb{C}^N$, that represents the complex response from the source at location $\mathbf{q}$ to the outputs of the N sensors. Then, for a given beamforming filter, $\mathbf{w}(\omega)$, the total system response is given by

$$r(\omega, \mathbf{q}) = \mathbf{w}(\omega)\mathbf{d}(\omega, \mathbf{q}) \,. \tag{12}$$

For a linear array with omni-directional microphones and a far-field source, the microphone response depends in good approximation only on the angle $\theta = \theta(\mathbf{q})$ between the source and the linear array

$$\mathbf{d}(\omega, \mathbf{q}) = \mathbf{d}(\omega, \theta) = e^{-j\omega \frac{p_i}{c} \sin(\theta)} \tag{13}$$

where p_i is the position of the ith microphone on the linear array and c is speed of sound.

Constraining the response to a particular orientation is simply expressed by the linear constraint, $r(\omega, \theta) = \mathbf{w}(\omega)\mathbf{d}(\omega, \theta) = const.$ This concept is used in the linearly constrained minimum variance (LCMV) algorithm [16].

3.3 Power vs. cross-power criteria

Most adaptive beamforming algorithms rely on a power criteria of a single output. Sometimes power is minimized such as in noise or sidelobe canceling.

There the aim is to adaptively minimize the response at the orientation of interfering signals [3]. Sometimes power is maximized such as in matched-filter approaches that seek to maximize the response of interest [17]. As outlined in section 2.2, blind source separation of non-stationary signals minimizes the off-diagonal elements of $R_{\mathbf{yy}}(t,\omega)$ rather than the diagonal terms as in conventional adaptive beamforming. It can thus identify proper beams for each source despite the fact that multiple sources are simultaneously active. Strict one-channel power criteria has a serious cross-talk or leakage problem, especially in reverberant environments.

4 Geometric Source Separation

We propose to combine blind source separation and geometric beamforming by minimizing cross-power spectra for multiple times while enforcing constraints used in conventional adaptive beamforming. This can be done explicitly by adding a geometric constraint to the optimization criteria, resulting in an algorithm we call Geometric Source Separation [4], or implicitly by embedding the constraint in the system architecture, resulting in the Generalized Sidelobe Decorrelator [5]. The former approach will be discussed in this section, and the latter in the next section.

4.1 Geometric constraints for source separation

To include geometric information we will assume that the sources we are trying to recover are localized at angles $\theta = [\theta_1, ..., \theta_M]$ and at sufficient distance for a far-field approximation to apply. Following section 3.2, the response of the M filters in W for the M directions in θ is given by $W(\omega)D(\omega,\theta)$, where $D(\omega,\theta) = [\mathbf{d}(\omega,\theta_1), ...\mathbf{d}(\omega,\theta_M)]$. In this section we consider linear constraints such as

$$\text{C1:} \qquad \text{diag}\,(W(\omega)D(\omega,\ \theta)) = 1\,, \tag{14}$$

$$\text{or C2:} \qquad W(\omega)D(\omega,\theta) = I\,. \tag{15}$$

Constraint (14) restricts each filter $\mathbf{w}_i(\omega)$—the ith row vector in $W(\omega)$—to have unit response in direction θ_i. Constraint (15) enforces in addition that they have zero response in the direction of interfering signals $\theta_j, i \neq j$.

Note that condition (15) requires that $D(\omega,\theta)$ is invertible for the given set of angles. This is however not always possible. At the frequencies where the grating lobes[1] of a beampattern cross the interfering angles, $D(\omega,\theta)$

[1]Periodic replica of the main lobe due to limited spatial sampling

is not invertible. It is therefore not reasonable to try to enforce (15) as a hard constraint. Rather, as we confirmed in our experiments, it is beneficial to enforce (15) as a soft constraint by adding a penalty term of the form $J_{C2}(\omega) = \|W(\omega)D(\omega,\theta) - I\|^2$ to the optimization criteria (6). Note also that power or cross-power minimization will try to minimize the response at the interference angles. This will lead to an equivalent singularity at those frequencies. It is therefore beneficial to enforce condition (14) also only as a soft constraint by using a penalty term of the form $J_{C1}(\omega) = \|\mathrm{diag}(W(\omega)D(\omega,\theta)) - 1\|^2$.

4.2 Constraints as penalty terms

We implemented the linear constraints (14) and (15) each as a soft constraint with a penalty term. We have further addressed the problem of non-invertibility discussed in section 4.1 by introducing a frequency dependent weighting of the penalty term. The idea is to eliminate the constraints from the optimization for those frequency bands for which $D(\omega,\theta)$ is not invertible. A rather straightforward metric for invertibility is the condition number. We therefore weight the penalty term with the inverse of the condition number of $\lambda(\omega) = \mathrm{cond}^{-1}(D(\omega,\theta))$, which converges to zero when $D(\omega,\theta)$ is not invertible and remains bounded otherwise, i.e. $0 \le \lambda(\omega) \le 1$. The total cost function including frequency dependent weighting of the geometric penalty term is given by

$$J(W) + \lambda \sum_{\omega} \lambda(\omega) J_{C1/2}(W(\omega)) \,. \tag{16}$$

In algorithm *GSS-C1* the penalty term J_{C1} will maximize the response of filters i in orientation θ_i. Note that the delay-sum beamformer ($\mathbf{w}(\omega) = \mathbf{d}(\omega,\theta)^H$) satisfies conditions C1 strictly. In algorithm *GSS-C2* the penalty term J_{C2} will in addition minimize the response for the orientations of the interfering sources. The filter structure that guarantees constraints C2 strictly can be computed with a least squares approach as the pseudo-inverse of $D^H(\omega,\theta)$, or including a regularization term βI for the non-invertibility problem the solution is given by $W(\omega) = D^H(\omega,\theta)\left(D(\omega,\theta)D^H(\omega,\theta) + \beta I)\right)^{-1}$. We denote this solution by *LS-C2*. All *GSS* algorithms reported here minimize cross-power using a straightforward gradient descent algorithm [2].

4.3 Performance evaluation and discussion

Examples of typical response patterns for the GSS algorithms are shown in Figure 1, which shows the beampatterns of the filter weights for a linear

array of 4 microphones with an aperture of 70 cm. There were two sources located at 0 and -40 degrees broadside to the array.

Algorithm *GSS-C1*, and *GSS-C2* place a zero at the angles of interfering sources while maintaining a main lobe in the directions of the corresponding source. For conflicting frequency bands, where a grating lobe coincides with the location of an interfering source, multiple cross-power minimization reduces the main lobe. Qualitatively, the results for the data independent *LS-C2* algorithm capture both main lobe and zeros at the correct angles. Its performance, however, is inferior to the data-adaptive algorithms.

A systematic performance evaluation of the algorithms for the case of two sources in a moderately reverberant room ($T_{30} = 50ms$) is presented in Figure 2. Signal to interference ratio (SIR) is used as a separation metric, which measures the ratio of power (dB) in the enhanced channel to the rejection channel during periods when only one speaker is active. We varied the locations of two speakers that were always at least 2 m from the array. The number of microphones was varied (2-8), but the array aperture was kept at 70 cm. The top row shows the results for some known beamforming algorithms (*del-sum*, *LS-C2*, *LCMV*).

The criteria (6) represents a non-convex optimization problem. The results for the optimization procedure therefore strongly depends on the initial conditions. For comparison, the center row in Figure 2 presents the results for unconstrained multiple cross-power minimization with different initializations of the filter structure. Initializations that have been considered are unit filters ($W(\omega) = I$), delay-sum beamformer (*del-sum*), and least squares (*LS-C2*). The results for unconstrained optimization with the different initializations are labeled *BSS*, *GSS-I2*, and *GSS-I1* respectively.

The last row shows the results for the geometrically constrained separation algorithms (*GSS-C1'*, *GSS-C1*, *GSS-C2*). Algorithm *GSS-C1'* is the same as *GSS-C1* only with constant penalty term λ. Within each row the algorithms are sorted by average performance. Comparison of the results for *GSS-C1'* and *GSS-C1* show the advantage of the frequency dependent weighting of the penalty term. Due to the limited angular resolution all algorithms perform poorly when the sources are too close.

We now present results obtained for the separation of three sources. Note that the permutation problem discussed in section 3.1 becomes worse as the number of sources increases. We show in Figure 3 the performance of separating two speakers and babble noise using a linear array of 8 microphones. The performance mirrors mostly the results obtained for the separation for two sources.

In these experiments the cross-power spectra were estimated at 5 time

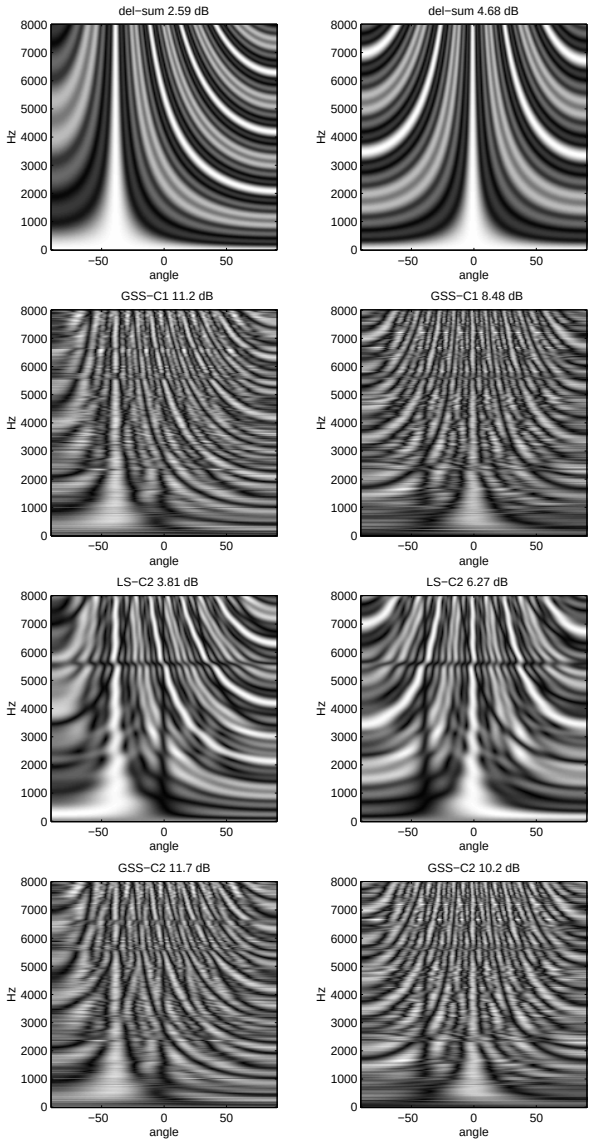

Figure 1: Response for geometrically constrained source separation. Algorithms *GSS-C1* and *GSS-C2* minimize (16) with constraints C1 and C2 respectively. *del-sum* and *LS-C2* satisfy the respective constraints explicitly and are shown for comparison.

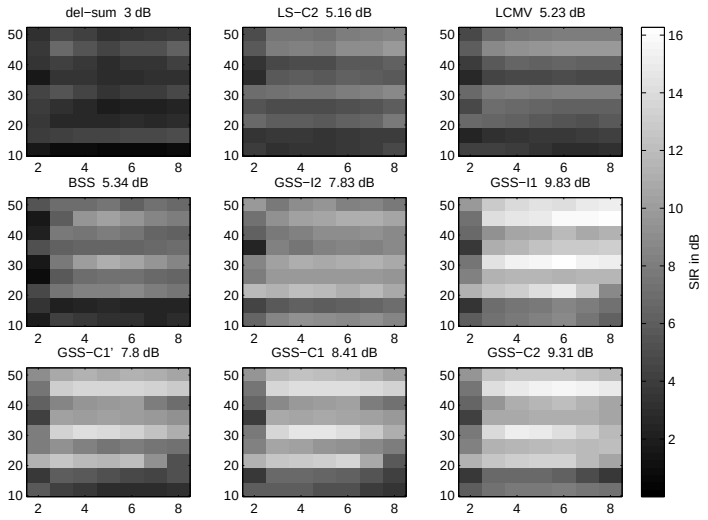

Figure 2: Performance comparison of the proposed algorithms and geometric beamforming for two sources. SIR performance in dB is encoded in grayscale as a function of number of microphones (horizontal axis) and angular separation (vertical axis: $12^o, 18^o, 19^o, 25^o, 33^o, 37^o, 38^o, 41^o, 50^o$). SIR performance averaged over all positions and number of microphones is also given.

instances with a time window of about 3s each, such that a total of about 15s of data is analyzed. In all experiments we used a linear array of cardioid condenser microphones. The user locations were identified acoustically [4].

5 Generalized Sidelobe Decorrelator

As we mentioned previously, one possibility for enhancing a point source while suppressing noise is the linearly constrained minimum variance (LCMV) algorithm, which adaptively filters the sensor signals so as to minimize power, subject to a constraint that a delay-sum beam points in the direction of the source of interest. An alternative but equivalent approach is the generalized sidelobe canceler (GSC) [18], shown in Figure 4(a). It also implements a power minimization criteria on the filtered sensor signals. However, unlike the LCMV, the requirement that a beam points in the direction of interest is enforced in the architecture rather than the criteria. Specifically, the GSC utilizes a delay-sum beam through the use of steering

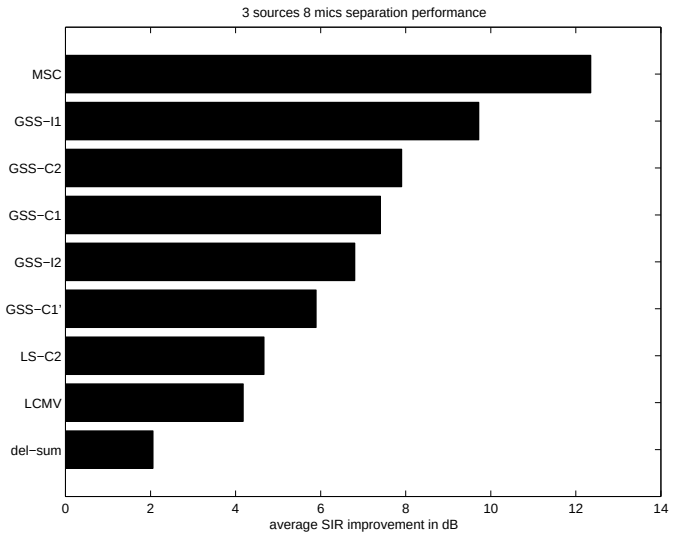

Figure 3: Performance for the separation of 3 sources using 8 microphones. SIR improvement averaged over three configurations with angles $-78^o, -41^o, 0^o; -60^o, 0^o, 60^o;$ and $-43^o, 0^o, 36^o$. The initial average SIR is about -3dB.

delays followed by a linear combiner. The linear combiner is a window that can be designed to vary the trade-off between main lobe width and sidelobe energy. After the steering delays but prior to the linear combiner, the signals are all in phase. This is exploited to form beams orthogonal to the primary beam through the use of a "blocking matrix" [3]. Each row of the blocking matrix is constrained to sum to zero to ensure that the resulting secondary beams will all have a null in the direction of the primary beam. During adaptive power minimization, the secondary beams are adapted out of the primary beam but are prevented by the blocking matrix from canceling any signal that exclusively resides in the primary beam. The GSC approach has the advantage that the resulting optimization can be carried out using unconstrained power minimization, such as the least mean squares (LMS) algorithm. Unlike the LCMV, the constraint is always enforced and no extra steps have to be taken to ensure that the filter weights don't stray from the constraint over time due to finite precision effects.

However, while the GSC exploits the available prior geometric information, it does not exploit the independence prior and is thus subject to the leakage problem associated with power minimization. That is, any leakage

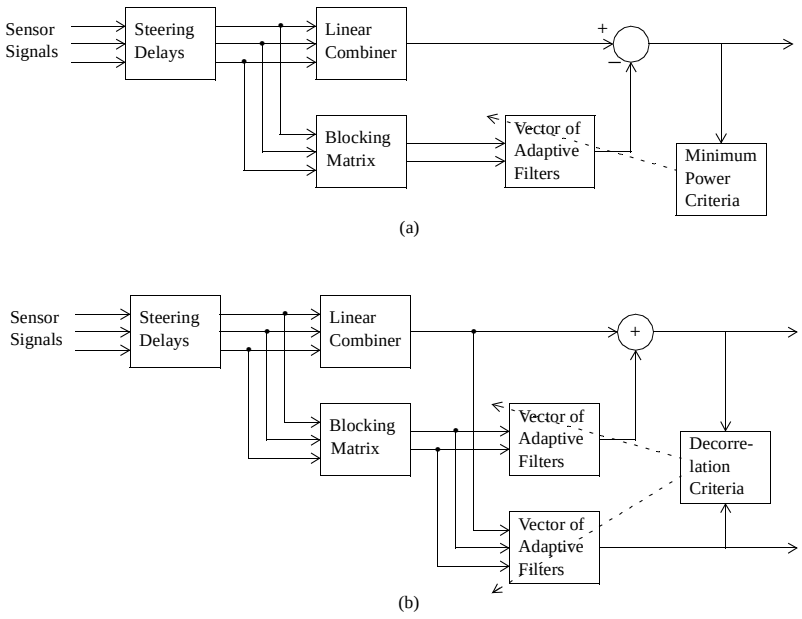

Figure 4: (a) generalized sidelobe canceller; (b) generalized sidelobe decorrelator.

of the primary source into the secondary beams will result in cancellation of the primary source and a degradation of the signal to noise ratio (SNR) improvement. This leakage can be due to any of several factors, including: (1) array calibration errors; (2) primary source location error; (3) a main beam that is narrower than the primary source, caused by a large array aperture; (4) spatial aliasing lobes, caused by an insufficiently spaced sensor array; (5) reverberation, caused by reflections of the primary source coming from directions outside the primary beam.

To overcome these deficiencies, we combine aspects of the generalized sidelobe canceler and blind source separation to create an algorithm we call the generalized sidelobe decorrelator (GSD) [5], shown in Figure 4(b). Like the GSC, it consists of steering delays that place all the sensor signals in-phase, a linear combiner that forms the primary beam, and a blocking matrix that forms secondary orthogonal beams. However, unlike the GSC, instead of adopting a power minimization criteria that adapts the secondary beams out of the primary beam, we adopt a cross-power minimization criteria, as described in Section 2.3, that decorrelates the secondary beams from the primary beam. This allows for removing leakage of the primary source into

Table 1: Real-room experiment

Algorithm	SNR	CER
none	1.2 dB	77.6%
fixed delay-sum beam	1.3 dB	19.4%
generalized sidelobe canceller	3.0 dB	73.9%
blind source separation	3.6 dB	100.0%
generalized sidelobe decorrelator	4.6 dB	5.4%

the secondary beams, while the blocking matrix guarantees the integrity of the primary beam independent of whether the sources are continually active.

5.1 Results in real rooms

We conducted an acoustic experiment designed to demonstrate the superior performance of the algorithm for noise reduction. A 2-D rectangular sensor array of dimension 10 cm x 7 cm was formed, corresponding to the dimensions of a personal digital assistant (PDA), using inexpensive omnidirectional lapel microphones (Audio-Technica ATR35S).

The array was located in a room of dimension 3.0 m x 3.6 m x 2.3 m. A loudspeaker was placed 0.5 m directly in front of the array, which was used to replay a quiet recording of a male speaking 300 short commands over a period of twenty minutes, with a pause of 2-3 seconds between commands. The recording was automatically segmented into speech/non-speech for the purpose of measuring signal to noise ratio (SNR), and the speaker had previously trained an automatic speech recognition system for the purpose of measuring speaker-dependent command error rate (CER). The recognizer and all algorithms operated at 11.025 kHz.

Also in the room but in the corner and facing the wall 2.5 m from the array, a loudspeaker played babble (the sounds of many voices). Outside the room, another loudspeaker played a recording of street noises. The nominal SNR at the microphones was 1.2 dB, which corresponded to a CER of 77.6%. We then applied four on-line adaptive algorithms to the array signals, each of which used FIR filter sizes of 512 taps. The results are shown in Table 1.

Because the source was directly in front of the array, the fixed delay-sum beam could be obtained by a simple averaging of the four sensors. Although the fixed beam does not provide much SNR improvement, it does provide significant CER improvement, primarily because it does not distort the speech.

Next, we implemented the GSC using a "Walsh" blocking matrix (see

[18]) to form three secondary beams orthogonal to the primary delay-sum beam. The secondary beams were adapted out of the primary beam using the frequency domain LMS algorithm. Although there is improvement in the SNR, there is degradation in the CER relative to the delay-sum beam, most likely due to spectral distortion of the speech.

Next, we applied BSS on the 4 raw inputs signals, using the algorithm of Section 2.3, with 2 outputs. Although BSS provides a small SNR improvement over GSC, the algorithm completely destroys the recognition performance. Part of the problem is that BSS requires that the sources be simultaneously active, and thus the filters start to degrade during the silent periods between commands. In addition, the frequency-domain permutation problem (Section 3.1) can distort the speech spectrum.

Finally, we applied our new hybrid GSD by performing BSS on the fixed delay-sum beam and blocking matrix outputs taps, and obtained very encouraging results. In addition to obtaining the largest SNR improvement of any of the algorithms, the CER was a very respectable 5.4%, approaching the single microphone CER of 2.0% in a quiet environment.

6 Summary

This chapter emphasizes the importance of second order criteria and the use of prior geometric information to solve the problem of separating multiple sources in an acoustic environment. It combines notion from adaptive beamforming and blind source separation resulting in semi-blind algorithms where at least microphone locations are known. The assumption is made that sources are reasonably well localized and that user location can be determined acoustically. The algorithms overcome the cross-talk problems of conventional adaptive beamforming and the ambiguity problems of convolutive blind source separation.

References

[1] E. Weinstein, M. Feder, and A.V. Oppenheim, "Multi-Channel Signal Separation by Decorrelation", *IEEE Trans. Speech Audio Processing*, vol. 1, no. 4, pp. 405–413, Apr. 1993.

[2] L. Parra and C. Spence, "Convolutive blind source separation of non-stationary sources", *IEEE Trans. on Speech and Audio Processing*, pp. 320–327, May 2000.

[3] B. Van Veen and K. Buckley, "Beamforming techniques for spatial filtering", in *Digital Signal Procesing Handbook*. CRC Press, 1997.

[4] L. Parra and C. Alvino, "Geometric Source Separation: Merging convolutive source separation with geometric beamforming", in *IEEE International Workshop on Neural Networks and Signal Processing*, 2001, pp. 273–282.

[5] C. Fancourt and L. Parra, "The generalized sidelobe decorrelator", in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[6] H.-L. N. Thi and C. Jutten, "Blind source separation for convolutive mixtures", *Signal Processing*, vol. 45, no. 2, pp. 209–229, 1995.

[7] D. Yellin and E. Weinstein, "Multichannel Signal Separation: Methods and Analysis", *IEEE Trans. Signal Processing*, vol. 44, no. 1, pp. 106–118, 1996.

[8] S. Shamsunder and G. Giannakis, "Multichannel Blind Signal Separation and Reconstruction", *IEEE Trans. Speech Audio Processing*, vol. 5, no. 6, pp. 515–528, Nov. 1997.

[9] F. Ehlers and H.G. Schuster, "Blind Separation for Convolutive Mixtures and an Application in Automatic Speech Recognition in a Noisy Environment", *IEEE Trans. Signal Processing*, vol. 45, no. 10, pp. 2608–2612, Oct. 1997.

[10] H. Sahlin and H. Broman, "Separation of real-world signals", *Signal Processing*, vol. 64, pp. 103–104, 1998.

[11] M. Kawamoto, "A method of blind separation for convolved non-stationary signals", *Neurocomputing*, vol. 22, no. 1-3, pp. 157–171, 1998.

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

[13] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 1996.

[14] V. Capdevielle, C. Serviere, and J.L. Lacoume, "Blind separation of wide-band sources in the frequency domain", in *Proc. ICASSP*, 1995, pp. 2080–2083.

[15] K. Diamantaras, A. Petropulu, and B. Chen, "Blind two-input-two-output FIR channel identification based on frequency domain second-order statistics", *IEEE Trans. on Signal Processing*, vol. 48, no. 2, pp. 534–542, Feb. 2000.

[16] E. Frost, "An algorithm for linearly constrained adaptive array processing", *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[17] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech", *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 425 – 437, Sept. 1997.

[18] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming", *IEEE Trans. Antennas Propagation*, vol. AP-30, no. 1, pp. 27–34, 1982.