Convolutive Blind Source Separation based on Multiple Decorrelation. *

Lucas Parra, Clay Spence, Bert De Vries
Sarnoff Corporation, CN-5300, Princeton, NJ 08543
lparra | cspence | bdevries @ sarnoff.com

Abstract

Acoustic signals recorded simultaneously in a reverberant environment can be described as sums of differently convolved sources. The task of source separation is to identify the multiple channels and possibly to invert those in order to obtain estimates of the underling sources. We tackle the problem by explicitly exploiting the non-stationarity of the acoustic sources. Changing cross-correlations at multiple times give a sufficient set of constraints for the unknown channels. A least squares optimization allows us to estimate a forward model, identifying thus the multi-path channel. In the same manner we can find an FIR backward model, which generates well separated model sources. Under certain conditions we obtain up to 14 dB signal enhancement in a real room environment.

1 Introduction

A growing number of researchers have published in recent years on the problem of blind source separation. For one, the problem seems of relevance in various application areas such as speech enhancement with multiple microphones, crosstalk removal in multichannel communications, multi-path channel identification and equalization, direction of arrival (DOA) estimation in sensor arrays, improvement over beam forming microphones for audio and passive sonar, and discovery of independent sources in various biological signals, such as EEG, MEG and others. Additional theoretical progress in our understanding of the importance of higher order statistics in signal modeling have generated new techniques to address the problem of identifying statistically independent signals - A problem which lays at the heard of source separation. This development has been driven not only by the signal processing community but also by machine learning research that has treated the issue mainly as a density estimation task.

*patent pending

The basic problem is simply described. Assume d_s statistically independent sources $\mathbf{s}(t) = [s_1(t), ..., s_{d_s}(t)]^T$. These sources are convolved and mixed in a linear medium leading to d_x sensor signals $\mathbf{x}(t) = [x_1(t), ..., x_{d_x}(t)]^T$ that may include additional sensor noise $\mathbf{n}(t)$,

$$\mathbf{x}(t) = \sum_{\tau=0}^{P} A(\tau)\mathbf{s}(t-\tau) + \mathbf{n}(t) \tag{1}$$

How can one identify the $d_x d_s P$ coefficients of the channels A and how can one find an estimate $\hat{\mathbf{s}}(t)$ for the unknown sources?

Alternatively one may formulate an FIR inverse model W,

$$\mathbf{u}(t) = \sum_{\tau=0}^{Q} W(\tau)\mathbf{x}(t-\tau) \tag{2}$$

and try to estimate W such that the model sources $\mathbf{u}(t) = [u_1(t), ..., u_{d_u}(t)]^T$ are statistically independent.

Most of the previous work has concentrated on the property of statistical independence, and ignores the additive noise [1, 2, 3, 4, 5, 4, 6, 7].

An alternative approach to the statistical independence condition in the convolutive case has been touched on by Weinstein et al. in [8]. For non-stationary signals a set of second order conditions can be specified that uniquely determines the parameters A. No algorithm has been given in [8] nor has there been to our knowledge any results reported on this approach. A resent paper by Ehlers and Schuster [9] carries that spirit in attempting to solve for the frequency components of A by extending prior work of Molgedey and Schuster [10] on instantaneous mixtures into the frequency domain. They fall short however in carefully considering the issues at hand and mistakenly confuse this idea with simple decorrelation of multiple taps in the time domain, which is known to be insufficient [11, 12]. Simultaneous work by Principe [13] suggests a similar approach for the time domain.

2 Convolutive Mixture

In previous work on instantaneous mixtures it has been demonstrated that one can separate signals by simultaneously diagonalizing the cross-correlation at different time lags [11, 10, 14, 8].

As suggested for other source separation algorithms our approach to the convolutive case is to transform the problem into the frequency domain and to solve simultaneously a separation problems for every frequency [15, 16, 5, 9]. The solution for each frequency would seem to have an arbitrary permutation. The main issues to be addressed here are how to obtain second order equations in the frequency domain, and how to choose the arbitrary

permutations for all individual problems consistently. We will take up these issues in the following sections.

2.1 Cross-correlations, circular and linear convolution

First consider the cross-correlations $R_x(t, t+\tau) = \left\langle \mathbf{x}(t)\mathbf{x}(t+\tau)^T \right\rangle$. For stationary signals the absolute time does not matter and the correlations depend on the relative time, i.e. $R_x(t, t+\tau) = R_x(\tau)$. Denote with $R_x(z)$ the z-transform of $R_x(\tau)$. We can then write

$$R_x(z) = A(z)\Lambda_s(z)A(z)^H + \Lambda_s(z) \tag{3}$$

where $A(z)$ represents the matrix of z-transforms of the FIR filters $A(\tau)$, and $\Lambda_s(z)$, and $\Lambda_n(z)$ are the z-transform of the auto-correlation of the sources and noise. Again they are diagonal due to the independence assumptions.

For practical purposes we have to restrict ourself to a limited number of sampling points of z. Naturally we will take T equidistant samples on the unit circle such that we can use the discrete Fourier transform (DFT). For periodic signals the DFT allows us to express circular convolutions as products such as in (3). However, in (1) and (2) we assumed linear convolutions. A linear convolution can be approximated by a circular convolution if $P \ll T$ and we can write approximately

$$\mathbf{x}(\omega, t) \approx A(\omega)\mathbf{s}(\omega, t) + \mathbf{n}(\omega, t), \text{ for } P \ll T \tag{4}$$

where $\mathbf{x}(\omega, t)$ represents the DFT of the frame of size T starting at t, $[\mathbf{x}(t), ..., \mathbf{x}(t+T)]$, and is given by $\mathbf{x}(\omega, t) = \sum_{\tau=0}^{T-1} e^{-\mathbf{i}2\pi\omega\tau}\mathbf{x}(t+\tau)$ and corresponding expressions for $\mathbf{s}(\omega, t)$ and $A(\omega)$.

For non-stationary signals the cross-correlation will be time dependent. Estimating the cross-correlation at the desired resolution of $1/T$ is difficult if the stationarity time of the signal is in the order of magnitude of T or smaller. We are content however with any cross-correlation average which diagonalizes for the source signals. One such sample average is,

$$\bar{R}_x(\omega, t) = \frac{1}{N}\sum_{n=0}^{N-1} \mathbf{x}(\omega, t+nT)\mathbf{x}^H(\omega, t+nT) \tag{5}$$

We can then write for such averages

$$\bar{R}_x(\omega, t) = A(\omega)\Lambda_s(\omega, t)A^H(\omega) + \Lambda_n(\omega, t) \tag{6}$$

If N is sufficiently large we can assume that $\Lambda_s(\omega, t)$ and $\Lambda_n(\omega, t)$ can be

modeled as diagonal again due to the independence assumption. For equations (6) to be linearly independent for different times t and different ω it will be necessary that $\Lambda_s(\omega, t)$ changes over time for a given frequency, i.e. the signal are non-stationary.

2.2 Backward model

Given a forward model A it is not guaranteed that we can find a stable inverse. In the two dimensional square case the inverse channel is easily determined from the forward model [8, 3]. It is however not apparent how to compute a stable inversion for arbitrary dimensions. In this present work we prefer to estimate directly a stable multi-path backward FIR model such as (2). In analogy to the discussion above we wish to find model sources with cross-power-spectra satisfying[1],

$$\Lambda_s(\omega, t) = W(\omega) \left(\bar{R}_x(\omega, t) - \Lambda_n(\omega, t)\right) W^H(\omega) \tag{7}$$

In order to obtain independent conditions for every time we choose the times such that we have non-overlapping averaging times for $\bar{R}_x(\omega, t_k)$, i.e. $t_k = kTN$. But if the signals vary sufficiently fast overlapping averaging times could have been chosen. A multi-path channel W that satisfies these equations for K times simultaneously can be found, again with an LS estimation[2]

$$\begin{aligned} E(\omega, k) &= W(\omega)(\bar{R}_x(\omega, k) - \Lambda_n(\omega, k))W^H(\omega) - \Lambda_s(\omega, k) \\ \hat{W}, \hat{\Lambda}_s, \hat{\Lambda}_n &= \underset{\substack{W, \Lambda_s, \Lambda_n, \\ W(\tau) = 0, \tau > Q, \\ W_{ii}(\omega) = 1}}{\arg\min} \sum_{\omega=1}^{T} \sum_{k=1}^{K} \|E(\omega, k)\|^2 \end{aligned} \tag{8}$$

Note the additional constraint on the filter size in the time domain. Up to that constraint it would seem the various frequencies $\omega = 1, ..., T$ represent independent problems. The solutions $W(\omega)$ however are restricted to those filters that have no time response beyond $\tau > Q \ll T$. Effectively we are parameterizing Td_sd_x filter coefficients in $W(\omega)$ with Qd_sd_x parameters $W(\tau)$. The LS solutions can again be found with a gradient descent algorithm. We will first compute the gradients with respect to the complex valued filter coefficients $W(\omega)$ and discuss their projections into the subspace of permissible solutions in the following section.

The gradients of the LS cost in (8) are,

[1] $W(\omega)$ represents the DFT with frame size T of the time domain $W(\tau)$. In what follows time and frequency domain are identified by their argument τ or ω.

[2] In short we write again $\Lambda_s(\omega, k) = \Lambda_s(\omega, t_k)$ and $\Lambda_s = \Lambda_s(\omega, t_1), ..., \Lambda_s(\omega, t_K)$ whenever possible. The same applies to $\Lambda_n(\omega, t)$ and $R_x(\omega, t)$

$$\frac{\partial E}{\partial W^*(\omega)} = 2\sum_{k=1}^{K} E(\omega,k)W(\omega)\left(\bar{R}_x(\omega,k) - \Lambda_n(\omega,k)\right) \tag{9}$$

$$\frac{\partial E}{\partial \hat{\Lambda}_s^*(\omega,k)} = -\operatorname{diag}\left(E(\omega,k)\right) \tag{10}$$

$$\frac{\partial E}{\partial \hat{\Lambda}_n^*(k)} = -\operatorname{diag}\left(W^H(\omega)E(\omega,k)W(\omega)\right) \tag{11}$$

With (10)=0 one can solve explicitly for parameters $\Lambda_s(\omega,k)$, while parameters $\Lambda_s(\omega,k), W(\omega)$ may be computed with a gradient descent rule.

2.3 Permutations and constraints

The above unconstrained gradients can not be used as such but have to be constrained to remain in the subspace of permissible solutions with $W(\tau) = 0$ for $\tau > Q \ll T$. This is important since it is a necessary condition for equations (7) to hold to a good approximation.

Additionally, and this is a crucial point that may have not been realized in previous literature, not all possible permutations of frequencies will lead to FIR filters which satisfy that constrain. Note that any permutation of the coordinates for every frequency will lead to exactly the same error $E(\omega,k)$. The total cost will therefore not change if we choose a different permutation of the solutions for every frequency ω. Obviously those solutions will not all satisfy the condition on the length of the filter. Effectively, requiring zero coefficients for elements with $\tau > Q$ will restrict the solutions to be smooth in the frequency domain, e.g., if $Q/T = 8$ the resulting DFT corresponds to a convolved version of the coefficients with a *sinc* function 8 times wider than the sampling rate.

It is therefore crucial to enforce that constraint by starting the gradient algorithm with an initial point that satisfies the constraints, and then following the constrained gradient. The normalization condition that avoid trivial solutions of the LS optimization have to be enforced simultaneously. The constrained gradients are obtained by applying the corresponding projection operators. The projection operator that zeros the appropriate delays for every channel $W_{ij} = [W_{ij}(0), ..., W_{ij}(\omega), ..., W_{ij}(T)]^T$ is

$$P^{(2)} = FZF^{-1} \tag{12}$$

where the DFT is given by $F_{ij} = 1/\sqrt{T}e^{-\mathrm{i}2\pi ij}$, and Z is diagonal with $Z_{ii} = 1$ for $i < Q$ and $Z_{ii} = 0$ for $i \geq Q$. The projection operator that enforces unit gains on diagonal filters $W_{ii}(\omega) = 1$ is applied simply by setting the diagonal terms of the gradients to zero. These projections are orthogonal and can be applied independently of each other. The so obtained constrained gradient can be used in a gradient update of the filter parameters.

The computational cost of the algorithm are dominated by the costs of estimating $\bar{R}_x(\omega, t)$ in (5), the gradient computation in (9), and the projection (12). Before the gradient descent starts one needs to evaluate (5) K times, resulting in a computational cost of $O(KNd_xT(\log T + dx))$. Every gradient step requires then a computation of $O(KTd_xd_s(2d_s + d_x))$ in (9) and $O(d_xd_sKT \log T)$ in (12).

3 Experimental results

The main difficulty in assessing the quality of a separation from real recordings is that the true sources are generally not available.

We define as the Signal to Signal Ratio (SSR) of a signal $\mathbf{s}(t)$ in a multipath channel $H(\omega)$ the total signal powers of the direct channel versus the signal power stemming from cross channels.

$$SSR[H, s] = \frac{\sum_\omega \sum_i |H_{ii}(\omega)|^2 \left\langle |s_i(\omega)|^2 \right\rangle}{\sum_\omega \sum_{i \neq j} \sum_j |H_{ij}(\omega)|^2 \left\langle |s_j(\omega)|^2 \right\rangle} \tag{13}$$

In the case of known channels and source signals we can compute the expressions directly by using a sample average over the available signal and multiplying the powers with the given direct and cross channel responses. In the case of unknown channel response and underlying signals we can estimate the direct powers (numerator) and cross-powers (denominator) by using alternating signals. We estimate the contributions of source j while source j is 'on' and all other sources are 'off'. During periods of silence, i.e. all sources are 'off' we can estimate background noise powers in all channels to subtract from the signal powers.

In first experiment we hand-segment the signal of alternating speakers into speech and non-speech to obtain the 'on' and 'off' labels. We recorded two speakers in a quiet office environment (12.6 dB) for 30 seconds at 8kHz. The two microphones were 50 cm apart facing the speakers situated at about 150 cm distance to the microphones and from each other. Half of the signal was alternating speech to allow the measurement of SSR as outlined above. In the second half the speakers talk simultaneously. The SSR of the recorded signal was about 0 dB. The resulting separation after 400 iterations with, $T = 4096, Q = T/8, N = 10, K = 5$, gave a SSR of 14.5 dB. The recordings and the separation can be heard at [17] along with other results on real room recordings. A systematic analysis with real recordings is still outstanding, in particular in order to determine the robustness of the algorithm to background noise (more sources than sensors) and various room responses (geometry of the room and location of sources and microphones)

We have used artificial random filters in order to determine the dependency of the algorithm on the various parameters such as number of channels, number of sources, filter size, and required signal length. All experiments used mixtures with an SSR of roughly 0 dB as input. The direct forward

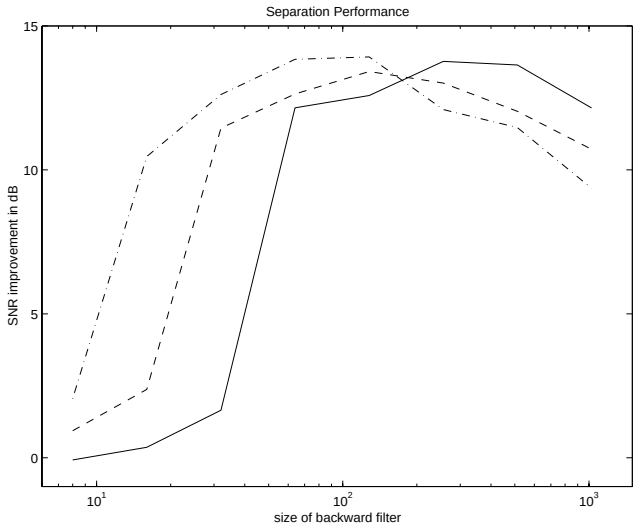

Figure 1: Separation performance as a function of separation filter size Q for forward filter sizes $P = 16, 32, 64$ (dash-doted, dashed, and solid lines respectively). Mean values over 15 runs with different random forward filters are shown. The deviation from that mean was in average 3.4 dB, 4.8 dB, and 1.3 dB respectively.

channels where constant gain ($A_{ii}(\omega) = 1$) and the cross-channels where set in the time domain to zero mean, normal, random numbers. The deviation was adjusted to produce in average a SSR of 0dB. We used $K = 5$ in all cases.

Figure 1 shows the dependency of the separation performance as measured with (13) on the size Q of the unmixing filters W for $P = 16, 32, 64$. Too few parameter are not sufficient to perform the separation, while too many parameters become harder to estimate. That tradeoff depends on the size P of the forward filter. We can see that the performance peaks as expected with a shifting maximum for different P.

Figure 2 shows the dependency of the separation performance on the length of the used signal. In order to keep $K = 5$ we had to use overlapping time windows and varying N. We see that the performance does not decay too rapidly, allowing reasonable performance even down to one second. This suggest that an online implementation of this algorithm might give reasonable result with an adaption time of a few seconds.

For the experiments in figure 1 and 2 a 15 seconds signal of continuous speech and music sampled at 8 kHz where used.

Another interesting question is how well the performance scales with the number of channels for the square case ($d_x = d_s$). Figure 3 shows the result up

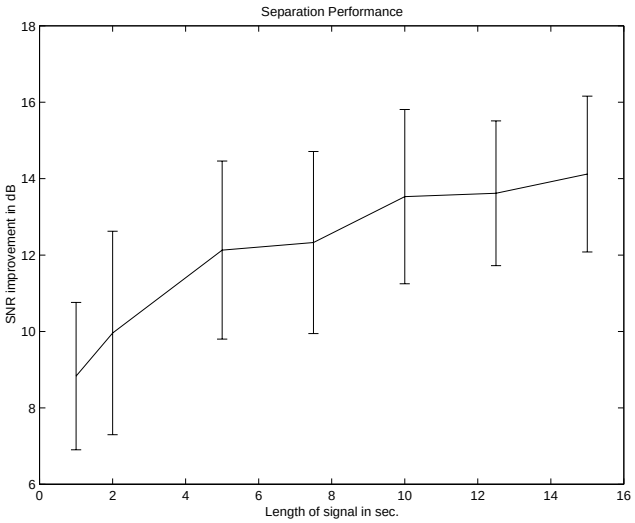

Figure 2: Separation performance as a function of signal length in seconds for random forward filters of size $P = 64$. Mean and standard deviation over 15 runs with different random forward filters are shown.

to 10 channels with five second of music per channel sampled at 8.8 kHz (CD recordings). Again, the separation performance does not decay seriously with the number of channels. We observed however that the algorithm converges slower with increasing number of channels. The results in figure 3 were obtained by increasing the number of iterations linearly with the number of channels. Additionally every gradient step is slower for increasing number of channels as the number of computations scales with $O(d_s d_x^2 + d_s^2 d_x)$.

Note that the separation quality obtained for the real room recording are similar those of the of the random filters. This is not a proof that the results obtained with the artificial mixtures apply fully to real recordings but gives at least some indication on its validity.

4 Conclusion

A large body of work has accumulated in the last two decades on the problem of blind source separation. We have concentrated on the rather general case of recovering convolutive mixtures of wideband signals for less or equal number of sources than sensors. Most of the concepts in this work where borrowed from previous work. The main contributions are: We explicitly use the property of non-stationary. Careful considerations of how to measure second order statistics in the frequency domain allows us to obtain a constraint LS cost that is optimal at the desired solutions. The constraint on the filter

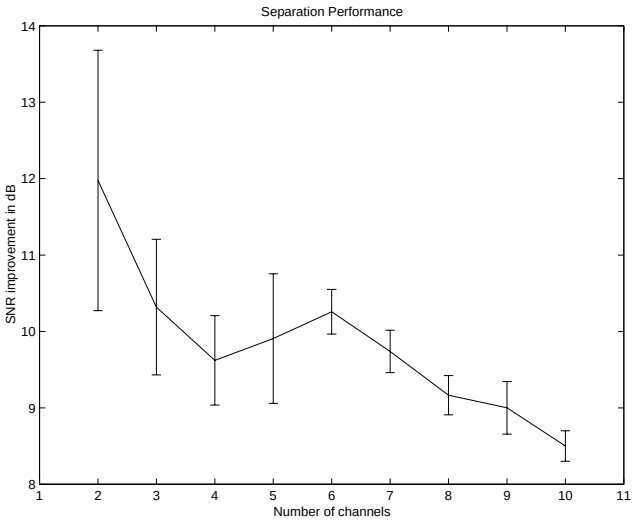

Figure 3: Separation performance as a function of number of channels ($d_s = d_x$) for random forward filters of size $P = 64$. Mean and standard deviation over 10 runs with different random forward filters are shown.

size solves the permutation problem of wideband signals. The current experimental results suggest that under proper conditions we can achieve for two channels a crosstalk reduction of up to 14 dB in a realistic environment.

References

[1] Daniel Yellin and Ehud Weinstein, "Multichannel Signal Separation: Methods and Analysis", *IEEE Transaction on Signal Processing*, vol. 44, no. 1, pp. 106–118, January 1996.

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

[3] Hoang-Lan Nguyen Thi and Christian Jutten, "Blind source separation for convolutive mixtures", *Signal Processing*, vol. 45, pp. 209–229, 1995.

[4] R. Lambert and A. Bell, "Blind Separation of Multiple Speakers in a Multipath Environment", in *ICASSP 97*. IEEE, 1997, pp. 423–426.

[5] T. Lee, A. Bell, and R. Lambert, "Blind separation of delayed and convolved sources", in *Advances in Neural Information Processing Systems 1996*, 1997, Authors provide signals and results at http://www.cnl.salk.edu/∼tewon/.

[6] S. Amari, S.C. Douglas, A. Cichocki, and A.A. Yang, "Multichannel blind deconvolution using the natural gradient", in *Proc. 1st IEEE Workshop on Signal Processing App. Wireless Comm.*, Paris, France, 1997, IEEE.

[7] Lucas Parra, *1997 International Summer School on Adaptive Processing of Sequences*, chapter Temporal Models in Blind Source Separation, Springer, 1998.

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

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

[10] L. Molgedey and G. Schuster, H., "Separation of a mixture of independent signals using time delayed correlations", *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994.

[11] Y. Bar-Ness, J. Carlin, and M. Steinberger, "Bootstrapping adaptive cross-pol canceller for satellite communications", in *IEEE Int. Conf. Comunications*, Philadelphia, PA, June 1982, pp. 4F.5.1–4F.5.5.

[12] R.L.L Tong and Y.H.V.C. Soon, "Interdeterminacy and Identifiability of blind identification", *IEEE Transaction on Circuits Systems*, vol. 38, pp. 499–509, 1991.

[13] H. Wee and J. Principe, "A criterion for BSS based on simultaneous diagonalization of time correlation matrices", in *Proc. IEEE Workshop NNSP'97*, 1997, pp. 496–508.

[14] Stefan Van Gerven and Van Compernolle Dirk, "Signal Separation by Symetric Adaptive Decorrelation: Stability, Convergence, and Uniqueness", *IEEE Transaction on Signal Processing*, vol. 43, no. 7, pp. 1602–1612, July 1995.

[15] C. Jutten, *Calcul neuromimetique et traitment du signal: Analyse en composantes independantes*, PhD thesis, UJF-INP Grenoble, 1987.

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

[17] Lucas Parra, "Separation results", http://www.sarnoff.com/Papers/BSS.html, 1998.