CONVOLUTIVE BLIND SOURCE SEPARATION FOR NOISY MIXTURES

Robert Aichner, Herbert Buchner, Walter Kellermann
Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg
{aichner,buchner,wk}@LNT.de

1. INTRODUCTION

The problem of separating convolutive mixtures of unknown time series arises in several application domains, a prominent example being the so-called cocktail party problem, where we want to recover the speech signals of multiple speakers who are simultaneously talking in a room. The room may be reverberant due to reflections on the walls, i.e., the original source signals $s_q(n)$, $q = 1, \dots, P$ are filtered by a multiple input and multiple output (MIMO) system before they are picked up by the sensors x_p (Fig. 1). Moreover, in most environments a noise term n_p (e.g.,

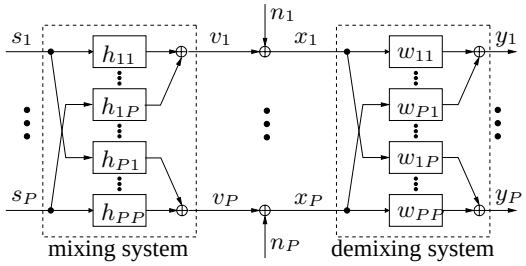

Fig. 1. Noisy BSS model.

sensor or background noise) will be picked up by each sensor x_p, $p = 1, \dots, P$. An M-tap mixing system is thus described by

$$x_p(k) = v_p(k) + n_p(k) = \sum_{q=1}^{P} \sum_{\kappa=0}^{M-1} h_{qp}(\kappa) s_q(k-\kappa) + n_p(k), \tag{1}$$

where $h_{qp}(\kappa)$, $\kappa = 0, \dots, M-1$ denote the coefficients of the filter from the q-th source to the p-th sensor.

In blind source separation (BSS), we are interested in finding a corresponding demixing system, where the output signals $y_q(n)$, $q = 1, \dots, P$ are described by

$$y_q(k) = \sum_{p=1}^{P} \sum_{\kappa=0}^{L-1} w_{pq}(\kappa) x_p(k-\kappa). \tag{2}$$

where $w_{pq}(\kappa)$, $\kappa = 0, \dots, L-1$ denotes the current weights of the MIMO filter taps from the p-th sensor channel to the q-th output channel. BSS is solely based on the fundamental assumption of mutual statistical independence of the different source signals ($|\gamma_{s_1 s_2}|^2 \approx 0$ in Fig. 2). Thus, the separation is achieved by forcing the output signals y_q to be mutually statistically decoupled up to joint moments of a certain order [1].

2. ROBUST BSS FOR NOISY SIGNALS

In [2, 3] a general BSS framework for convolutive mixtures was presented for the noise-less case. Starting from a generic algorithm various efficient algorithms in the time and frequency domain were introduced. To investigate the robustness against noise, we exemplarily choose a narrowband frequency-domain algorithm derived from this framework. It is based on second-order statistics simultaneously utilizing the nonwhiteness and the nonstationarity of the source signals. The narrowband approach allows a bin-wise processing so that the time-domain cost function introduced in [2] can be reformulated for each frequency bin $\nu = 0, \dots, 4L-1$

$$\mathcal{J}^{(\nu)}(m) = \sum_{i=0}^{\infty} \beta(i,m) \left\{ \log\det \operatorname{diag} \mathbf{S}_{\mathbf{yy}}^{(\nu)}(i) - \log\det \mathbf{S}_{\mathbf{yy}}^{(\nu)}(i) \right\}, \tag{3}$$

where m denotes the block index and $\mathbf{S}_{\mathbf{yy}}^{(\nu)}$ is the $P \times P$ cross-power spectral density matrix in the ν-th frequency bin. Here we choose the normalized weighting function $\beta(i,m) = (1-\lambda)\lambda^{m-i}$ leading to an efficient on-line version allowing for tracking of time-varying environments. It should be noted that as shown in [2, 3] at least one time-domain constraint has to be included to prevent permutations among the output signals in each frequency bin. The natural gradient (e.g., [1]) derivation of (3) with respect to the $P \times P$ demixing matrix $\underline{\mathbf{W}}^{(\nu)}$ leads to an iterative algorithm with the following coefficient update

$$\Delta\underline{\mathbf{W}}^{(\nu)} = 2 \sum_{i=0}^{\infty} \beta(i,m) \underline{\mathbf{W}}^{(\nu)} \left\{ \mathbf{S}_{\mathbf{yy}}^{(\nu)} - \operatorname{diag} \mathbf{S}_{\mathbf{yy}}^{(\nu)} \right\} \operatorname{diag}^{-1} \mathbf{S}_{\mathbf{yy}}^{(\nu)}. \tag{4}$$

As in [3] we are initializing the demixing filter matrix $\underline{\mathbf{W}}^{(\nu)}$ for each frequency bin ν with $\underline{W}_{pp}^{(\nu)} = 1$ and $\underline{W}_{pq}^{(\nu)} = 0$, $p \neq q$. A similar update equation based on the stochastic gradient which includes additional approximations can be found in [5].

As discussed below, the noise can be decomposed in coherent and incoherent contributions. To examine the noise-robustness of the iterative algorithm (4) we can approximate (3) by a Taylor series as shown in [2], to obtain

$$\mathcal{J}^{(\nu)}(m) = \sum_{i=0}^{\infty} \beta(i,m) \left\{ 1 - \frac{\det \mathbf{S}_{\mathbf{yy}}^{(\nu)}(i)}{\prod_{p=1}^{P} S_{y_p y_p}^{(\nu)}(i)} \right\}, \tag{5}$$

where the term in brackets denotes the generalized coherence introduced in [4]. Now it can be seen that the cost function (5) becomes zero if and only if the *cross*-power spectral densities of the output signals $S_{y_p y_q}^{(\nu)}$, i.e., the off-diagonal elements of $\mathbf{S}_{\mathbf{yy}}^{(\nu)}$ are zero. Thus, the iterative algorithm (4) tries to minimize the coherence between the output channels $\gamma_{y_1 y_2}$ (see Fig. 2). To evaluate

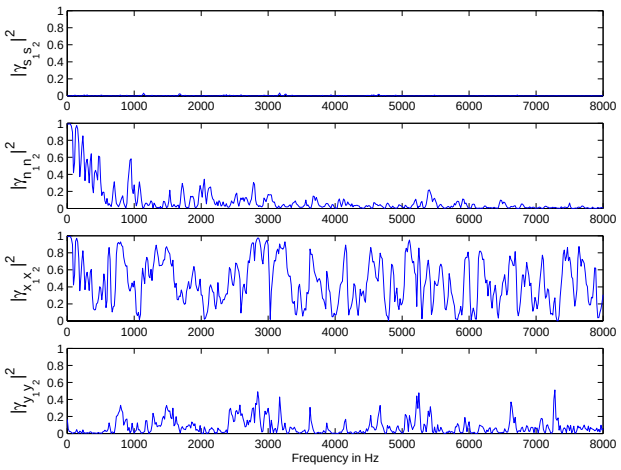

Fig. 2. Magnitude squared coherence function $|\gamma|^2$ of source signals, car noise, microphone and output signals.

This work was partly supported by the ANITA project funded by the European Commission under contract IST-2001-34327.

the influence of noise we can express $\mathbf{S}_{\mathbf{yy}}^{(\nu)}$ in terms of $\mathbf{S}_{\mathbf{xx}}^{(\nu)}$ and decompose this matrix according to (1) into its speech signal and noise components

$$\begin{aligned}\mathbf{S}_{\mathbf{yy}}^{(\nu)} &= \underline{\mathbf{W}}^{(\nu)^H}\mathbf{S}_{\mathbf{xx}}^{(\nu)}\underline{\mathbf{W}}^{(\nu)} \\ &= \underline{\mathbf{W}}^{(\nu)^H}\left(\mathbf{S}_{\mathbf{vv}}^{(\nu)}+\mathbf{S}_{\mathbf{nn}}^{(\nu)}\right)\underline{\mathbf{W}}^{(\nu)}. \qquad (6)\end{aligned}$$

For noise which is uncorrelated between the channels (e.g., sensor noise), $\mathbf{S}_{\mathbf{nn}}^{(\nu)}$ corresponds to a diagonal matrix whereas for correlated noise (e.g., diffuse noise) the matrix $\mathbf{S}_{\mathbf{nn}}^{(\nu)}$ is not sparse. Moreover, it can be shown [3] that (4) only affects the cross-power spectral densities $S_{y_py_q}^{(\nu)}$ without modifying the auto-power spectral densities $S_{y_py_p}^{(\nu)}$. Thus, with the initialization given above, the diagonal noise term of $\mathbf{S}_{\mathbf{nn}}^{(\nu)}$ of the initial block leads to a bias of $S_{y_py_p}^{(\nu)}$ which cannot be removed by (4). For a given SNR this bias will be more severe for uncorrelated noise whereas for correlated noise the initial bias is distributed among all elements of $\mathbf{S}_{\mathbf{yy}}^{(\nu)}$. The noise components appearing at the cross-power spectral densities $S_{y_py_q}^{(\nu)}$ will be minimized by (4) leading to an SNR gain at the outputs (see $\gamma_{n_1n_2}$ and $\gamma_{y_1y_2}$ in Fig. 2).

3. BIAS REMOVAL UTILIZING MINIMUM STATISTICS

To increase robustness of BSS algorithms against uncorrelated noise, bias removal techniques have been introduced, mainly consisting in the estimation and subtraction of the diagonal matrix $\mathbf{S}_{\mathbf{nn}}^{(\nu)}$ from $\mathbf{S}_{\mathbf{yy}}^{(\nu)}$ [1]. To deal with correlated and slowly time-varying noise, we propose to use the minimum statistics approach [6] for the estimation of the noise characteristics. This method is based on the observation that the power of a noisy speech signal frequently decays to the power of the background noise. Hence by tracking the minima we obtain the auto-power spectral density of the noise. However, not only the auto- but also the cross-power spectral densities of the noisy signal x_p and the background noise n_p are required. They are estimated and averaged recursively for each frequency bin whenever we detect a minimum (i.e. speech pause) of the noisy speech signals. Thus, for slowly time-varying noise statistics this method gives an accurate estimate of the noise spectral density matrix used for the bias removal. Note that for multiple active speakers this estimation problem is more difficult than for a single speaker due to less speech pauses. This technique was also used for beamforming in diffuse noise fields [7].

4. EXPERIMENTS

The data was recorded with a two-element microphone array with a spacing of 20cm. The array was mounted at the rear mirror of a Skoda Felicia car which was directed towards the driver. The reverberation time was $T_{60} = 50$ms. As source signals we used two speech signals from the TIMIT database which were convolved with the measured impulse responses of the car from the driver and codriver position. Car noise was recorded while driving through a suburban area at a speed of 60km/h. Moreover, uncorrelated white noise was used. Both noise types were additively mixed with speech at an SNR of -5 dB. To evaluate the performance, the signal-to-interference ratio (SIR) was used which is defined as the ratio of the signal power of the target *speech* signal to the signal power from the jammer *speech* signal. The SIR was averaged over both channels. The upper plot of Fig. 3 shows the influence of *uncorrelated noise* on the BSS algorithm. Compared to the noiseless case (dashed) the separation performance of the noisy case deteriorates (solid). This is due to the bias of the auto-power spectral

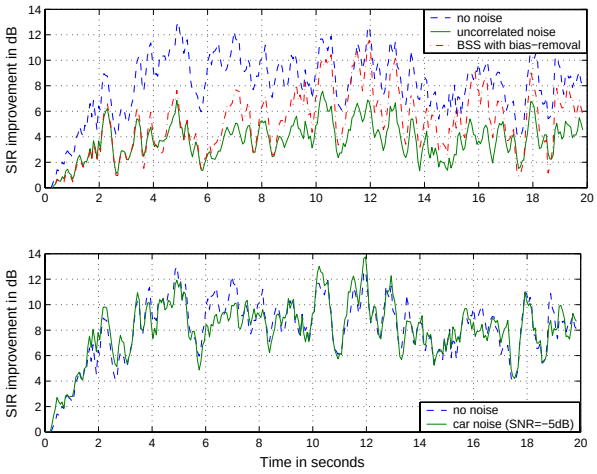

Fig. 3. Comparison of SIR improvement for different noisy data.

density matrices introduced by the noise term. By using a bias-removal technique utilizing minimum statistics the SIR can be improved (dash-dotted). It should be noted that proper regularization has to be ensured in (4), when using bias-removal techniques.

The lower plot of Fig. 3 shows that *diffuse car noise* does almost not affect the BSS algorithm and results in similar SIR performance as the noiseless case. This is due to the lower bias of $S_{y_py_p}^{(\nu)}$ as the noise terms are distributed among all elements of $\mathbf{S}_{\mathbf{yy}}^{(\nu)}$.

In addition to the SIR improvement, the BSS algorithm achieves also an SNR improvement of 5 dB and 6.5 dB for uncorrelated and diffuse car noise, respectively. This is due to the minimization of the cross-power spectral densities $S_{y_py_q}^{(\nu)}$ which contain also a noise term as shown in (6).

5. CONCLUSIONS

We presented a robust BSS algorithm which exhibits good performance for noisy signals. If required it can be complemented with a bias-removal technique. This efficient algorithm has been implemented in real-time on a regular laptop. (see `www.LNT.de/~aichner/bss_video.html`)

6. REFERENCES

[1] A. Hyvaerinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of a class of blind source separation algorithms for convolutive mixtures," in *Proc. ICA*, 2003, pp. 945–950.

[3] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: A unified treatment," in *Audio Signal Processing*, J. Benesty and Y. Huang, Eds., pp. 255–293. Kluwer Academic Publishers, Boston, April 2004.

[4] H. Gish and D. Cochran, "Generalized coherence," in *Proc. ICASSP*, April 1988, vol. 5, pp. 2745–2748.

[5] C.L. Fancourt and L. Parra, "The coherence function in blind source separation of convolutive mixtures of non-stationary signals," in *Proc. NNSP*, 2001.

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 5, pp. 504–512, 2001.

[7] R. Aichner et al., "Least-squares error beamforming using minimum statistics and multichannel frequency-domain adaptive filtering," in *Proc. IWAENC*, 2003, pp. 223–226.