Residual Cross-talk and Noise Suppression for Convolutive Blind Source Separation

Robert Aichner, Meray Zourub, Herbert Buchner, and Walter Kellermann
Multimedia Communications and Signal Processing, University Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
Email: {aichner, zourub, buchner, wk}@LNT.de

Introduction

Blind source separation (BSS) refers to the problem of recovering signals from several observed linear mixtures (e.g., [1]). In this paper we deal with the convolutive mixing case as encountered, e.g., in acoustic environments, and aim at finding a corresponding demixing system, whose output signals $y_q(n)$, $q = 1, \ldots, P$ are described by $y_q(n) = \sum_{p=1}^{P} \sum_{\kappa=0}^{L-1} w_{pq,\kappa} x_p(n-\kappa)$, and where $w_{pq,\kappa}$, $\kappa = 0, \ldots, L-1$ denote the current weights of the MIMO filter taps from the p-th sensor channel $x_p(n)$ to the q-th output channel (Fig. 1). We assume that the number of *active source signals* Q is less or equal to the number of microphones P. BSS algorithms are solely based on the assumption of mutual statistical independence of the different source signals. The separation is achieved by forcing the output signals y_q to be mutually statistically decoupled up to joint moments of a certain order. In

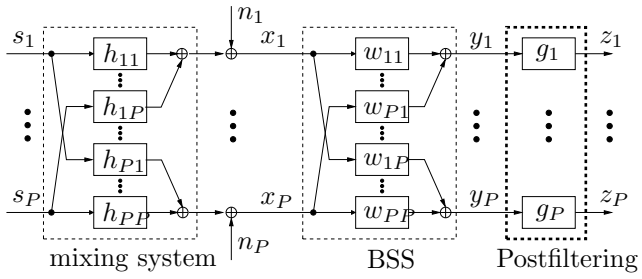

Figure 1: Noisy BSS model combined with post-filtering.

noisy scenarios additional background noise denoted by n_p is picked up by each sensor x_p (Fig. 1). In practice the noise fields often have spatially correlated as well as spatially white components. In general, convolutive BSS algorithms aim at separating spatially correlated point sources s_p, $p = 1, \ldots, P$ and thus, in noisy scenarios the spatially white component of the noise signal n_p cannot efficiently be suppressed. Moreover, due to the existence of noise, moving point sources, or long reverberation, the BSS algorithm is often unable to converge to the optimum solution. To achieve additional suppression of residual cross-talk stemming from the interfering point sources and of the background noise, it is possible, similarly as in adaptive beamforming or acoustic echo cancellation, to apply single-channel post-processing methods (Fig. 1), see, e.g., [2]. In this paper we will present a novel technique to estimate the power spectral densities of the residual cross-talk which are necessary for subsequently determining the proposed post-filters.

Residual Crosstalk and Noise Suppression

The frequency-domain representations $\underline{Y}_q^{(\nu)}(m)$ of the BSS output signals y_q, $q = 1, \ldots, P$ are obtained by a weighting by, e.g., a Hann window and a transformation using the discrete Fourier transform (DFT). $\underline{Y}_q^{(\nu)}(m)$ can be decomposed as

$$\underline{Y}_q^{(\nu)}(m) = \underline{Y}_{\mathrm{s},q}^{(\nu)}(m) + \underline{Y}_{\mathrm{c},q}^{(\nu)}(m) + \underline{Y}_{\mathrm{n},q}^{(\nu)}(m), \qquad (1)$$

where $\underline{Y}_{\mathrm{s},q}^{(\nu)}$ is the desired source component, $\underline{Y}_{\mathrm{c},q}^{(\nu)}$ denotes the residual cross-talk containing both, the remaining point sources that could not be suppressed by the BSS algorithm, and the *spatially correlated* background noise at the BSS outputs. The *spatially white* background noise components at the BSS outputs are denoted as $\underline{Y}_{\mathrm{n},q}^{(\nu)}$, $\nu = 0, \ldots, R$ is the index of the discrete frequency bin, and m denotes the block time index.

It is assumed that the desired signal, the residual cross-talk components, and the spatially white background noise in the q-th channel are all mutually uncorrelated. Then, the ν-th bin of a Wiener filter for the q-th channel and the m-th block, $\underline{G}_{\mathrm{c+n},q}^{(\nu)}$, which simultaneously suppresses residual cross-talk and background noise, is given by

$$\underline{G}_{\mathrm{c+n},q}^{(\nu)}(m) \approx \frac{\hat{E}\{|\underline{Y}_q^{(\nu)}(m)|^2\} - \hat{E}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\} - \hat{E}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)|^2\}}{\hat{E}\{|\underline{Y}_q^{(\nu)}(m)|^2\}}, \qquad (2)$$

where $\hat{E}\{|\underline{Y}_q^{(\nu)}|^2\}$, $\hat{E}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}|^2\}$, and $\hat{E}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}|^2\}$ are the power spectral density estimates.

To obtain a good estimate of $\underline{Y}_{\mathrm{c},q}^{(\nu)}$ needed for the post-filter we first need to set up an appropriate model. The cross-talk in the q-th channel stemming from point source interferers can be modeled as filtered versions of the other separated point sources $\underline{Y}_{\mathrm{s},i}^{(\nu)}$, which are estimated at all other output channels $i = 1, \ldots, P$ with $i \neq q$. It was shown in [2] that this is a valid model also for reverberant acoustic environments. However, the drawback is that the quantities $\underline{Y}_{\mathrm{s},i}^{(\nu)}$ are not observable in a practical system. Therefore, the desired signal component $\underline{Y}_{\mathrm{s},i}^{(\nu)}$ for the i-th channel is replaced by the observable BSS output signal of the i-th channel $\underline{\tilde{Y}}_{i,q}^{(\nu)}$, where the tilde and the subscript q express that the cross-talk component from the q-th point source (i.e., desired source s_q) to the i-th channel ($i = 1, \ldots, P$; $i \neq q$) is assumed to be zero. In practice this condition is fulfilled by determining time-frequency points where the desired source s_q is inactive

as described later in this section. Moreover, replacing $\underline{Y}_{\mathrm{s},i}^{(\nu)}$ by $\underline{\tilde{Y}}_{i,q}^{(\nu)}$ has the benefit that also the spatially correlated background noise is incorporated into the model. Thus, in the frequency domain the model for the residual cross-talk in the q-th channel is expressed as

$$\underline{Y}_{\mathrm{c},q}^{(\nu)}(m) = \sum_{i=1, i\neq q}^{P} \underline{\tilde{Y}}_{i,q}^{(\nu)}(m)\underline{B}_{i,q}^{(\nu)}(m) = \underline{\tilde{\mathbf{y}}}_{q}^{(\nu)^T}(m)\underline{\mathbf{b}}_{q}^{(\nu)}(m). \tag{3}$$

Here, $\underline{\tilde{\mathbf{y}}}_{q}^{(\nu)}$ is the column vector containing $\underline{\tilde{Y}}_{i,q}^{(\nu)}$ for $i = 1, \ldots, P$, $i \neq q$ and $\underline{\mathbf{b}}_{q}^{(\nu)}$ is the column vector containing the unknown filter weights $\underline{B}_{i,q}^{(\nu)}$ for $i = 1, \ldots, P$, $i \neq q$. The model is illustrated in Fig. 2 exemplarily for the first channel $q = 1$.

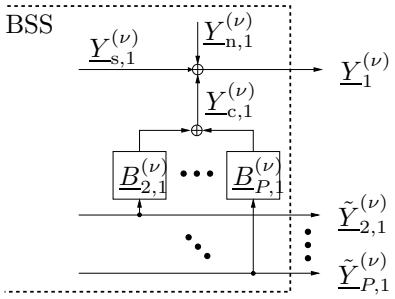

Figure 2: Model of the residual cross-talk component $\underline{Y}_{\mathrm{c},1}^{(\nu)}$ contained in the first BSS output channel $\underline{Y}_{1}^{(\nu)}$.

In [3] it has been shown that based on the model (3) the residual cross-talk for the q-th channel can be estimated in each frequency bin $\nu = 0, \ldots, R-1$ as

$$\hat{E}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}|^2\} = \underline{\mathbf{s}}_{\tilde{\mathbf{y}}_q Y_q}^{(\nu)^H} \left(\underline{\mathbf{S}}_{\tilde{\mathbf{y}}_q \tilde{\mathbf{y}}_q}^{(\nu)}(m)\right)^{-1} \underline{\mathbf{s}}_{\tilde{\mathbf{y}}_q Y_q}^{(\nu)}, \tag{4}$$

where the $P-1 \times P-1$ cross-power spectral density matrix $\underline{\mathbf{S}}_{\tilde{\mathbf{y}}_q \tilde{\mathbf{y}}_q}^{(\nu)}$ and the $P-1 \times 1$ cross-power spectral density matrix vector $\underline{\mathbf{s}}_{\tilde{\mathbf{y}}_q Y_q}^{(\nu)}$ are given as

$$\underline{\mathbf{S}}_{\tilde{\mathbf{y}}_q \tilde{\mathbf{y}}_q}^{(\nu)}(m) = \gamma \underline{\mathbf{S}}_{\tilde{\mathbf{y}}_q \tilde{\mathbf{y}}_q}^{(\nu)}(m-1) + (1-\gamma)\underline{\tilde{\mathbf{y}}}_{q}^{(\nu)^*}(m)\underline{\tilde{\mathbf{y}}}_{q}^{(\nu)^T}(m),$$

$$\underline{\mathbf{s}}_{\tilde{\mathbf{y}}_q Y_q}^{(\nu)}(m) = \gamma \underline{\mathbf{s}}_{\tilde{\mathbf{y}}_q Y_q}^{(\nu)}(m-1) + (1-\gamma)\underline{\tilde{\mathbf{y}}}_{q}^{(\nu)^*}(m)\underline{Y}_{q}^{(\nu)^T}(m).$$

To estimate the power spectral density of the background noise $\hat{E}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}|^2\}$ in the q-th BSS output channel the minimum statistics method [4] is used.

As pointed out before, the estimation of the residual cross-talk power spectral density in the q-th channel is only possible at time instants when the desired point source at the q-th channel is inactive. Speech signals can be assumed to be sufficiently sparse in the time-frequency domain so that even in environments with moderate reverberation, regions can be found where one or more sources are inactive. For a BSS system with two output channels $P = 2$, we can determine time instants where the desired source in the first or second channel is inactive by comparing the powers of both BSS output channels. E.g., if $\hat{E}\{|\underline{Y}_{1}^{(\nu)}|^2\} < \Upsilon \cdot \hat{E}\{|\underline{Y}_{2}^{(\nu)}|^2\}$, then it is assumed that the desired source in the first channel is inactive and thus, the residual cross-talk $\hat{E}\{|\underline{Y}_{\mathrm{c},1}^{(\nu)}|^2\}$ is estimated for the ν-th frequency bin. The parameter Υ with $0 < \Upsilon < 1$ is used to introduce a safety margin to prevent misdetections. Moreover, to reduce artifacts (e.g., musical noise) additional measures such as the adaptive oversubtraction factor have been implemented as explained in [3].

Experimental results

The experiments were conducted using an array of two omnidirectional sensors with spacing 20 cm ($T_{60} = 50$ms, array mounted to the interior mirror of a car). Recorded car noise with 0dB long-term SNR has been added to speech data convolved with measured impulse responses of a driver and co-driver. To evaluate the performance two measures have been used: The signal-to-interference ratio (SIR) defined as the ratio of the signal power of the desired signal to the signal power of the residual cross-talk stemming from point source interferers. Moreover, the segmental SNR defined as the ratio of the signal power of the desired signal to the signal power of the possibly diffuse background noise was calculated. To assess the desired signal distortion, the segmental signal-to-distortion ratio (SDR) between the desired signal at the input of the post-filter and the processed desired signal was used. For the BSS stage the algorithm and the parameters described in [5] have been used, and for the post-processing algorithm, $\gamma = 0.9$, $\Upsilon = 0.9$ and a block length of $N = 1024$ were chosen. In Fig. 3 it can be seen

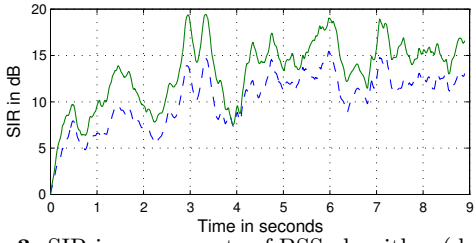

Figure 3: SIR improvements of BSS algorithm (dashed) and of BSS combined with post-processing (solid).

that the novel method (solid) improves the BSS performance (dashed) in terms of SIR. Also the background noise was further attenuated leading to a segmental SNR gain of 2.3 dB and the obtained SDR value of 17.4 dB shows that the quality of the desired signal is preserved.

Conclusions

We proposed a novel BSS post-processing scheme containing a robust estimation of the residual cross-talk power spectral densities and simultaneously addressing the suppression of background noise.

References

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

[2] R. Mukai, S. Araki, H. Sawada, and S. Makino. Removal of residual cross-talk components in blind source separation using LMS filters. In *Proc. NNSP*, pages 435–444, 2002.

[3] R. Aichner, M. Zourub, H. Buchner, and W. Kellermann. Post-processing for convolutive blind source separation. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2006.

[4] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech and Audio Proc.*, Vol. 9, No. 5, pages 504–512, 2001.

[5] R. Aichner, H. Buchner, F. Yan, and W. Kellermann. A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments. *Signal Processing*, 2006.