POST-PROCESSING FOR CONVOLUTIVE BLIND SOURCE SEPARATION

Robert Aichner, Meray Zourub, Herbert Buchner, and Walter Kellermann

Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
{aichner, zourub, buchner, wk}@LNT.de

ABSTRACT

Convolutive blind source separation (BSS) aims at separating point sources from mixtures picked up by several sensors. In real-world environments moving speakers, background noise and long reverberation are encountered which often degrade the performance of BSS algorithms. In such cases, the application of a post-filter can improve the output signal quality by suppression of residual cross-talk and of background noise. In this paper we propose a novel technique to estimate the necessary power spectral densities of the cross-talk components and present a robust system which allows to further suppress both, the remaining interference from point sources and the background noise. Experimental results show the benefit of this post-processing method in realistic environments.

1. INTRODUCTION

Blind source separation (BSS) refers to the problem of recovering signals from several observed linear mixtures [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 fundamental 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 noisy scenarios as depicted in Fig. 1 additional background noise denoted by n_p is picked up by each sensor x_p. In practice often the noise fields 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 signals 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 and thus only achieves partial separation suppressing the interfering sources by, e.g., 10 to 15dB. To achieve additional suppression of residual cross-talk stemming from the interfering point sources and of the background noise, it is possible, similarly to adaptive beamforming or acoustic echo cancellation, to apply single-channel post-processing methods (Fig. 1), see [2, 3, 4, 5]. In this paper we will first 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. In contrast to the methods in [2, 3, 4] we also investigate the simultaneous suppression of background noise by the postfilter.

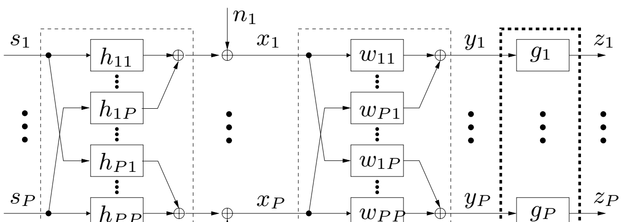

Fig. 1. Noisy BSS model combined with post-filtering.

In general, any BSS algorithm in the time or frequency domain can be combined with the proposed post-processing scheme. In [6] a general framework has been proposed covering both, frequency-domain, i.e., purely narrowband and time-domain, i.e., broadband BSS approaches. Furthermore, novel algorithms resulting in a combination of both approaches have been derived. In this paper we use in the experiments an efficient BSS algorithm originating from this framework, which shows good separation performance even in noisy environments and allows for real-time implementation [7]. For consistency we use the same notation as in [6, 7].

2. RESIDUAL CROSSTALK AND NOISE SUPPRESSION

The output signals $y_q(n)$, $q = 1, \ldots, P$ of the BSS algorithm can be decomposed as

$$y_q(n) = y_{\mathrm{s},q}(n) + y_{\mathrm{c},q}(n) + y_{\mathrm{n},q}(n), \tag{1}$$

where $y_{\mathrm{s},q}$ is the desired source component, $y_{\mathrm{c},q}$ 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 $y_{\mathrm{n},q}$. N samples are combined to an output signal block, which is weighted by, e.g., a Hann window and is then transformed by the discrete Fourier transform (DFT) of length $R \geq N$. Thus, we obtain a frequency-domain representation of the output signals given by

$$\underline{Y}_q^{(\nu)}(m) = \underline{Y}_{\mathrm{s},q}^{(\nu)}(m) + \underline{Y}_{\mathrm{c},q}^{(\nu)}(m) + \underline{Y}_{\mathrm{n},q}^{(\nu)}(m) \tag{2}$$

where $\nu = 0, \ldots, R$ is the index of the discrete frequency bin and m denotes the block time index. Frequency-domain quantities are

1-4244-0469-X/06/$20.00 ©2006 IEEE

denoted by underlining the respective variables analogously to [6, 7].

2.1. Wiener filter for suppression of residual cross-talk and background noise

In the following it is assumed that the desired signal component, the interfering signal components and the 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 components, is given by

$$\underline{G}_{\mathrm{c+n},q}^{(\nu)}(m) = \frac{E\{|\underline{Y}_{\mathrm{s},q}^{(\nu)}(m)|^2\}}{E\{|\underline{Y}_{q}^{(\nu)}(m)|^2\}}, \tag{3}$$

where $E\{\cdot\}$ denotes statistical expectation. To realize the Wiener filter in a practical system, the ensemble average has to be estimated and thus, it is usually replaced by a time average $\hat{E}\{\cdot\}$. Thereby, the Wiener filter is approximated by

$$\underline{G}_{\mathrm{c+n},q}^{(\nu)}(m) \approx \frac{\hat{E}\{|\underline{Y}_{q}^{(\nu)}(m)|^2\} - \hat{E}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}(m)|^2\} - \hat{E}\{|\underline{Y}_{\mathrm{n},q}^{(\nu)}(m)|^2\}}{\hat{E}\{|\underline{Y}_{q}^{(\nu)}(m)|^2\}}, \tag{4}$$

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 of the BSS output signal, residual cross-talk and background noise, respectively. The main difficulty is to obtain reliable estimates of the residual cross-talk components and the background noise. A novel method for this estimation process leading to high noise reduction with little signal distortion will be shown in the next sections. It should be noted that the estimates of residual cross-talk and background noise can also be used to implement other spectral weighting algorithms as described, e.g., in [8] instead of the Wiener filter (4).

2.2. Modelling the residual cross-talk components

In Sect. 1 we restricted our scenario to the case that the number of microphones equals the maximum number of simultaneously active point sources. Therefore, the BSS algorithm is able to provide an estimate of one separated point source at each output y_q. Due to movement of sources or long reverberation, the BSS algorithm might not converge fast enough to the optimum solution and thus some residual cross-talk from point source interferers remains in the BSS output. Additionally, spatially correlated background noise at the BSS outputs is contained in the residual cross-talk components $\underline{Y}_{\mathrm{c},q}^{(\nu)}$. To obtain a good estimate of $\underline{Y}_{\mathrm{c},q}^{(\nu)}$ needed for the post-filter in the q-th channel 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 $\tilde{\underline{Y}}_{i,q}^{(\nu)}$, 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. A detailed discussion of this procedure can be found in Sect. 2.4. Moreover, replacing $\underline{Y}_{\mathrm{s},i}^{(\nu)}$ by $\tilde{\underline{Y}}_{i,q}^{(\nu)}$ 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

$$\begin{aligned} \underline{Y}_{\mathrm{c},q}^{(\nu)}(m) &= \sum_{i=1, i\neq q}^{P} \tilde{\underline{Y}}_{i,q}^{(\nu)}(m) \underline{B}_{i,q}^{(\nu)}(m) && (5) \\ &= \underline{\tilde{\mathbf{y}}}_{q}^{(\nu)^T}(m) \underline{\mathbf{b}}_{q}^{(\nu)}(m). && (6) \end{aligned}$$

Here, $\underline{\tilde{\mathbf{y}}}_{q}^{(\nu)}$ is the column vector containing $\tilde{\underline{Y}}_{i,q}^{(\nu)}$ 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 $i = 1, \ldots, P$, $i \neq q$. In this paper, column vectors are denoted by bold lower case and matrices are written using bold upper case. The model given in (5) is illustrated in Fig. 2 exemplarily for the first channel $q = 1$. Note that in contrast [4, 5] where only spectral magnitudes are used, (5) uses complex spectra to cancel residual cross-talk. In the following Sect. 2.3 a method to estimate the residual cross-talk based on $\underline{\tilde{\mathbf{y}}}_{q}^{(\nu)}$ is derived. Subsequently, in Sect. 2.4 a procedure to determine $\underline{\tilde{\mathbf{y}}}_{q}^{(\nu)}$ is discussed.

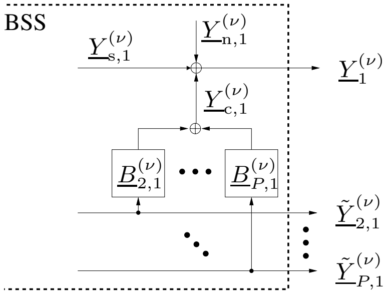
BSS
$\underline{Y}_{\mathrm{s},1}^{(\nu)}$
$\underline{Y}_{\mathrm{n},1}^{(\nu)}$
$\underline{Y}_{1}^{(\nu)}$
$\underline{Y}_{\mathrm{c},1}^{(\nu)}$
$\underline{B}_{2,1}^{(\nu)}$
$\underline{B}_{P,1}^{(\nu)}$
$\tilde{\underline{Y}}_{2,1}^{(\nu)}$
$\tilde{\underline{Y}}_{P,1}^{(\nu)}$

Fig. 2. Model of the residual cross-talk component $\underline{Y}_{\mathrm{c},q}^{(\nu)}$ contained in the q-th BSS output channel $\underline{Y}_{q}^{(\nu)}$ illustrated for the first channel, i.e., $q = 1$.

2.3. Estimation of residual cross-talk and background noise power spectral densities

After introducing the residual cross-talk model (5) we need to estimate the power spectral densities $|\underline{Y}_{\mathrm{c},q}^{(\nu)}|^2$ of residual cross-talk and $|\underline{Y}_{\mathrm{n},q}^{(\nu)}|^2$ of the background noise for evaluating (4). To obtain an estimation procedure based on observable quantities we first calculate the cross-power spectral density vector $\underline{\mathbf{s}}_{\tilde{\mathbf{y}}_q Y_{\mathrm{c},q}}^{(\nu)}$ between $\underline{\tilde{\mathbf{y}}}_{q}^{(\nu)}$ and $\underline{Y}_{\mathrm{c},q}^{(\nu)}$ in the residual cross-talk model depicted in Fig. 2:

$$\begin{aligned} \underline{\mathbf{s}}_{\tilde{\mathbf{y}}_q Y_{\mathrm{c},q}}^{(\nu)} &= \hat{E}\{\underline{\tilde{\mathbf{y}}}_{q}^{(\nu)^*}(m) \underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\} && (7) \\ &= \hat{E}\{\underline{\tilde{\mathbf{y}}}_{q}^{(\nu)^*}(m) \underline{\tilde{\mathbf{y}}}_{q}^{(\nu)^T}(m)\} \underline{\mathbf{b}}_{q}^{(\nu)}(m) && (8) \\ &=: \underline{\mathbf{S}}_{\tilde{\mathbf{y}}_q \tilde{\mathbf{y}}_q}^{(\nu)}(m) \underline{\mathbf{b}}_{q}^{(\nu)}(m), && (9) \end{aligned}$$

where in the step from (7) to (8) $\underline{\mathbf{b}}_{q}^{(\nu)}$ was assumed to be slowly time-varying. Using (6) the power spectral density estimate $\hat{E}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}|^2\}$ can be expressed as

$$\begin{aligned} \hat{E}\{|\underline{Y}_{\mathrm{c},q}^{(\nu)}|^2\} &= \hat{E}\{\underline{Y}_{\mathrm{c},q}^{(\nu)^H}(m) \underline{Y}_{\mathrm{c},q}^{(\nu)}(m)\} && (10) \\ &= \underline{\mathbf{b}}_{q}^{(\nu)^H}(m) \underline{\mathbf{S}}_{\tilde{\mathbf{y}}_q \tilde{\mathbf{y}}_q}^{(\nu)}(m) \underline{\mathbf{b}}_{q}^{(\nu)}(m). && (11) \end{aligned}$$

Solving (9) for $\underline{\mathbf{b}}_q^{(\nu)}$ and inserting it into (11) leads to

$$\hat{E}\{|\underline{Y}^{(\nu)}_{\mathrm{c},q}|^2\} = \underline{\mathbf{s}}^{(\nu)^H}_{\tilde{\mathbf{y}}_q Y_{\mathrm{c},q}} \left(\underline{\mathbf{S}}^{(\nu)}_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q}(m)\right)^{-1} \underline{\mathbf{s}}^{(\nu)}_{\tilde{\mathbf{y}}_q Y_{\mathrm{c},q}}. \qquad (12)$$

As $\underline{Y}^{(\nu)}_{\mathrm{c},q}$, $\underline{Y}^{(\nu)}_{\mathrm{s},q}$, and $\underline{Y}^{(\nu)}_{\mathrm{n},q}$ in Fig. 2 are assumed to be mutually uncorrelated, $\underline{\mathbf{s}}^{(\nu)}_{\tilde{\mathbf{y}}_q Y_{\mathrm{c},q}}$ can also be estimated as the cross-power spectral density $\underline{\mathbf{s}}^{(\nu)}_{\tilde{\mathbf{y}}_q Y_q}$ between $\tilde{\underline{\mathbf{y}}}_q^{(\nu)}$ and q-th output of the BSS system $\underline{Y}_q^{(\nu)}$ leading to the final estimation procedure:

$$\hat{E}\{|\underline{Y}^{(\nu)}_{\mathrm{c},q}|^2\} = \underline{\mathbf{s}}^{(\nu)^H}_{\tilde{\mathbf{y}}_q Y_q} \left(\underline{\mathbf{S}}^{(\nu)}_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q}(m)\right)^{-1} \underline{\mathbf{s}}^{(\nu)}_{\tilde{\mathbf{y}}_q Y_q}. \qquad (13)$$

One possible implementation for estimating this expectation is given by an exponentially weighted average $\hat{E}\{a(m)\} = (1-\gamma)\sum_i \gamma^{m-i} a(i)$, where $a(m)$ is the quantity to be averaged. The advantage is that this can also be formulated recursively leading to

$$\underline{\mathbf{S}}^{(\nu)}_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q}(m) = \gamma \underline{\mathbf{S}}^{(\nu)}_{\tilde{\mathbf{y}}_q\tilde{\mathbf{y}}_q}(m-1) + (1-\gamma)\tilde{\underline{\mathbf{y}}}_q^{(\nu)^*}(m)\tilde{\underline{\mathbf{y}}}_q^{(\nu)^T}(m), \quad (14)$$

$$\underline{\mathbf{s}}^{(\nu)}_{\tilde{\mathbf{y}}_q Y_q}(m) = \gamma \underline{\mathbf{s}}^{(\nu)}_{\tilde{\mathbf{y}}_q Y_q}(m-1) + (1-\gamma)\tilde{\underline{\mathbf{y}}}_q^{(\nu)^*}(m)\underline{Y}_q^{(\nu)^T}(m). \quad (15)$$

In summary, the power spectral density of the residual cross-talk for the q-th channel can be efficiently estimated in each frequency bin $\nu = 0, \ldots, R-1$ using (13) together with the recursive calculation of the $P-1 \times P-1$ cross-power spectral density matrix (14) and the $P-1 \times 1$ cross-power spectral density matrix vector (15). It should be noted that such an estimation technique has also been used to determine a post-filter for residual echo suppression in the context of acoustic echo cancellation (AEC) [9]. However, the method presented in [9] is different in two ways: Firstly, in contrast to BSS where several interfering point sources may be active, the AEC post-filter was derived for a single channel, i.e., the residual echo originates from only one point source and thus all quantities in (13) boil down to scalar values. Secondly, in the AEC problem a reference signal for the echo is available. In BSS however, $\tilde{\underline{\mathbf{y}}}_q^{(\nu)}$ is not immediately available as it can only be estimated if the desired source signal in the q-th channel is currently inactive. Strategies how to determine such time intervals are discussed in the next section.

To estimate the power spectral density of the background noise $\hat{E}\{|\underline{Y}^{(\nu)}_{\mathrm{n},q}|^2\}$ in the q-th BSS output channel the minimum statistics method [10] is used. 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 power spectral density of the noise. In [10] a recursive estimation of the noise power spectral density based on an optimal smoothing parameter is proposed and is applied in this paper to estimate $\hat{E}\{|\underline{Y}^{(\nu)}_{\mathrm{n},q}|^2\}$.

2.4. Strategies to determine $\tilde{\underline{\mathbf{y}}}_q^{(\nu)}$ and increase robustness of the post-processing

As pointed out in the previous sections 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 (e.g., experiments in Sect. 3: $T_{60} \approx 250$ms) regions can be found where one or more sources are inactive. This property is often exploited in underdetermined blind source separation where there are more simultaneously active sources than sensors (see, e.g., [11] for an examination of the sparseness of speech signals in reverberant environments).

if $\hat{E}\{|\underline{Y}_1^{(\nu)}|^2\} < \Upsilon \cdot \hat{E}\{|\underline{Y}_2^{(\nu)}|^2\}$
estimate residual cross-talk $\hat{E}\{|\underline{Y}^{(\nu)}_{\mathrm{c},1}|^2\}$ according to (13)
compute post-filters $\underline{G}^{(\nu)}_{\mathrm{c+n},1}$ and $\underline{G}^{(\nu)}_{\mathrm{n},2}$ according to (4) and (16)
elseif $\Upsilon \cdot \hat{E}\{|\underline{Y}_1^{(\nu)}|^2\} > \hat{E}\{|\underline{Y}_2^{(\nu)}|^2\}$
estimate residual cross-talk $\hat{E}\{|\underline{Y}^{(\nu)}_{\mathrm{c},2}|^2\}$ according to (13)
compute post-filters $\underline{G}^{(\nu)}_{\mathrm{n},1}$ and $\underline{G}^{(\nu)}_{\mathrm{c+n},2}$ according to (4) and (16)
else
compute post-filters $\underline{G}^{(\nu)}_{\mathrm{n},1}$ and $\underline{G}^{(\nu)}_{\mathrm{n},2}$ according to (16)

Table 1. Decision mechanism for $P = 2$ and resulting application of the postfilters.

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 $\hat{E}\{|\underline{Y}^{(\nu)}_{\mathrm{c},1}|^2\}$ 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. A similar decision mechanism has also been applied successfully in [2, 3]. An extension of this mechanism to $P > 2$ is to compare the power of the q-th channel $\hat{E}\{|\underline{Y}_q^{(\nu)}|^2\}$ to the maximum power of the remaining channels $\Upsilon\hat{E}\{|\underline{Y}_i^{(\nu)}|^2\}$, $i \neq q$. A careful selection of Υ for $P > 2$ is important but has not yet been thorougly investigated.

It can be seen that by using the proposed decision mechanism there will be several frequency bins in each block where an update of the residual cross-talk estimate is not possible for the q-th channel due to activity of the desired source. This means that for these frequency bins the residual cross-talk estimate from the previous block has to be used. As speech is a nonstationary process and therefore the statistics of the residual cross-talk are quickly time-varying, this would deteriorate the performance of the postfilter $\underline{G}^{(\nu)}_{\mathrm{c+n},q}$. On the other hand, background noise is often slowly time-varying so that the minimum statistics algorithm can provide good estimates of the noise power spectral density. Therefore, for those time instants where the estimate of residual cross-talk can not be updated, we propose to apply a postfilter $\underline{G}^{(\nu)}_{\mathrm{n},q}$ which aims only at suppression of the background noise

$$\underline{G}^{(\nu)}_{\mathrm{n},q}(m) = \frac{\hat{E}\{|\underline{Y}_q^{(\nu)}(m)|^2\} - \hat{E}\{|\underline{Y}^{(\nu)}_{\mathrm{n},q}(m)|^2\}}{\hat{E}\{|\underline{Y}_q^{(\nu)}(m)|^2\}}. \qquad (16)$$

In Table 1 the decision mechanism and the resulting application of the postfilters is outlined for a BSS system with $P = 2$ output channels together with the postfilters given as $\underline{G}^{(\nu)}_{\mathrm{c+n},q}$ and $\underline{G}^{(\nu)}_{\mathrm{n},q}$ for both channels $q = 1, 2$.

To reduce artifacts such as, e.g., musical noise, the postfilters (4) and (16) are calculated using an adaptive oversubtraction factor $\xi_q^{(\nu)}$ as proposed in [12]. Moreover negative gains of the postfilters are set to zero. Here, exemplarily the equation for the robust postfilter $\underline{G}^{(\nu)}_{\mathrm{n},q}$ is given

$$\underline{G}^{(\nu)}_{\mathrm{n},q}(m) = \frac{\max\left[\left(\hat{E}\{|\underline{Y}_q^{(\nu)}(m)|^2\} - \xi_q^{(\nu)}\hat{E}\{|\underline{Y}^{(\nu)}_{\mathrm{n},q}(m)|^2\}\right), 0\right]}{\hat{E}\{|\underline{Y}_q^{(\nu)}(m)|^2\}}. \qquad (17)$$

3. EXPERIMENTAL RESULTS

The experiments were conducted using an array of two omnidirectional sensors with spacing 20 cm and speech data convolved with measured impulse responses of (a) speakers in a real room with reverberation time $T_{60} = 250$ms at ±45° and 2m distance of the sources to the array and (b) impulse responses of a driver and co-driver in a car ($T_{60} = 50$ms, array mounted to the rear mirror). In the reverberant room scenario artificially generated diffuse babble noise with 10dB long-term signal-to-noise ratio (SNR) and in the car scenario recorded car noise with 0dB long-term SNR has been added. The sampling frequency was $f_s = 16$kHz. To evaluate the performance two measures have been used: The signal-to-interference ratio (SIR) which is defined as the ratio of the signal power of the desired signal to the signal power from 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 using a blocklength of 16ms. To assess the desired signal distortion, the Itakura distance [13] and the segmental signal-to-distortion ratio (SDR) with 16ms blocklength have been used. For the BSS algorithm the parameters described in the experimental section in [7] have been used, and for the post-processing algorithm, $\gamma = 0.9$, $\Upsilon = 0.9$ and a block length of $N = 1024$ was chosen.

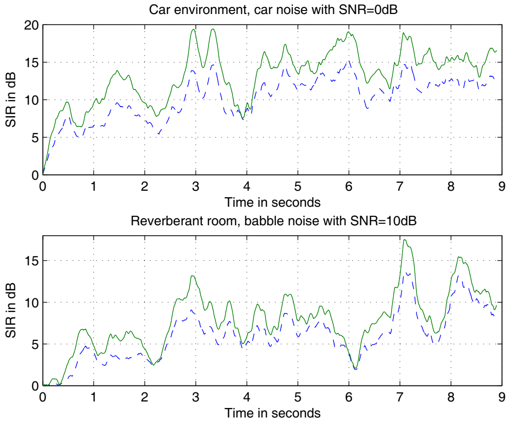

Fig. 3. SIR improvements of BSS algorithm (dashed) and of BSS combined with post-processing (solid).

In Fig. 3 it can be seen that for both scenarios the novel post-processing method (solid) improves the BSS performance (dashed) in terms of SIR by further suppressing the residual-cross talk. Moreover, in the car scenario also the background noise was further attenuated leading to a segmental SNR gain of 2.3 dB. The reduced absolute SIR of the BSS algorithm in the reverberant room is due to longer reverberation and especially due to the background babble noise which exhibits speech-like long-term spectrum. The post-filter increases the SIR (Fig. 3) even in such adverse environments but suppresses the babble noise only by 0.9 dB as the estimation of the background noise components with the minimum statistics algorithm fail due to the nonstationarity of the babble noise. To assess the speech quality, the Itakura distance and the SDR between the desired signal at the input of the post-filter and the processed desired signal was calculated and averaged over both output channels. The Itakura distance yields 0.10 for the car environment and 0.11 for the reverberant room. Using the SDR the values 17.4dB and 15.8dB are obtained, respectively. This shows that the quality of the desired signal is preserved, which is also verified by audio examples available in [14].

4. 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. Experimental results exemplarily given for a Wiener gain function show that this leads to superior performance in terms of SIR and SNR without introducing audible distortion. Moreover, an application of the proposed estimation method to other spectral gain functions and BSS algorithms is easily possible.

5. ACKNOWLEDGMENTS

We thank Sriram Srinivasan of the Royal Institute of Technology (KTH) for fruitful discussions.

6. 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 time-delayed spectral subtraction," in *Proc. ICASSP*, Orlando, FL, USA, May. 2002, vol. 2, pp. 1789–1792.

[3] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Removal of residual cross-talk components in blind source separation using lms filters," in *Proc. Neural Networks for Signal Processing (NNSP)*, Martigny, Switzerland, Sept. 2002, pp. 435–444.

[4] C. Choi, G.-J. Jang, Y. Lee, and S.R. Kim, "Adaptive cross-channel interference cancellation on blind source separation outputs," in *Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, Sept. 2004, pp. 857–864.

[5] J.-M. Valin, J. Rouat, and F. Michaud, "Microphone array post-filter for separation of simultaneous non-stationary sources," in *Proc. ICASSP*, May 2004, vol. 1, pp. 221–224.

[6] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 1, pp. 120–134, Jan. 2005.

[7] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "Real-time convolutive blind source separation based on a broadband approach," in *Int. Symp. Independent Component Analysis and Blind Signal Separation (ICA)*, Sept. 2004, pp. 833–840.

[8] R. Martin, "Statistical methods for the enhancement of noisy speech," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds., pp. 43–65. Springer, Berlin, 2005.

[9] V. Turbin, A. Gilloire, P. Scalart, and C. Beaugeant, "Using psychoacoustic criteria in acoustic echo cancellation algorithms," in *Int. Workshop Acoustic Echo Noise Control (IWAENC)*, London, UK, Sept. 1997, pp. 53–56.

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

[11] A. Blin, S. Araki, and S. Makino, "Blind source separation when speech signals outnumber sensors using a sparseness-mixing matrix estimation (SMME)," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, Sep. 2003, pp. 211–214.

[12] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *ICASSP*, April 1979, pp. 208–211.

[13] J.R. Deller, J.H.L. Hansen, and J.G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, New York, 2000.

[14] `http://www.LNT.de/~aichner/icassp06.html`