REDUCING REVERBERATION EFFECTS IN CONVOLUTIVE BLIND SOURCE SEPARATION

Radoslaw Mazur and Alfred Mertins

Signal Processing Group, Department of Physics
University of Oldenburg
26111 Oldenburg, Germany

ABSTRACT

In this paper, we propose a new method for reducing the reverberation effects in convolutive blind source separation which lead to reduced intelligibility of the separated sources (e.g., speech signals). The existing methods mainly try to maximize the separating performance without paying much attention to the linear distortion in the separated signals. We propose a modification to the existing algorithms that reduces the distortions introduced by the demixing filters. In particular we investigate the possibilities of modifying the frequency responses of the demixing filters to have no spectral peaks, which leads to near allpass character of the overall system. The good performance of the modified algorithm will be demonstrated on real-world data.

1. INTRODUCTION

In recent years, many algorithms for blind source separation of both instantaneous and convolutive mixtures have been proposed. While the algorithms often perform quite well on white sources and short mixing channels, which are usually encountered in data transmission, the separation of convolutive mixtures in acoustic environments with long reverberation times and nonwhite sources still remains challenging. A serious problem in acoustic settings is the additional linear distortion (reverberation) introduced by the demixing filters. One therefore aims at recovering the source signals with approximately the same power spectral densities as observed at the microphones. For example, the recurrent network setup introduced by Jutten and Herault [1] achieves this goal. Weinstein et al. [2] proposed to use postfilters which should be the inverse of the demixing filters. In practice the postfilters often drastically reduce the separating performance. Ikeda and Murata [3] proposed another setup which aims to recover the signals as they have been received by the microphones. They used the instantaneous case in the frequency domain. After separating, he applied the inverse of the separating matrix to each frequency bin of the separated signals, so that the scaling ambiguity was resolved. However, the same practical problems occur as with the method in [2]. A new approach was proposed by Huang et al. [4]. They proposed to first identify the mixing channels, and based on this information, build the demixing system. Again, system inversion can cause numerical problems, and the method works only when the mixing channels do not share common zeros. Under some strong conditions, the method in [5], which is also based on mixing-system identification, even allows for the recovery of the original source without distortion. However, the convergence of the method in practical applications could only be shown for extremely short channels, which renders the method inapplicable for acoustic scenarios.

In this paper we propose a modification to the update rule of convolutive blind source separation that reduces the reverberation while keeping the separating performance almost constant. We study this modification for the algorithm in [6], which uses the integrated frequency-domain Kullback-Leibler divergence as its objective function and minimizes it with respect to the time-domain coefficients of the demixing filters.

Notation. Convolution is denoted by $*$, and $(\cdot)^T$ is the transpose. $\boldsymbol{I}$ is the identity matrix. The operator diag$(\cdot)$ turns a vector into a diagonal matrix and vice versa. Time-domain matrices and vectors are set in boldface italic, and the frequency or z-domain correspondents are set in regular boldface letters. A matrix $\mathbf{W}(z)$ is the z-transform of a matrix sequence $\boldsymbol{W}(n)$, where $\mathbf{W}(z) = \sum_n \boldsymbol{W}(n) z^{-n}$.

2. PROBLEM STATEMENT

In real-world acoustic scenarios, the mixing channels can be modeled by FIR filters of length L, where L can be 2000 or more, depending on the reverberation time and sampling rate. In the following, we assume an equal number of sources and sensors. Given the source vector $\boldsymbol{s}(n) = [s_1(n), \ldots, s_N(n)]^T$, the vector of observation signals denoted by $\boldsymbol{x}(n) = [x_1(n), \ldots, x_N(n)]^T$ can be described as

$$\boldsymbol{x}(n) = \boldsymbol{H}(n) * \boldsymbol{s}(n) = \sum_{l=0}^{L-1} \boldsymbol{H}(l)\boldsymbol{s}(n-l) \qquad (1)$$

where $\boldsymbol{H}(n)$ is a sequence of $N \times N$ matrices containing the impulse responses of the mixing channels. For the separation, we use FIR filters of length M and obtain

$$\boldsymbol{y}(n) = \boldsymbol{W}(n) * \boldsymbol{x}(n) = \sum_{l=0}^{M-1} \boldsymbol{W}(l)\boldsymbol{x}(n-l) \qquad (2)$$

with $\boldsymbol{y}(n) = [y_1(n), \ldots, y_N(n)]^T$ being the vector of separated outputs and $\boldsymbol{W}(n)$ containing the unmixing coefficients. Fig. 1 shows the scenario for two sources and sensors.

This work has been supported by the German Science Foundation (DFG) under Grant No. ME 1170/1.

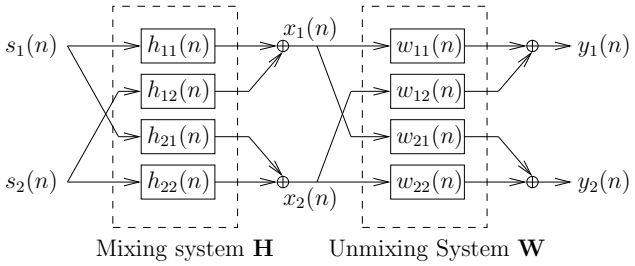

Figure 1: BSS model with two sources and sensors.

The overall system can be described by

$$\boldsymbol{y}(n) = \boldsymbol{W}(n) * \boldsymbol{H}(n) * \boldsymbol{s}(n) = \boldsymbol{G}(n) * \boldsymbol{s}(n), \quad (3)$$

which reduces to a multiplication in the z-domain:

$$\mathbf{Y}(z) = \mathbf{W}(z)\mathbf{H}(z)\mathbf{S}(z) = \mathbf{G}(z)\mathbf{S}(z). \quad (4)$$

The aim of BSS is to find $\mathbf{W}(z)$ from the observed process $\boldsymbol{x}(n)$ so that

$$\mathbf{G}(z) = \mathbf{PD}(z) \quad (5)$$

where $\mathbf{P}$ is a permutation matrix and $\mathbf{D}(z)$ an arbitrary diagonal matrix. These matrices represent the two ambiguities of BSS:

- There is no way to determine the order of the sources.
- The separated signals are scaled and filtered versions of the sources.

3. BLIND SEPARATION ALGORITHM

We here follow the method in [6], which uses the integrated Kullback-Leibler divergence in the frequency domain as the objective function and minimizes it with respect to the time-domain matrices $\boldsymbol{W}(n)$. This allows us to overcome the local permutation problem known from pure frequency-domain methods. The resulting update rule is given by

$$\boldsymbol{W}^{l+1}(n) = \boldsymbol{W}^{l}(n) - \mu \frac{\partial f(\boldsymbol{\mathcal{W}}^l)}{\partial \boldsymbol{W}^l(n)} \quad (6)$$

with $\boldsymbol{\mathcal{W}} = [\boldsymbol{W}(0), \boldsymbol{W}(1), \ldots, \boldsymbol{W}(M-1)]$, l being the iteration index and $f(\cdot)$ denoting the integrated Kullback-Leibler divergence. The expression for the gradient derived in [6] is given by

$$\frac{\partial f(\boldsymbol{\mathcal{W}}^l)}{\partial \boldsymbol{W}^l(n)} = \int_{-\pi}^{\pi} [\boldsymbol{I} - \mathbf{D}^{-1}(l,\omega)\mathbf{P}(l,\omega)]\mathbf{W}^l(e^{j\omega})e^{j\omega n}d\omega \quad (7)$$

where

$$\mathbf{D}(l,\omega) = \mathrm{diag}\left([\sigma_1^{r_1}(l,\omega), ..., \sigma_N^{r_N}(l,\omega)]^{\mathrm{T}}\right) \quad (8)$$

and

$$\mathbf{P}(l,\omega) = \mathbf{Y}^{r-1}(l,e^{j\omega})\mathbf{Y}^{\mathrm{H}}(l,e^{j\omega}), \quad (9)$$

$$\mathbf{Y}^{r-1}(l,e^{j\omega}) = \Big[\left|Y_1(l,e^{j\omega})\right|^{r_1-1} e^{j\theta(Y_1(l,e^{j\omega}))}, ... \\ ..., \left|Y_N(l,e^{j\omega})\right|^{r_N-1} e^{j\theta(Y_N(l,e^{j\omega}))}\Big]^{\mathrm{T}}$$

with $Y_i(e^{j\omega})$ being the short-time Fourier transforms of $y_i(n)$, $i = 1, 2, \ldots, N$ and

$$\sigma_p^{r_p}(l,\omega) = \beta\sigma_p^{r_p}(l,\omega) + (1-\beta)\left|y_p(l,e^{j\omega})\right|^{r_p}. \quad (10)$$

The parameter β with $0 < \beta < 1$ is a moving-average parameter, and r_p is the order of an assumed generalized Gaussian source model.

The method allows for long filters and is suitable for separating real-room recordings, but it suffers from linear distortions which are introduced by the demixing filters and will be discussed in the next section.

4. REDUCING REVERBERATION EFFECTS OF THE DEMIXING FILTERS

4.1 General considerations

The mixing filters $h_{ij}(n)$ and demixing filters $w_{ij}(n)$ introduce, in general, a linear distortion to the signals, because only filtered versions of the sources can be recovered with blind techniques. Depending on the objective function used to measure the independence of the outputs and on the power spectra of the sources, the spectral shaping of the outputs can be very strong, and certain frequencies can be emphasized significantly. Fig. 2 shows a typical frequency response of one of the demixing filters designed to separate two competing voices in a reverberant environment. Apparently, this filter has a number of significant spectral peaks that help to enhance the measure of independence and the signal-to-interference ratio (SIR) [7], but decrease the intelligibility of the separated voices. The perceived effect is an added reverberation through the unmixing system. In this particular example, the signals have been separated with an SIR of almost 20 dB, but they are nearly unintelligible. Therefore, although there is a filtering ambiguity, one is interested in recovering the sources with approximately the same power spectral densities as the ones observed at the sensors. As mentioned in the introduction, a number of techniques have been proposed, which often introduce their own problems such as the need to find stable inverses to MIMO systems with long impulse responses.

4.2 The postfilter method

Ikeda and Murata proposed in [3] to apply the inverse of the unmixing matrix to the individual separated sources, in order to recover the sources as observed at one of the microphones. Starting from a separated output $Y_i(e^{j\omega})$, filtered versions of this source are computed as the entries of the vector

$$\mathbf{v}_i(e^{j\omega}) = \mathbf{W}^{-1}(e^{j\omega}) \cdot [0 \ldots 0, Y_i(e^{j\omega}), 0 \ldots 0]^T. \quad (11)$$

The postfilter $Q_{ji}(e^{j\omega})$ to obtain the separated source from output i, but as observed at the jth microphone, is thus given by the jith element of the inverse unmixing matrix: $Q_{ji}(e^{j\omega}) = [\mathbf{W}^{-1}(e^{j\omega})]_{j,i}$.

The standard technique is to apply such postfilters after the separation algorithm has converged. In Section 6 we will present experimental results for this method. In addition, we study the behavior when the postfilter is applied during the blind unmixing-filter update.

4.3 The new proposed methods

In this paper, we combat the reverberation problem by demanding unmixing filters that nearly have an allpass character.[1] Thus, unlike the method in [3], we do not aim at recovering the sources with the same phase as observed at the microphones, but only try to obtain similar power spectra.

Demanding the individual filters to be allpass does, in general, not ensure that the overall system is allpass. However, we will show in the following that it ensures that the overall system has only spectral gaps and no sharp spectral peaks. To show this, we consider the postfilters $Q_{ij}(z)$ that would allow us to reconstruct the signals as they were recorded at the microphones, as proposed in [3]. Thus, the inverses of the postfilters, given by $U_{ij}(z) = 1/Q_{ij}(z)$, describe the transmission of individual sources from the microphones to the outputs y_1 and y_2 of the separation network. For a 2×2 system, these filters are given by

$$U_{ij}(z) := [W_{11}(z)W_{22}(z) - W_{21}(z)W_{12}(z)]/W_{ij}(z). \tag{12}$$

Assuming allpass demixing filters with $|W_{ij}(e^{j\omega})| = 1$ and evoking the triangle inequality, it is easy to see that $0 \leq |U_{ij}(e^{j\omega})| \leq 2$. This means that the demixing filters can have spectral gaps, but no large peaks. When we move from a 2×2 to an $M\times M$ system, the generalization of the above property reads $0 \leq |U_{ij}(e^{j\omega})| \leq (M!)$ where "!" stands for the factorial.

In the following, two concepts to achieve allpass-like unmixing filters will be proposed and investigated.

Method 1. In the first method, we amend the integrated KLD $f(\boldsymbol{\mathcal{W}})$ by a term $\varrho_1(\boldsymbol{\mathcal{W}})$ with

$$\varrho_1(\boldsymbol{\mathcal{W}}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=0}^{M-1}(|W_{ij}(k)|^2 - 1)^2, \tag{13}$$

where $W_{ij}(k)$ is the discrete Fourier transform (DFT) of $w_{ij}(n)$. The function $\varrho_1(\boldsymbol{\mathcal{W}})$ becomes zero when all filters $W_{ij}(z)$ are allpass. The modified update rule reads

$$\boldsymbol{W}^{l+1}(n) = \boldsymbol{W}^{l}(n) - \mu\frac{\partial f(\boldsymbol{\mathcal{W}}^l)}{\partial \boldsymbol{W}^l(n)} - \lambda\frac{\partial \varrho_1(\boldsymbol{\mathcal{W}}^l)}{\partial \boldsymbol{W}^l(n)} \tag{14}$$

Method 2. The second concept is to define an objective function in the logarithmized frequency domain in such a way that it is minimized when the unmixing filters are allpass. Here we choose the following:

$$\varrho_2(\boldsymbol{\mathcal{W}}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=0}^{M-1}[\widetilde{W}_{ij}(k)]^2 \tag{15}$$

with

$$\widetilde{W}_{ij}(k) = \log(|W_{ij}(k)|) \tag{16}$$

The gradient becomes $\frac{\partial \varrho_2(\boldsymbol{\mathcal{W}})}{\partial \widetilde{W}_{ij}(k)} = 2\widetilde{W}_{ij}(k)$, and we obtain the following update rule:

$$\widetilde{W}_{ij}^{l+1}(k) = \widetilde{W}_{ij}^{l}(k) - \epsilon\ \widetilde{W}_{ij}^{l}(k). \tag{17}$$

Hence the update rule in the linear frequency domain:

$$W_{ij}^{l+1}(k) = |W_{ij}^{l}(k)|^{\gamma} \cdot e^{j\angle W_{ij}^{l}(k)} \tag{18}$$

where $\gamma = 1 - \epsilon$. The overall procedure is as follows:

- Calculate the first gradient as in (7) and make the update according to (6):

$$\boldsymbol{W}^{tmp}(n) = \boldsymbol{W}^{l}(n) - \mu\frac{\partial f(\boldsymbol{\mathcal{W}})}{\partial \boldsymbol{W}^l(n)} \tag{19}$$

- Set $W_{ij}^{l}(k) = DFT\{w_{ij}^{tmp}(n)\}$
- Calculate $W_{ij}^{l+1}(k)$ as in (18)
- Set $w^{l+1}(n) = IDFT\{W_{ij}^{l+1}(k)\}$

The objective functions $\varrho_1(\cdot)$ and $\varrho_2(\cdot)$ of methods 1 and 2, although being both minimized when all filters are allpass, have slightly different properties, because they show different asymmetries around the ideal case where $|W_{ij}(k)| = 1$. The objective function $\varrho_1(\cdot)$ is sensitive to very large values of $|W_{ij}(k)|$, whereas $\varrho_2(\cdot)$ is sensitive to both very large and very small values of $|W_{ij}(k)|$.

5. PERFORMANCE MEASURES

5.1 Measurement of separating performance

When the original sources or at least the individual components of the mixtures are available then the separating performance can be measured by the signal-to-interference ratio (SIR) defined as [7]:

$$SIR_{y_i} = 10log_{10}\frac{E[(g_{ii}(n) * s_i(n))^2]}{E[(\sum_{j=1,j\neq i}^{N} g_{ij}(n) * s_j(n))^2]} \tag{20}$$

5.2 Measurement of distortion

As stated in the last section, the demixing filters $w_{ij}(n)$ introduce a distortion to the separated signals that results in reverberation and can drastically reduce the intelligibility of the separated signals although the separation was successful.

To quantify this type of distortion, the Spectral Flatness Measure (SFM) can be used [8]. The SFM is defined as

$$SFM = \frac{G_m}{A_m} \tag{21}$$

with G_m being the geometric and A_m the arithmetic mean of the power spectrum of a signal. To measure the

[1] True all-pass filters would have uniform frequency responses but would not be demixing filters, in general.

distortion introduced by a filter $w(n)$ we assume a white noise signal filtered with $w(n)$. The power spectrum of the output depends only on the frequency response $W(e^{j\omega})$. Therefore the SFM can be computed as:

$$SFM = \frac{\sqrt[N]{\prod_{k=0}^{N-1} |W(k)|^2}}{\frac{1}{N}\sum_{k=0}^{N-1} |W(k)|^2} \tag{22}$$

with $W(k)$, $k = 0, 1, \ldots, N-1$ being the DFT of $w(n)$.

The values of SFM range between 0 and 1. A value of 1 means that the filter is an allpass. Low values indicate high linear distortions introduced by the filter.

6. EXPERIMENTAL RESULTS

To test the modified update rules, simulations were performed on the ICA99 data [9]. This data set contains real room recordings with individual contributions of the sources to the microphones, so that the separating performance can be calculated using (20).

We used separation filters of length 1024 and a stepsize parameter of $\mu = 0.01$. The parameters of (7) were set to the values proposed in [6].

Table 1 shows the SIR measure of separation performance and the spectral flatness measure for the original algorithm (denoted as plain) and the two postfilter modifications discussed in Section 4.2. We see that for the plain method, the separation in terms of the SIR is very good, but the linear distortion introduced by the demixing filters is very high, as seen from the low SFM values. The frequency response of the filter $w_{11}(n)$, designed with the plain method is depicted in Fig. 2. It clearly shows a number of strong peaks. Applying the postfilter minimizes the linear distortions introduced by the filters, but it also drastically reduces the separation performance. Applying the postfilter correction during every iteration yields filters that perform better in terms of separation and distortion. This indicates that the algorithm converges to a different and better final solution when the postfilter-normalization step is carried out during the blind coefficient-update iteration.

Table 2 shows how the separation performance and the distortion for our proposed first algorithm based on the measure $\varrho_1(\cdot)$ for different values of λ. With $\lambda = 0$ we get the original plain setup. As the influence of λ grows, the separation performance decreases, but the achieved spectral flatness and the intelligibility increase. Hearing tests confirm this behavior. With the original rule, the separated signals are barely intelligible, while with increasing λ, the audible distortions vanish. The perceived effect is a reduction of the room size from a long tunnel to a small, but empty living room. For comparison, Fig. 3 shows the frequency response of an optimized filter. As one can see, the frequency response is much smoother than the one in Fig. 2.

Table 3 shows results for the second proposed method based on the measure $\varrho_2(\cdot)$ for different choices of γ. To allow for a better comparison of the different algorithms, Fig. 4 depicts the average SIR's over the average SFM. From this figure, one can clearly see that the method based on $\varrho_2(\cdot)$ performs better than the one based on $\varrho_1(\cdot)$. Compared to the plain output of the demixing system without shaping the spectra, both techniques significantly enhance the SFM and the intelligibility at a slight reduction of the SIR. Fig. 4 also shows that both proposed methods reduce the linear distortions of the individual unmixing filters to the level of the postfilter method and still show up to 7 dB better separating performance.

Fig. 5 depicts the power spectrum of an original signal, as observed at one of the microphones, and the corresponding spectra of the separated source, using the plain method and the one based on $\varrho_2(\cdot)$. As one can see, the spectrum of the separated signal using the new update rule is much closer to the original speech spectrum than the one for the plain method.

Table 1: Comparison of the separating performance of postfilter methods. "postfilter" stands for applying the postfilter once at the end of the iteration, and "postfilter+" means using it in every iteration.

$Method$	$Plain$	$Postfilter$	$Postfilter+$
SIR_1	12.20	4.37	5.84
SIR_2	19.84	6.87	7.17
SFM_{11}	0.19	0.81	0.90
SFM_{12}	0.19	0.38	0.56
SFM_{21}	0.27	0.54	0.61
SFM_{22}	0.33	0.81	0.90

Table 2: Comparison of the separating performance with update rule based on $\varrho_1(\cdot)$.

$\lambda \cdot 10^{-3}$	0	0.01	0.02	0.04	0.08	0.15
SIR_1	12.20	9.92	9.46	9.36	9.56	9.18
SIR_2	19.84	16.73	15.60	14.94	14.73	13.25
SFM_{11}	0.19	0.31	0.38	0.48	0.61	0.78
SFM_{12}	0.19	0.29	0.34	0.42	0.54	0.67
SFM_{21}	0.27	0.36	0.42	0.47	0.56	0.63
SFM_{22}	0.33	0.48	0.54	0.61	0.72	0.80

Table 3: Comparison of the separating performance with update rule based on $\varrho_2(\cdot)$.

γ	1	.9998	.9996	.9994	.9992	.9990
SIR_1	12.20	10.32	9.38	9.70	10.04	9.56
SIR_2	19.84	19.43	16.54	15.85	14.79	12.39
SFM_{11}	0.19	0.48	0.64	0.72	0.77	0.81
SFM_{12}	0.19	0.43	0.60	0.67	0.73	0.76
SFM_{21}	0.27	0.44	0.56	0.64	0.68	0.73
SFM_{22}	0.33	0.57	0.68	0.76	0.79	0.82

7. CONCLUSIONS

In this paper, we have proposed a new update rule for blind source separation in reverberant environments. The results obtained with real-world data clearly show that the new update rule generates demixing filters which have much smoother frequency responses than the

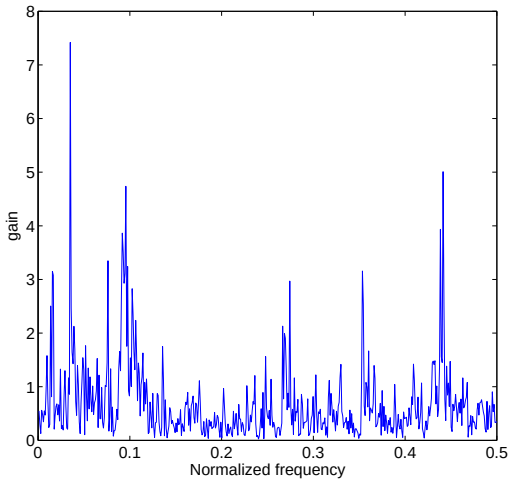

Figure 2: Magnitude of $W_{11}(e^{j\omega})$ obtained with original algorithm.

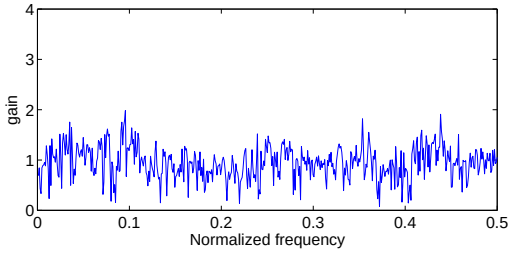

Figure 3: Magnitude of $W_{11}(e^{j\omega})$ obtained with the modified algorithm.

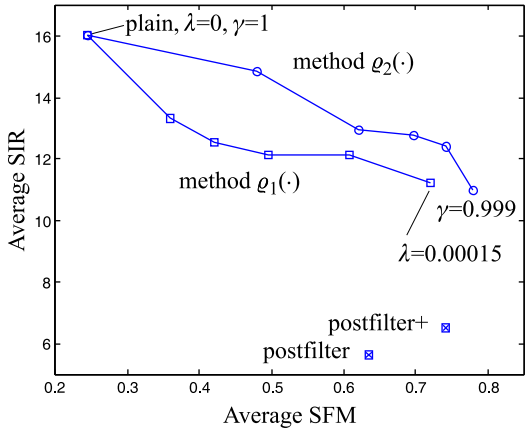

Figure 4: Average SIR versus average SFM for different separation methods and varying parameters. The free parameters are λ in case of $\varrho_1(\cdot)$ and γ in case of $\varrho_2(\cdot)$).

ones produced with a plain blind separation algorithm. The smoother frequency responses result in less reverberation in the separated signals, and, as a consequence, the intelligibility of speech is significantly enhanced with the new method. Although the proposed modifications have only been studied for the use with Kullback-Leibler divergence based blind source separation, they are also applicable to source separation using other techniques

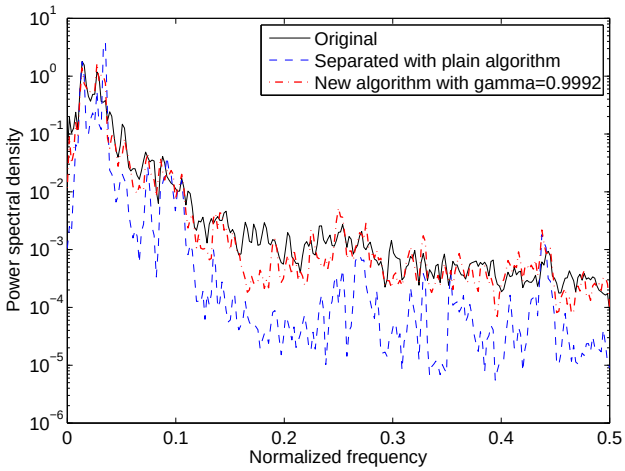

Figure 5: Comparison of the power spectra of original and separated signals.

such as joint diagonalization of second-order correlation matrices.

REFERENCES

[1] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, pp. 1–10, Feb 1991.

[2] E. Weinstein, M. Feder, and A.V. Oppenheim, "Multi-channel signal separation by decorrelation," in *IEEE Trans. Speech Audio Processing*, Apr 1993, pp. 405–413.

[3] Shiro Ikeda and Noburo Murata, "A method of blind separation based on temporal structure of signals.," in *ICONIP*, 1998, pp. 737–742.

[4] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," in *IEEE Trans. Speech Audio Processing*, Sept 2005, pp. 882–895.

[5] Y. Hua, S. An, and Y. Xiang, "Blind identification of FIR MIMO channels by decorrelating subchannels," *IEEE Transactions on Signal Processing*, vol. 51, no. 5, pp. 1143–1155, May 2003.

[6] T. Mei, J. Xi, F. Yin, A. Mertins, and J. F. Chicharo, "Blind source separation based on time-domain optimizations of a frequency-domain independence criterion," in *IEEE Trans. Speech Audio Processing*, in press.

[7] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proc. Int. Workshop Independent Component Analysis and Blind Signal Separation*, Aussois, France, Jan. 1999.

[8] James D. Johnston, "Transform coding of audio signals using perceptual noise criteria.," *IEEE Journal on Selected Areas in Communication*, vol. 6, no. 2, pp. 314–232, Feb. 1988.

[9] http://www2.ele.tue.nl/ica99/realworld.html