



ELSEVIER

Speech Communication 39 (2003) 79–95

**SPEECH**  
COMMUNICATION

www.elsevier.com/locate/specom

# Adaptive separation of acoustic sources for anechoic conditions: A constrained frequency domain approach

Jörn Anemüller <sup>\*,1</sup>, Birger Kollmeier

*Medizinische Physik, Universität Oldenburg, 26111 Oldenburg, Germany*

---

## Abstract

Blind source separation represents a signal processing technique with a large potential for noise reduction. However, its application in modern digital hearing aids poses high demands with respect to computational efficiency and speed of adaptation towards the desired solution. In this paper, an algorithm is presented which fulfills these goals under the idealized assumption that the superposition of sources in rooms can be approximated as a superposition under anechoic conditions. Specifically, attenuation, the signals' finite propagation speed, and diffuse noise are accounted for, whereas reflections and reverberation are considered as negligible effects. This approximation is referred to as the 'free field' assumption. Starting from a general blind source separation algorithm for Fourier transformed speech signals, the free field assumption is incorporated into the framework, yielding a simple, fast and adaptive algorithm that is able to track moving sources. Implementation details are given which were found to be indispensable for fast and robust signal separation. Performance is evaluated both by simulations and experimentally, including separation of a moving and a fixed speaker in a recorded real anechoic environment. The potential benefits and shortcomings of this algorithm are discussed with regard to its inclusion into the signal processing framework of digital hearing aids for real reverberant acoustic situations.

© 2002 Elsevier Science B.V. All rights reserved.

*Keywords:* Blind source separation; Independent component analysis; Noise reduction; Speech signal processing; Hearing aids

---

## 1. Introduction

The need to separate some sound sources from others is ubiquitous in acoustic signal processing. A typical example is the field of signal processing for the hearing impaired, where speech intelligibility

needs to be enhanced in situations with multiple simultaneous speakers or with speech embedded in a background of noise. Similar problems are encountered in the field of automatic speech recognition where recognition rates still drastically degrade in the presence of interfering sources.

Blind source separation (BSS) and the related field of independent component analysis (Jutten and Héroult, 1991; for an introduction refer to, e.g., Hyvärinen et al., 2001) represent a relatively novel approach to this problem which has gained some attention over the past years. In contrast to other noise reduction schemes, BSS techniques aim

---

\* Corresponding author.

E-mail address: [jorn@salk.edu](mailto:jorn@salk.edu) (J. Anemüller).

<sup>1</sup> Present address: Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, P.O. Box 85800, San Diego, CA 92186-5800, USA.

at incorporating as little prior knowledge as possible into the algorithms, hence the term ‘blind’. The key assumptions made incorporate basic knowledge about the (second-order or higher-order) statistics of the different sources and about the principles of the mixing process by which the sound source signals are superimposed to form the recorded microphone signals. However, explicit knowledge about, e.g., typical source or noise spectra, or spatial locations of microphones or sources are *not* made which distinguishes BSS from such techniques as beam-forming, directional filtering and spectral subtraction.

The lack of a priori knowledge opens a great potential of BSS techniques, with some remarkable results for separating speech from interfering sounds. However, the generality of the assumed demixing filters also results in a large number of free parameters which need to be determined to achieve separation, and in the related problem of finding the optimal parameters fast, with modest computational requirements, and adaptively to compensate for changes in the acoustic environment. Therefore, the general problem of separating sources that have been mixed in real rooms with realistic reverberation is still an active area of research.

Recently proposed algorithms for convolutively mixed sources that have been shown to perform well with real-room sound recordings include Lee et al. (1998), Sahlin and Broman (1998), Murata et al. (2001) and Anemüller et al. (2000). In particular, the algorithm of Parra and Spence (2000a) has gained attention, since the algorithm performs successful separation in some difficult acoustic situations. An adaptive version of this algorithm has been presented by the same authors (Parra and Spence, 2000b), showing good separation after as little as 1 s of signal time and reaching its optimum separation after about 6 s time. However, evaluation of the algorithm was done for spatially fixed sources, only.

One area of application for BSS algorithms is automatic speech recognition, results on which have been reported by several authors (e.g., Anemüller et al., 2000; van der Kouwe et al., 2001). This field appears to be promising for preprocessing by BSS algorithms since the acoustic en-

vironment is relatively stationary, the delay due to preprocessing is not problematic, and today’s desktop computers offer fast computation.

In contrast, the field of signal processing for digital hearing aids poses much stronger constraints on algorithms. Here, the acoustic environment can change rapidly due to head turns of the subject, the processing delay should be on the order of only few tens of milliseconds, and the computational cost of algorithms should be modest. Therefore, potential BSS algorithms for hearing aids should be fast, simple and adaptive. It might not be of greatest importance to aim at the optimal solution in terms of quality of separation, but to simplify the problem at hand by introducing additional constraints and assumptions, hence making the algorithms ‘semi-blind’.

Following this idea, the approach presented in this paper is based on the assumption that the most prominent effects induced by real rooms’ transfer functions are attenuation and delay of the propagating sounds. This ‘free field’ assumption describes sound propagation in an ideal anechoic chamber. Note, however, that it is approximately met in real rooms if the direct sound is much larger than any acoustic reflections from nearby surfaces such as, e.g., the walls, the floor or the ceiling of the respective room. This is primarily the case for a small distance between sound sources and microphones and for a room with a short reverberation time, respectively. Conversely, the assumption is less fulfilled with an increasing reverberation time and, in addition, the larger the distance between the sound sources and the microphones gets in comparison to the distance to any reflecting surfaces. The diffuse, spatially uncorrelated ‘tail’ of the reverberation process, however, does not limit the performance in the same way as the early reflections referred to above, because an uncorrelated noise at both recording microphones does not interfere with the assumptions under which the algorithm still operates.

It should also be noted that BSS algorithms for delayed and attenuated sources have been proposed previously in the literature. Platt and Faggin (1992) report results on an adaptive time-domain algorithm that achieves separation after 2.5 s signal time for digitally delayed and mixed signals.

Torkkola (1996) proposes a time-domain algorithm which adapts from 15 ms long signal blocks and achieves separation after 1.5–3 s. The algorithm is also evaluated using digitally mixed signals, only, and local minima of the proposed algorithm are found. Algorithms for delayed sources have also been investigated by Emile and Comon (1998) and by Yeredor (2001).

In contrast, the algorithm presented in this paper is based on a frequency domain approach to the BSS problem, that could in principle be used to separate sources that have been mixed by an arbitrary convolution operation (including reverberation). By incorporating the free field constraint into this framework, an adaptive algorithm is derived that separates sources within  $\approx 250$  ms of signal time and is easily implemented in real-time. Due to its adaptive nature, separation of mixtures of moving speakers in anechoic environment is also possible. Since the algorithm works entirely in the frequency domain, it is particularly well suited for incorporation into the filterbank-based noise reduction schemes of modern hearing aids.

The outline of the present paper is as follows. In Section 2 the unconstrained and constrained acoustic mixing and the corresponding demixing models are introduced. Based on the maximum likelihood principle, a BSS algorithm for Fourier transformed speech signals is derived in Section 3. Section 4 is devoted to the incorporation of the free field constraint into the algorithm. Implementation details are given in Section 5, and evaluation is performed in Section 6.

Throughout the paper, vectors and matrices are denoted by bold font; time-domain signals are denoted by, e.g.,  $x(t)$  and the corresponding frequency domain signals by  $x(T, f)$ ; the imaginary unit  $\sqrt{-1}$  is denoted as  $i$ . Transposition is denoted by  $x^T$ , complex conjugation by  $x^*$ , transposition and complex conjugation by  $x^H$ .

## 2. Acoustic mixing and demixing

Mixing of sound sources in air is linear and involves finite propagation speed and reverberation. The signal component originating from source  $s_j(t)$ ,  $j = 1, \dots, N$ , and recorded by micro-

phone  $i$ ,  $i = 1, \dots, N$ , is therefore obtained as the convolution of  $s_j(t)$  with the room's impulse response  $a_{ij}(t)$  from the place of the source to the place of the microphone. The microphone signals  $x_i(t)$  stemming from simultaneously active sources are composed as the sum over the individual source components, together with some small measurement noise  $n_i(t)$ ,

$$x_i(t) = \sum_j \int dt' a_{ij}(t') s_j(t - t') + n_i(t). \quad (1)$$

In the free field, sound propagating from source to microphone is attenuated by a gain factor  $a_{ij}$  and delayed by a time  $\tau_{ij}$ . The corresponding impulse response simplifies to  $a_{ij}(t) = a_{ij} \delta(t - \tau_{ij})$ , where  $\delta(t)$  denotes the Dirac delta function. Therefore, the free field mixing system is

$$x_i(t) = \sum_j a_{ij} s_j(t - \tau_{ij}) + n_i(t). \quad (2)$$

If no prior knowledge is assumed to be known about the sources or the mixing system, an arbitrary gain factor  $\tilde{a}_j$  and time delay  $\tilde{\tau}_j$  can be interchanged between each source and the corresponding column of the mixing system  $a_{ij}(t)$  without altering the microphone signals. Specifically, setting

$$a'_{ij}(t) = \frac{a_{ij}}{\tilde{a}_j} \delta(t - \tau_{ij} + \tilde{\tau}_j), \quad (3)$$

$$s'_j(t) = \tilde{a}_j s_j(t + \tilde{\tau}_j), \quad (4)$$

leaves the mixed signals invariant. Furthermore, any permutation  $\pi(j)$  of the sources  $s_j(t)$  and of the corresponding columns of  $a_{ij}(t)$  leaves the mixed signals unchanged. The corresponding rescaling- and permutation-ambiguities for linear, memory-less mixtures of sources are well-known in the field of BSS (Tong et al., 1991).

Since the absolute gain factors and propagation times from the sources to the microphones are in principle unidentifiable, we are only concerned with the level- and time *differences* between the source components received at different microphones and normalize the diagonal elements of  $a_{ij}(t)$  to unity. The corresponding mixing system for the situation of two sources recorded by two microphones in the free field is therefore

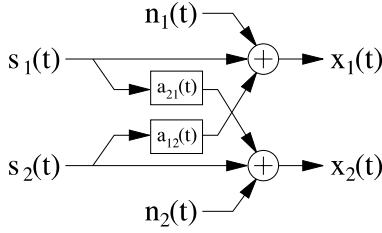


Fig. 1. The mixing system assumed for the current approach.

$$\begin{aligned} x_1(t) &= s_1(t) + a_{12}s_2(t - \tau_{12}) + n_1(t), \\ x_2(t) &= s_2(t) + a_{21}s_1(t - \tau_{21}) + n_2(t), \end{aligned} \quad (5)$$

which is illustrated in Fig. 1.

### 2.1. Frequency domain formulation

The approach pursued in the present paper is to separate the sources in the frequency domain. To this end, spectrograms are computed from the time domain signals using the windowed short time Fourier transformation (windowed STFT). The spectrogram  $x_j(T, f)$  corresponding to signal  $x_j(t)$  is defined as

$$x_i(T, f) = \sum_{t=0}^{2K-1} x_i(T+t)h(t)e^{-i\pi ft/K}. \quad (6)$$

Indices  $t = 0, 1, \dots$  and  $f = 1, \dots, K$  denote time and frequency, respectively. The short-time spectra are computed at times  $T = 0, \Delta T, 2\Delta T, \dots$  using the window function  $h(t)$ , e.g., the hanning window. Similarly,  $a_{ij}(f)$ ,  $s_j(T, f)$  and  $n_i(T, f)$  denote the spectrograms of  $a_{ij}(t)$ ,  $s_j(t)$  and  $n_i(t)$ , respectively. Note that for the sake of deriving the algorithms' update equations,  $a_{ij}(t)$  is assumed to be short and stationary over time, and therefore its STFT does not depend on time  $T$ . The algorithm's ability to operate in non-stationary environments is implemented by using the adaptive optimization scheme described in Section 4.1.

In the frequency domain formulation, the convolution in the acoustic mixing model (1) factorizes, provided the window-length is larger than the length of the impulse responses  $a_{ij}(t)$ , yielding the mixing model

$$x_i(T, f) = \sum_j a_{ij}(f)s_j(T, f) + n_i(T, f). \quad (7)$$

Under the free field assumption, model (7) is a good approximation to the acoustic mixing, and the transfer functions  $a_{ij}(f)$  are computed from the corresponding level- and time differences (2) as

$$a_{ij}(f) = a_{ij}e^{-i2\pi f\tau_{ij}}. \quad (8)$$

In the remainder of the paper, the focus is on the case of two microphones and two sources. However, the discussion directly carries over to the  $N \times N$ -case. The frequency domain formulation of the mixing system (5) therefore is

$$\begin{pmatrix} x_1(T, f) \\ x_2(T, f) \end{pmatrix} = \begin{pmatrix} 1 & a_{12}(f) \\ a_{21}(f) & 1 \end{pmatrix} \begin{pmatrix} s_1(T, f) \\ s_2(T, f) \end{pmatrix} + \begin{pmatrix} n_1(T, f) \\ n_2(T, f) \end{pmatrix}, \quad (9)$$

and the unmixed signals' spectrograms  $\hat{u}_i(T, f)$  are obtained as

$$\hat{u}_i(T, f) = \sum_j \hat{w}_{ij}(f)x_j(T, f), \quad (10)$$

where  $\hat{w}_{ij}(f)$  denotes the separation filters. Without noise, the perfect solution for the parameters  $\hat{w}_{ij}(f)$  would be

$$\begin{pmatrix} \hat{w}_{11}(f) & \hat{w}_{12}(f) \\ \hat{w}_{21}(f) & \hat{w}_{22}(f) \end{pmatrix} = c(f) \begin{pmatrix} 1 & -a_{12}(f) \\ -a_{21}(f) & 1 \end{pmatrix}, \quad (11)$$

$$c(f) = (1 - a_{12}(f)a_{21}(f))^{-1},$$

which recovers the first source as recorded at the first microphone if the second source was silent and similarly the second source as recorded at the second microphone.

In the presence of noise  $n_i(T, f)$ , however, the complex factor  $c(f)$  results in the amplification of the noise energy at harmonic frequencies since the magnitudes  $|a_{12}(f)|$  and  $|a_{21}(f)|$  of the off-diagonal elements are in practice close to unity (if a distance between microphones of 35 cm or less is assumed, which is reasonable for hearing aids; cf. to Section 6 for experimentally obtained parameter values). Therefore, it is advisable to set  $\hat{w}_{11}(f) = \hat{w}_{22}(f) = 1$  resulting in the separating system

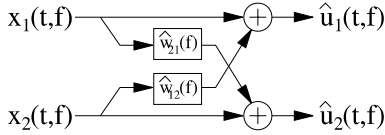


Fig. 2. The separating system assumed to unmix the signals from the mixing system depicted in Fig. 1.

$$\begin{pmatrix} \hat{u}_1(T, f) \\ \hat{u}_2(T, f) \end{pmatrix} = \begin{pmatrix} 1 & \hat{w}_{12}(f) \\ \hat{w}_{21}(f) & 1 \end{pmatrix} \begin{pmatrix} x_1(T, f) \\ x_2(T, f) \end{pmatrix}, \quad (12)$$

which is depicted in Fig. 2. Note that after this normalization the filters  $\hat{w}_{ij}(f)$  do not correspond to the inverse of  $a_{ij}(f)$  and, hence, filtered versions of the original sources will be recovered. However, the noise energy gets limited to

$$E\{|\hat{u}_i(T, f) - s_i(T, f)|\} \approx E\{|n_1(T, f)|^2\} + E\{|n_2(T, f)|^2\}, \quad (13)$$

where the level differences between the microphones,  $|a_{12}(f)|$  and  $|a_{21}(f)|$ , have been approximated by unity.

### 3. Blind source separation algorithm for Fourier transformed speech

The superposition of sources in the frequency domain (7) has the form of a matrix vector product in each frequency channel  $f$ . In contrast to the time domain representation (5), which contains coupling across different time-points, Eq. (7) can be regarded as a set of  $K$  decoupled instantaneous BSS problems, albeit with complex valued variables. Several algorithms (e.g., Bell and Sejnowski, 1995; Cardoso and Laheld, 1996; Pham et al., 1992) have been proposed in the literature to solve the instantaneous BSS problem. However, most are concerned with real valued variables, whereas few consider complex valued signals (Back and Tsoi, 1994; Bingham and Hyvärinen, 2000; Cardoso and Laheld, 1996; Fiori, 2000; Moreau and Macchi, 1994; Smaragdis, 1998; cf. also to the discussion at the end of this section).

In this section, the standard method of maximum likelihood estimation is applied to the

problem of separating Fourier transformed speech signals to obtain an adaptation algorithm for the complex valued separating parameters  $\hat{w}_{ij}(f)$ . It is noted that the derivation given in this section applies to the general frequency domain mixing model (7). The combination of this section’s learning rule with the prior knowledge about the free field constraint (8) for the mixing model is given in Section 4.

#### 3.1. Maximum likelihood estimation

Speech signals, both in the time and in the frequency domain, exhibit a non-Gaussian histogram with positive kurtosis, i.e., small signal amplitudes occur with higher probability than for a Gaussian distribution of equal variance, and also large amplitudes tend to be more likely than for a Gaussian (e.g., Brehm and Stammer, 1987; Zelinski and Noll, 1977, and reference therein). Intermediate amplitudes, in contrast, occur with lower probability than it would be the case for a Gaussian distribution.

This property allows to distinguish between a speech signal originating from a single source and a mixture of speech signals from multiple independent sources, since the mixture’s histogram is more Gaussian, due to the central limit theorem. A large class of algorithms for blind source separation, those which are based on higher-order statistics (e.g. Comon, 1994), exploit this principle by aiming to reconstruct unmixed signals whose histogram resembles the non-Gaussian histogram of the original source signals.

The maximum likelihood principle (e.g. Bishop, 1995) represents a general statistical tool for the estimation of optimal parameter values. As such, it can be employed to derive algorithms for estimating the separation parameters in BSS tasks, as has first been shown by Pham et al. (1992) for the separation of real-valued time-domain signals. To give a brief outline, under the maximum likelihood approach it is aimed to find parameters of the mixing system  $\mathcal{A}$  which maximize the probability  $\mathcal{P}(\mathbf{x}|\mathcal{A})$  that measured data  $\mathbf{x}$  has been generated by this particular  $\mathcal{A}$ . Assuming that the sources  $s(T, f)$  can be recovered using the demixing system  $\mathcal{W}(f) = \mathcal{A}^{-1}(f)$ , it can be shown (MacKay, 1996)

that for a single observation  $\mathbf{x}(T, f)$  the log-likelihood  $L(\mathbf{W}(f), \mathbf{x}(T, f))$  of matrix  $\mathbf{W}(f)$  being the desired unmixing system is

$$\begin{aligned} L(\mathbf{W}(f), \mathbf{x}(T, f)) &= \log \mathcal{P}(\mathbf{x}(T, f) | \mathbf{W}(f)) \\ &= \log \det(\mathbf{W}(f)) \\ &\quad + \log \mathcal{P}(\mathbf{W}(f) | \mathbf{x}(T, f)). \end{aligned} \quad (14)$$

The separating system  $\mathbf{W}(f)$  is obtained by maximizing the expectation of  $L(\mathbf{W}(f), \mathbf{x}(T, f))$  with respect to  $\mathbf{W}(f)$ ,

$$\mathbf{W}(f) = \operatorname{argmax}_{\mathbf{W}(f)} E\{L(\mathbf{W}(f), \mathbf{x}(T, f))\}. \quad (15)$$

### 3.2. Model density for $\mathcal{P}(s(T, f))$

In order to use the log-likelihood (14) to build an optimization algorithm based on it, the sources probability density function (pdf)  $\mathcal{P}(\mathbf{W}(f) | \mathbf{x}(T, f)) = \mathcal{P}(s(T, f))$  needs to be modeled. Due to the sources' mutual independence it follows that their joint pdf  $\mathcal{P}(s(T, f))$  factorizes into the product of the individual source pdfs,  $\mathcal{P}(s(T, f)) = \prod_j \mathcal{P}(s_j)$ , so that a model for  $\mathcal{P}(s_j(T, f))$  is needed. Since the Fourier transformed speech signal  $s_j(T, f)$  is complex, the model for  $\mathcal{P}(s_j(T, f))$  must be a two-dimensional pdf, taking into account real and imaginary part of  $s_j(T, f)$ .

First, it is noted that the phase  $\arg(s_j(T, f))$  depends on two quantities: the speech signal  $s_j(t)$  and the position of the window  $h(t)$  relative to the speech signal. Since the window position is chosen independently of the signal, and since the signal itself is non-periodic (at least for time-scales larger than 100 ms), it immediately follows that all values of  $\arg(s_j(T, f))$  have equal probability and, moreover, that  $\mathcal{P}(s_j(T, f))$  must necessarily be circularly symmetric, i.e.,  $\mathcal{P}(s_j(T, f))$  only depends on the magnitude  $|s_j(T, f)|$  and can be written as

$$\mathcal{P}(s_j(T, f)) = g(|s_j(T, f)|) \quad (16)$$

for some properly chosen function  $g(\cdot)$  which models the dependence of  $\mathcal{P}(s_j(T, f))$  on the source amplitude.

In accordance with time-domain BSS algorithms, which frequently model the pdf of real

valued source signals  $s$  as  $\mathcal{P}(s) = \cosh^{-1}(s)$  (MacKay, 1996), the function  $g(\cdot)$  is chosen to be

$$g(x) = c^{-1} \cosh^{-1}(x), \quad c = \int dx g(|x|). \quad (17)$$

Eq. (17) is not intended to be a precise model for the pdf of speech signals. Rather, (17) represents a compromise between a faithful approximation to the sources' pdf and a function  $g(\cdot)$  that results in an adaptation rule with good convergence properties. It is acknowledged that speech signals exhibit a higher kurtosis than is accounted for by (17). On the other hand, choosing  $g(\cdot)$  to model the true pdf of speech results in the non-linear term (20) for the gradient (19) being divergent at  $u_i = 0$ . This compromise is justified by the findings of many researchers (e.g. Lee, 1998, and references therein) that an approximation to the true pdf is in practice sufficient, which has also been justified by theoretical results (Yang and Amari, 1997). It is important, however, that both true and model pdfs have the same sign of kurtosis (Lee, 1998), which is fulfilled in the present situation. Applicability of (17) is also confirmed by the results obtained with the proposed algorithm.

Note that from the non-Gaussianity and circular symmetry of  $\mathcal{P}(s_j(T, f))$  it follows immediately, that the real- and imaginary-part of  $s_j(T, f)$  are *not* independent, since for any two independent random variables with circular symmetric distribution it follows that their pdfs are Gaussian (see Papoulis, 1991).

### 3.3. Adaptation rule for BSS in the frequency domain

In order to obtain an adaptive algorithm, stochastic gradient ascent optimization is used to maximize the log-likelihood. Since the searched parameters  $w_{ij}$  are complex valued, optimization is based on the complex stochastic gradient  $\delta w_{ij}(T, f)$ ,

$$\begin{aligned} \delta w_{ij}(T, f) &= \left( \frac{\partial}{\partial \Re w_{ij}(f)} + i \frac{\partial}{\partial \Im w_{ij}(f)} \right) \\ &\quad \times L(\mathbf{W}(f), \mathbf{x}(T, f)), \end{aligned} \quad (18)$$

where  $\partial/\partial \Re w_{ij}(f)$  denotes differentiation with respect to the real-part of  $w_{ij}(f)$  and  $\partial/\partial \Im w_{ij}(f)$

differentiation with respect to the imaginary-part.

As the result of the derivation, the matrix  $\nabla \mathbf{W}(T, f)$  with elements  $\delta w_{ij}(T, f)$  is given by

$$\nabla \mathbf{W}(T, f) = (\mathbf{I} + \mathbf{v}(T, f)\mathbf{u}^H(T, f))\mathbf{W}^{-H}(f), \quad (19)$$

where  $\mathbf{I}$  is the identity matrix and the unmixed signals are denoted as  $\mathbf{u}(T, f) = \mathbf{W}(f)\mathbf{x}(T, f) = (u_1(T, f), u_2(T, f))^T$ . The vector  $\mathbf{v}(T, f) = (v_1(T, f), v_2(T, f))^T$  is computed as a non-linear function of  $\mathbf{u}(T, f)$ ,

$$v_i(T, f) = -\frac{u_i(T, f)}{|u_i(T, f)|} \frac{g'(|u_i(T, f)|)}{g(|u_i(T, f)|)} \quad (20)$$

$$= -\frac{u_i(T, f)}{|u_i(T, f)|} \tanh(|u_i(T, f)|), \quad (21)$$

where  $g'(\cdot)$  is the derivative of  $g(\cdot)$ .

It is well known for BSS algorithms that the gradient (19) leads to a rather slow convergence to the separating solution. Speed of convergence can be improved by orders of magnitude by using the modified gradient

$$\begin{aligned} \tilde{\nabla} \mathbf{W}(T, f) &= (\nabla \mathbf{W}(T, f))\mathbf{W}^H(f)\mathbf{W}(f) \\ &= (\mathbf{I} + \mathbf{v}(T, f)\mathbf{u}^H(T, f))\mathbf{W}(f), \end{aligned} \quad (22)$$

which has been denoted as the ‘natural’ or ‘relative’ gradient by Amari et al. (1996) and Cardoso and Laheld (1996), respectively.

We note that in contrast to the unmixing system proposed in (12), the parameters  $w_{11}(f)$  and  $w_{22}(f)$  will not converge to 1. Rather their optimum values will be such that the variance of the unmixed signals matches the variance specified by choice of the sources’ pdf  $g(\cdot)$ . This fact simply corresponds to a different scaling of the rows of  $w_{ij}(f)$  with respect to the rows of  $\hat{w}_{ij}(f)$  in (12). The relationship between the two is given by

$$\hat{w}_{ij}(f) = w_{ij}(f)/w_{ii}(f), \quad (23)$$

or, in terms of the unmixed signals,

$$\hat{u}_i(T, f) = u_i(T, f)/w_{ii}(f). \quad (24)$$

Since  $\mathcal{P}(s(T, f))$  is assumed to be circularly symmetric, there is no preferred complex phase of the unmixed signals. Hence, each row of  $\mathbf{W}(f)$  can be multiplied by a complex number of magnitude one

without altering the likelihood  $L(\mathbf{W}(f), \mathbf{x}(T, f))$ . To fix this invariance, we require that  $w_{ii}(f)$  is normalized to be real and positive for all  $i$ ,

$$w_{ii}(f) \in \mathbb{R} \quad \text{and} \quad w_{ii}(f) \geq 0. \quad (25)$$

The learning rule (22) should be compared to the corresponding equation for real variables. In the case of real valued signals, the only difference is in the definition of  $v_i$  (20), which simplifies to

$$v_i = -\frac{g'(u_i)}{g(u_i)}. \quad (26)$$

i.e., in the case of complex signals, the non-linearity is simply computed from the magnitude and the result acquires the original complex phase.

It is noted that the non-linearity (20) for circular symmetric source distributions coincides with the non-linearity given (albeit without explanation) by Cardoso and Laheld (1996) for the generalization of their separation algorithm from real-valued sources to the complex case. However, for sources without circular symmetry, the simple form of (20) does not hold (for a discussion of complex sources with non-symmetric distributions encountered in digital communications, see Torkkola, 1998). E.g., the non-linearity proposed by Smaragdis (1998) for the separation of Fourier transformed speech signals cannot be written in the form of (20) and therefore implies source signals without circular symmetry which, for the reasons given above, appears to be unrealistic.

Since the unmixing (10) takes the form of a matrix-vector product for each frequency  $f$ , a straight-forward solution would be to maximize the likelihood function (14) for each separating matrix  $\mathbf{W}(f)$  separately. This procedure results in a set of separating matrices  $\mathbf{W}(f)$ , one for each frequency  $f$ . However, since each of the separating matrices is derived independently, the source signals’ components are in general reconstructed in (unknown) disparate order in different frequency channels, making a time-domain reconstruction of the unmixed signals impossible, as depicted in Fig. 3. To deal with such permutations, supplementary methods for sorting them need to be employed (e.g. Murata et al., 2001). A further disadvantage of working in each frequency separately is, that

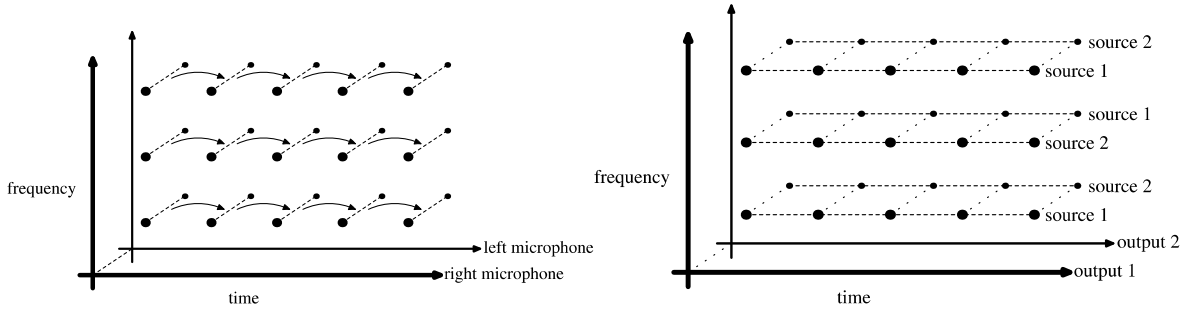


Fig. 3. Performing separation *independently* in each frequency (depicted on the left) results in unmixed signals components whose order with respect to the corresponding source components is permuted in different frequencies (see right).

relatively long signal-segments need to be known in order to achieve descent separation (Smaragdis, 1998, reported signal lengths of at least 2s).

Rather than performing separation in each frequency independently, we are pursuing the aim of incorporating the prior knowledge of free field mixing into the algorithm. By exploiting this knowledge, a constrained adaptive algorithm is derived which avoids local permutations, which is easily implemented in real-time, and which exhibits rapid convergence.

#### 4. Constrained optimization

Due to the free field assumption (8) and (25), separation can be achieved by the matrix

$$\begin{aligned} \mathbf{W}(f) &= \begin{pmatrix} w_{11}(f) & w_{12}(f) \\ w_{21}(f) & w_{22}(f) \end{pmatrix} \\ &= \begin{pmatrix} w_{11} & -w_{12}e^{-i2\pi f\tau_{12}} \\ -w_{21}e^{-i2\pi f\tau_{21}} & w_{22} \end{pmatrix}, \end{aligned} \quad (27)$$

where  $w_{ij}$  is real and positive for all  $i, j$ . Hence, the quantities which need to be known to perform separation are the  $w_{ij}$  and  $\tau_{ij}$ .

The parameters  $w_{ij}$  are readily computed as  $w_{ij} = |w_{ij}(f)|$ . Hence, if  $|w_{ij}(f)|$  is known for some frequency  $f$ , the corresponding magnitudes  $|w_{ij}(f')|$  for all other frequencies  $f' \neq f$  are known, as well. Therefore, improving on the estimate of  $w_{ij}(f)$  for some frequency  $f$  using the algorithm presented in Section 3, results in improved estimates of  $|w_{ij}(f')|$  for all  $f'$ .

However, the situation is more complex for the phase factors  $-\exp(-i2\pi f\tau_{12})$  and  $-\exp(-i2\pi f\tau_{21})$ . Due to the  $2\pi$ -ambiguity of the complex phase, it is in general not possible to obtain  $\tau_{ij}$  from  $-\exp(-i2\pi f\tau_{21})$ . In contrast, the  $2\pi$ -ambiguity does not exist for the corresponding *change* of parameters  $\tau_{ij}$  during update steps (22).

Therefore, we change from the complex parameter  $w_{ij}(f)$  to the (real) parameters of magnitude and time-delay,  $w_{ij}$  and  $\tau_{ij}$ , respectively. The stochastic gradient for the new parameters  $(\delta w_{ij}, \delta \tau_{ij})$  is obtained from (18) and (27) as

$$\begin{aligned} \tilde{\delta} w_{ij}(T, f) &= \frac{1}{w_{ij}} \Re(w_{ij}(f) \delta w_{ij}^*(T, f)), \\ \tilde{\delta} \tau_{ij}(T, f) &= \frac{1}{2\pi f w_{ij}^2} \Im(w_{ij}(f) \delta w_{ij}^*(T, f)), \end{aligned} \quad (28)$$

where  $\Re(\cdot)$  and  $\Im(\cdot)$  denote real- and imaginary-part, respectively, and  $\tilde{\delta} w_{ij}(T, f)$  is the  $(i, j)$ -element of  $\tilde{\nabla} \mathbf{W}(T, f)$ , calculated from (20) and (22) as

$$\tilde{\nabla} \mathbf{W}(T, f) = (\mathbf{I} + \mathbf{v}(T, f) \mathbf{u}^H(T, f)(T, f)) \mathbf{W}(f). \quad (22)$$

Given some initial estimate  $(w_{ij}, \tau_{ij})$  for magnitudes and time-delays, any measurement  $\mathbf{x}(T, f)$  for arbitrary  $(T, f)$  can be used to calculate improved estimates  $(w'_{ij}, \tau'_{ij})$  by the following steps:

1. Using (27), calculate  $\mathbf{W}(f)$  from  $(w_{ij}, \tau_{ij})$ .
2. From (22), calculate the complex gradient  $\delta w_{ij}(T, f)$  of the parameter  $w_{ij}(f)$ .



3. From (28), calculate the corresponding gradient  $(\delta w_{ij}, \delta \tau_{ij})$  of the magnitude and time-delay parameters  $(w_{ij}, \tau_{ij})$ .
4. The improved estimates for  $w_{ij}$  and  $\tau_{ij}$  are given by
 
$$w'_{ij} = w_{ij} + \eta \delta w_{ij}, \quad \tau'_{ij} = \tau_{ij} + \eta \delta \tau_{ij}, \quad (29)$$
 where  $0 < \eta \ll 1$  is the adaptation rate.

#### 4.1. Adaptation scheme

Using this update procedure, the data at arbitrary points in the time-frequency plane can be used to iteratively improve the estimate of  $w_{ij}$  and  $\tau_{ij}$ . In particular, it is possible to first use data  $\mathbf{x}(T, f)$  from *all* frequencies at a particular time  $T$  before moving to the next time point  $T + 1$ . We propose the following adaptation scheme:

1. Start with some initial guess for  $(w_{ij}, \tau_{ij})$ , and with  $T = 1$  and  $f = 1$ .
2. Based on the signal  $\mathbf{x}(T, f)$ , calculate improved estimates  $(w'_{ij}, \tau'_{ij})$  for  $(w_{ij}, \tau_{ij})$ , using the procedure described above.
3. Compute the algorithm's output signals  $\hat{\mathbf{u}}_i(T, f)$  from (24).
4. If  $f$  is *not* the highest possible frequency, set  $f' = f + 1$  and  $T' = T$ .
5. If  $f$  is the highest frequency, set  $f' = 1$  and  $T' = T + 1$ .
6. Use  $(T', f')$  and  $(w'_{ij}, \tau'_{ij})$  as the new values for  $(T, f)$  and  $(w_{ij}, \tau_{ij})$ .
7. Continue with step 2.

Using this adaptation scheme, the algorithm iterates in 'loops' across the spectrogram, as de-

picted in Fig. 4. Since the parameter  $w_{ij}$  and  $\tau_{ij}$  'tie' together the different frequencies, the source components are reconstructed in the same order in all frequencies, making a reconstruction of the time-domain signals by, e.g., the overlap-add technique possible (cf. Fig. 4).

### 5. Implementation

Adaptive algorithms pose additional problems compared to their non-adaptive counterparts, in particular if the signals to be processed are as non-stationary as speech signals are. In this section, three implementation techniques are described which have been found indispensable in order to ensure that the algorithm converges fast and reliably to the separating solution, and to ensure that it remains, with small variance, in the vicinity of the solution while still being adaptive.

#### 5.1. Variable adaptation rate for different frequencies

As in any on-line adaptation algorithm with fixed adaptation rate, the parameter estimate is biased by data which was presented most recently to the algorithm. This effect is to some extent desirable, since it enables the algorithm to adapt to changing environments. However, the scheme proposed in Section 4.1 involves the presentation of data sequentially in both time- and frequency-dimension (cf. also to Fig. 4, left panel). Presenting the data sequentially in time (data from earlier frames is presented prior to data from later

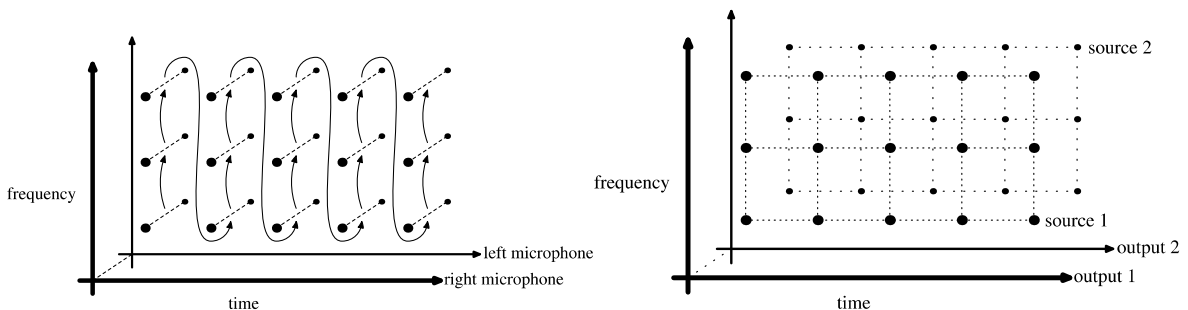


Fig. 4. Iterating the separation algorithm *across* frequencies (left) results in the same order of unmixed components with respect to the corresponding sources for all frequencies (right).

frames) results in the desirable effect, that the parameter estimates are biased towards data that occurred most recently, i.e., the algorithm can adapt to changing environments. Presentation of the data sequentially in frequency (i.e., within each frame data from low frequencies is used for parameter estimation prior to data from high frequencies) results in a bias of the estimated separation parameters towards high frequencies; an effect that we have observed to be undesirable.

In our investigations, we found that the bias towards high frequencies reduces the stability of the algorithm and should be avoided. Therefore, different methods have been examined to compensate for this effect. The scheme which yielded the best results, both in terms of speed of convergence and robustness, is a simple  $1/f$ -decay in the adaptation rate for the magnitudes  $w_{ij}$ . Hence, (29) should be replaced by

$$w'_{ij} = w_{ij} + \frac{\eta}{f} \delta w_{ij}, \quad \tau'_{ij} = \tau_{ij} + \eta \delta \tau_{ij}. \quad (30)$$

This is justified by the classic result from stochastic approximation theory (Robbins and Monro, 1951) that weighting parameter updates proportionally to  $1/n$ , where  $n$  denotes the order of sequential presentation, results (under certain conditions) in an unbiased parameter update. A result which is, e.g., widely applied to the learning rate in neural network training (e.g., Sompolinsky et al., 1995). Hence, with (30) the estimates for  $w_{ij}$  are *not* biased by the samples which occurred at high frequencies. However, the bias with respect to samples most recent in time remains, so that the algorithm can still adapt.

We also experimented with a  $1/f$ -decay in the adaptation rate for the time-delay  $\tau_{ij}$ , but it was found to decrease the speed of convergence too much while the robustness of the  $\tau_{ij}$  was already sufficient without the decay. This can be explained by the fact that a decay is already inherent in  $\delta \tau_{ij}$  of (28) through the factor  $1/f$ , and therefore an additional decay of the adaptation rate for  $\tau_{ij}$  is not necessary.

The  $1/f$ -decay introduced here can intuitively be interpreted as follows: The low frequencies may be forced to a rapid convergence at high adaptation rates to the vicinity of the correct solution

because it is more difficult to find an exact solution than for higher frequencies. The higher frequencies, from which a time-delay can be better estimated, provide improved accuracy at a lower adaptation rate.

## 5.2. Preemphasis

Convergence of the algorithm was further improved by applying a preemphasis filter to the original microphone signals  $x_i^{(o)}(t)$ , resulting in input signals  $x_i(t) = x_i^{(o)}(t+1) - x_i^{(o)}(t)$  for the algorithm. It is easily verified that the free field mixing and demixing models (2) and (12) still apply if the original sources  $s_j^{(o)}(t)$  are replaced by filtered sources  $s(t) = s_j^{(o)}(t+1) - s_j^{(o)}(t)$ . After separation has been performed, the unmixed signals must be low-pass filtered to compensate for the effect of the preemphasis.

Two reasons can be regarded to account for the beneficial effect of the preemphasis on the algorithms' performance.

First, the preemphasis has the effect of reducing the source signals' kurtosis considerably, as shown in Table 1. Due to the low signal energy towards high frequencies, the original kurtosis is very high, and by approximately flattening the spectrum the preemphasis results in a more uniformly distributed variance across frequencies, thereby reducing the kurtosis and improving the match between the true and the assumed model pdf (for a discussion of the effects of non-stationarity on a signal's pdf, see, e.g., Parra et al., 2001).

Furthermore, the preemphasis operation results in a larger effect of high frequencies on the adaptation steps. However, it should be noted that according to the update Eq. (22), the preemphasis is not equivalent to a higher adaptation rate for high

Table 1  
Kurtosis of speech in the time-domain, in the frequency domain and the kurtosis of differentiated (high-pass filtered) speech in the frequency domain

	Kurtosis
Time domain	5.5
Frequency domain	289.8
Frequency domain, high-pass filtered	21.2

frequencies. Therefore, it is advisable to use both preemphasis and decay of the adaptation rate.

### 5.3. *Speech pause detection*

Speech pauses in one source which, in the examples of Section 6, last up to 700 ms, can be a problem for the adaptive algorithm. Without additional precautions, the algorithm would diverge during these intervals, since it would attempt to find an alternative source to be separated. One possibility to account for this effect could be to preset a fixed energy threshold for each source, below which no parameter adaptation is performed in order to avoid divergence. However, a fixed threshold is inconsistent with the framework of blind separation where no assumptions are made about the sources' level. Therefore, we have opted to introduce a relative threshold for the power of the sources. If the energy of any reconstructed signal in the current FFT-frame is less than 15% of the energy of the other reconstructed signal, then solely separation but no parameter update is performed.

## 6. Evaluation

Results from experiments with artificially mixed sources and with real-world recordings in an anechoic chamber are reported. In the first experiment, we verify the proposed algorithm using speech signals which have been mixed digitally in the time-domain with time- and level-differences. In the second experiment, source separation is performed on real-world recordings of two speakers in an anechoic chamber. Finally, it is demonstrated that the proposed algorithm successfully separates moving speakers by applying it to anechoic recordings where one speaker is standing while the second is moving.

In all experiments the following preprocessing was used in order to obtain the input spectrograms: The signals were recorded using a sampling rate of 48 kHz and a preemphasis was applied. Speech pauses were not removed. Spectrograms were computed using a hanning-window of length 30 ms and a window-shift of 10 ms. The resulting frames were padded with zeros to 2048 samples

before a fast-Fourier-transform was applied. Spectral components from 23 Hz to 10 kHz were used for adaptation, since the main energy of the signals occurs in this range.

The parameters of the algorithm were initialized to  $w_{11} = w_{22} = 1$ ,  $w_{12} = w_{21} = 0$ ,  $\tau_{12} = \tau_{21} = 0$ , i.e., the algorithm started off from the (wrong) assumption that no mixing occurs. The initial adaptation rate was set to  $\eta = 0.4$  in order to pass first transients. It was then lowered proportionally to  $1/T$  until it reached  $\eta = 0.001$  after 4 s.  $\eta = 0.001$  was then kept constant for the remaining time.

Finally, the separated signals were transformed back to the time-domain, using the overlap-add method (e.g., Oppenheim and Schaefer, 1975), and the effect of the preemphasis was compensated by low-pass filtering the separated signals.

The entire processing, including spectral decomposition, source separation and overlap-add reconstruction, was implemented as a C++ program which performed processing approximately in real-time on a Silicon Graphics workstation with computing power equivalent to a Pentium 133 PC.

Sound files corresponding to all experiments can be downloaded from the internet-address <http://medi.uni-oldenburg.de/demo/ane/specocom>.

### 6.1. *Artificially mixed sources*

Two mono speech signals were digitally mixed in the time-domain according to the mixing system (5), using time- and level-differences of  $\tau_{21} = 0.5$  ms and  $a_{21} = 0.95$ , respectively, for the first source, and  $\tau_{12} = 1.0$  ms and  $a_{12} = 0.90$ , respectively, for the second source.

Fig. 5 displays the time-course of estimated time- and level-differences assumed by the demixing system for both reconstructed signals. The estimates of the time differences have converged to the correct solution after only 0.2 s, already resulting in very good separation. It takes up to  $\approx 1$  s, unless the level differences have also adapted to their optimum, which results in a small improvement of the separation. Due to the non-stationary nature of speech signals, the parameters remain to

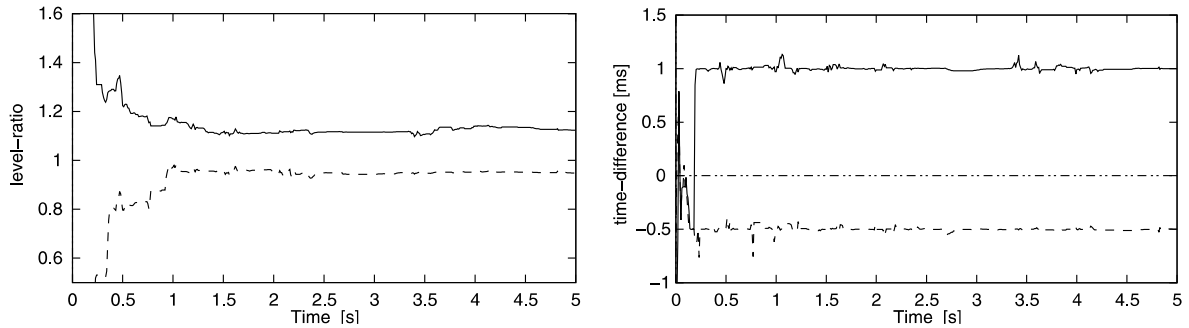


Fig. 5. Time-course of estimated level- (left) and time-differences (right) assumed by the demixing system for the separation of artificially mixed sources. For better visual presentation,  $1/w_{21}$  and  $\tau_{21}$  correspond to the solid lines, whereas  $w_{12}$  and  $-\tau_{12}$  correspond to the dashed lines. Therefore, parameter values corresponding to a source in the right hemisphere are found in the upper half of the figures, and vice versa. The optimum is attained at  $1/w_{21} = 1.11$ ,  $\tau_{21} = 1$  ms,  $w_{12} = 0.95$  and  $-\tau_{12} = -0.5$  ms.

Table 2  
Signal separation caused by the algorithm

Situation	Signal separation (dB)
Synthetic delay and gain	26.5
Anechoic chamber	15.5

fluctuate slightly during the remaining time of the recording.

Informal listening to the reconstructed signals reveals that separation is almost perfect and the remaining cross-talk is nearly inaudible. The improvement in signal separation is displayed in Table 2. It was measured as the increase of direct-to-cross-talk energy from before separation to after separation. The fast and almost perfect separation demonstrates that the proposed algorithm operates successfully under optimal conditions.

## 6.2. Stationary sources in anechoic environment

Recordings for this experiment were performed in the anechoic chamber of the University of Oldenburg, so that the free field assumption was fulfilled to a first approximation.

Two microphones were placed 35 cm apart. Stereo recordings were performed of one male speaker talking from two positions of approximately  $60^\circ$  to the left and  $60^\circ$  to the right of the mid-perpendicular of the microphones, respectively. The recordings were of moderate quality, in particular, recording noise is clearly audible. The dis-

tance between speakers and microphones was 3 m (cf. Fig. 6). The two stereo recordings were digitally added in the time-domain to obtain the mixed signals, a procedure that is justified by the linearity of sound superposition in air. Since with this recording method the source signals as recorded at the position of the microphones are known, direct-to-cross-talk energy ratios can be computed both for the mixed signals and for the unmixed signals obtained by the proposed algorithm.

Using the parameters as described above, the mixed signals were processed by the algorithm. The improvement of the direct-to-cross-talk ratio was determined to be 15.5 dB. Analysis of the separation parameters' time-course again revealed the rapid convergence of the algorithm within less than 1 s. In informal listening tests, only a very soft cross-talk of the unmixed signals was audible.

The result of 15.5 dB is compared to the results obtained by another algorithm ('AMDecor algorithm') which has been proposed by the authors for the *non*-adaptive separation of convolutive mixtures (including reverberation) of speech signals (see Anemüller and Kollmeier, 2000). The AMDecor algorithm has been shown to result in very good separation which is close to the physical limits imposed by the length of the separation filters. In the same anechoic situation, the AMDecor algorithm caused an improvement in direct-to-cross-talk energy of 15.3 dB, though with a window length of 85 ms. Since the longer windows favor the AMDecor algorithm by allowing for

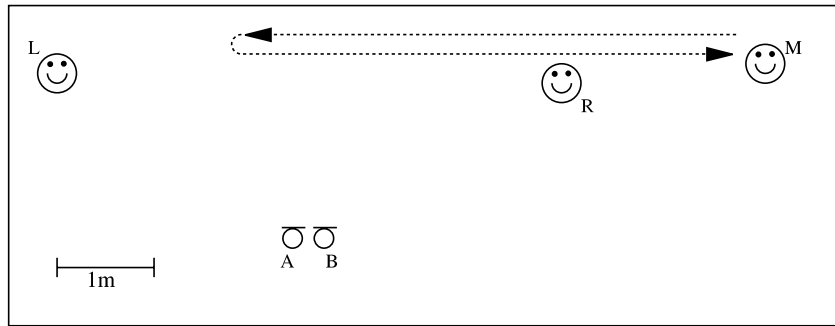


Fig. 6. Setup for the recordings performed for evaluation. Microphones are located at positions A and B. Speaker positions for the experiment from Section 6.2 are L and R, respectively. For the experiments of Section 6.3, the moving speaker started at position M, followed the indicated route and returned to position M, while the standing speaker was at position L.

longer separation filters, it is concluded that the adaptive algorithm proposed in this paper performs excellent. Even though it is adaptive, and even though it uses shorter separation filters, it obtains a slightly better signal separation than its non-adaptive counterpart.

6.3. *Moving sources in anechoic environment*

In the final experiment, signals from a moving and a stationary speaker in anechoic environment were separated, demonstrating that the adaptation of the separation algorithm is sufficient to track moving sources.

With the exception of the moving speaker, the experimental setup was the same as in the previous experiment. The moving speaker started in a dis-

tance of 4.7 m at a position at 70° to the right, walked in a straight line parallel to the microphones until he reached a position at about 30° left of the microphones’ mid-perpendicular, and then returned to his original position (cf. Fig. 6).

Fig. 8 displays the source signals, the mixed signals, and the unmixed signals obtained by the algorithm. Time-courses of the time- and level-difference parameters estimated by the algorithm are displayed in Fig. 7.

Again, it is observed that the timing parameters  $\tau_{12}$  and  $\tau_{21}$  assumed by the demixing system converge rapidly to the separating solution. Their time-course clearly displays the movement of one speaker from the right to the left and back, while the second speaker remains stationary. The convergence of the level difference parameters is again slower, however the separation solution is also

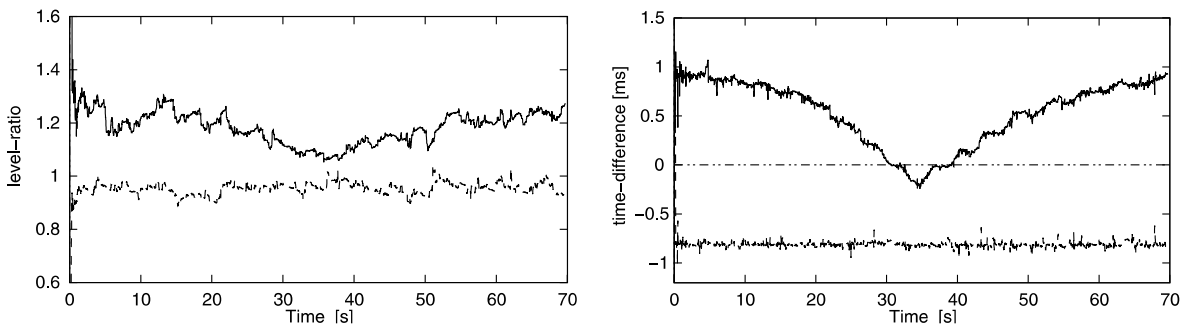


Fig. 7. Time-course of estimated level- (left) and time-differences (right) assumed by the demixing system for the separation of a moving and a standing speaker in anechoic environment. As in Fig. 5,  $1/w_{21}$  and  $\tau_{21}$  correspond to the solid lines, whereas  $w_{12}$  and  $-\tau_{12}$  correspond to the dashed lines. Therefore, parameter values corresponding to a source in the right hemisphere are found in the upper half of the figures, and vice versa.

attained in less than one second. Comparing in Fig. 8 the first ten seconds of the source signals with the algorithm's output signals shows that separation is already very good after less than 0.2 s, since the individual sources' waveforms are clearly recognizable in the unmixed signals.

Informal listening reveals that very good signal separation is achieved almost instantly. However, quality of separation is slightly lower for the position reached at about 35 s signal time where both sources are at their closest distance. In this position, source separation is most difficult to achieve

since the transfer functions are almost identical for both sources, making the inversion of the mixing system an almost ill-posed inverse problem. As a side effect, recording noise contained in the signals (cf. Section 6.2) is slightly amplified. However, this does not affect the algorithm's convergence.

## 7. Discussion

In this paper, an algorithm for the blind separation of acoustically mixed sources was proposed.

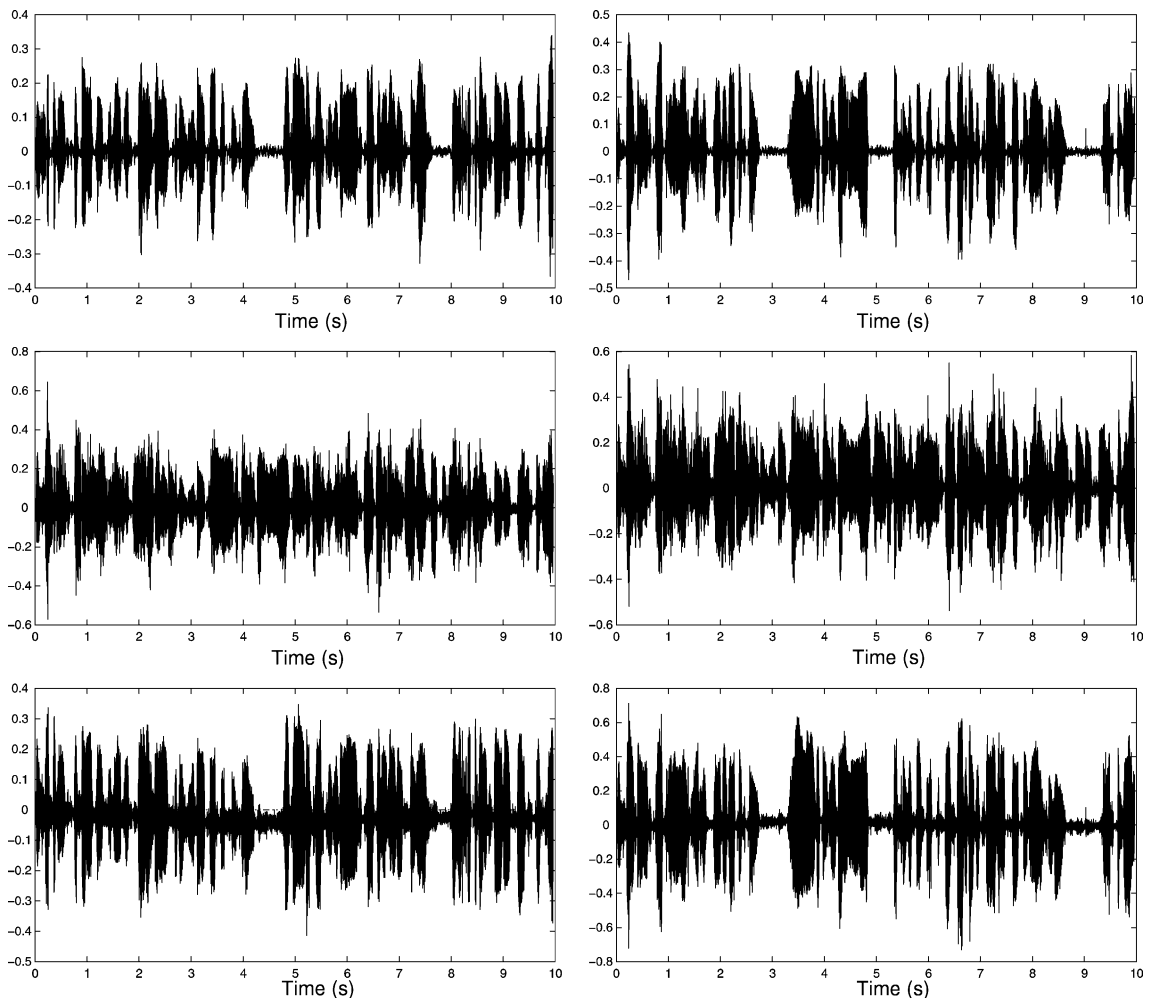


Fig. 8. First ten seconds of speech recordings from the separation of moving sources. Top row: original signals of the moving speaker (left) and the standing speaker (right). Center row: left and right channel of the mixed signals. Bottom row: unmixed signals obtained by the algorithm.

Based on a general algorithm for the separation of Fourier transformed speech, constraints derived from the free field assumption were incorporated in order to obtain an adaptive algorithm with good convergence properties. Effectiveness was investigated using both digitally mixed signals and recordings from anechoic environment, including the situation of spatially moving sources. In conclusion, methods from the fields of acoustics, digital signal processing, BSS and neural network theory have contributed to the fast and robust convergence of the presented algorithm, which, to the authors' knowledge, represents the first algorithm described in the literature that performs the separation of real recordings of moving speakers (intermediate results presented in Anemüller and Gramß, 1999).

In comparison with previous algorithms for the separation of delayed and attenuated sources (for references, cf. Section 1), the main differences are the implementation in the frequency domain, the evaluation with real-world signals, the fact that the algorithm does not get trapped in local minima, and the rapid convergence. In particular, it is surprising that the convergence towards the correct time-delay parameters is so fast and stable for the present algorithm, whereas for the time-domain algorithm of Torrkola (1996) convergence problems involving local minima were reported for the delay parameters. While the frequency domain implementation introduces a processing delay that is larger than the time-delays  $\tau_{12}$  and  $\tau_{21}$ , it should be noted that the processing delay depends only on the length of the FFT windows (30 ms in our experiments), but not on the convergence time.

For the goal of fast adaptation, the frequency domain formulation allows the use of the improved gradient expression (22) which results in much faster convergence than the standard gradient (19). Furthermore, the frequency domain is beneficial for the algorithm's applicability within more complex processing schemes. Since many other noise reduction schemes, in particular spectral approaches, work in the frequency domain, as well, it is possible to combine them with the presented algorithm at a low computational cost. Taking into account that the C++ implementation used for this paper performed the spectral

decomposition at 48 kHz, source separation for frequencies up to 10 kHz, and overlap-add reconstruction at 48 kHz approximately in real-time with computing power equivalent to a 133 MHz Pentium computer, it is obvious that much faster implementations are possible for lower sampling rates and, in particular, if the data at hand is already split into spectral components.

Since the frequency domain implementation allows for fractional delays, it appears to be well suited for applications with closely spaced microphones, as in modern multi-microphone hearing aids. For truly binaural hearing aids, where head related transfer functions replace the delay-and-gain assumption of Eq. (2), it is in principle possible to include this prior knowledge into the algorithm by parameterizing the unmixing system by the azimuth, i.e., using certain combinations of interaural time- and level-differences instead of tracking them independently.

It is expected that the algorithm also achieves some degree of source separation in real rooms if sources and microphones are placed at a small distance, i.e., within the radius of reverberation (e.g. Heckl and Müller, 1994), and if only diffuse noise is present. Late reflections, which are decorrelated at the microphones, can be regarded as diffuse noise. In contrast, early reflections with correlated components at both microphones, effectively constitute a third signal source which violates the assumed mixing model and therefore might hinder convergence. Within the radius of reverberation, the algorithm might also be used as a preprocessing step for unconstrained BSS algorithms which separate convolutive (reverberant) mixtures: The direct sound can be separated by means of the current free field algorithm, whereas the reverberant signal components are separated by an unconstrained BSS algorithm. By splitting the problem into two parts, the overall adaptation speed might be increased since the convolutive algorithm can be implemented with shorter separation filters.

For the application in digital hearing aids, the presented 'blind' algorithm will have to be combined with a 'non-blind' control algorithm which incorporates additional prior knowledge. The control algorithm should activate the algorithm

only in those acoustical situations in which the assumptions of the current source separation algorithm are approximately fulfilled. This analysis of room acoustics could be performed, e.g., based on a measure like the degree of diffusiveness (Wittkop, 2001) which characterizes the reverberation in the present acoustic environment. Furthermore, the control algorithm should identify which of the separated signals represents the signal of interest for the listener. This decision could be based on, e.g., speech activity detection. Alternatively, the time difference parameters  $\tau_{12}$  and  $\tau_{21}$  could be compared to reference values corresponding to directions where signals of interest are expected (such as the frontal incidence direction).

## 8. Conclusion

The current algorithm has been shown to separate two sound sources fast, with a small processing delay (about 30 ms in the current overlap-add-implementation) and with a moderate computational effort. However, since a satisfactory suppression of one of two sound sources only takes place if the free field assumption is approximately met, the current approach is limited to certain acoustical situations that are characterized by small distances between the two sound sources and the recording positions and negligible acoustical reflections. For a broader application of the current approach in hearing aids, a combination with other algorithms appears to be necessary.

## Acknowledgements

Supported by Deutsche Forschungsgemeinschaft (DFG) and Bundesministerium für Bildung und Forschung (BMBF, PT-AUG, Kompetenzzentrum HörTech). The authors would like to thank all members of the ‘Medizinische Physik’ Group as well as the ‘Graduiertenkolleg Psychoakustik’ for their support and help. The helpful comments of three anonymous reviewers on a former draft of the manuscript are gratefully acknowledged.

## References

- Amari, S., Cichocki, A., Yang, H.H., 1996. A new learning algorithm for blind signal separation. In: Touretzky, D., Mozer, M., Hasselmo, M. (Eds.), *Advances in Neural Information Processing Systems* 8, pp. 757–763.
- Anemüller, J., Gramß, T., 1999. On-line blind separation of moving sound sources. In: Cardoso, J.F., Jutten, C., Loubaton, P. (Eds.), *Proceedings of the First International Workshop on Independent Component Analysis and Blind Signal Separation*, Aussois, France, pp. 331–334.
- Anemüller, J., Kollmeier, B., 2000. Amplitude modulation decorrelation for convolutive blind source separation. In: Pajunen, P., Karhunen, J. (Eds.), *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation*, Helsinki, Finland, pp. 215–220.
- Anemüller, J., Kleinschmidt, M., Kollmeier, B., 2000. Blinde Quellentrennung als Vorverarbeitung zur robusten Spracherkennung. In: Mellert, V. (Ed.), *Fortschritte der Akustik: DAGA 2000*. Deutsche Gesellschaft für Akustik (DEGA), Oldenburg, Germany, pp. 364–365.
- Back, A.D., Tsoi, A.C., 1994. Blind deconvolution of signals using a complex recurrent network. In: Vlontzos, J., Hwang, J., Wilson, E. (Eds.), *Neural Networks for Signal Processing*, Vol. 4, pp. 565–574.
- Bell, A.J., Sejnowski, T.J., 1995. An information maximization approach to blind separation and blind deconvolution. *Neural Computation* 7, 1129–1159.
- Bingham, E., Hyvärinen, A., 2000. A fast fixed-point algorithm for independent component analysis of complex valued signals. *International Journal of Neural Systems* 10, 1–8.
- Bishop, C.M., 1995. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford.
- Brehm, H., Stammler, W., 1987. Description and generation of spherically invariant speech-model signals. *Signal Processing* 12, 119–141.
- Cardoso, J.-F., Laheld, B.H., 1996. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing* 44, 3017–3030.
- Comon, P., 1994. Independent component analysis, a new concept? *Signal Processing* 36, 287–314.
- Emile, B., Comon, P., 1998. Estimation of time delays between unknown colored signals. *SIGPROC* 69, 93–100.
- Fiori, S., 2000. Blind separation of circularly distributed sources by neural extended APEX algorithm. *Neurocomputing* 34, 239–252.
- Heckl, M., Müller, H.A. (Eds.), 1994. *Taschenbuch der Technischen Akustik*, second ed. Springer, Berlin.
- Hyvärinen, A., Karhunen, J., Oja, E., 2001. *Independent Component Analysis*. Wiley, New York.
- Jutten, C., Héroult, J., 1991. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing* 24, 1–10.
- Lee, T.-W., 1998. *Independent Component Analysis: Theory and Applications*. Kluwer Academic Publishers, Boston.



- Lee, T.-W., Ziehe, A., Orglmeister, R., Sejnowski, T.J., 1998. Combining time-delayed decorrelation and ICA: Towards solving the cocktail party problem. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA. Vol. 2, pp. 1249–1252.
- MacKay, D.J.C., 1996. Maximum likelihood and covariant algorithms for independent component analysis. Technical Report, Department of Physics, Cambridge University, England. URL <ftp://wol.ra.phy.cam.ac.uk/pub/mackay/ica.ps.gz>.
- Moreau, E., Macchi, O., 1994. Complex self-adaptive algorithms for source separation based on higher order contrasts. In: *Proceedings of EUSIPCO'94*, Edinburgh, Scotland, pp. 1157–1160.
- Murata, N., Ikeda, S., Ziehe, A., 2001. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing* 41, 1–24.
- Oppenheim, A.V., Schaefer, R.W., 1975. *Digital Signal Processing*. Prentice-Hall, Englewood Cliffs.
- Papoulis, A., 1991. *Probability, Random Variables, and Stochastic Processes*, third ed. McGraw-Hill, New York.
- Parra, L., Spence, C., 2000a. Convolutional blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing* 8, 320–327.
- Parra, L., Spence, C., 2000b. On-line blind source separation of non-stationary signals. *Journal of VLSI Signal Processing Systems for Signal Image and Video Technology* 26, 39–46.
- Parra, L., Spence, C., Sajda, P., 2001. Statistical properties arising from the non-stationarity of natural signals. In: Leen, T.K., Dietterich, T.G., Tresp, V. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 13. MIT Press, Cambridge, MA.
- Pham, D.T., Garat, P., Jutten, C., 1992. Separation of a mixture of independent sources through a maximum likelihood approach. In: Vandewalle, J., Boite, R., Moonen, M., Oosterlinck, A. (Eds.), *Signal Processing VI: Theories and Applications*, pp. 771–774.
- Platt, J.C., Faggin, F., 1992. Networks for the separation of sources that are superimposed and delayed. In: Moody, J., Hansen, S., Lippmann, R. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 4. MIT Press, Cambridge, MA, pp. 730–737.
- Robbins, H., Monro, S., 1951. A stochastic approximation method. *Annals of Mathematical Statistics* 22, 400–407.
- Sahlin, H., Broman, H., 1998. Separation of real-world signals. *Signal Processing* 64, 103–104.
- Smaragdis, P., 1998. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing* 22, 21–34.
- Sompolinsky, H., Barkai, N., Seung, H.S., 1995. On-line learning of dichotomies: algorithms and learning curves. In: Oh, J.-H. (Ed.), *Neural Networks: The Statistical Mechanics Perspective*, pp. 105–130.
- Tong, L., Liu, R.-w., Soon, V.C., Huang, Y.-F., 1991. Indeterminacy and identifiability of blind identification. *IEEE Transactions on Circuits and Systems* 38, 499–509.
- Torkkola, K., 1996. Blind separation of delayed sources based on information maximization. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, GA, pp. 3509–3512.
- Torkkola, K., 1998. Blind signal separation in communications: making use of known signal distributions. In: *Proceedings of the 1998 IEEE Digital Signal Processing Workshop*, Bryce Canyon, UT.
- van der Kouwe, A.J.W., Wang, D.L., Brown, G.J., 2001. A comparison of auditory and blind separation techniques for speech segregation. *IEEE Transactions on Speech Audio Processing* 9, 189–195.
- Wittkop, T., 2001. Two-channel noise reduction algorithms motivated by models of binaural interaction. Ph.D. thesis, Fachbereich Physik, Universität Oldenburg.
- Yang, H.H., Amari, S.-i., 1997. Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information. *Neural Computation* 9, 1457–1482.
- Yeredor, A., 2001. Blind source separation with pure delay mixtures. In: Lee, T.-W., Jung, T.-P., Makeig, S., Sejnowski, T.J. (Eds.), *Proceedings of ICA 2001*, San Diego, CA.
- Zelinski, R., Noll, P., 1977. Adaptive transform coding of speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-25*, 299–309.