

# **Across-Frequency Processing in Convolutional Blind Source Separation**

Jörn Anemüller  
geboren am 21. Mai 1971  
in Lippstadt

Vom Fachbereich Physik der Universität Oldenburg zur Erlangung des Grades eines  
Doktors der Naturwissenschaften (Dr. rer. nat.) angenommene Dissertation.

Erstreferent: Prof. Dr. Dr. Birger Kollmeier  
Korreferent: Prof. Dr. Volker Mellert  
Tag der Disputation: 30. Juli 2001

# **Across-Frequency Processing in Convolutional Blind Source Separation**

Jörn Anemüller  
geboren am 21. Mai 1971  
in Lippstadt

Vom Fachbereich Physik der Universität Oldenburg zur Erlangung des Grades eines  
Doktors der Naturwissenschaften (Dr. rer. nat.) angenommene Dissertation.

Erstreferent: Prof. Dr. Dr. Birger Kollmeier  
Korreferent: Prof. Dr. Volker Mellert  
Tag der Disputation: 30. Juli 2001

# Contents

<b>1</b>	<b>General Introduction</b>	<b>5</b>
<b>2</b>	<b>Adaptive separation of acoustic sources in the free field</b>	<b>11</b>
2.1	Introduction . . . . .	11
2.2	Acoustic Mixing and Demixing . . . . .	13
2.3	BSS algorithm for Fourier transformed speech . . . . .	16
2.4	Constrained optimization . . . . .	20
2.4.1	Adaptation scheme . . . . .	22
2.5	Implementation . . . . .	22
2.6	Evaluation . . . . .	24
2.6.1	Artificially mixed sources . . . . .	25
2.6.2	Stationary sources in anechoic environment . . . . .	26
2.6.3	Moving sources in anechoic environment . . . . .	28
2.7	Discussion . . . . .	29
2.8	Conclusion . . . . .	31
<b>3</b>	<b>Amplitude Modulation Decorrelation for Convolutional Blind Source Separation</b>	<b>33</b>
3.1	Introduction . . . . .	33
3.2	Problem Formulation . . . . .	35
3.2.1	Frequency domain formulation . . . . .	36
3.2.2	Invariances . . . . .	36
3.3	Amplitude Modulation Correlation . . . . .	38
3.3.1	Structure in Speech . . . . .	38
3.3.2	Amplitude Modulation Correlation . . . . .	39
3.4	Source Separation by Amplitude Modulation Decorrelation . . . . .	41
3.4.1	Separation . . . . .	41
3.4.2	Effect of permutations . . . . .	43
3.4.3	AM decorrelation algorithm . . . . .	44
3.4.4	Optimization scheme . . . . .	47
3.5	Experimental evaluation . . . . .	48
3.5.1	Synthetic data . . . . .	48
3.5.2	Separation in different acoustic situations . . . . .	52

3.5.3	Performance on benchmark data . . . . .	58
3.6	Conclusion . . . . .	62
<b>4</b>	<b>Separation of multidimensional sources</b>	<b>65</b>
4.1	Introduction . . . . .	65
4.2	Multidimensional sources and mixing . . . . .	67
4.3	Solution based on correlations across frequency . . . . .	69
4.3.1	Conditions for identifiability . . . . .	70
4.3.2	Solving the permutation problem . . . . .	70
4.3.3	More than two frequencies . . . . .	71
4.3.4	Time-delayed correlations . . . . .	72
4.4	Evaluation . . . . .	73
4.4.1	Synthetic signals . . . . .	73
4.4.2	Color image data . . . . .	75
4.4.3	Speech signals . . . . .	78
4.5	Discussion . . . . .	81
<b>5</b>	<b>Summary and Conclusion</b>	<b>83</b>
<b>A</b>	<b>Technical Appendix</b>	<b>87</b>
A.1	Optimization under unitary matrix constraint . . . . .	87
A.2	Determination of the SIR . . . . .	88
A.3	Non-blind correction of local permutations . . . . .	89
A.4	Non-blind MMSE separation . . . . .	90
<b>B</b>	<b>Blinde Quellentrennung als Vorverarbeitung zur robusten Spracherken-</b>	
	<b>nung</b>	<b>91</b>
B.1	Einleitung . . . . .	91
B.2	Blinde Quellentrennung . . . . .	92
B.3	Robuste Spracherkennung . . . . .	93
B.4	Methoden . . . . .	93
B.5	Ergebnisse . . . . .	94
B.6	Zusammenfassung . . . . .	96
	<b>References</b>	<b>108</b>

# Chapter 1

## General Introduction

Every cocktail-party makes great demands on the visitors' 'neural processors' (Strube, 1981; von der Malsburg and Schneider, 1986). Extracting a single voice from a babble of multiple speakers and background noise is a highly non-trivial task, and the human ear's performance is still unsurpassed in this situation. However, this capability degrades in persons with hardness of hearing, creating the need for smart hearing aids which can mimic the signal processing performed by the healthy auditory system.

Very similar problems are encountered when automatic speech recognition systems are required to operate under noisy conditions. Even though recognition on undisturbed signals can be almost perfect, additional noise still results in a drastic decrease of the performance. Therefore, capabilities similar to those of the human ear are also desirable for automatic speech recognition.

In an attempt to mimic the auditory system's abilities, several noise reduction schemes have been developed which try to suppress signal components corresponding to 'noise' and enhance the 'speech' components by exploiting their respective characteristics. For instance, in the application of spectral noise suppression schemes (Ephraim and Malah, 1984; Cappé, 1994) to speech enhancement it is assumed that the signal of interest is speech with its typical speech pauses, whereas the noise signal is regarded as stationary and uninterrupted. Therefore, it is possible to estimate the noise spectrum during speech pauses and subsequently subtract it from the spectrum of the noise contaminated speech segments in order to obtain the enhanced speech signal. Similarly, a clear notion of 'speech' and 'noise' is also built into the binaural directional filter (Wittkop et al., 1997; Wittkop, 2001), where speech is assumed to impinge from the frontal direction, whereas noise is assumed to originate laterally. Accordingly, it is attempted to suppress signal components that have been identified as lateral.

An alternative point of view is to regard the acoustic scene as being generated by several simultaneously active signal sources at different spatial positions. By decomposing the recorded sound into its components corresponding to the different sources, and by subsequently picking out the particular source which is of interest (e.g. a speech source), it is also possible to suppress the unwanted 'noise' sources. However, in this

approach a distinction between ‘speech’ and ‘noise’ needs only to be made in the last step, where a particular source of interest is selected. In the first, and presumably more difficult step of decomposing the acoustic scene into its underlying sources, the notion of physically separated, i.e., ‘independent’ sources suffices.

Blind source separation (BSS) constitutes an approach which tries to achieve this decomposition with as little prior-knowledge as possible, hence the term ‘blind’. The formulation of the task as a source separation problem points to many more possible applications than the example of noise reduction, since in many situations it is not possible to measure ‘pure’ signals, corresponding to a single source, only. Rather, a superposition of several sources is measured in many applications. Examples are the areas of wireless communications where signals from multiple cellular phones are received by a transmitter, analysis of biomedical signals obtained by electroencephalography (EEG, e.g. Jung et al., 2000), magnetoneurography (MNG, e.g. Ziehe et al., 2000) and functional magnetic resonance imaging (fMRI, e.g. McKeown et al., 1998) where each sensor picks up signals from several neural generators, and text analysis (e.g. Kabl  n and Girolami, 2000) where words from several topics are found in a single text document.

Also in application where it is not known a-priori that the measured data is composed of mutually independent parts, one might attempt to perform such a decomposition in order to facilitate further analysis of the signals. Areas where such attempts have been pursued are, e.g., the analysis of small patches from natural images (Bell and Sejnowski, 1997), short sound segments (Bell and Sejnowski, 1996) and financial data (Back and Weigend, 1997).

Several choices exist for the definition of mutually ‘different’ or ‘independent’ sources, as will be discussed below.

In the first place the question is, which transformation should be employed to obtain the independent signals from the measurements. In general, an arbitrarily complex function might be chosen which maps a number of sensor signals to a (possibly different) number of independent components. However, without any additional assumptions, the resulting problem is ill-determined. To make the problem tractable, it is assumed that a linear transformation suffices to map  $N$  measured signals  $x_1(t), \dots, x_N(t)$  to  $M$  independent signals  $u_1(t), \dots, u_M(t)$ ,

$$\mathbf{u}(t) = \mathbf{W} \mathbf{x}(t), \quad (1.1)$$

where the vectors  $\mathbf{u}(t) = [u_1(t), \dots, u_M(t)]^T$  and  $\mathbf{x}(t) = [x_1(t), \dots, x_N(t)]^T$  contain the independent components and the measured signals, respectively,  $\mathbf{W}$  denotes the  $M \times N$  transformation matrix, and  $t$  numbers the observations. Hence, the task is to find  $\mathbf{W}$  and  $\mathbf{u}(t)$  from knowledge of  $\mathbf{x}(t)$ , only. Since the matrix  $\mathbf{W}$  can also be regarded as a linear (and in general non-orthogonal) coordinate transform to coordinates in which the signals are independent, the sources must be characterized by different directions in  $N$ -dimensional space in order to allow the decomposition (1.1), i.e., the sources must be spatially separated. It is noted that from (1.1) the reconstructed signals  $u_i(t)$  are only determined up to an arbitrary rescaling and permutation, since any (invertible) rescaling and permutation of independent signals is again independent.



Obviously, the requirement to obtain independent signals implies that  $M \leq N$ , since otherwise the components of  $\mathbf{u}(t)$  would be linearly dependent. However, also this simplified version of the original problem was regarded as unsolvable in the early 1980s (see the reference to Bienvenu and Kopp (1983) in Jutten and Taleb (2000)). Extensions of (1.1) involving non-linear transformations and the possibility to obtain more independent components than the number of mixed signals are investigated by several researchers, but no solution has been shown to work for real-world problems, yet.

A solution for (1.1) was obtained by Héroult and Jutten (1986) based on non-linear correlations of the  $u_i(t)$ , and elaborated by Jutten and Héroult (1991). Within this approach, the notion of ‘different’ or ‘independent’ reconstructed signals  $u_i(t)$  is defined as ‘statistical independence’, and the method is generally referred to as independent component analysis (ICA). ICA is based on higher order statistics, i.e., the underlying source signals are assumed to have a non-Gaussian probability density function (at most one source may be Gaussian) which can be exploited for separation. The theory of ICA was further developed by, e.g., Comon (1994), Bell and Sejnowski (1995), Cardoso and Laheld (1996), Amari et al. (1996) and Lee (1998a).

Alternatively, the criterion of ‘different’ reconstructed signals may be based on shifted correlations, as proposed by, e.g., Molgedey and Schuster (1994). This approach exploits information originating from the sources’ spectra which must be different for all sources.

Non-stationarity has been proposed by Matsuoka et al. (1995) as another criterion for blind source separation. Note that non-stationarity and higher-order statistics are closely related (Parra et al., 2001).

Hence, if the data can be split into independent components using the transformation (1.1), several sufficiently well elaborated algorithms exist for finding the solution. In the case of acoustic mixing, however, the simple model (1.1) is not sufficient to separate signals. Since the superposition of sound sources by the acoustic medium involves time-delays, echoes and reverberation, it constitutes a *convolutive* mixing system, which requires algorithms of convolutive BSS for its inversion. That is, instead of a multiplication as in (1.1), filters have to be employed in order to obtain independent components  $u_i(t)$  from the mixed signals.

Several solutions have been proposed for this problem, too, and again they can be classified with respect to their notion of independence. Furthermore, they differ with regard to their implementation of the filtering operation, which can be performed either in the time- or in the frequency domain.

Time domain methods were proposed by several authors, e.g., by Weinstein et al. (1993), Gerven and Compernelle (1995), Chan et al. (1996) and Lindgren and Broman (1998) using second-order statistics; by Yellin and Weinstein (1994), Bell and Sejnowski (1995), Torkkola (1996a), Yellin and Weinstein (1996) and Lee et al. (1997) using higher-order statistics; and by Kawamoto et al. (1998) exploiting the non-stationarity of the source signals.

Frequency domain methods were proposed by, e.g., Capdevielle et al. (1995) using higher-order statistics, Murata et al. (1998) using second order statistics and Parra

et al. (1998) exploiting non-stationarity.

Several methods may be classified as ‘hybrids’, performing some computations in the time domain and some in the frequency domain (e.g. Ehlers and Schuster, 1997; Amari et al., 1997; Lee et al., 1998).

First results for the separation of real room recordings were reported by Yellin and Weinstein (1996), using recordings from a large laboratory room with a short distance of 20 cm to 30 cm between speakers and microphones, and 16ms long separating filters. The algorithms of Ehlers and Schuster (1997), Lee et al. (1998) and Murata et al. (1998) also employed small distances between speakers and microphones or short filters of up to 15 ms length, a filter size that is too small to obtain separation in rooms with considerable reverberation.

Improved quality for the separation of existing data sets, and the ability to separate sources in more difficult acoustical situations, was attained by the frequency domain algorithm of Parra et al. (1998), which is still the benchmark algorithm in the field of acoustic source separation (an extended paper is published as Parra and Spence (2000a)).

The convolutive blind source separation algorithms presented in this thesis work entirely in the frequency domain, where both second-order statistics (cf. chapter 4) and higher-order statistics (cf. chapters 2 and 3) are employed for separation. Since the convolution in the time domain factorizes into a product in the frequency domain, the Fourier transformation permits an elegant formulation of the problem. However, this procedure results in the drawback of recovering the source signals in disparate order in different frequency bands, making a time domain reconstruction of the original sources impossible without additional precautions (Capdevielle et al., 1995; Ikram and Morgan, 2000). Therefore, three different methods to avoid such ‘local permutations’ are presented in this thesis, and it is attempted to shed some light on the origin of local permutations.

Across-frequency interactions serve as the means to avoid permutations. The first algorithm (cf. chapter 2) employs interactions of the *filter parameters* across frequencies, whereas the remaining algorithms (cf. chapters 3 and 4) make use of statistical dependencies of the *source signals’ components* at different frequencies. In all three approaches, the across-frequency interactions are not used solely to sort permutations, but they are also utilized to improve quality of separation; a feature which distinguishes the algorithms from most of the literature.

Chapter 2 presents an algorithm which separates acoustic sources under the idealized assumption that the superposition of sources in rooms can be approximated as a superposition in the free field, involving time- and level differences and diffuse noise, but only negligible reflections and reverberation. After deriving 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.

Chapter 3 approaches the problem from the opposite direction, not imposing any constraints on the separating filters and thereby being applicable also in rooms with reverberation. Rather it is assumed that the source signals exhibit a modulation struc-

---

ture similar to speech. Since the modulation in different frequency channels of speech signals is highly interrelated, envelope correlations across different frequencies are employed to solve the source separation task. The resulting ‘AMDecor’ algorithm is evaluated in different acoustical situations, including strong reverberation, and compared to other source separation algorithms. Performance is further analyzed in appendix B by applying the AMDecor algorithm as a preprocessing stage in an automatic speech recognition system and comparing the resulting recognition rates to the performance of other noise reduction algorithms on the same task.

Motivated by the previous chapter’s results, chapter 4 expands the concept of across-frequency interactions to applications in other domains, such as color images, by introducing the notion of multidimensional sources. In addition, an algorithm based on second order statistics is given which leads to a solution in closed form for the separating system. The permutation problem is solved by a condition on the order of eigenvalues corresponding to the separating system.



## Chapter 2

# Adaptive separation of acoustic sources in the free field:

## A constrained frequency domain approach

### 2.1 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) 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 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,

---

<sup>1</sup>This chapter has been submitted for publication in *Speech Communication*.

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. (1998) and Anemüller and Kollmeier (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). This field appears to be promising for preprocessing by BSS algorithms since the acoustic environment 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 time- and level differences between microphones are the most prominent effects of sound superposition in real rooms that can be used for source separation. Note that this 'free field' assumption is only approximately met in real rooms with short reverberation time and a small distance between sound sources and microphones, respectively.

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 (1996b) proposes a time-domain algorithm which adapts from 15 ms long signal blocks and achieves separation after 1.5 s to 3 s. The algorithm is also evaluated using digitally mixed signals, only, and local minima of the proposed algorithm are found.

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 msec 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 2.2 the unconstrained and constrained acoustic mixing and the corresponding demixing models are introduced. Based on the maximum likelihood principle, a blind source separation algorithm for Fourier transformed speech signals is derived in section 2.3. Section 2.4 is devoted to the incorporation of the free field constraint into the algorithm. Implementation details are given in section 2.5, and evaluation is performed in section 2.6.

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  $\mathbf{x}^T$ , complex conjugation by  $x^*$ , transposition and complex conjugation by  $\mathbf{x}^H$ .

## 2.2 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 microphone  $i$ ,  $i = 1, \dots, N$ , 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  $n_i(t)$ ,

$$x_i(t) = \sum_j \int dt' a_{ij}(t') s_j(t - t') + n_i(t). \quad (2.1)$$

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

$$x_i(t) = \sum_j a_{ij} s_j(t - \tau_{ij}) + n_i(t). \quad (2.2)$$

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

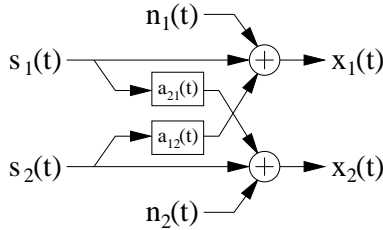


Figure 2.1: The mixing system assumed for the current approach.

microphone signals. Specifically, setting

$$a'_{ij}(t) = \frac{a_{ij}}{\tilde{a}_j} \delta(t - \tau_{ij} + \tilde{\tau}_j) \quad (2.3)$$

$$s'_j(t) = \tilde{a}_j s_j(t + \tilde{\tau}_j) \quad (2.4)$$

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, memoryless mixtures of sources are well-known in the field of blind source separation (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

$$\begin{aligned} x_1(t) &= s_1(t) + a_{12} s_2(t - \tau_{12}) + n_1(t) \\ x_2(t) &= s_2(t) + a_{21} s_1(t - \tau_{21}) + n_2(t), \end{aligned} \quad (2.5)$$

which is illustrated in figure 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

$$x_i(T, f) = \sum_{t=0}^{2K-1} x_i(T+t) h(t) e^{-i\pi f t / K}. \quad (2.6)$$

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 since  $a_{ij}(t)$  is assumed to be short and stationary over time, its STFT does not depend on time  $t$ .

In the frequency domain formulation, the convolution in the acoustic mixing model (2.1) factorizes, provided the window-length is larger than the length of the impulse responses  $a_{ij}(t)$ , yielding the mixing model

$$x_i(T, f) = \sum_j a_{ij}(f) s_j(T, f) + n_i(T, f). \quad (2.7)$$

Under the free field assumption, model (2.7) is a good approximation to the acoustic mixing, and the transfer functions  $a_{ij}(f)$  are computed from the corresponding level- and time differences (2.2) as

$$a_{ij}(f) = a_{ij} e^{-i2\pi f \tau_{ij}}. \quad (2.8)$$

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 (2.5) therefore is

$$\begin{pmatrix} x_1(T, f) \\ x_2(T, f) \end{pmatrix} = \begin{pmatrix} 1 & a_{12}(f) \\ a_{21}(f) & 1 \end{pmatrix} \begin{pmatrix} s_1(T, f) \\ s_2(T, f) \end{pmatrix} + \begin{pmatrix} n_1(T, f) \\ n_2(T, f) \end{pmatrix}. \quad (2.9)$$

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

$$\hat{u}_i(T, f) = \sum_j \hat{w}_{ij}(f) x_j(T, f). \quad (2.10)$$

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

$$\begin{pmatrix} \hat{w}_{11}(f) & \hat{w}_{12}(f) \\ \hat{w}_{21}(f) & \hat{w}_{22}(f) \end{pmatrix} = c(f) \begin{pmatrix} 1 & -a_{12}(f) \\ -a_{21}(f) & 1 \end{pmatrix} \quad (2.11)$$

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

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 (cf. section 2.6 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

$$\begin{pmatrix} \hat{u}_1(T, f) \\ \hat{u}_2(T, f) \end{pmatrix} = \begin{pmatrix} 1 & \hat{w}_{12}(f) \\ \hat{w}_{21}(f) & 1 \end{pmatrix} \begin{pmatrix} x_1(T, f) \\ x_2(T, f) \end{pmatrix} \quad (2.12)$$

which is depicted in figure 2.2. 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

$$E \{ |\hat{u}_i(T, f) - s_i(T, f)|^2 \} \approx E \{ |n_1(T, f)|^2 \} + E \{ |n_2(T, f)|^2 \}, \quad (2.13)$$

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

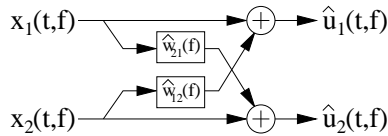


Figure 2.2: The separating system assumed to unmix the signals from the mixing system depicted in figure 2.1.

## 2.3 BSS algorithm for Fourier transformed speech

The superposition of sources in the frequency domain (2.7) has the form of a matrix vector product in each frequency channel  $f$ . In contrast to the time domain representation (2.5), which contains coupling across different time-points, equation (2.7) can be regarded as a set of  $K$  decoupled instantaneous blind source separation problems, albeit with complex valued variables. Several algorithms (e.g. Pham et al., 1992; Bell and Sejnowski, 1995; Cardoso and Laheld, 1996) have been proposed in the literature to solve the instantaneous BSS problem, however, most are concerned with real valued variables, only.

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 (2.7). The combination of this section's learning rule with the prior knowledge about the free field constraint (2.8) for the mixing model is given in section 2.4.

### 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. Zelinski and Noll, 1977; Brehm and Stammers, 1987, 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 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  $\mathbf{A}$  which maximize the probability  $\mathcal{P}(\mathbf{x}|\mathbf{A})$  that measured data  $\mathbf{x}$  has been generated by this particular  $\mathbf{A}$ . Assuming that the sources  $\mathbf{s}(T, f)$  can be recovered using the demixing system  $\mathbf{W}(f) = \mathbf{A}^{-1}(f)$ , 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

$$\begin{aligned} L(\mathbf{W}(f), \mathbf{x}(T, f)) &= \log \mathcal{P}(\mathbf{x}(T, f) | \mathbf{W}(f)) \\ &= \log \det(\mathbf{W}(f)) + \log \mathcal{P}(\mathbf{W}(f) \mathbf{x}(T, f)). \end{aligned} \quad (2.14)$$

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

$$\mathbf{W}(f) = \underset{\mathbf{W}(f)}{\operatorname{argmax}} E \{L(\mathbf{W}(f), \mathbf{x}(T, f))\}. \quad (2.15)$$

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

In order to use the log-likelihood (2.14) to build an optimization algorithm based on it, the sources probability density function (pdf)  $\mathcal{P}(\mathbf{W}(f) \mathbf{x}(T, f)) = \mathcal{P}(\mathbf{s}(T, f))$  needs to be modeled. Due to the sources' mutual independence it follows that their joint pdf  $\mathcal{P}(\mathbf{s}(T, f))$  factorizes into the product of the individual source pdfs,  $\mathcal{P}(\mathbf{s}(T, f)) = \prod_j \mathcal{P}(s_j)$ , 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 probability density function, 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 100msec), 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

$$\mathcal{P}(s_j(T, f)) = g(|s_j(T, f)|) \quad (2.16)$$

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 blind source separation algorithms, which frequently model the probability density function (pdf) of real valued source signals  $s$  as  $\mathcal{P}(s) = \cosh^{-1}(s)$  (MacKay, 1996), the function  $g(\cdot)$  is chosen to be

$$g(x) = c^{-1} \cosh^{-1}(x), \quad c = \int dx g(|x|). \quad (2.17)$$

Equation (2.17) is not intended to be a precise model for the pdf of speech signals. Rather, (2.17) 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 (2.17). On the other hand, choosing  $g(\cdot)$  to model the true pdf of speech results in the nonlinear term (2.20) for the gradient (2.19) being divergent at  $u_i = 0$ . This compromise is justified by the finding of many researchers (e.g. Lee, 1998a, and references therein) that an approximation to the true pdf is in practice sufficient, and this finding 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, 1998a), which is fulfilled in the present situation. Applicability of (2.17) 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).

### 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  $\delta w_{ij}(T, f)$ ,

$$\delta w_{ij}(T, f) = \left( \frac{\partial}{\partial \Re w_{ij}(f)} + \mathbf{i} \frac{\partial}{\partial \Im w_{ij}(f)} \right) L(\mathbf{W}(f), \mathbf{x}(T, f)), \quad (2.18)$$

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

$$\nabla \mathbf{W}(T, f) = (\mathbf{I} + \mathbf{v}(T, f) \mathbf{u}^H(T, f)) \mathbf{W}^{-H}(f), \quad (2.19)$$

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 nonlinear function of  $\mathbf{u}(T, f)$ ,

$$v_i(T, f) = - \frac{u_i(T, f)}{|u_i(T, f)|} \frac{g'(|u_i(T, f)|)}{g(|u_i(T, f)|)} \quad (2.20)$$

$$= - \frac{u_i(T, f)}{|u_i(T, f)|} \tanh(|u_i(T, f)|), \quad (2.21)$$

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

It is well known in for BSS algorithms that the gradient (2.19) 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

$$\tilde{\nabla} \mathbf{W}(T, f) = (\nabla \mathbf{W}(T, f)) \mathbf{W}^H(f) \mathbf{W}(f) = (\mathbf{I} + \mathbf{v}(T, f) \mathbf{u}^H(T, f)) \mathbf{W}(f), \quad (2.22)$$

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

We note that in contrast to the unmixing system proposed in (2.12), 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 (2.12). The relationship between the two is given by

$$\hat{w}_{ij}(f) = w_{ij}(f)/w_{ii}(f), \quad (2.23)$$

or, in terms of the unmixed signals,

$$\hat{u}_i(T, f) = u_i(T, f)/w_{ii}(f). \quad (2.24)$$

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  $i$ ,

$$w_{ii}(f) \in \mathbb{R} \quad \text{and} \quad w_{ii}(f) \geq 0. \quad (2.25)$$

The learning rule (2.22) 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  $v_i$  (2.26), which simplifies to

$$v_i = -\frac{g'(u_i)}{g(u_i)}. \quad (2.26)$$

I.e., in the case of complex signals, the nonlinearity is simply computed from the magnitude and the result acquires the original complex phase.

It is noted that the nonlinearity (2.20) for circular symmetric source distributions coincides with the nonlinearity 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 (2.20) does not hold (for a discussion of complex sources with non-symmetric distributions encountered in digital communications, see Torkkola, 1998). E.g., the nonlinearity proposed by Smaragdis (1998) for the separation of Fourier transformed speech signals cannot be written in the form of (2.20) and therefore implies source signals without circular symmetry which, for the reasons given above, appears to be unrealistic.

Since the unmixing (2.10) takes the form of a matrix-vector product for each frequency  $f$ , a straight-forward solution would be to maximize the likelihood function (2.14) 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 figure 2.3. To deal with such permutations, supplementary methods for sorting them need to be

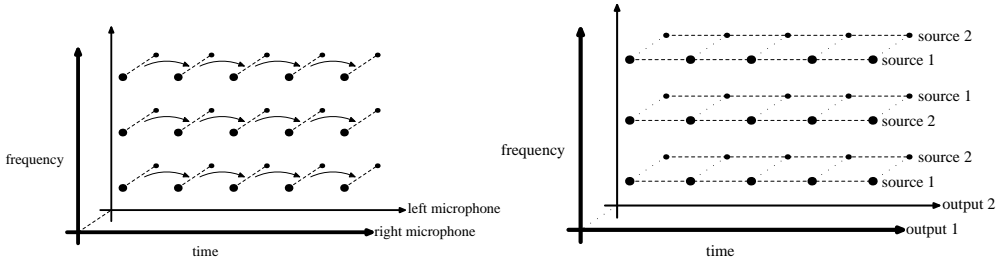


Figure 2.3: Performing separation *independently* in each frequency (depicted on the right) results in unmixed signals components whose order with respect to the corresponding source components is permuted in different frequencies (see right).

employed (e.g. Murata et al., 1998). A further disadvantage of working in each frequency separately is, that 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.

## 2.4 Constrained optimization

Due to the free field assumption (2.8) and (2.25), separation can be achieved by the matrix

$$\mathbf{W}(f) = \begin{pmatrix} w_{11}(f) & w_{12}(f) \\ w_{21}(f) & w_{22}(f) \end{pmatrix} = \begin{pmatrix} w_{11} & -w_{12} e^{-i2\pi f \tau_{12}} \\ -w_{21} e^{-i2\pi f \tau_{21}} & w_{22} \end{pmatrix} \quad (2.27)$$

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 2.3, 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 (2.22).

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

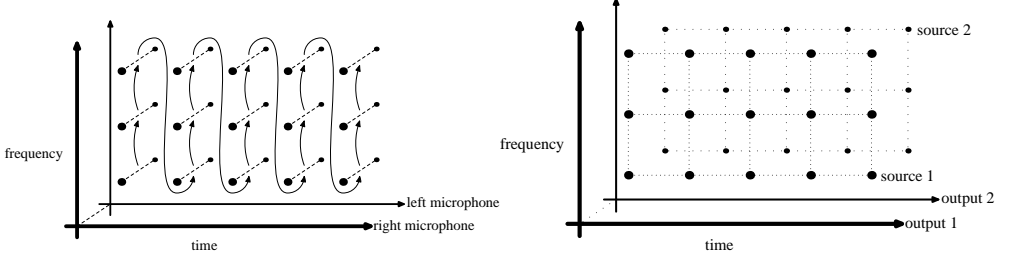


Figure 2.4: Iterating the separation algorithm *across* frequencies (left) results in the same order of unmixed components with respect to the corresponding sources for all frequencies (right).

new parameters  $(\delta w_{ij}, \delta \tau_{ij})$  is obtained from (2.18) and (2.27) as

$$\begin{aligned}\tilde{\delta} w_{ij}(T, f) &= \frac{1}{w_{ij}} \Re(w_{ij}(f) \delta w_{ij}^*(T, f)) \\ \tilde{\delta} \tau_{ij}(T, f) &= \frac{1}{2\pi f w_{ij}^2} \Im(w_{ij}(f) \delta w_{ij}^*(T, f)),\end{aligned}\quad (2.28)$$

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 (2.20) and (2.22) as

$$\tilde{\nabla} \mathbf{W}(T, f) = (\mathbf{I} + \mathbf{v}(T, f) \mathbf{u}^H(T, f)(T, f)) \mathbf{W}(f). \quad (2.22)$$

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 (2.27), calculate  $\mathbf{W}(f)$  from  $(w_{ij}, \tau_{ij})$ .
2. From (2.22), calculate the complex gradient  $\delta w_{ij}(T, f)$  of the parameter  $w_{ij}(f)$ .
3. From (2.28), 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

$$w'_{ij} = w_{ij} + \eta \delta w_{ij} \quad \tau'_{ij} = \tau_{ij} + \eta \delta \tau_{ij} \quad (2.29)$$

where  $0 < \eta \ll 1$  is the adaptation rate.

### 2.4.1 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  $\hat{u}_i(T, f)$  from (2.24).
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 depicted in figure 2.4. 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. figure 2.4).

## 2.5 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.

### Variable adaptation rate for different frequencies

As in any on-line adaptation algorithm with fixed adaptation rate, the estimate of the parameters 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, for the proposed scheme of iterating the algorithm also across frequencies, the estimates for  $(w_{ij}, \tau_{ij})$  are not only biased towards the most recent samples in time, but also to samples at nearby lower frequencies. And, as can be seen from figure 2.4, the estimates obtained at low frequencies are biased by samples from high frequencies at the previous time-step.



In our investigations, we found that this effect 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, (2.29) should be replaced by

$$w'_{ij} = w_{ij} + \frac{\eta}{f} \delta w_{ij} \quad \tau'_{ij} = \tau_{ij} + \eta \delta \tau_{ij} \quad (2.30)$$

This is justified by the theoretical result from neural network theory that a  $1/t$  decay in the learning rate yields a parameter estimate which is *not* biased towards the samples which occurred most recent in time (Sompolinsky et al., 1995). Hence, with (2.30) the estimates for  $w_{ij}$  are *not* biased by the samples which occurred at the most recent 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 (2.28) 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.

### 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 (2.2) and (2.12) 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 2.1. 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

equation (2.22), the preemphasis is not equivalent to a higher adaptation rate for high frequencies. Therefore, it is advisable to use both preemphasis and decay of the adaptation rate.

	kurtosis
time domain	5.5
frequency domain	289.8
frequency domain, high-pass filtered	21.2

Table 2.1: Kurtosis of speech in the time-domain, in the frequency domain and the kurtosis of differentiated (high-pass filtered) speech in the frequency domain.

### Speech pause detection

Speech pauses in one source which, in the examples of section 2.6, last up to 700 milliseconds, 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.

## 2.6 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 23Hz to 10kHz 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 seconds.  $\eta = 0.001$  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 pre-emphasis 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/specom>.

### 2.6.1 Artificially mixed sources

Two mono speech signals were digitally mixed in the time-domain according to the mixing system (2.5), 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.

Figure 2.5 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 fluctuate slightly during the remaining time of the recording.

Informal listening to the reconstructed signals reveals that separation is almost perfect and the remaining crosstalk is nearly inaudible. The improvement in signal separation is displayed in table 2.2. 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.

situation	signal separation (dB)
synthetic delay and gain	26.5
anechoic chamber	15.5

Table 2.2: Signal separation caused by the algorithm.

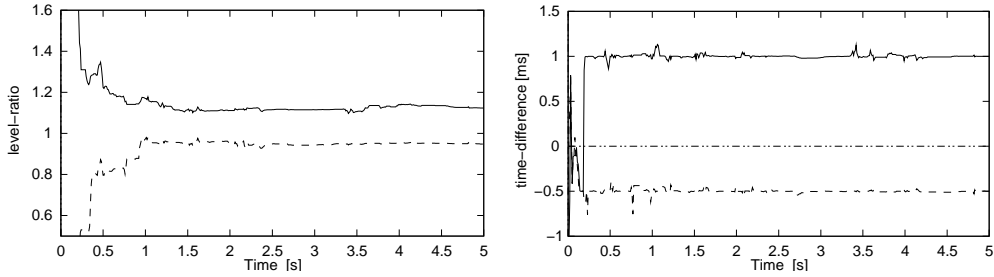


Figure 2.5: Time-course of estimated level- (left) and time differences (right) assumed by the demixing system for the separation of artificially mixed sources. For better visual presentation,  $1/w_{21}$  and  $\tau_{21}$  correspond to the solid lines, whereas  $w_{12}$  and  $-\tau_{12}$  correspond 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  $1/w_{21} = 1.11$ ,  $\tau_{21} = 1\text{ms}$ ,  $w_{12} = 0.95$ , and  $-\tau_{12} = -0.5\text{ms}$ .

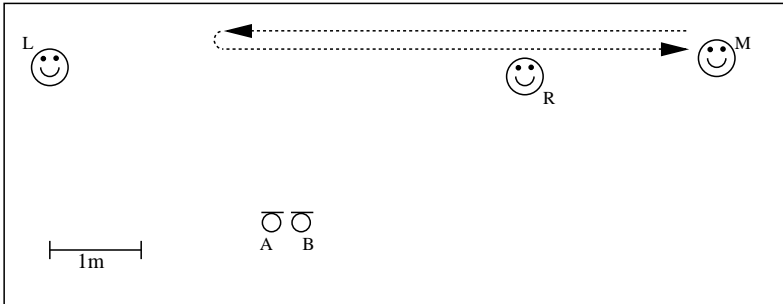


Figure 2.6: Setup for the recordings performed for evaluation. Microphones are located at positions  $A$  and  $B$ . Speaker positions for the experiment from section 2.6.2 are  $L$  and  $R$ , respectively. For the experiments of section 2.6.3, the moving speaker started at position  $M$ , followed the indicated route and returned to position  $M$ , while the standing speaker was at position  $L$ .

## 2.6.2 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 degrees to the left and 60 degrees to the right of the mid-perpendicular of the microphones, respectively. The recordings were of moderate quality, in particular, recording noise is clearly audible.

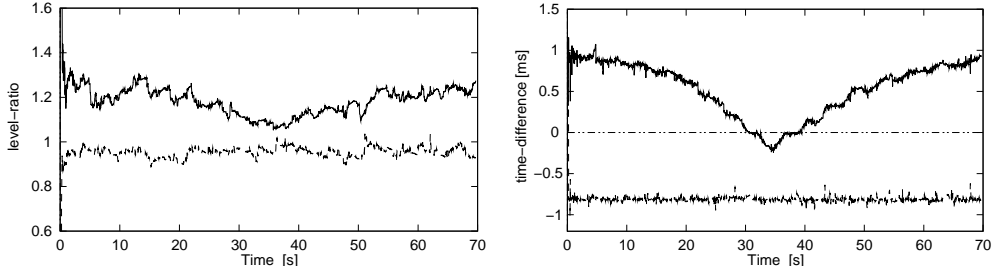


Figure 2.7: Time-course of estimated level- (left) and time differences (right) assumed by the demixing system for the separation of a moving and a standing speaker in anechoic environment. As in figure 2.5,  $1/w_{21}$  and  $\tau_{21}$  correspond to the solid lines, whereas  $w_{12}$  and  $-\tau_{12}$  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.

The distance between speakers and microphones was 3 m (cf. figure 2.6). 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-crosstalk 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-crosstalk 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 crosstalk 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-crosstalk energy of 15.3 dB, though with a window length of 85 ms. Since the longer windows favor the AMDecor algorithm by allowing for 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.

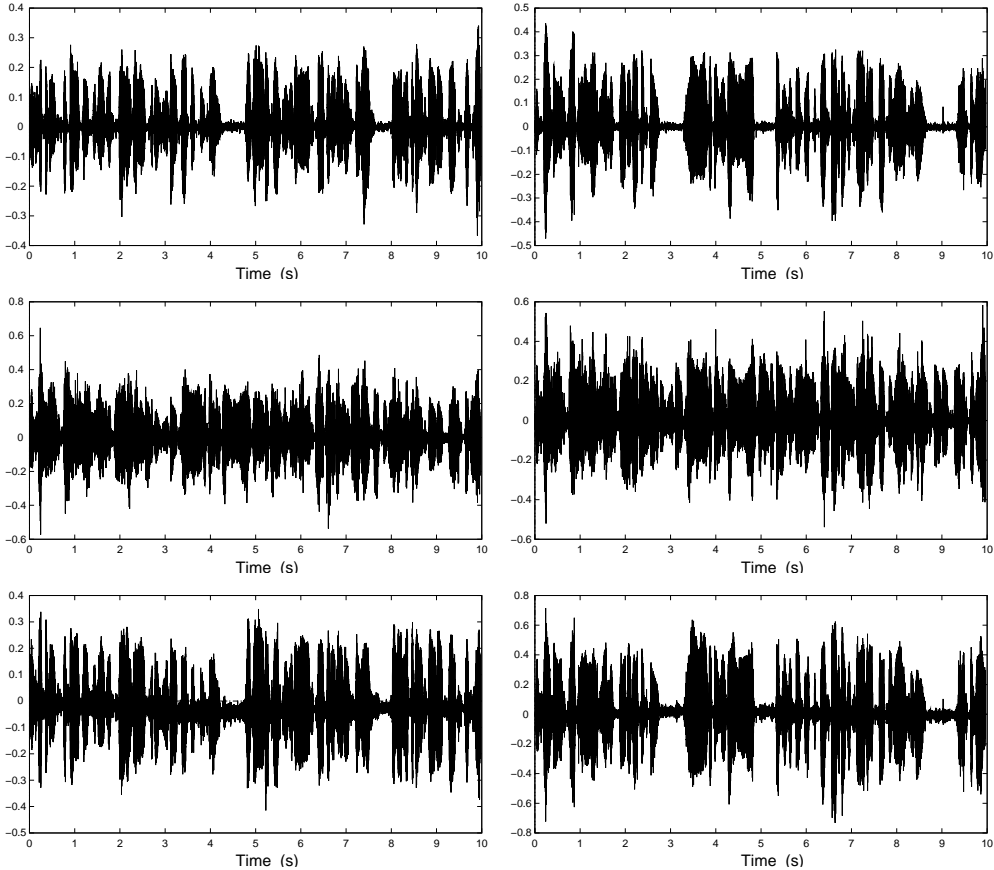


Figure 2.8: First ten seconds of speech recordings from the separation of moving sources. Top row: Original signals of the moving speaker (left) and the standing speaker (right). Center row: left and right channel of the mixed signals. Bottom row: unmixed signals obtained by the algorithm.

### 2.6.3 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 distance of 4.7 m at a position at 70 degrees to the right, walked in a straight line parallel to the microphones until he reached a position at about 30 degrees left of the microphones' mid-perpendicular, and then returned to his original position (cf. figure 2.6).

Figure 2.8 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 figure 2.7.

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 attained in less than one second. Comparing in figure 2.8 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 2.6.2) is slightly amplified. However, this does not affect the algorithm's convergence.

## 2.7 Discussion

In this paper, an algorithm for the blind separation of acoustically mixed sources was proposed. 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, blind source separation 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 2.1), 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 Torkkola (1996b) 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 (30ms 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 (2.22) which results in much faster convergence than the standard gradient (2.19). 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 equation (2.2), 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 blind source separation 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).



## 2.8 Conclusion

The current algorithm has been shown to separate two sound sources fast, with a small processing delay (about 30 msec) 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, a combination of the current approach with other algorithms appears to be necessary in hearing aid applications.



## Chapter 3

# Amplitude Modulation Decorrelation for Convolutive Blind Source Separation

### 3.1 Introduction

The problem of blind source separation (BSS) is encountered in various applications where it is desired to reconstruct multiple original source signals while only mixtures of them can be observed. Lack of additional information, e.g., about spatial locations of the sources, is indicated by the term ‘blind’. One example is the area of noise reduction algorithms where the aim is to separate out a speech signal from a background of noise or competing speech signals, in order to enhance speech intelligibility for hearing aid users or to improve the recognition rate of automatic speech recognition systems. Many further applications exist in domains such as image processing, biomedical data analysis and document analysis.

In its simplest form, the BSS setting assumes that  $M$  source signals are superimposed by a linear and instantaneous transformation to form  $N$  mixed signals, where the number of observed signals is larger or equal to the number of sources,  $N \geq M$ . A vast number of algorithms has been proposed in the literature to find estimates of the original sources (e.g. Lee, 1998a, and references therein). Their common goal is to find an unmixing matrix which transforms the mixed signals into separated signals that are by some measure as distinct as possible and resemble the original sources. Principles on which the algorithms are based rely on the sources’ second-order statistics (e.g. Molgedey and Schuster, 1994; Belouchrani et al., 1997), on their higher-order statistics (e.g. Jutten and Hérault, 1991; Comon, 1994; Bell and Sejnowski, 1995; Cardoso and Laheld, 1996) or on non-stationarity of the sources (e.g. Matsuoka et al., 1995). It is well known (Tong et al., 1991) that the original sources can only be reconstructed upto an unknown permutation and rescaling operation since independent sources remain

independent if their order is permuted or they are rescaled.

The blind source separation problem in the field of acoustics is more intricate due to the propagation in the acoustic medium. While the acoustic superposition of sound signals is still linear at normal sound pressure levels, it involves finite propagation speed and reverberation which gives rise to a convolutive mixing. Fewer algorithms have been proposed in the literature for the case of convolutive mixing, and the search for methods which are capable of signal separation for a wide range of real-world situations is still being carried on.

To separate convolutively mixed source signals, filtering of the microphone signals must be performed—instead of a multiplication in the case of non-convolutive mixing. Depending on the domain in which the filters are implemented, algorithms from the literature fall into the classes of time domain or frequency domain based algorithms. Some algorithms can be regarded as ‘hybrid’ algorithms since they implement the separation structure and the optimization cost-function in the time-domain but switch to the frequency domain during parameter adaptation (e.g. Lambert, 1996; Amari et al., 1997). Time-domain algorithms (e.g. Weinstein et al., 1993; Yellin and Weinstein, 1996; Lee et al., 1997) have to solve a non-trivial optimization problem in which all coefficients of the unmixing filters are coupled. Lindgren and Broman (1998) report that this leads to local minima which make it difficult to find the global optimum. Existence of local minima is also indicated by Ehlers and Schuster (1997) using a Monte-Carlo optimization of time-domain parameters.

Frequency domain algorithms (e.g. Capdevielle et al., 1995; Murata et al., 1998; Parra and Spence, 2000a), in contrast, are based on the property of the Fourier transformation that the convolution in the time domain results in a multiplication in the frequency domain. Thereby, the convolutive source separation problem in the time domain is transformed into  $K$  decoupled instantaneous source separation problems in the frequency domain, one for each frequency  $f = 1, \dots, K$ . After separation has been performed in the frequency domain, the separated sources are transformed back to time domain signals using, e.g., the overlap-add technique (e.g. Oppenheim and Schaefer, 1975).

The drawback of frequency domain methods is that in general local permutations arise, i.e., the sources’ spectral components are recovered in a different (unknown) order in different frequency channels, thereby making a time domain reconstruction of the source signals impossible. Several approaches, as discussed in section 3.2.2, have been proposed in the literature to deal with the problem of local permutations.

It is common to all frequency domain based algorithms found in the literature that two processing stages are used to obtain separated signals. In the first stage, a solution for the blind source separation problem in a single frequency channel is searched by taking into account signal components at the same frequency, only. In a consecutive stage, it is aimed at reordering the unmixing filters and the separated signal components such that local permutations do not occur.

In contrast, the algorithm proposed in the present paper for the separation of speech signals introduces a novel cost-function which integrates information across different frequencies in order to perform separation. Different methods for taking into account

across-frequency information have been proposed by a few authors (see Gramss, 1995; Shamsunder and Giannakis, 1997; Diamantaras et al., 2000). Unlike existing methods, the proposed algorithm employs correlations of signal envelopes at different frequencies. It is shown that this approach solves the problem of local permutations and results in a good quality of signal separation.

The outline of the paper is as follows. In section 3.2 the convolutive blind source separation problem and its invariances are formulated. Section 3.3 introduces the amplitude modulation correlation property of speech signals and section 3.4 explains its application to blind source separation. Experimental evaluation is performed in section 3.5.

Throughout the paper, the following notation is used. Vectors and matrices are printed in bold font;  $[\mathbf{A}]_{ij}$  denotes the  $(i, j)$ -element of matrix  $\mathbf{A}$ ;  $x(T, f)$  denotes the spectrogram of signal  $x(t)$ ; complex conjugation is denoted by  $x^*$ ; the expectation operator is denoted by  $E\{\cdot\}$ ; transposition of vector  $\mathbf{x}$  is denoted by  $\mathbf{x}^T$ ; transposition and complex conjugation by  $\mathbf{x}^H$ ; the imaginary unit  $\sqrt{-1}$  is denoted as  $i$ .

## 3.2 Problem Formulation

Superposition of sound sources in the acoustic medium involves echoes and time-delays and is linear at the sound pressure levels normally encountered in conversations. Hence, if  $N$  independent sound sources  $s_j(t)$ ,  $j = 1, \dots, N$  in a room are recorded by  $M$  microphones, the relation between sources and microphone signals  $x_i(t)$ ,  $i = 1, \dots, M$ , is

$$x_i(t) = \sum_{j=1}^N \sum_{t'} a_{ij}(t') s_j(t - t'), \quad (3.1)$$

where  $a_{ij}(t)$  denotes the room's impulse response from the location of source  $j$  to microphone  $i$ .

The goal of blind source separation is to recover the source signals  $s_j(t)$  from knowledge of the  $x_i(t)$  only, by approximating them with unmixed signals  $u_j(t)$ . Ideally, the unmixed signals would be identical to the source signals. However, if only the sources' mutual independence and the linearity of the mixing system (3.1) are known a priori, it is at best possible to reconstruct signals which resemble the sources upto an unknown filtering. This invariance is due to the fact that independent signals remain independent if they are transformed by invertible filters. Furthermore, the reconstructed signals  $u_j(t)$  might be arranged in a different order than the source signals  $s_j(t)$  since also the permutation of their order leaves independent signals independent.

If at least as many microphones as sources are present ( $M \geq N$ ), the reconstruction can be performed by the linear unmixing system

$$u_j(t) = \sum_{i=1}^M \sum_{t'} w_{ij}(t') x_i(t - t'). \quad (3.2)$$

In the present paper, the approach is pursued to implement the unmixing system (3.2) in the frequency domain.

### 3.2.1 Frequency domain formulation

The standard time-frequency representation used for the analysis and filtering of speech signals is the spectrogram. The spectrogram  $x_i(T, f)$  corresponding to signal  $x_i(t)$  is obtained by computing the windowed short time Fourier transformation (STFT) of  $x_i(t)$  which is defined as

$$x_i(T, f) = \sum_{t=0}^{2K-1} x_i(t + T) h(t) e^{-2\pi i f t / (2K)}. \quad (3.3)$$

Here,  $h(t)$  denotes the windowing function,  $f = 1, 2, \dots, K$  denotes frequency and  $T = 0, \Delta T, 2\Delta T, \dots$  is the time-index of the spectrogram. Similarly,  $s_j(T, f)$ ,  $u_j(T, f)$ ,  $a_{ij}(f)$  and  $w_{ij}(f)$  are the STFTs of  $s_j(t)$ ,  $u_j(t)$ ,  $a_{ij}(t)$  and  $w_{ij}(t)$ , respectively. Note that  $a_{ij}(f)$  and  $w_{ij}(f)$  do not depend on time since the mixing system in (3.1) is stationary.

The convolutional blind source separation problem (3.1) can be recast in the frequency domain using the spectrogram. Provided that the frames for computing the short-time spectra are sufficiently long, the linear convolution in (3.1) can be approximated by the circular convolution in (3.3). In consequence, (3.1) factorizes into a set of  $K$  equations, each corresponding to a matrix multiplication in a single frequency band,

$$\mathbf{x}(T, f) = \mathbf{A}(f) \mathbf{s}(T, f). \quad (3.4)$$

$\mathbf{x}(T, f) = [x_1(T, f), \dots, x_N(T, f)]^T$  denotes the vector of the mixed spectrograms and  $\mathbf{s}(T, f) = [s_1(T, f), \dots, s_M(T, f)]^T$  is the corresponding vector for the source signals. The  $(i, j)$ -element  $a_{ij}(f)$  of matrix  $\mathbf{A}(f)$  denotes the room transfer function from source  $j$  to microphone  $i$ .

Hence, the convolutional source separation problem (3.1) has been transformed to a set of  $K$  linear instantaneous source separation problems for complex variables.

Unmixed spectrograms  $\mathbf{u}(T, f) = [u_1(T, f), \dots, u_M(T, f)]^T$ ,

$$\mathbf{u}(T, f) = \mathbf{W}(f) \mathbf{x}(T, f), \quad (3.5)$$

are obtained by finding matrices  $\mathbf{W}(f)$  with  $(i, j)$ -elements  $w_{ij}(f)$  such that the unmixed signals are independent. It is well known (Tong et al., 1991) that it is only possible to reconstruct the source signals subsequent to a permutation and rescaling,

$$\mathbf{u}(T, f) = \mathbf{P}(f) \mathbf{D}(f) \mathbf{s}(T, f), \quad (3.6)$$

where  $\mathbf{P}(f)$  and  $\mathbf{D}(f)$  denote a permutation and diagonal matrix, respectively. After separation has been performed in the frequency domain, separated time-domain signals are obtained by transforming the separated spectrograms back to the time-domain using the overlap-add technique (e.g. Oppenheim and Schaefer, 1975).

### 3.2.2 Invariances

If (3.5) is interpreted as  $K$  independent source separation problems, one for each frequency  $f = 1, \dots, K$ , then the matrices  $\mathbf{P}(f)$  and  $\mathbf{D}(f)$  are obtained independently

for each frequency. Hence, both will in general be different at different frequencies. However, if  $\mathbf{P}(f)$  is different for frequencies  $f$  and  $f'$ , this results in a different ordering of the sources' spectral components in the two frequency channels. We denote such different ordering in different frequencies as 'local permutations', for an illustration see figure 3.1. If reconstructed time-domain signals are computed from separated spectrograms with inherent local permutations, the inverse Fourier transformation mixes spectral components belonging to different sources. Even if perfect separation has been accomplished within each frequency channel, the effect of local permutations therefore is very poor or no separation at all.

Therefore, it must be ensured that the reconstructed signals' ordering with respect to the original signals is the same in every frequency channel, i.e.,

$$\mathbf{P} = \mathbf{P}(f_1) = \mathbf{P}(f_2) = \dots = \mathbf{P}(f_K). \quad (3.7)$$

Note that even after all local permutations have been eliminated, the 'global permutation'  $\mathbf{P}$ , which is independent of frequency  $f$ , is still present and remains by principle unknown. However, since it is constant over frequency, it does not hinder time-domain reconstruction of the separated signals, as illustrated in figure 3.1.

Several methods have been proposed to correct local permutations in the unmixed signals. Some authors propose to exploit source signal properties to correct local permutations. Murata et al. (1998) minimize the overlap between the unmixed signals' broad-band and narrow-band envelopes to find the correct order of reconstructed spectral components. Mejuto et al. (2000) compute fourth order cross cumulants to realign the unmixed components of wireless communication signals.

Alternatively, constraints on the unmixing filters have been proposed to avoid local permutations. Attias and Schreiner (1998) and Parra and Spence (2000a) propose to restrict the allowed separating filters to those with a limited length of their impulse response, resulting in smooth transfer functions in the frequency domain. However, this goes at the expense of possibly lower signal separation due to the limited filter length.

It has been proposed by Capdevielle et al. (1995) and Servi re (1999) to use properties of the discrete Fourier transform to correct permutations. However, this method is computationally expensive since it requires computation of short-time-spectra with a window shift of one sample.

Finally, some authors have combined the good convergence properties of frequency domain methods with the observed robustness of time domain methods with respect to local permutations. Ehlers and Schuster (1997) and Lee et al. (1998) use a sequential scheme of frequency-domain and time-domain algorithms to perform separation.

In contrast to previous methods, the approach pursued in the present paper is to define a cost function that inherently avoids local permutations.

To fix the invariance with respect to arbitrary rescaling  $\mathbf{D}(f)$  in each frequency channel, two different approaches exist in the literature.

The first is back-projection of the (rescaled) separated signals to the microphone signals, as proposed by Murata et al. (1998). As a result, it is possible to reconstruct the separated source signals as they were recorded at the microphones in the absence

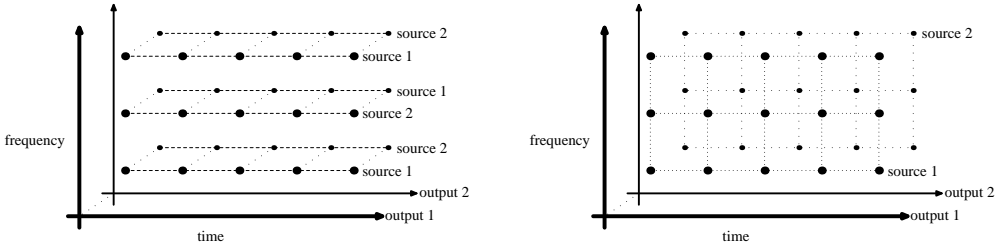


Figure 3.1: Effect of local permutations. Left: At the medium frequency band, the ordering of the source components is permuted relative to the other frequency bands. Right: Same ordering of source components in all frequency channels.

of the other sources. Under the assumption of an ideal situation this method obtains separated signals with little distortion. However, the step of back-projecting the signals leads to problems in the presence of even small levels of recordings noise, since it involves the inversion of the matrix  $\mathbf{W}(f)$  which might be ill-conditioned at several frequencies. In some experiments Anemüller et al., this approach has the side-effect of adding additional reverberation to the signals and slightly enhancing recording noise. The second approach with respect to the rescaling problem consists of imposing constraints on the matrix  $\mathbf{W}(f)$ . The simplest and probably most popular constraint is to simply set the diagonal elements of  $\mathbf{W}(f)$  equal to one. This leads to less problems with noise. However, it is still possible that recording noise gets amplified at several frequencies. Alternatively, the norm of the rows of  $\mathbf{W}(f)$  can be constrained to unity. This guarantees that noise does not get amplified, however at the expense of slightly distorting the signals. In the case of a complex matrix  $\mathbf{W}(f)$  this constraint still leaves undetermined a factor for the complex phase in each row of  $\mathbf{W}(f)$ . This indeterminacy can be fixed by, e.g., setting the imaginary part of the diagonal elements of  $\mathbf{W}(f)$  to zero.

### 3.3 Amplitude Modulation Correlation

#### 3.3.1 Structure in Speech

Speech, as an important means of communication, is generally believed to contain information in the time-frequency distribution of the signal energy, a fact that gives rise to the strong amplitude modulation which is present in speech. It has been found by several researchers that neither the transmitted information nor the amplitude modulation are independent in different spectral bands.

From the speech processing literature, it is known that the amplitudes in different frequencies are interrelated. In the context of speech enhancement, Kollmeier and Koch (1994) employ an analysis based on the amplitude modulation spectrogram and observe that vowels are characterized by clusters of signal energy at different positions with respect to frequency and modulation-frequency. The property of acoustical ob-



jects to be distributed across frequency is confirmed by Tchorz and Kollmeier (2000) and exploited for the robust estimation of the signal-to-noise ratio of noise contaminated speech signals. Michaelis et al. (1997) find high correlations between Hilbert envelopes computed from different frequency bands of speech signals in the context of speech quality assessment.

For speech signals, semantic structure and the physiology of speech production are regarded as the origin for the observed similarities in different frequency. The composition of speech from small elements — phonemes, syllables and words — which are separated by minima in the signal amplitude directly contributes to interrelated modulation in different frequencies, most prominent at, but not limited to the typical modulation frequency of four Hertz. Vowels, in turn, are themselves characterized by simultaneous spectral peaks at the formant frequencies (Paulus, 1998). The main energy source for speech production is the glottis which emits a broadband sound with spectral peaks at the harmonics of the speaker's pitch frequency. Therefore, any modulation of the glottis excitation affects all frequencies simultaneously. Subsequent filtering by the vocal tract involves a smooth transfer function so that any change in the shape of the vocal tract also alters the signal amplitude at multiple frequencies simultaneously.

The human auditory system appears to be tuned to such interrelated activity in different spectral bands, as evidence from psychoacoustic experiments suggests. The effect of 'comodulation masking release' (Hall et al., 1984) may in part be explained on the basis of across-frequency interactions in the auditory system (Verhey et al., 1999). Furthermore, the improvement in the prediction of speech intelligibility accomplished by taking into account redundant information at different frequencies (Steeneken and Houtgast, 1999) may be regarded as an indication for across-frequency processing in the auditory system, as well.

The basis for the algorithm presented in this paper is formed by the described speech signal property of highly interrelated amplitude modulation in different and even distant frequency channels. This property is termed *amplitude modulation (AM) correlation* (Anemüller and Kollmeier, 2000). Quantitative analysis of AM correlation is based on the amplitude spectrogram which is obtained from the complex valued spectrogram by preserving only the amplitude and discarding the phase information. To illustrate AM correlation, figure 3.2 displays the amplitude spectrogram of a speech sample. Note that many elements of this image change smoothly over both time and — more important for the present purpose — frequency, and that even distant frequency channels exhibit related changes in amplitude.

### 3.3.2 Amplitude Modulation Correlation

A natural way to measure the synchrony of the amplitude modulation in two frequency channels of two (possibly different) signals is to compute the correlation between the corresponding frequency specific signals envelopes. Due to the low-pass filtering property of the magnitude operation, the envelope correlation can be computed as the correlation of the time-courses in two frequency channels of amplitude spectrograms.

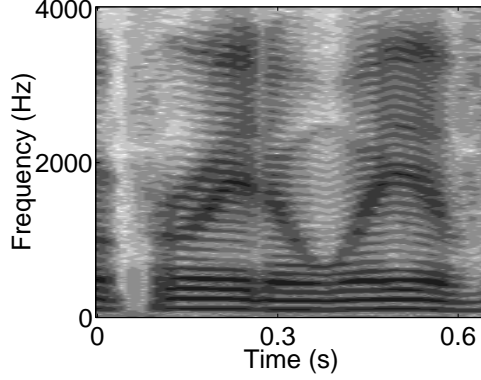


Figure 3.2: Amplitude spectrogram of a speech signal. Time is plotted on the ordinate and frequency on the abscissa. The signal's intensity for each combination of time and frequency is given on a grayscale where white denotes minimum values and black denotes maximum values. The respective value on the grayscale is proportional to the logarithm of the intensity which spans a total range of 80 dB.

The amplitude modulation correlation (*AMCor*)  $c(x(T, f_k), y(T, f_l))$  between the frequency channel  $f_k$  of spectrogram  $x(T, f)$  and frequency channel  $f_l$  of spectrogram  $y(T, f)$  is defined as

$$c(x(T, f_k), y(T, f_l)) = E\{|x(T, f_k)| |y(T, f_l)|\} - E\{|x(T, f_k)|\} E\{|y(T, f_l)|\} \quad (3.8)$$

Note that the AMCor is a real valued quantity since it is computed from the magnitude spectrogram. In this respect it differs from the notion of coherence, which is computed from the complex valued spectrum. Computing the complex correlation of two different frequency channels results in very low correlations since the STFT has the property of approximately decorrelating the Fourier coefficients at different frequencies. Therefore, the complex correlation is not appropriate to capture the properties of speech signals discussed above.

By computing the AMCor for all possible pairs of frequencies  $(f_k, f_l)$  of a single signal  $s(T, f)$ , the AM auto-covariance matrix  $\mathbf{C}(s)$  is obtained which is of size  $K \times K$  and whose  $(k, l)$ -element is

$$[\mathbf{C}(s, s)]_{kl} = c(s(T, f_k), s(T, f_l)). \quad (3.9)$$

We use  $\mathbf{C}(s)$  as short-hand notation for  $\mathbf{C}(s, s)$ .

The AM auto-covariance matrix corresponding to the first signal from figure 3.3 is displayed in figure 3.3 (bottom row, left). As expected, particularly high values of AMCor are reached for nearby frequencies (i.e., near the diagonal), and high values of AMCor can also be found for distant frequencies.

It is expected that the amplitude modulation of two independent speech signals is unrelated and therefore the corresponding amplitude modulation correlation is zero.

To this end the AM correlation is computed between the amplitudes  $|s_1(T, f_k)|$  at frequency channel  $f_k$  of source  $s_1$  and the amplitudes  $|s_2(T, f_l)|$  at frequency channel  $f_l$  of source  $s_2$ . Performing this operation for all possible pairs of frequencies  $(f_k, f_l)$  yields the AM cross-covariance matrix  $\mathbf{C}(s_1, s_2)$  which is of size  $K \times K$  with  $(k, l)$ -element

$$[\mathbf{C}(s_1, s_2)]_{kl} = c(s_1(T, f_k), s_2(T, f_l)). \quad (3.10)$$

The AM cross-covariance matrix computed from the two speech signals in figure 3.3 is displayed in figure 3.3 (bottom row, right). As expected, the AM correlation across the two different signals is close to zero compared to the AM auto-covariance matrix from figure 3.3 (bottom row, left).

It could be argued that in the case of two sentences spoken by the same speaker or in the same language, the AM cross-covariance might not vanish due to speaker- or language-characteristics. Since the two speech signals in figure 3.3 are spoken by the same speaker in the same language, the corresponding AM cross-covariance matrix in figure 3.3 (bottom row, right) also demonstrates that, apart from small residual correlations, this is not the case.

Similarities in two related, but not identical signals result in high values in the corresponding AM cross-covariance matrix. However, the correlations found are not as high as for the corresponding auto-covariance matrices computed from the individual signals. This is displayed in figure 3.11, corresponding to signals used for evaluation in section 3.5.3, which shows the auto- and cross-covariance matrices computed from two microphone signals of two simultaneously speaking persons.

## 3.4 Source Separation by Amplitude Modulation Decorrelation

Since amplitude modulation correlation provides a measure of the similarity of speech signals, it can be used as a criterion for blind source separation by requiring that the AM cross-covariance matrix of the reconstructed signals vanishes. It is demonstrated that this requirement of amplitude modulation *decorrelation* of the unmixed signals solves both the blind source separation problem and the problem of local permutations simultaneously.

### 3.4.1 Separation

For the proposed algorithm to be applicable, the source signals are assumed to have the AM auto-covariance property

$$[\mathbf{C}(s_i)]_{kl} \neq 0 \quad \forall i \forall k, l. \quad (3.11)$$

Since different sources are independent, the AM cross-covariance for any pair  $(s_i, s_j)$ ,  $i \neq j$ , of different sources vanishes,

$$[\mathbf{C}(s_i, s_j)]_{kl} = 0 \quad \forall i \neq j \forall k, l. \quad (3.12)$$

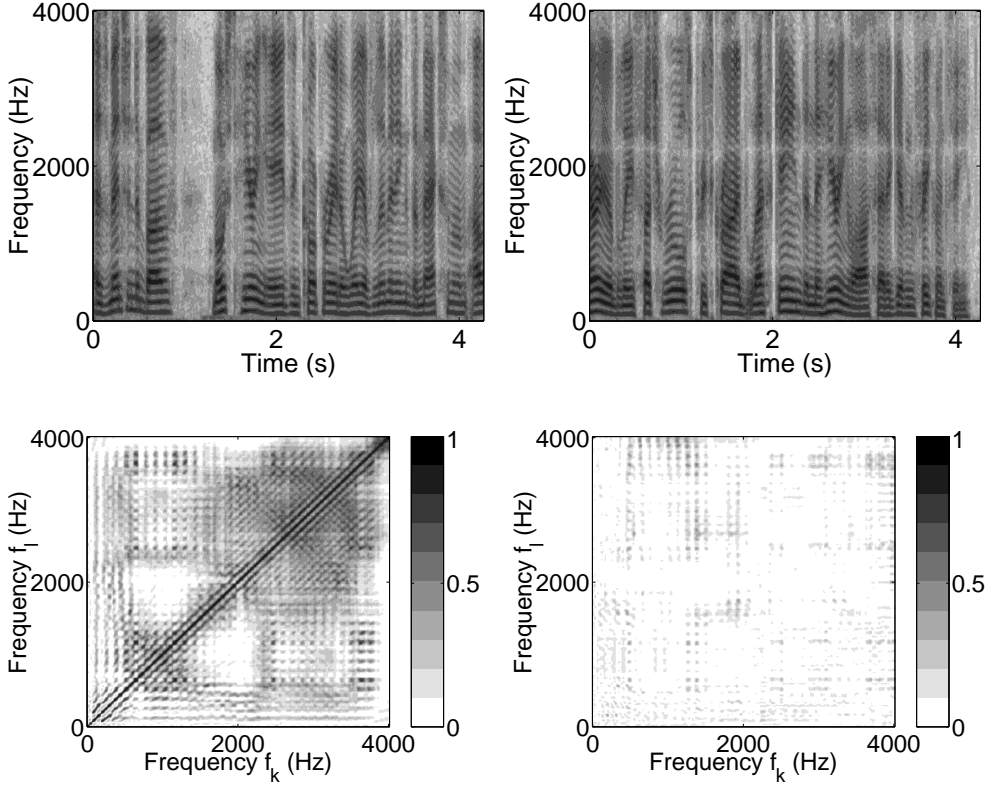


Figure 3.3: Spectrograms and the corresponding amplitude modulation auto- and cross-covariance matrices. Top row: Spectrograms of two different speech signals, spoken by the same male speaker. Bottom row, left: AM *auto*-covariance matrix of the first (top row, left) source spectrogram. Bottom row, right: AM *cross*-covariance matrix of both spectrograms. The frequencies for which the correlations have been computed are plotted on the ordinate and abscissa. Hence, each point in the diagram displays the correlation for the corresponding pair of frequencies. For better visualization the normalized correlation is displayed.

Following the mixing by the acoustic mixing model (3.4), contributions from each source are in general recorded at all microphones. Hence it follows from (3.11) that the AM cross-covariance for all possible pairs  $(x_i, x_j)$  of microphone signals is non-zero,

$$[\mathbf{C}(x_i, x_j)]_{kl} \neq 0 \quad \forall i, j \forall k, l. \quad (3.13)$$

For the unmixed signals to be independent, the source signals' AM decorrelation property (3.12) must be fulfilled by the unmixed signals,

$$[\mathbf{C}(u_i, u_j)]_{kl} = 0 \quad \forall i \neq j \forall k, l, \quad (3.14)$$

corresponding to zero AM cross-covariance for all possible pairs  $(u_i, u_j)$ ,  $i \neq j$ , of *different* unmixed signals.

Of course, the AM *auto*-covariance for each reconstructed signal  $u_i$  is non-zero,

$$[\mathbf{C}(u_i)]_{kl} \neq 0 \quad \forall i \forall k, l. \quad (3.15)$$

Equation (3.14) is termed the 'AM decorrelation principle'.

If in practice complete separation is not possible, the AM cross-covariance matrices (3.14) are minimized, making the reconstructed signals as independent as possible. Clearly, (3.14) constitutes a necessary condition for source separation. Intuitively, any signal components from a particular source are characterized by their specific AM auto-covariance (3.11). If components originating from source  $s_i$  are present in two different unmixed signals,  $u_i$  and  $u_j$ ,  $i \neq j$ , the source's AM auto-covariance results in the non-zero AM cross-covariance of the unmixed signals. Hence, the cross-talk of source  $s_i$  to two unmixed signals results in the violation of the AM decorrelation principle. While a rigorous prove of the intuitive explanation is difficult due to the non-linear magnitude operation in (3.8), experiments performed with synthetic data and speech signals demonstrate that in practice the AM decorrelation principle (3.14) is sufficient in order to achieve a good quality of source separation.

### 3.4.2 Effect of permutations

In this section it is shown that the AM decorrelation principle (3.14) also has the desirable property of avoiding local permutations.

Without loss of generality, it is assumed that a local permutation at frequency  $f'$  relative to all other frequencies  $f \neq f'$  occurs. Ignoring the scaling ambiguity which is irrelevant at the moment, the local permutation results in

$$u_i(T, f) = \begin{cases} s_i(T, f) & \text{if } f \neq f' \\ s_{\pi(i)}(T, f) & \text{if } f = f' \end{cases} \quad (3.16)$$

where  $\pi(i)$  denotes the permutation of the indices  $i$ .

Computing the AM cross-covariance for any pair  $(u_i, u_j)$ ,  $i \neq j$ , of *different* outputs gives

$$\forall i \neq j : \quad [\mathbf{C}(u_i, u_j)]_{kl} = \begin{cases} 0 & \text{if } f_k = f_l = f' \\ 0 & \text{if } f_k \neq f' \text{ and } f_l \neq f' \\ \delta_{\pi(i),j} [\mathbf{C}(s_i)]_{kl} & \text{if } f_k = f' \text{ and } f_l \neq f' \\ \delta_{\pi(j),i} [\mathbf{C}(s_i)]_{kl} & \text{if } f_k \neq f' \text{ and } f_l = f' \end{cases} \quad (3.17)$$

where  $\delta_{ij}$  denotes the Kronecker delta. Result (3.17) is in contradiction to the AM decorrelation principle (3.14) which requires a vanishing cross-covariance for all  $k, l$  in (3.17).

Furthermore, computing the AM *auto*-covariance for any single reconstructed output  $u_i$  gives

$$\forall i : \quad [\mathbf{C}(u_i)]_{kl} = \begin{cases} [\mathbf{C}(s_{\pi(i)})]_{kk} & \text{if } f_k = f_l = f' \\ [\mathbf{C}(s_i)]_{kl} & \text{if } f_k \neq f' \text{ and } f_l \neq f' \\ \delta_{\pi(i),i} [\mathbf{C}(s_i)]_{kl} & \text{if } f_k = f' \text{ and } f_l \neq f' \\ \delta_{\pi(i),i} [\mathbf{C}(s_i)]_{kl} & \text{if } f_k \neq f' \text{ and } f_l = f'. \end{cases} \quad (3.18)$$

In contrast, the auto-covariance (3.15) is non-zero for all  $i, k, l$ .

Hence, the local permutation violates both conditions (3.14) and (3.15) simultaneously. Conversely, if any of conditions (3.14) and (3.15) is fulfilled, local permutations are *not* present. It is noted that conditions (3.14) and (3.15) apply to any number of sources and are not limited to  $M = 2$ . The effect of local permutations on the covariance matrices  $\mathbf{C}(u_i)$  and  $\mathbf{C}(u_i, u_j)$  is visualized in figure 3.4.

While (3.15) can solely be used to detect local permutations, (3.14) can serve as a criterion for both separation and absence of local permutations. Therefore, (3.14) is used as the single condition which needs to be fulfilled in order to achieve both separation and correct ordering of unmixed components simultaneously. It is observed experimentally that the additional use of criterion (3.15) does not have a significant effect on the results.

### 3.4.3 AM decorrelation algorithm

To achieve separation without local permutations, the AM decorrelation algorithm is proposed. It is based on the definition of a cost-function whose minimization with respect to the separating matrices results in a solution that fulfills the AM decorrelation principle (3.14).

Clearly, for a particular pair of *different* unmixed signals  $(u_i, u_j)$ ,  $i \neq j$ , the corresponding condition in equation (3.14) is fulfilled if the squared Frobenius norm

$$\|\mathbf{C}(u_i, u_j)\|_{\text{Fro}}^2 = \sum_{k,l} [\mathbf{C}(u_i, u_j)]_{kl}^2 \quad (3.19)$$

of the cross-covariance matrix  $\mathbf{C}(u_i, u_j)$  is minimized, since the Frobenius norm is bounded from below by zero and acquires its minimum if and only if  $\mathbf{C}(u_i, u_j) = 0$ .

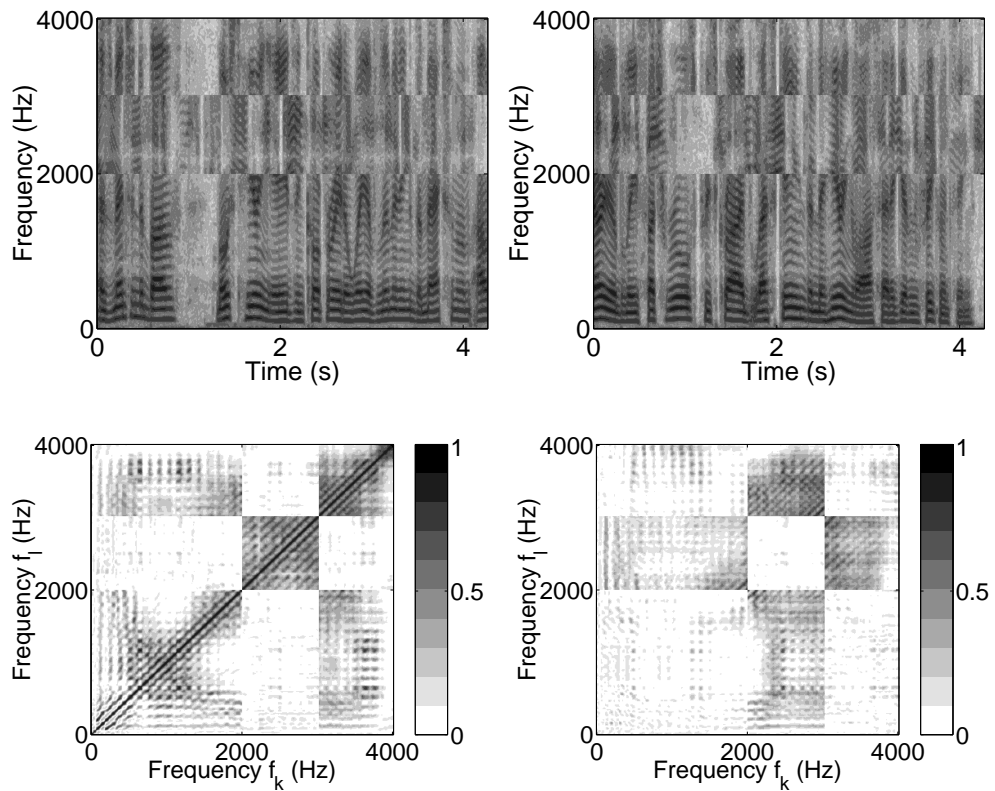


Figure 3.4: The effect of local permutations on the AM covariance matrices. Top row: The two spectrograms from figure 3.3, but with the frequency range from 2kHz to 3kHz permuted between the two signals. Bottom row, left: Normalized AM *auto*-covariance matrix of the first (top row, left) source spectrogram. Bottom row, right: Normalized AM *cross*-covariance matrix of both spectrograms.

Since the AM decorrelation principle (3.14) requires AM decorrelation for all possible pairs of different outputs  $(u_i, u_j)$ ,  $i \neq j$ , condition (3.14) as a whole is fulfilled if and only if the sum  $H$  of the squared Frobenius norms of all outputs pairs with  $i \neq j$ ,

$$H = \sum_{i \neq j} \|\mathbf{C}(u_i, u_j)\|_{\text{Fro}}^2 = \sum_{i \neq j} \sum_{k,l} [\mathbf{C}(u_i, u_j)]_{kl}^2, \quad (3.20)$$

is minimized.

It is noted that equation (3.20) cannot be written as a sum of energy terms which are computed from each frequency channel separately, as is the case in other approaches proposed in the literature. Rather, the contributions of all frequencies to the total energy  $H$  are coupled.

Since the unmixed signals  $u_i(T, f)$  are determined from the known mixed signals  $x_i(T, f)$  by the linear demixing model

$$\mathbf{u}(T, f) = \mathbf{W}(f) \mathbf{x}(T, f), \quad (3.5)$$

the parameters to be optimized in order to minimize  $H$  are the matrices  $\mathbf{W}(f)$ ,  $f = 1, \dots, K$ .

A gradient based optimization algorithm is employed which relies on the gradient of  $H$  with respect to the matrices  $\mathbf{W}(f)$ . Note that while the cost-function (3.20) is real-valued (since it is defined via the amplitudes of the unmixed signals), all variables in the demixing model (3.5) are complex. Therefore, the derivatives with respect to the parameters  $w_{ij}(f)$  are also complex-valued and are computed as

$$\delta w_{ij}(f) = \frac{\partial}{\partial \Re w_{ij}(f)} H + i \frac{\partial}{\partial \Im w_{ij}(f)} H, \quad (3.21)$$

where  $\partial/\partial \Re w_{ij}(f)$  and  $\partial/\partial \Im w_{ij}(f)$  denote differentiation with respect to the real- and imaginary-part of  $w_{ij}(f)$ , respectively. The  $(i, j)$ -element of the gradient matrix  $\nabla \mathbf{W}(f)$  is

$$[\nabla \mathbf{W}(f)]_{ij} = \delta w_{ij}(f) \quad (3.22)$$

Evaluation of (3.21) and (3.22) yields the expression

$$\nabla \mathbf{W}(f) = 2 E \{ \boldsymbol{\theta}(T, f) \mathbf{x}^H(T, f) \} \quad (3.23)$$

with the abbreviations

$$\begin{aligned} \boldsymbol{\theta}(T, f) &= [\theta_1(T, f), \dots, \theta_M(T, f)]^T \\ \theta_i(T, f) &= \frac{u_i(T, f)}{|u_i(T, f)|} \sum_{i \neq j} \sum_{f'} c(u_i(T, f), u_j(T, f')) \xi_j(T, f') \\ \xi_j(T, f') &= |u_j(T, f')| - E \{ |u_j(T, f')| \}. \end{aligned}$$

It is noted that the gradient  $\nabla \mathbf{W}(f)$  at frequency  $f$  depends on the unmixed signals  $u_i(T, f')$  for all frequencies  $f' = 1, \dots, K$  and hence also on the unmixing matrices



$\mathbf{W}(f')$  for frequencies  $f' \neq f$ . Consequently, the separation parameters for all frequencies are coupled.

The optimization has to be done subject to one of the constraints for  $\mathbf{W}(f)$  mentioned in section 3.2.2. Otherwise the algorithm would converge to the trivial solution  $\mathbf{W}(f) = 0$  for all  $f$ . The simplest solution is to set the diagonal entries  $\mathbf{W}(f)$  to one and the corresponding entries of the gradient (3.23) to zero. Another possibility is to normalize the rows of  $\mathbf{W}(f)$  to unit norm and the imaginary parts of the diagonal elements to zero. Both normalizations have been tested experimentally and lead to successful separation. Most robust separation performance over a variety of situations was obtained by employing a variant of the whitening preprocessing step which is standard in several blind source separation algorithms (e.g. Comon, 1994; Cardoso and Souloumiac, 1996; Murata et al., 1998). Details are given in appendix A.1.

### 3.4.4 Optimization scheme

Minimization of (3.20) constitutes an optimization problem in a high-dimensional space. In general, gradient based optimization methods converge to the nearest local minimum of the cost-function, which is not necessarily the global optimum. Analysis of the convergence on the error surface is therefore beneficial in order to assure convergence to the vicinity of the global minimum for a variety of situations.

Experiments have shown that ‘direct’ optimization, i.e., optimization with respect to  $\mathbf{W}(f)$  for all frequencies  $f = 1, \dots, K$  simultaneously, in many situations leads to convergence to local minima which are far from the solution of signal separation. The local minima partly seem to be due to separating solutions with local permutations, which are locally optimal, but not globally.

Therefore, a particular optimization scheme is used in which the  $\mathbf{W}(f)$  are sequentially optimized, one frequency after another. Sequential optimization of the  $\mathbf{W}(f)$  for different frequencies corresponds to holding most of the parameters constant and moving only in the direction of a few coordinates in parameter space. If convergence to the nearest local minimum is required, this is an optimization strategy which can lead to poor convergence speed (see Press et al., 1992). In the present case, however, the purpose of this procedure is to ensure convergence to the vicinity of the global optimum and our experiments prove that it is the appropriate method for the problem at hand.

In detail, the procedure is as follows:

1. Determine the frequency channel,  $f_{\text{start}}$ , with the highest signal energy. It is assumed that in the presence of recording noise, which is always present in real-world recordings, we find a particularly high signal-to-noise ratio at this frequency. Set the current frequency to  $f_{\text{curr}} = f_{\text{start}}$ .
2. Minimize (3.20) with respect to  $\mathbf{W}(f_{\text{curr}})$ , holding all other  $\mathbf{W}(f)$ ,  $f \neq f_{\text{curr}}$ , constant.
3. Iteratively, increase  $f_{\text{curr}}$  to the next higher frequency and minimize (3.20) with respect to  $\mathbf{W}(f_{\text{curr}})$ , holding all other  $\mathbf{W}(f)$ ,  $f \neq f_{\text{curr}}$ , constant.

4. If  $f_{\text{curr}}$  has reached the highest frequency, set  $f_{\text{curr}} = f_{\text{start}} - 1$  and perform optimization for  $\mathbf{W}(f_{\text{curr}})$ .
5. Iteratively, decrease  $f_{\text{curr}}$  by one and perform optimization for that frequency until the lowest frequency is reached.
6. Three iterations (*sweeps*) through all frequencies usually suffice to achieve good separation.
7. Convergence is further improved if during the first sweep energy (3.20) and gradient (3.23) are evaluated taking into account only correlations with those frequencies which have already been optimized. Using this procedure, the first frequency to be optimized converges to some random permutation and, due to the cross frequency terms in (3.20) and (3.23), the consecutively optimized frequencies are driven to converge to the same permutation.

## 3.5 Experimental evaluation

To demonstrate the capabilities of the proposed AMDecor algorithm, several experiments are performed. Synthetic source signals are employed to show that the algorithm successfully separates signals which are inseparable for algorithms working in isolated frequency channels. Quality of separation and ability to avoid local permutations are addressed using real-room recordings from several acoustic situations. Performance on publicly available benchmark datasets is compared with results from previous blind source separation algorithms. Sound files corresponding to the experiments presented can be obtained from <http://medi.uni-oldenburg.de/demo/ane/diss>.

### 3.5.1 Synthetic data

The aim of this section is to demonstrate that the proposed algorithm with its coupling across frequencies has the ability to separate signals which cannot be separated by algorithms which attempt to perform separation within each frequency separately.

To this end, synthetic source spectrograms  $s_i(T, f)$  are constructed which have the property that within each frequency channel they contain purely independent and identically distributed (i.i.d.) noise with Gaussian distribution. However, looking across different frequencies, the signals exhibit a common amplitude modulation. Therefore, the AM auto-covariance matrix for a single synthetic source spectrogram is non-zero. The signals cannot be separated by looking at a single frequency channel only, since neither cues from higher order statistics (Comon, 1994), nor from autocorrelation information (Molgedey and Schuster, 1994), nor from non-stationarity in the data (Matsuoka et al., 1995) that could be used to perform separation are present.

The construction of the synthetic source signals is performed as follows. Random Gaussian i.i.d. data  $\zeta_i(T, f)$  of variance one and mean zero is generated for each source  $s_i(T, f)$ . Amplitude modulation correlation is introduced by multiplying all  $\zeta_i(T, f)$

for a particular source  $i$  with a modulator  $\mu_i(T) \geq 0$  which is constant over frequency  $f$  but uncorrelated for different sources,

$$E \{ \mu_i(T) \mu_j(T) \} - E \{ \mu_i(T) \} E \{ \mu_j(T) \} = 0 \quad \text{for } i \neq j. \quad (3.24)$$

Since the multiplication with  $\mu_i(T)$  alters the probability density function (pdf) of  $\zeta_i(T, f)$ , a non-linear function  $g_{i,f}(\cdot)$  is applied for each source  $i$  and frequency channel  $f$  which transforms the data in each frequency channel such that the result has Gaussian pdf with variance one and mean zero.

Hence, the synthetic source spectrograms are defined as

$$s_i(T, f) = g_{i,f}(\mu_i(T) \zeta_i(T, f)).$$

The non-linear function  $g_{i,f}(\cdot)$  is implemented by a ranking transformation of the product time-series  $\mu_i(T) \zeta_i(T, f)$  to order numbers and a subsequent transformation of the order numbers to a Gaussian random variable with unit variance and mean zero. Since for each source  $i$  and frequency  $f$  the auto-correlation function of  $s_i(T, f)$  is a delta-peak and since the pdf is Gaussian, the signals do not contain cues related to auto-correlation information or higher-order statistics that could be used for separation.

To ensure that also information from non-stationarity of the signals is not available as a cue for signal separation, the modulator has to be chosen properly. It could, e.g., be chosen to be a slowly varying, possibly periodic function, say,  $\mu_i(T) = 1 + \sin(iT)$ . However, this would result in fluctuations in signal power on a slow time-scale which could be exploited by algorithms that use non-stationarity in the data to separate sources (e.g. Matsuoka et al., 1995; Parra and Spence, 2000a). Therefore, we choose a random variable with uniform distribution in the interval  $[0, 1]$  as modulator. This results in AM correlation across different frequencies since the same value of the modulator function is applied to all frequencies at a given time instance. However, the choice for the modulator does not result in non-stationarity of the data on a slow time-scale since the modulator fluctuates as quickly as the signal itself.

Note that the time-series in two different frequencies  $f_k \neq f_l$  of a single source  $s_i$  are second order uncorrelated. However, they are not mutually independent since the same amplitude modulation has been applied to both frequencies. The resulting statistical dependency is of higher than second order and can be detected by computing the correlation of the amplitudes, i.e., the AM auto-covariance  $\mathbf{C}(s_i)$ .

Figure 3.5 displays an example illustrating the construction of synthetic spectrograms from Gaussian noise and the random modulator function, together with the corresponding AM auto-covariance matrices.

To demonstrate that the AMDecor algorithm successfully separates mixes of synthetic source spectrograms, a simulation was carried out with two synthetic sources, each with  $K = 20$  frequencies and 10000 time-points. The source spectrograms were mixed using  $2 \times 2$  mixing matrices  $\mathbf{A}(f)$ , which were chosen at random and independently for each frequency. The total signal-to-interference ratio (SIR, as defined in appendix A.2) before separation was approximately 0 dB. The SIR after separation by the AMDecor

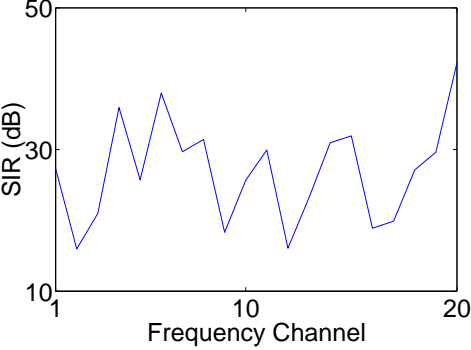
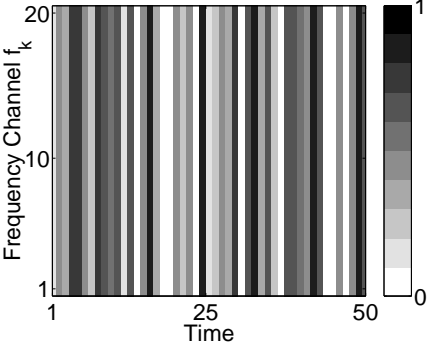
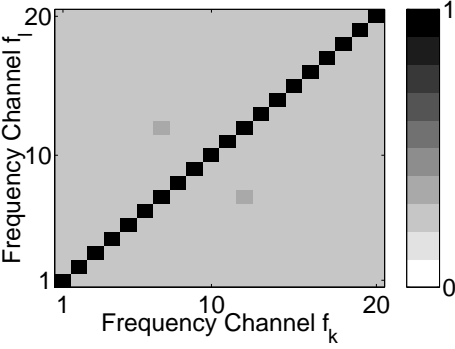
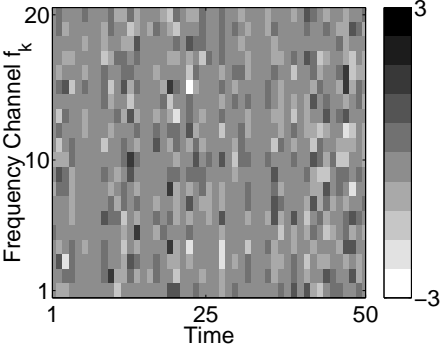
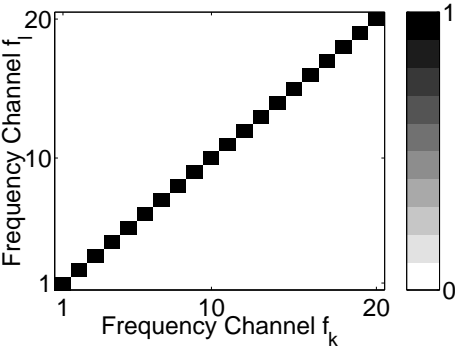
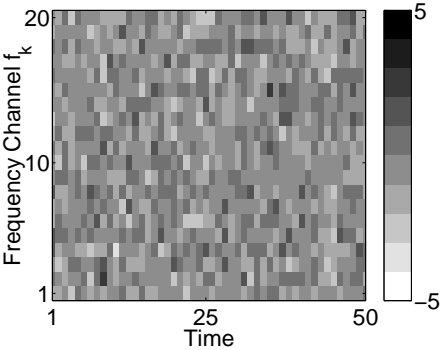


Figure 3.5: (Facing page) Construction and separation of synthetic spectrogram data. Top row: Gaussian i.i.d. noise  $\zeta(T, f)$  in each frequency channel (left) and the corresponding AM auto-covariance matrix (right). Since the Gaussian noise is unmodulated, the AM auto-covariance is zero. Middle row, left: Synthetic source spectrogram  $s(T, f)$  obtained by multiplying the Gaussian noise  $\zeta(T, f)$  with the modulator  $\mu(T)$  and transforming the result by the non-linear function  $g_f(\cdot)$  to make the pdf Gaussian. Middle row, right: AM auto-covariance matrix corresponding to the synthetic source spectrogram  $s(T, f)$ . Since the same amplitude modulation is applied to each frequency channel, the AM correlation across different frequencies is non-zero. Bottom row, left: The random modulator function  $\mu(T)$  which is independent of frequency. Bottom row, right: Frequency-dependent signal-to-interference-ratio (SIR) after separation of a mixture of two synthetic source spectrograms with  $K = 20$  frequencies and 10000 time-points.

algorithm is also displayed in figure 3.5, showing that the algorithm has successfully separated the mixed signals. Since the SIR after separation has the same sign for all frequencies, local permutations did not occur and the source signals have been reconstructed with the same order in each frequency channel.

The successful separation, together with the fact that this data cannot be separated by looking at isolated frequency channels separately, shows that taking into account information from AM correlation across different frequencies constitutes a distinct and novel criterion for source separation.

### 3.5.2 Separation in different acoustic situations

In this section, we present results obtained in different acoustic situations which include recordings in reverberant and non-reverberant environment and data obtained by digitally convolving speech signals with impulse responses of a real room. Since the recordings were performed such that the original source signals are available, and since the corresponding room impulse responses have also been recorded, the recordings allow for a detailed evaluation and analysis of the AM decorrelation algorithm.

The issues addressed in the evaluation concern performance in different acoustic situations, comparison with the performance of a non-blind reference method (MMSE method), analysis of permutations present in the unmixed signals and analysis of importance of across-frequency interactions in the AM decorrelation algorithm.

Speech signals from a total of four acoustic situations were recorded in two rooms. In the first room, a medium-sized seminar room at University of Oldenburg, speech was recorded from three different distances between speakers and microphones, ranging from 0.5 m to 3.5 m. The details of the setup are shown in figure 3.6. Note that the room contained a large window front to the right of the microphones and a blackboard behind the microphones. These surfaces contribute very strong reflections to the room acoustics, resulting in a reverberation time  $T_{60}$  of 0.5 s.

The signals used as original source signals were two different speech segments, each of length 5 s, spoken by the same male speaker. Therefore, the long-time spectrum of both signals was approximately identical, making spectral suppression of either of the signals impossible. The recordings were performed as stereo recordings of the separate source signals transmitted via a loudspeaker. The first stereo recording was performed with the first speech signal coming from the location of source one, and subsequently the second stereo recording was performed with the second speech signal coming from the location of source two. To obtain the mixed signals, both stereo signals were digitally mixed in the computer. This recording procedure has the advantage that the original source signals from their respective positions are available and make it possible to calculate signal-to-interference-ratios. The described procedure is justified by the fact that sound propagation in air at normal acoustic levels is fully linear, as are the microphones. All recordings were performed with omnidirectional microphones at 48kHz sampling rate.

The fourth acoustic situation was recorded in the anechoic chamber at University of Oldenburg. In first approximation this situation contains only propagation time

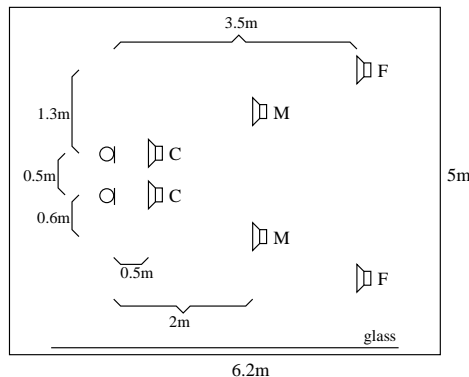


Figure 3.6: The setup for the real-room recordings with microphones to the left and speakers at close (C), medium (M) and far (F) positions to the right.

differences, but no reverberation. The relative locations of microphones and speakers were approximately those of position ‘far’ in figure 3.6. The source signals were two texts, length 5 s, read by the same male speaker from two positions. Recordings were done at 48 kHz sampling rate with the procedure of separate stereo recordings for each source signal, as described above.

In addition to the sound recordings, room impulse responses were measured in the seminar room setup from all speaker positions (with approx. 5 cm realignment error) to the microphones. Measurement was performed with maximum length sequences of length 65535 samples at a sampling rate of 48kHz. This amounts to about 1.35 s length of the measured impulse responses which is sufficiently long considering the given reverberation time of 0.5 s. Determination of the room impulse responses aims at two purposes. First, simulated stereo signals of the sources were computed by convolving the original source signals with the room’s impulse responses. Thereby, all factors due to non optimal recording conditions, such as recording noise, were eliminated.

Furthermore, the impulse responses were used to quantify acoustic characteristics of the room. By splitting the impulse responses into their first part, which corresponds to the path of direct sound propagation from speakers to microphones, and their second part, which corresponds to the reverberation, the direct-to-reverberation-energy-ratio (DRR) was computed. For close spacing of microphones and speakers, the DRR is positive, showing their placement within the radius of reverberation (e.g. Heckl and Müller, 1994), while for medium and large distance between microphones and speakers it is negative, indicating that more energy from signal reflections than from the direct path arrives at the microphones. For the latter case, mixing involves longer impulse responses and therefore makes separation harder, as is demonstrated below. The measured DRR coefficients are given in table 3.1.

Spectrograms were computed from the mixed signals using a Hanning window of length

Position	DRR
close	4.7 dB
medium	-1.0 dB
far	-6.9 dB

Table 3.1: Ratio of direct to reverberant energy contributions (DRR) at the microphones for different distances to the speakers.

		situation			
		nonrev	close	medium	far
SIR prior to separation (dB)		0.86	3.35	0.28	0.89
SIR gain (dB)	MMSE	19.68	7.07	5.78	3.05
	AMDecor	15.29	4.37	5.96	-0.25
	AMDecor + PC	15.30	5.08	6.57	3.88
	AMDSF	-0.11	-1.74	0.51	-0.36
	AMDSF + PC	9.14	3.83	4.46	3.28

Table 3.2: Summary of separation results for sound recordings from different acoustic situations. Situation ‘nonrev’ denotes non-reverberant environment while ‘close’, ‘medium’ and ‘far’ correspond to the respective speaker position in figure 3.6. Separation was performed with the MMSE method from appendix A.4 (‘MMSE’), the proposed AMDecor algorithm (‘AMDecor’) and amplitude modulation in single frequency channels (‘AMDSF’, see text). For the blind methods, performance has also been evaluated using non-blind correction of permutations as described in appendix A.3 (‘AMDecor+PC’ and ‘AMDSF+PC’, respectively).

4096 samples (which amounts to 85 ms), window-shift of 1024 samples and FFT-length of 8192 samples. Separation with all methods described below was done using frequencies up to 4 kHz since the main energy of the speech signals falls into this range.

### Separation by AM Decorrelation algorithm

In the first experiment, it is investigated how separation varies with the acoustical situation and whether the AM Decorrelation algorithm leads to separation without local permutations.

Data from all four acoustic situations and from the simulated mixed signals obtained by the room impulse responses was processed using the proposed AM decorrelation algorithm. The signal-to-interference-ratios (SIR) prior to and after separation are displayed in figure 3.7 and tables 3.2 and 3.3. For details about the computation of the SIR values, refer to appendix A.2. It is concluded from the results that the



		situation		
		close	medium	far
SIR prior to separation (dB)		2.98	0.41	1.12
SIR gain (dB)	MMSE	5.80	5.55	4.16
	AMDecor	5.39	5.73	3.30
	AMDecor + PC	5.70	5.86	4.03
	AMDSF	-1.16	0.12	-0.52
	AMDSF + PC	4.29	3.91	3.18

Table 3.3: Summary of separation results for signals mixed by room impulse responses. For explanation of the situations and the algorithm abbreviations, refer to table 3.2.

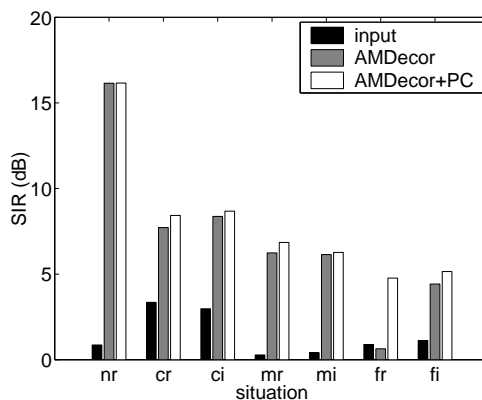


Figure 3.7: The signal-to-interference-ratios (SIR) prior to separation (‘input’), after separation by the AM decorrelation algorithm (‘AMDecor’) and with remaining local permutations corrected by the non-blind method from appendix A.3 (‘AMDecor+PC’). The different acoustic situations are non-reverberant (‘nr’), close (‘cr’, ‘ci’), medium (‘mr’, ‘mi’) and far (‘fr’, ‘fi’). Data for ‘nr’, ‘cr’, ‘mr’ and ‘fr’ was obtained by sound recordings in a room, while data for ‘ci’, ‘mi’ and ‘fi’ was obtained by convolving the original source signals with impulse responses measured in the room.

AMDecor algorithm successfully improves the SIR. The general trend is that SIR after separation is highest for the most ‘simple’ acoustic situation, the non-reverberant condition, and monotonically drops down towards more ‘complex’ situations where the impulse responses become longer. The monotonic decrease does not hold for the *gain* in SIR accomplished by the algorithm. In the ‘close’ position, the SIR prior to separation is already quite large, reducing the corresponding gain in SIR even below the gain that can be achieved for the more difficult ‘medium’ position.

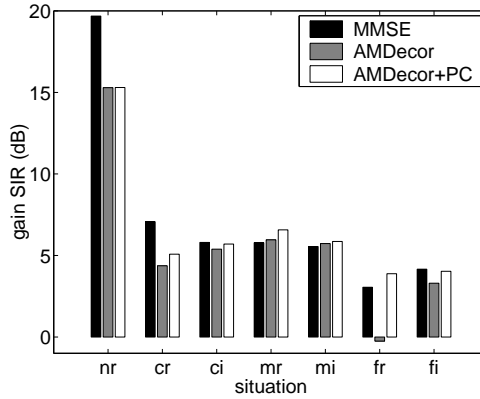


Figure 3.8: Comparison of the gain in SIR accomplished by AM decorrelation algorithm (‘AMDecor’) and MMSE method (‘MMSE’). Gain in SIR with the AM Decor algorithm and corrected permutations (‘AMDecor+PC’) is also shown. Acoustic situations (‘nr’, ‘cr’, ‘ci’, ‘mr’, ‘mi’, ‘fr’, ‘fi’) are denoted as in figure 3.7.

An important question is whether the overall gain in SIR could be further increased if possibly remaining local permutations were sorted correctly. To this end, we have applied a method for computing the SIR that is obtained if no local permutations occur. The method exploits the availability of the source signals to correct local permutations and is outlined in appendix A.3. It is displayed in figure 3.7 that the possible gain from correcting local permutations is below 0.73 dB in all cases except one. Therefore, it is concluded that the AM decorrelation algorithm does avoid local permutations to a very good degree. The only exception occurs for the room recording in the ‘far’ position. For this position, the AM decorrelation appears to converge to a local minimum of the cost-function in which local permutations reduce the gain in SIR to less than zero dB. Notably, if the permutations are corrected, the accomplished gain in SIR rises to a decent value.

### Separation by MMSE method

The quality of separation is limited by the length of the unmixing filters, i.e., by the length of the window function used for computing the spectrograms. Therefore, perfect separation of the signals cannot be accomplished, even in the case of computing simulated mixed signals from the room impulse responses. By using a blind method only, it is not possible to determine how close the attained separating solution is to the best possible solution. Therefore, a non-blind method based on the minimum mean squared error (MMSE) method is employed to find the optimal linear reconstruction of the source signals from the mixed signals’ spectrograms. Details of the method are given in appendix A.4.

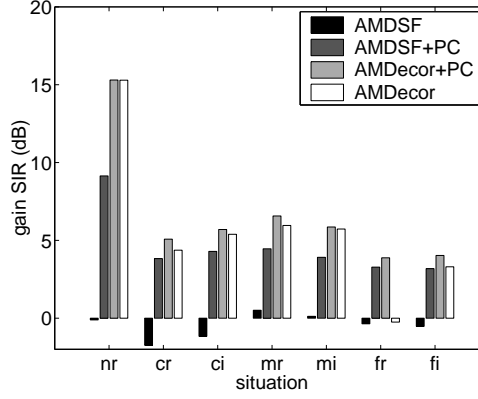


Figure 3.9: Comparison of separation by the proposed AM decorrelation algorithm (‘AMDecor’) based on cost-function (3.20) with separation obtained by decorrelation in single frequency bands (‘AMDSF’) based on cost-function (3.25). For both algorithms, the gain in SIR is also given for remaining permutations corrected using the non-blind method from appendix A.3 (‘AMDecor+PC’ and ‘AMDSF+PC’, respectively). Acoustic situations (‘nr’, ‘cr’, ‘ci’, ‘mr’, ‘mi’, ‘fr’, ‘fi’) are denoted as in figure 3.7.

Figure 3.8 compares the separation results obtained by the AM decorrelation algorithm with separation by the MMSE method. In the case of simulated mixing, the result of the AMDecor separation is close to the MMSE result and in one case even better. In the case of real recordings, AMDecor performs on average slightly worse, however still close to the MMSE result. In the ‘medium’ situation, AMDecor outperforms MMSE and in the problematic case of the ‘far’ situation, the AMDecor result with corrected permutations is also better than the MMSE result. In conclusion, the separation obtained by AM decorrelation is on average in the vicinity of the optimum.

### Separation by AM decorrelation *without* across-frequency interactions

Finally, the question is addressed, whether the across-frequency terms in the AM decorrelation cost function (3.20) actually improve quality of separation or whether they merely serve to avoid local permutations. To elucidate this point, the alternative cost function  $H_{\text{sf}}$ ,

$$H_{\text{sf}} = \sum_{i \neq j} \sum_k [\mathbf{C}(u_i, u_j)]_{kk}^2, \quad (3.25)$$

is investigated which is similar to (3.20) but contains single-frequency terms only and lacks the across-frequency interactions of (3.20). Therefore,  $H_{\text{sf}}$  does not have the ability to avoid local permutations.

Figure 3.9 compares the separation based on cost-functions (3.20) and (3.25), with and without local permutations being corrected by the method from appendix A.3.

As expected, the results from  $H_{\text{sf}}$  *without* permutations corrected are very poor since the local permutations result in an average gain in separation of around 0 dB. If the local permutations are corrected, it becomes clear that  $H_{\text{sf}}$  does result in some degree of signal separation. However, comparing with the results obtained with the cost function  $H$ , the latter performs significantly better. In particular,  $H$  *without* permutations corrected is in all situations (with the exception of the pathological ‘far’ situation) better than  $H_{\text{sf}}$  *with* permutations corrected. If, in addition, the remaining permutations in the results with  $H$  are corrected, then  $H$  outperforms  $H_{\text{sf}}$  even further. Therefore, it is concluded that the across-frequency terms in (3.20) are not only needed in order to avoid local permutations, but that they also improve on the quality of separation.

### 3.5.3 Performance on benchmark data

In this section, the AM decorrelation algorithm is applied to publicly available real room recordings of speech signals. The quality of separation is evaluated and compared to the quality accomplished by previous algorithms on the same data.

#### Separation of data provided by Lee

The first data set was obtained from Lee (1998b). It consists of two speakers counting from zero to ten in English and Spanish language, respectively. According to Lee (1998a), the recording was performed in rectangular order of speakers and microphones in an office room, with a distance of 40cm between the microphones and 60cm between microphones and speakers.

This dataset can be regarded as relatively easy to separate and successful separation has been performed by several researchers with different algorithms. Since the original source signals are not available, it is difficult to compare the performance of the different algorithms on this dataset quantitatively. E.g., it is not possible to compute the improvement in signal-to-interference-ratio accomplished by the different algorithms. In an attempt to apply some measure of how ‘independent’ the unmixed signals are, we have computed the value of the AM decorrelation cost-function (3.20) for the recorded signals and for the unmixed signals obtained by different algorithms. The result is displayed in figure 3.10 (left). It shows that the AM decorrelation algorithm has the best performance and improves slightly on the algorithm of Parra and Spence (2000a) which is generally regarded as exhibiting high-quality output signals and excellent separation results. Admittedly, the value of the AM decorrelation cost-function is a very crude measure for the performance. Also, it may be argued that quite naturally our algorithm achieves the best performance using this measure since it directly aims at minimizing the AM correlation. However, from informal listening to the unmixed signals we conclude that the quality of separation as perceived by human listeners coincides quite well with the numeric results. In particular, the perceived difference in separation quality is quite high between the results of Lee and the AM decorrelation algorithm, and it is relatively small between the results of Parra and the AM decorrelation algorithm, while the improvement due to the AMDecor algorithm is still clearly

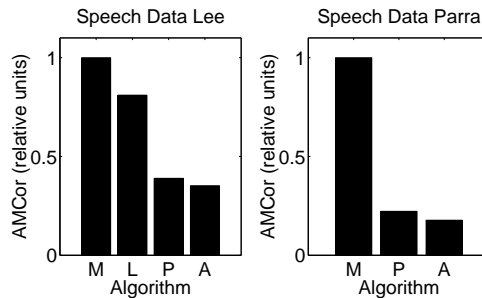


Figure 3.10: Separation of real-room recordings of speech by different algorithms. Left side refers to speech data recorded by Lee (1998b), right side to speech data recorded by Parra (1998). The value of the AMCor cost-function (3.20) (‘AMCor’) is shown for the amplitude modulation decorrelation algorithm (‘A’), Lee’s algorithm (‘L’, Lee et al. (1998), results for Lee’s speech data only) and Parra’s algorithm (‘P’, Parra and Spence (2000a)), relative to the mixed signals (‘M’). Total energy of the signals was normalized prior to computing the cumulative AMCor.

audible.

To illustrate minimization of AM cross-covariance by the AM decorrelation algorithm, figures 3.11 and 3.12 display spectrograms of the recorded and the unmixed signals and their corresponding auto- and cross-covariance matrices. Separation was performed using a 384 samples long Hanning window, DFT length 512 samples and window shift 64 samples at a sampling rate of 16 kHz.

### Separation of data provided by Parra

The second dataset has been provided by Parra (1998), accompanied with the separation results obtained by the algorithm of Parra and Spence (2000a) on this data. It consists of a person talking and a TV set in the background. The signals were recorded in a hotel room with microphones attached to a laptop computer and are of relatively poor sound quality. This dataset is known by several researchers in the field and regarded as difficult to separate. To the authors’ knowledge, only two algorithms—the algorithm of Parra and Spence (2000a) and the proposed AM decorrelation algorithm—have been shown to perform successful separation on this dataset.

Since the original source signals are not available, we again resort to use the value of the cost-function (3.20) as an indication for the algorithms’ performance and use listening tests to assess subjective quality of separation. The corresponding values of the cost-function are displayed in figure 3.10 (right). Both algorithms achieve a strong reduction of the AM correlation relative to the mixed signals. Again, the AM decorrelation algorithm attains slightly better values than the algorithm of Parra and Spence (2000a). Listening to the results, quality of separation is very good for both algorithms with the AM decorrelation accomplishing slightly better separation. In

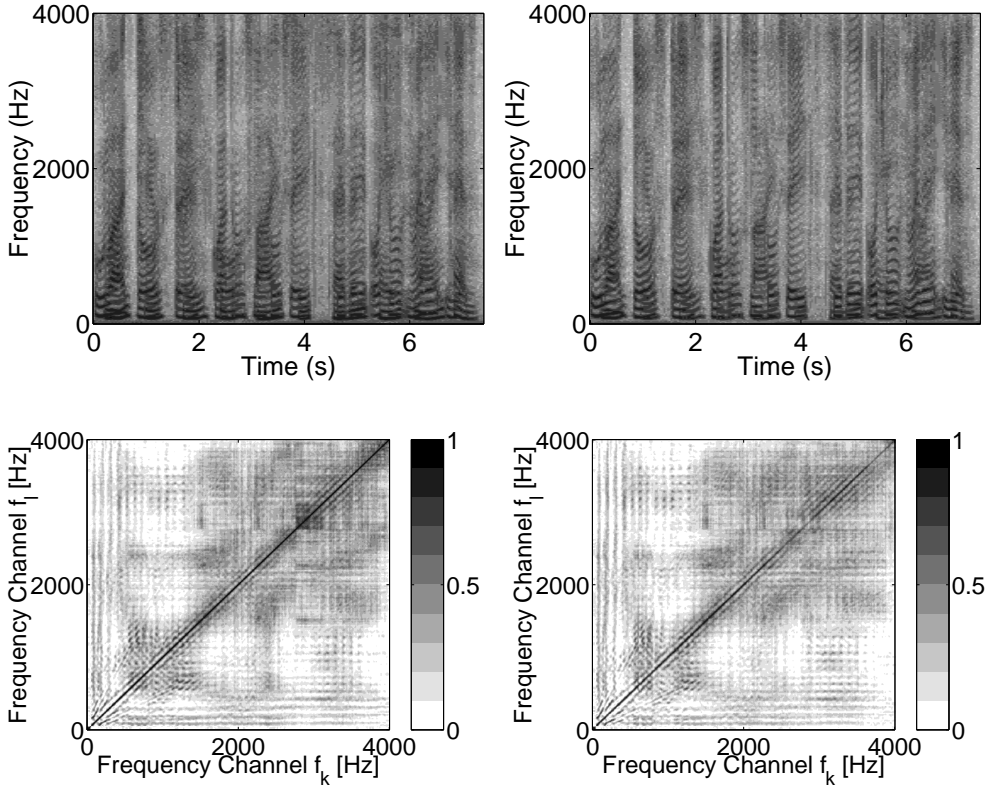


Figure 3.11: Spectrograms and corresponding AM covariance matrices for the dataset provided by Lee (1998b). Top row: Spectrograms of the left and right microphone signal, respectively. Bottom row, left: Normalized AM *auto*-covariance matrix of the first microphone signal. Clearly, the speech signals exhibit similar amplitude modulation even at very distant frequencies. Bottom row, right: Normalized AM *cross*-covariance matrix of the both microphone signals. Since both microphones receive contributions from both sources, the AM cross-covariance is almost as high as the AM auto-covariance displayed on the left.

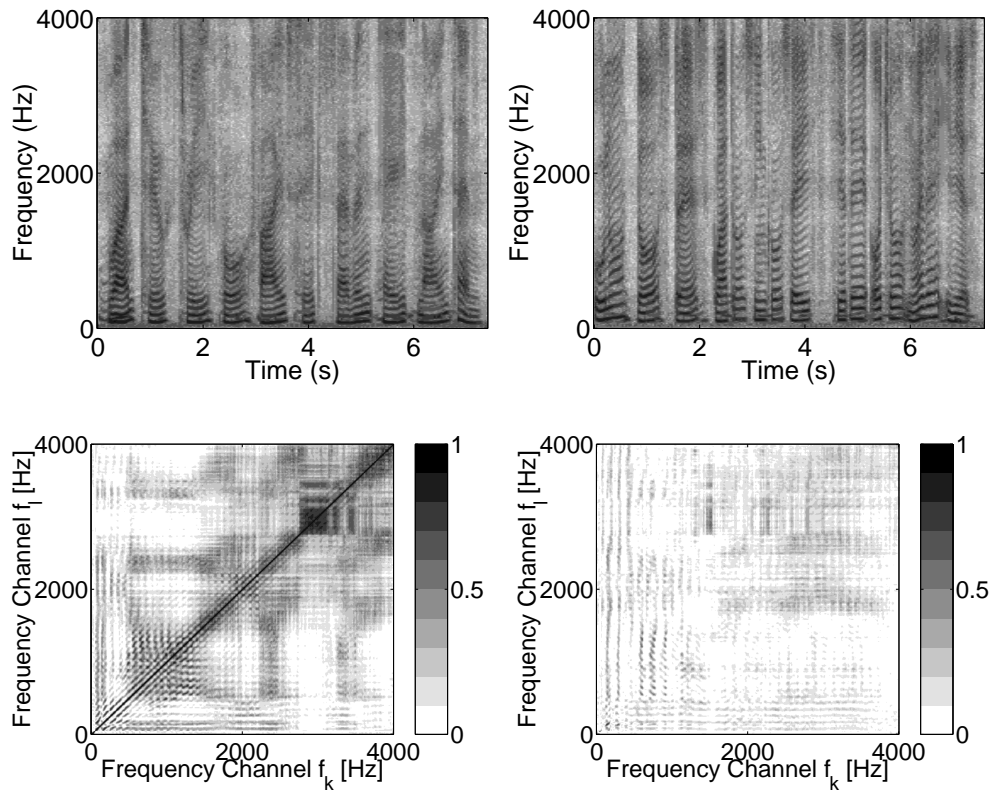


Figure 3.12: Top row: Separated signals obtained by the proposed algorithm from the mixed signals of figure 3.11. Bottom row, left: Normalized AM *auto*-covariance matrix of the first (top row, right) separated signal's spectrogram. Since correlations within each signal remain after separation, the AMCor within each signal is non-zero. Bottom row, right: Normalized AM *cross*-covariance matrix of both separated signals' spectrograms. Obviously, the algorithm has caused a reduction of AMCor to low residual levels.

particular the AM decorrelation algorithm almost completely eliminates the speech signal in the output corresponding to the TV set while some crosstalk from the TV set is still audible in the output corresponding to the single talker. This result is attributed to the speaking person being closer to microphones than the interfering TV set. Therefore, the impulse responses from the speaker to the microphones can be assumed to exhibit the better direct- to reverberation-energy ratio which may make them easier to be canceled out by the algorithm.

For separation of the data, very long filters were used. Length of the Hanning window was 3584 samples, DFT length 4096 samples, window shift 1024 samples at a sampling rate of 8 kHz.

### 3.6 Conclusion

Amplitude modulation correlation in speech arises naturally from the structure of speech and from the physiology of speech production. Quantification of AM correlation is possible by computing auto- and cross-covariance matrices from amplitude spectrograms. A novel cost-function and algorithm for blind separation of convolutively mixed speech signals were proposed, based on minimization of AM cross-covariance matrices for all pairs of different unmixed signals. Its key feature, which distinguishes it from algorithms in the existing literature, is the incorporation of envelope correlations across different frequencies. It was shown analytically and experimentally that the introduction of across-frequency interactions into the cost-function solves the problem of local permutations which arises in frequency domain blind source separation algorithms. Because of its cross-frequency terms, the algorithm successfully separates mixtures of synthetic signals with Gaussian i.i.d. statistics which cannot be separated by algorithms working in single frequency bands. Furthermore, the experimental evaluation showed that the across-frequency interactions improve on the quality of separation of real-room recordings of speech.

The observation of AM correlation in speech signals is closely related to the underlying statistical model of speech signals which is (implicitly or explicitly) assumed. E.g., the assumption of statistically independent Fourier transform coefficients in different frequency channels is stated explicitly by Attias and Schreiner (1998). The same assumption is made implicitly by any algorithm that attempts to solve the blind source separation problem in each frequency channel separately. In contrast, it is shown in the present paper that for speech signals the assumption of independency of different frequency channels does not hold. Furthermore, the results presented permit the conclusion that the problem of local permutations arises if different frequency channels are assumed to be statistically independent. However, if statistical dependencies between signal components in different frequency channels are taken into account, the permutation problem can be solved.

Measuring the correlations between amplitude time-courses in different frequency channels might appear as a very crude measure of statistical dependency, and indeed it has been proposed based on empirical observations and knowledge from speech processing. However, it should be noted that AM correlation is closely related to the notion



of fourth order cross cumulants (Nikias and Petropulu, 1993) which, for zero-mean random variables  $x(f)$  and  $y(f)$ , can be defined as

$$c_4(x(f_k), y(f_l)) = E\{|x(f_k)|^2 |y(f_l)|^2\} - E\{|x(f_k)|^2\} E\{|y(f_l)|^2\} \\ - |E\{x(f_k) y^*(f_l)\}|^2 - |E\{x(f_k) y(f_l)\}|^2. \quad (3.26)$$

For speech spectrogram data, the fourth term on the r.h.s. of (3.26) is zero, and the third term on the r.h.s. is essentially a diagonal contribution for  $f_k = f_l$ . Hence, the fourth order cross cumulant expression (3.26) is very similar to the proposed measure of AM correlation (3.8).

This analogy permits the interpretation of AM correlation as a quantity which measures higher-order statistical dependencies between Fourier transform coefficients in different frequency channels. The present paper has shown that taking into account this higher-order structure in speech signals results in an improved algorithm for blind source separation.



## Chapter 4

# Separation of multidimensional sources

### 4.1 Introduction

The aim of blind source separation (BSS, Jutten and Héroult, 1991) is to recover independent source signals from knowledge of their superpositions, only. One typical example is the ‘Cocktail-Party’ situation, i.e., mixed signals of several speakers are recorded with multiple microphones whereas the signals of interest are the individual speaker signals, which BSS tries to reconstruct. Methods based on different principles have been proposed to achieve this goal. Their common basis is the assumption that the sources are independent systems. By reconstructing signals which are as independent as possible, an attempt is made to recover the original sources. Since little additional knowledge is assumed to be known, the methods are termed ‘blind’.

The class of the probably most widely employed methods relies on the notion of ‘independence’ in the sense of statistical independence, and is also referred to as independent component analysis (ICA). This method decomposes mixed signals into statistically independent source signals by exploiting the assumed non-Gaussian probability density functions of the sources (e.g. Jutten and Héroult, 1991; Comon, 1994; Bell and Sejnowski, 1995; Cardoso and Laheld, 1996).

Another group of algorithms employs methods based on second order statistics and recovers the sources by requiring that the cross-correlation functions of different unmixed signals must vanish (e.g. Weinstein et al., 1993; Molgedey and Schuster, 1994; Belouchrani et al., 1997).

Finally, algorithms have been proposed that separate mixed signals based on the non-stationarity of the underlying sources (e.g. Matsuoka et al., 1995; Parra and Spence, 2000a).

An assumption common to all of the algorithms mentioned is that the  $N$  underlying sources  $s_i(t)$ ,  $i = 1, \dots, N$ , are essentially one-dimensional, i.e., they depend on a single variable  $t$ , only, where  $t$  may denote time. Even in applications where the raw

data is higher dimensional, it is rearranged into a one-dimensional feature vector. For blind source separation of two-dimensional images, for example, the data is reordered into a one-dimensional vector which contains the concatenation of all pixel values (Bell and Sejnowski, 1997; Wachtler et al., 2001). This is justified by the assumption of a translation invariant mixing process so that the sources' pixels are superimposed in the same way at all spatial positions. Furthermore, the data is assumed to be stationary with respect to two-dimensional space.

In the context of blind source separation, the case of multidimensional signals (e.g. Priestley, 1981) is encountered in frequency-domain based approaches to the separation of acoustically mixed sound sources. Due to time-delays and reverberation, the acoustic medium causes the convolutive mixing of sources. By computing consecutive short-time spectra, the data is transformed into the time-frequency spectrogram representation, and the convolutive mixing in the time domain factorizes into instantaneous mixing (i.e., mixing without time-delays) in each frequency band (for details cf. section 4.4.3). Hence, each source  $i$ ,  $i = 1, \dots, N$ , is no longer represented by the one-dimensional signal  $s_i(t)$ , but rather by the two-dimensional spectrogram  $s_i(t, f)$ , where the coordinates  $t$  and  $f$  correspond to the time- and frequency dimension, respectively. In contrast to the situation for image data outlined above, two aspects of the problem prohibit its simplification to a single one-dimensional problem. First, the signal mixing varies with frequency, which is a result of the convolution operation in the time domain. Second, the data is non-stationary with respect to the frequency dimension, since the power of, e.g., speech signals varies considerably across frequency. Since both mixing and data are non-stationary, this problem is truly multidimensional.

Several researchers have proposed to split the multidimensional problem into a set of  $K$  independent one-dimensional BSS problems, one for each frequency  $f = 1, \dots, K$ , and separate source components independently in each frequency (e.g. Capdevielle et al., 1995; Ehlers and Schuster, 1997; Murata et al., 1998; Parra and Spence, 2000a). However, this approach causes in particular the problem that the sources' components are recovered in disparate (unknown) order in different frequencies, which makes the direct assignment of unmixed components to the corresponding sources impossible. The need to sort the source components' permutations has led to additional post-processing steps which incorporate further prior knowledge that had not been assumed for the sake of separation. In contrast, it has recently been shown by Anemüller and Kollmeier (2000) for the case of separating convolutively mixed speech signals that taking into account the multidimensional nature of the source signals by modeling the statistical dependencies between different frequency channels resolves the permutation problem without post-processing and results in improved quality of separation.

The aim of the present paper is to suggest that proper consideration of multidimensional sources can be beneficial also in applications other than the separation of acoustically mixed sound signals. Furthermore, a novel solution to the separation of mixed multidimensional sources is given, which is based on second-order statistics.

The outline of the paper is as follows. In section 4.2 the notion of multidimensional source signals and the assumed mixing model is specified. Examples for situations exhibiting multidimensional source signals are given. Section 4.3 is dedicated to a

particular algebraic solution, together with conditions for identifiability and resolving permutations. Evaluation of the proposed algorithm is presented in section 4.4.

Throughout the paper, vectors and matrices are printed in bold font.  $x^*$  denotes complex conjugation. Transposition is denoted by  $\mathbf{x}^T$ , and transposition and complex conjugation by  $\mathbf{x}^H$ . The expectation operator is denoted by  $E\{\cdot\}$ . The exposition is given for complex variables.

## 4.2 Multidimensional sources and mixing

The multidimensional signal generated by source  $s$  is denoted by  $s(t_1, \dots, t_L, f_1, \dots, f_K)$ . The parameters  $(t_1, \dots, t_L, f_1, \dots, f_K)$  denote coordinates in  $L+K$ -dimensional space, and for each dimension the source signal is defined at coordinate values  $t_{l,1}, \dots, t_{l,T_l}$ ,  $l = 1, \dots, L$ , and  $f_{k,1}, \dots, f_{k,F_k}$ ,  $k = 1, \dots, K$ . Without loss of generality, we limit our treatment to zero mean sources.

The  $t$ -dimensions, corresponding to coordinates  $t_1, \dots, t_L$ , are denoted as ‘stationary dimensions’ since it is assumed that both the mixing system and the source signals’ statistics are stationary with respect to a shift in the  $t$ -dimensions. Hence, expectations can be computed by averaging over the  $t$ -parameters, assuming ergodicity. It is noted that due the stationarity of mixing and data, the  $t$ -dimensions can be regarded as essentially one-dimensional, as explained in section 4.1.

In contrast, the  $f$ -dimensions, corresponding to coordinates  $f_1, \dots, f_K$ , are denoted as ‘non-stationary dimensions’ since it is assumed that the mixing system varies with respect to a shift in each coordinate  $f_k$ , and the source signals may be non-stationary with respect to each  $f_k$ . Therefore, it is not possible to compute expectations as averages over the  $f$ -parameters.

Let  $N$  independent sources  $s_i$ , with signals  $s_i(t_1, \dots, t_L, f_1, \dots, f_K)$ , be mixed linearly by an invertible mixing system with coefficients  $a_{ij}(f_1, \dots, f_K)$  which depend on the  $f$ -dimensions’ coordinates, only. An equal number of mixed signals  $x_i(t_1, \dots, t_L, f_1, \dots, f_K)$ ,  $i = 1, \dots, N$ , is obtained as

$$x_i(t_1, \dots, t_L, f_1, \dots, f_K) = \sum_{j=1}^N a_{ij}(f_1, \dots, f_K) s_j(t_1, \dots, t_L, f_1, \dots, f_K). \quad (4.1)$$

For the sake of a concise notation, the exposition in the remainder of the paper is given for two-dimensional sources  $s(t, f)$ . However, the derivation directly carries over to the case of  $L+K$ -dimensional sources.

Combining all sources in one vector yields, in the two-dimensional case, the source vector  $\mathbf{s}(t, f) = [s_1(t, f), \dots, s_N(t, f)]^T$ . Denoting in analogy the mixed signals’ vector as  $\mathbf{x}(t, f) = [x_1(t, f), \dots, x_N(t, f)]^T$ , and the  $N \times N$  mixing matrix with  $(i, j)$ -element  $a_{ij}(f)$  by  $\mathbf{A}(f)$ , the mixing system in matrix-vector notation reads

$$\mathbf{x}(t, f) = \mathbf{A}(f) \mathbf{s}(t, f). \quad (4.2)$$

From knowledge of the mixed signals  $\mathbf{x}(t, f)$ , only, it is aimed to find an estimate  $\hat{\mathbf{A}}(f)$

of the mixing matrix so that unmixed signals  $\mathbf{u}(t, f) = \hat{\mathbf{A}}^{-1}(f) \mathbf{x}(t, f)$  can be obtained which resemble the source signals.

Choice of the symbols  $t$  and  $f$  is motivated by previous work on convolutive signal separation using time-frequency representations (Anemüller and Kollmeier, 2000). For this reason, the parameters  $t$  and  $f$  are also referred to as ‘time’ and ‘frequency’, respectively. However, from the exposition of multidimensional sources above and examples given below it should be clear that the characteristic properties of  $t$ - and  $f$ -dimension are their stationarity and non-stationarity, respectively. Therefore, examples of applications are given below where  $t$  and  $f$  do *not* denote ‘time’ and ‘frequency’, respectively.

In image processing, multidimensional signals similar to the time-frequency spectrogram of acoustic sources are encountered. For the analysis of spectral image data, the sources are represented by 3-dimensional signals  $s_i(t_1, t_2, f)$  where  $t_1$  and  $t_2$  represent the spatial  $(x, y)$ -coordinates, and  $f$  denotes the spectral wavelength. We assume that the mixing matrix  $\mathbf{A}(f)$  is frequency-dependent, which may be due to, e.g., different surface reflectance at different wavelengths. With respect to the spatial position  $(t_1, t_2)$ , however, the mixing is regarded as stationary, as is the statistics of the sources. For the algorithm presented in this paper to be applicable, it is furthermore required that each source’s components at different wavelengths are interrelated, such that non-vanishing correlations between different spectral components are exhibited. This assumption is fulfilled for many visual scenes since natural objects are hardly ever monochromatic in color. Such correlations also exist between broad spectral bands, as is illustrated in section 4.4.2 for RGB encoded color images.

As in the case of spectrogram data, source components could be separated independently for each wavelength  $f$ , which would, however, lead again to the problem of recovering the source components in permuted order in different spectral bands. In contrast, the proposed algorithm performs separation by making use of correlations between different spectral bands, and thereby leads to reconstruction of consistently ordered source components. Furthermore, it opens the possibility to unmix signals which cannot be separated by taking into account information at individual spectral bands, only, as is demonstrated in section 4.4.1.

As a final example, the analysis of signals mixed by a time-varying mixing system involves multidimensional source signals, as well. Consider, e.g., the time-varying mixture of an image sequence. Again, the mixing is regarded as spatially invariant and, therefore, the spatial  $(x, y)$ -coordinates are denoted by the parameters  $(t_1, t_2)$  which correspond to the ‘stationary’ dimensions. Since the mixing changes over time, time is regarded as the ‘non-stationary’ dimension and, hence, denoted by the parameter  $f$ . Without averaging over time, the proposed algorithm makes it possible to estimate the mixing system at each time by taking into account signal values at other times, as well. The condition which needs to be fulfilled by the source signals to make this possible is that their auto-correlation function is non-zero also at non-zero time-lag.

### 4.3 Solution based on correlations across frequency

Having introduced the basic concepts underlying the algorithm, we now turn to a quantitative description. Since the sources are assumed to be independent systems, all correlations computed from two different sources  $s_i$  and  $s_j$ ,  $i \neq j$ , vanish. In particular, if correlations are computed from source components at two frequencies, the result is zero for all pairs of frequencies  $(f, f')$ ,

$$E \{s_i(t, f) s_j^*(t, f')\} = 0 \quad \forall i \neq j, \forall f, f'. \quad (4.3)$$

However, this does not hold for correlations computed from two frequencies of the same source, since data generated by a single source cannot be assumed to be independent. Therefore, correlations within a single source are in general non-zero,

$$E \{s_i(t, f) s_i^*(t, f')\} \neq 0. \quad (4.4)$$

Defining the sources' cross-covariance matrix  $\mathbf{R}_s(f, f')$  computed from frequencies  $f$  and  $f'$  as

$$\mathbf{R}_s(f, f') = E \{ \mathbf{s}(t, f) \mathbf{s}^H(t, f') \}, \quad (4.5)$$

equations (4.3) and (4.4) can be restated such that  $\mathbf{R}_s(f, f')$  is diagonal for all  $(f, f')$ ,

$$[\mathbf{R}_s(f, f')]_{ij} = \delta_{ij} E \{s_i(t, f) s_i^*(t, f')\}, \quad (4.6)$$

where  $\delta_{ij}$  is the Kronecker symbol.

Since the mixed signals are not independent, their covariance matrix  $\mathbf{R}_x(f, f')$ ,

$$\mathbf{R}_x(f, f') = E \{ \mathbf{x}(t, f) \mathbf{x}^H(t, f') \}, \quad (4.7)$$

is not diagonal. It can be expressed in terms of the sources' covariance matrix as

$$\mathbf{R}_x(f, f') = \mathbf{A}(f) \mathbf{R}_s(f, f') \mathbf{A}^H(f'). \quad (4.8)$$

If the mixing system was identical in both frequencies,  $\mathbf{A}(f) = \mathbf{A}(f')$ , then an eigenvalue equation could be derived in exactly the same manner as presented by Molgedey and Schuster (1994). However, since in general  $\mathbf{A}(f) \neq \mathbf{A}(f')$ , the analog derivation is not possible.

It is observed that by forming the products

$$\mathbf{Q}_s(f, f') = \mathbf{R}_s(f, f') \mathbf{R}_s^{-1}(f', f') \mathbf{R}_s(f', f) \quad (4.9)$$

$$\mathbf{Q}_x(f, f') = \mathbf{R}_x(f, f') \mathbf{R}_x^{-1}(f', f') \mathbf{R}_x(f', f) \quad (4.10)$$

the algebraic relation between the sources'  $\mathbf{Q}_s(f, f')$  and the mixed signals'  $\mathbf{Q}_x(f, f')$  involves matrix  $\mathbf{A}(f)$ , but not  $\mathbf{A}(f')$ ,

$$\mathbf{Q}_s(f, f') = \mathbf{A}^{-1}(f) \mathbf{Q}_x(f, f') \mathbf{A}^{-H}(f). \quad (4.11)$$

Hence,  $\mathbf{A}^{-1}(f)$  diagonalizes  $\mathbf{Q}_x(f, f')$  for all  $f'$ .

An eigenvalue equation for  $\mathbf{A}(f)$  can be derived from (4.11) by forming the product

$$\mathbf{Q}_x(f, f') \mathbf{Q}_x^{-1}(f, f), \quad (4.12)$$

yielding

$$\mathbf{A}(f) \mathbf{A}(f, f') = \mathbf{Q}_x(f, f') \mathbf{Q}_x^{-1}(f, f) \mathbf{A}(f), \quad (4.13)$$

where

$$\mathbf{A}(f, f') = \mathbf{Q}_s(f, f') \mathbf{Q}_s^{-1}(f, f) \quad (4.14)$$

is diagonal and contains the eigenvalues of  $\mathbf{Q}_x(f, f') \mathbf{Q}_x^{-1}(f, f)$ .

Similarly,  $\mathbf{A}(f')$  is obtained from the Eigenvalue equation

$$\mathbf{A}(f') \mathbf{A}(f', f) = \mathbf{Q}_x(f', f) \mathbf{Q}_x^{-1}(f', f') \mathbf{A}(f'). \quad (4.15)$$

### 4.3.1 Conditions for identifiability

Equation (4.13) has a unique solution if all eigenvalues on the diagonal of  $\mathbf{A}(f, f')$  are different. Similarly, for (4.15) it must hold that the diagonal elements of  $\mathbf{A}(f', f)$  are different. Since  $\mathbf{R}_s(f, f')$  is diagonal and  $\mathbf{R}_s(f, f') = \mathbf{R}_s^H(f', f)$ , we obtain

$$\mathbf{A}(f, f') = \mathbf{A}(f', f) = \mathbf{R}_s(f, f') \mathbf{R}_s^H(f, f') \mathbf{R}_s^{-1}(f, f) \mathbf{R}_s^{-1}(f', f'). \quad (4.16)$$

Hence, together with (4.6) it follows that for  $\mathbf{A}(f)$  and  $\mathbf{A}(f')$  to be identifiable it must be fulfilled that

$$\frac{|E\{s_i(t, f) s_i^*(t, f')\}|^2}{E\{|s_i(t, f)|^2\} E\{|s_i(t, f')|^2\}} \neq \frac{|E\{s_j(t, f) s_j^*(t, f')\}|^2}{E\{|s_j(t, f)|^2\} E\{|s_j(t, f')|^2\}} \quad \forall i \neq j. \quad (4.17)$$

### 4.3.2 Solving the permutation problem

Since the eigenvectors corresponding to the solution of (4.13) are unambiguous only upto their order and a scale factor, the mixing matrix  $\mathbf{A}(f)$  cannot be determined uniquely. Rather, any matrix  $\mathbf{A}'(f)$  which can be expressed as

$$\mathbf{A}'(f) = \mathbf{A}(f) \mathbf{D}(f) \mathbf{P}(f), \quad (4.18)$$

where  $\mathbf{D}(f)$  is a diagonal matrix and  $\mathbf{P}(f)$  a permutation matrix, represents a solution of (4.13). Hence, it is only possible to determine  $\mathbf{A}(f)$  upto an unknown rescaling and permutation of its columns by  $\mathbf{D}(f)$  and  $\mathbf{P}(f)$ , respectively. This corresponds to the well-known invariances inherent to all blind source separation algorithms (see Tong et al., 1991).

For one-dimensional source signals this is usually not a problem. With multidimensional sources, however, the components belonging to a single source are reconstructed with disparate (unknown) order and scale in different frequencies  $f \neq f'$  if the corresponding frequency-specific permutation and diagonal matrices differ, i.e.,

$$\mathbf{P}(f) \neq \mathbf{P}(f') \quad \mathbf{D}(f) \neq \mathbf{D}(f'). \quad (4.19)$$



Thus, a coherent picture of each source's activity cannot be obtained.

No solution is given for the invariance with respect to varied scaling in different frequencies. Instead, each row of the estimated unmixing matrix  $\hat{\mathbf{A}}^{-1}(f)$  is rescaled to have unit norm.

The solution to the permutation problem is based on the observation that transformation (4.18) results in rearranged eigenvalues  $\mathbf{\Lambda}'(f, f')$ ,

$$\mathbf{\Lambda}'(f, f') = \mathbf{P}^T(f) \mathbf{\Lambda}(f, f') \mathbf{P}(f). \quad (4.20)$$

That is, the column permutation of  $\mathbf{A}(f)$  results in a corresponding permutation of the eigenvalues' order on the diagonal of  $\mathbf{\Lambda}(f, f')$ .

Denote by  $\hat{\mathbf{A}}(f)$  and  $\hat{\mathbf{A}}(f')$  the estimates of the true mixing matrices  $\mathbf{A}(f)$  and  $\mathbf{A}(f')$ , respectively. Without loss of generality, we assume

$$\hat{\mathbf{A}}(f) = \mathbf{A}(f) \quad \hat{\mathbf{A}}(f') = \mathbf{A}(f') \mathbf{P}, \quad (4.21)$$

so that the estimates  $\hat{\mathbf{\Lambda}}(f, f')$  and  $\hat{\mathbf{\Lambda}}(f', f)$  of the true eigenvalue matrices  $\mathbf{\Lambda}(f, f')$  and  $\mathbf{\Lambda}(f', f)$ , respectively, are

$$\hat{\mathbf{\Lambda}}(f, f') = \mathbf{\Lambda}(f, f') \quad (4.22)$$

$$\hat{\mathbf{\Lambda}}(f', f) = \mathbf{P}^T \mathbf{\Lambda}(f', f) \mathbf{P}. \quad (4.23)$$

Since, according to (4.16) we have  $\mathbf{\Lambda}(f', f) = \mathbf{\Lambda}(f, f')$ , it follows

$$\hat{\mathbf{\Lambda}}(f', f) = \mathbf{P}^T \mathbf{\Lambda}(f, f') \mathbf{P} = \mathbf{P}^T \hat{\mathbf{\Lambda}}(f, f') \mathbf{P}. \quad (4.24)$$

Therefore, the permutation matrix  $\mathbf{P}$  can be directly read from the relative ordering of the eigenvalues on the diagonals of  $\hat{\mathbf{\Lambda}}(f, f')$  and  $\hat{\mathbf{\Lambda}}(f', f)$ . Permutations are corrected by forming the matrix  $\hat{\mathbf{A}}'(f') = \hat{\mathbf{A}}(f') \mathbf{P}^T$  whose columns are ordered in accordance with  $\hat{\mathbf{A}}(f)$ .

### 4.3.3 More than two frequencies

#### Separation

If frequencies  $f = 1, \dots, F$ ,  $F \geq 2$ , are to be used for separation, the mixing matrix  $\mathbf{A}(f)$  is obtained as the matrix which simultaneously solves the  $F$  diagonalization equations

$$\begin{aligned} \mathbf{Q}_s(f, 1) &= \mathbf{A}^{-1}(f) \mathbf{Q}_x(f, 1) \mathbf{A}^{-H}(f) \\ \mathbf{Q}_s(f, 2) &= \mathbf{A}^{-1}(f) \mathbf{Q}_x(f, 2) \mathbf{A}^{-H}(f) \\ &\vdots \\ \mathbf{Q}_s(f, F) &= \mathbf{A}^{-1}(f) \mathbf{Q}_x(f, F) \mathbf{A}^{-H}(f). \end{aligned} \quad (4.25)$$

The solution can be obtained by using numerical techniques for simultaneous diagonalization (Bunse-Gerstner et al., 1993; Cardoso and Souloumiac, 1996).

### Identifiability

Equations (4.25) have a unique solution (up to rescaling and permutation) if, analogous to equation (4.17), for each  $f = 1, \dots, F$  there exists at least one frequency  $f'$  for which it is fulfilled that

$$\frac{|E\{s_i(t, f) s_i^*(t, f')\}|^2}{E\{|s_i(t, f)|^2\} E\{|s_i(t, f')|^2\}} \neq \frac{|E\{s_j(t, f) s_j^*(t, f')\}|^2}{E\{|s_j(t, f)|^2\} E\{|s_j(t, f')|^2\}} \quad \forall i \neq j. \quad (4.26)$$

### Permutations

The permutations must be sorted for each pair of frequencies  $(f, f')$  by using the method outlined in section 4.3.2.

#### 4.3.4 Time-delayed correlations

If in addition to (4.4) not only the source correlations at equal time are non-zero, but also the time-delayed source correlations do not vanish, i.e., if in general

$$E\{s_i(t, f) s_i^*(t + \tau, f')\} \neq 0, \quad (4.27)$$

then this additional information can be used to derive further diagonalization equations which correspond to time-delayed versions of (4.11). Since time-delayed correlations are also the basis of other source separation algorithms (e.g. Molgedey and Schuster, 1994; Belouchrani et al., 1997), the incorporation of time-delays into the algorithm presented above can be regarded as the combination with existing methods of blind source separation. Time-delayed correlations must vanish for all pairs of different sources due to the assumption of independent sources. Therefore we have

$$E\{s_i(t, f) s_j^*(t + \tau, f')\} = 0 \quad \forall i \neq j, \forall \tau, \forall f, f', \quad (4.28)$$

which is a generalization of (4.3).

Defining, in analogy to (4.5), (4.7), (4.9), (4.10),

$$\mathbf{R}_s(\tau, f, f') = E\{\mathbf{s}(t, f) \mathbf{s}(t + \tau, f')^H\} \quad (4.29)$$

$$\mathbf{R}_x(\tau, f, f') = E\{\mathbf{x}(t, f) \mathbf{x}(t + \tau, f')^H\} \quad (4.30)$$

$$\mathbf{Q}_s(\tau, f, f') = \mathbf{R}_s(\tau, f, f') \mathbf{R}_s(\tau, f', f')^{-1} \mathbf{R}_s(\tau, f', f) \quad (4.31)$$

$$\mathbf{Q}_x(\tau, f, f') = \mathbf{R}_x(\tau, f, f') \mathbf{R}_x(\tau, f', f')^{-1} \mathbf{R}_x(\tau, f', f), \quad (4.32)$$

the inverse mixing matrix  $\mathbf{A}^{-1}(f)$  diagonalizes  $\mathbf{Q}_x(\tau, f, f')$  since

$$\mathbf{Q}_s(\tau, f, f') = \mathbf{A}^{-1}(f) \mathbf{Q}_x(\tau, f, f') \mathbf{A}^{-H}(f). \quad (4.33)$$

Eigenvalue equations corresponding to (4.13) and (4.15), and equations for simultaneous diagonalization corresponding to (4.25) follow immediately. In analogy to (4.26),

a unique solution exists if for each frequency  $f$  there is some frequency  $f'$  and some time-delay  $\tau$  for which it is fulfilled that

$$\frac{|E\{s_i(t, f) s_i^*(t + \tau, f')\}|^2}{E\{|s_i(t, f)|^2\} E\{|s_i(t + \tau, f')|^2\}} \neq \frac{|E\{s_j(t, f) s_j^*(t + \tau, f')\}|^2}{E\{|s_j(t, f)|^2\} E\{|s_j(t + \tau, f')|^2\}} \quad \forall i \neq j. \quad (4.34)$$

Since in contrast to (4.26) also  $\tau \neq 0$  is permitted, condition (4.34) is weaker than (4.26). Permutations are corrected homologously to the method of section 4.3.3, with triples  $(f, f', \tau)$  instead of pairs  $(f, f')$  being used.

## 4.4 Evaluation

To evaluate the algorithm, results are presented for the separation of three data sets. By separating multidimensional data with independent and identically distributed (i.i.d.) Gaussian noise in each frequency, it is demonstrated that the proposed algorithm has the ability to separate data that is inseparable for other algorithms. Next, the correlation structure of color images is analyzed and the proposed algorithm is used to separate mixtures of the images, where each frequency channel has been mixed with a different mixing matrix. Finally, the algorithm is applied to spectrogram data of speech signals from a standard data set and the result is compared to unmixed signals obtained by a different, particularly good separation algorithm on the same data.

### 4.4.1 Synthetic signals

In the first evaluation, a synthetic data set of Gaussian i.i.d. noise in two frequency channels is separated. Since the data in each frequency channel is purely Gaussian, this data cannot be separated by looking at a single frequency only. The actual values of the relevant quantities are given in order to demonstrate in detail the processing steps of the algorithm.

The data consisted of four sources  $s_1(t, f), \dots, s_4(t, f)$ , with time-points  $t = 1, \dots, 10000$  and two frequencies  $f = 1, 2$ . Within each frequency channel of each source, the data was chosen to be i.i.d. noise with Gaussian distribution. To enable separation by the proposed algorithm, correlations were introduced between the data in different frequency channels of each source by composing the signals as the sum

$$s_i(t, f) = \xi_i(t, f) + \zeta_i(t) \quad (4.35)$$

of frequency-dependent and frequency-independent Gaussian random variables  $\xi_i(t, f)$  and  $\zeta_i(t)$ , respectively.

Since the data within each frequency contained neither cues related to higher-order statistics, nor cues related to auto-correlation information or non-stationarity, it is inseparable for any algorithm looking at isolated frequency channels, only. However, taking into account correlations across different frequencies, it can be separated as is demonstrated below.

The correlations within each source and the independence of the different sources are reflected by the covariance matrices<sup>1</sup>  $\mathbf{R}_s(f, f')$ ,

$$\mathbf{R}_s(1, 1) = \begin{pmatrix} 1.99 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.89 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.20 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.04 \end{pmatrix} \quad \mathbf{R}_s(1, 2) = \begin{pmatrix} 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.64 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.16 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.04 \end{pmatrix} \quad (4.36)$$

$$\mathbf{R}_s(2, 1) = \begin{pmatrix} 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.64 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.16 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.04 \end{pmatrix} \quad \mathbf{R}_s(2, 2) = \begin{pmatrix} 2.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.89 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.20 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.04 \end{pmatrix}. \quad (4.37)$$

Since the different sources are independent, the off-diagonal terms of all covariance matrices are zero. The diagonals of  $\mathbf{R}_s(1, 2)$  and  $\mathbf{R}_s(2, 1)$  are non-zero due to the correlations across frequency within each source.

The eigenvalues of equation (4.16) are computed as

$$\text{diag } \mathbf{\Lambda}(1, 2) = (\Lambda_1(1, 2), \dots, \Lambda_4(1, 2)) = (0.25, 0.51, 0.64, 1.00). \quad (4.38)$$

Since all eigenvalues are different, the condition for identifiability (4.17) is fulfilled and mixtures of the sources can be separated by the proposed algorithm.

The mixing matrices  $\mathbf{A}(1)$  and  $\mathbf{A}(2)$  were randomly chosen as

$$\mathbf{A}(1) = \begin{pmatrix} -1.66 & 0.18 & 2.48 & 0.82 \\ -2.53 & -0.51 & -1.42 & -0.30 \\ 0.47 & 0.51 & 0.50 & 2.40 \\ 0.75 & -2.54 & -0.81 & 0.05 \end{pmatrix} \quad \mathbf{A}(2) = \begin{pmatrix} -0.71 & 0.59 & -0.79 & -0.60 \\ 0.61 & -0.99 & -0.85 & -1.02 \\ 0.13 & 0.27 & -0.87 & 1.06 \\ -1.56 & -1.13 & -0.45 & 0.24 \end{pmatrix}. \quad (4.39)$$

Mixing the data by the mixing system (4.2) results in covariance matrices of the mixed signals of

$$\mathbf{R}_x(1, 1) = \begin{pmatrix} 6.78 & 7.55 & -1.15 & -3.26 \\ 7.55 & 13.31 & -2.77 & -2.37 \\ -1.15 & -2.77 & 0.95 & -0.52 \\ -3.26 & -2.37 & -0.52 & 6.96 \end{pmatrix} \quad \mathbf{R}_x(1, 2) = \begin{pmatrix} 0.91 & -1.50 & -0.49 & 2.28 \\ 1.79 & -1.03 & -0.22 & 4.41 \\ -0.27 & -0.19 & 0.18 & -1.12 \\ -1.38 & 2.17 & -0.24 & 0.73 \end{pmatrix} \quad (4.40)$$

$$\mathbf{R}_x(2, 1) = \begin{pmatrix} 0.91 & 1.79 & -0.27 & -1.38 \\ -1.50 & -1.03 & -0.19 & 2.17 \\ -0.49 & -0.22 & 0.18 & -0.24 \\ 2.28 & 4.41 & -1.12 & 0.73 \end{pmatrix} \quad \mathbf{R}_x(2, 2) = \begin{pmatrix} 1.46 & -1.23 & 0.07 & 1.70 \\ -1.23 & 1.81 & 0.02 & -0.85 \\ 0.07 & 0.02 & 0.30 & -0.58 \\ 1.70 & -0.85 & -0.58 & 6.08 \end{pmatrix} \quad (4.41)$$

which are processed by the proposed algorithm, using the eigenvalue method. The separating matrices computed by the algorithm are

$$\hat{\mathbf{A}}^{-1}(1) = \begin{pmatrix} 0.47 & 0.87 & -0.06 & -0.14 \\ -0.77 & 0.54 & 0.33 & -0.10 \\ -0.33 & -0.03 & 0.12 & -0.93 \\ -0.01 & 0.23 & 0.96 & 0.15 \end{pmatrix} \quad \hat{\mathbf{A}}^{-1}(2) = \begin{pmatrix} 0.56 & 0.34 & -0.74 & -0.18 \\ -0.50 & 0.50 & 0.32 & -0.63 \\ 0.71 & -0.49 & 0.07 & -0.50 \\ 0.50 & 0.47 & 0.73 & 0.02 \end{pmatrix}, \quad (4.42)$$

---

<sup>1</sup>All numbers are rounded to two significant digits.

so that the combined mixing-unmixing system  $\hat{\mathbf{A}}^{-1}\mathbf{A}$  is given by

$$\begin{aligned} (\hat{\mathbf{A}}^{-1}\mathbf{A})(1) &= \begin{pmatrix} \boxed{-3.11} & -0.03 & 0.01 & -0.02 \\ 0.00 & 0.01 & \boxed{-2.43} & -0.02 \\ -0.01 & \boxed{2.39} & 0.04 & -0.01 \\ 0.00 & 0.00 & 0.00 & \boxed{2.23} \end{pmatrix} \\ (\hat{\mathbf{A}}^{-1}\mathbf{A})(2) &= \begin{pmatrix} 0.00 & 0.00 & 0.00 & \boxed{-1.50} \\ \boxed{1.69} & 0.01 & -0.03 & -0.02 \\ -0.01 & \boxed{1.49} & 0.02 & 0.02 \\ 0.00 & 0.00 & \boxed{-1.43} & -0.01 \end{pmatrix} \end{aligned} \quad (4.43)$$

Since each row of the combined system contains only one non-zero element, the algorithm has successfully separated the signals. The increase in signal-to-signal from before to after separation amounts to 37.8 dB. However, as can be seen from the different positions of the non-zero elements of  $(\hat{\mathbf{A}}^{-1}\mathbf{A})(1)$  and  $(\hat{\mathbf{A}}^{-1}\mathbf{A})(2)$ , the sources' components are reconstructed in a different order in the two frequency channels.

Therefore, the method for sorting permutations described in section 4.3.2 must be employed. To this end, the estimated eigenvalue matrices  $\hat{\mathbf{\Lambda}}(1, 2)$  and  $\hat{\mathbf{\Lambda}}(2, 1)$  obtained from solving the eigenvalue problems (4.13) and (4.15), respectively, are

$$\hat{\mathbf{\Lambda}}(1, 2) = \begin{pmatrix} \boxed{0.25} & 0.00 & 0.00 & 0.00 \\ 0.00 & \boxed{0.64} & 0.00 & 0.00 \\ 0.00 & 0.00 & \boxed{0.51} & 0.00 \\ 0.00 & 0.00 & 0.00 & \boxed{1.00} \end{pmatrix} \quad \hat{\mathbf{\Lambda}}(2, 1) = \begin{pmatrix} \boxed{1.00} & 0.00 & 0.00 & 0.00 \\ 0.00 & \boxed{0.25} & 0.00 & 0.00 \\ 0.00 & 0.00 & \boxed{0.51} & 0.00 \\ 0.00 & 0.00 & 0.00 & \boxed{0.64} \end{pmatrix}. \quad (4.44)$$

By permuting the eigenvalues on the diagonals of  $\hat{\mathbf{\Lambda}}(1, 2)$  and  $\hat{\mathbf{\Lambda}}(2, 1)$  to occur in the same order in both matrices, and by performing the same permutations for the rows of  $\hat{\mathbf{A}}^{-1}(1)$  and  $\hat{\mathbf{A}}^{-1}(2)$ , respectively, it is ensured that the sources' components are reconstructed in the same order in both frequencies.

#### 4.4.2 Color image data

In this section it is demonstrated that the proposed algorithms can be applied to RGB ('red-green-blue') coded color image data and that it successfully separates images which have been mixed with a different mixing matrix for each color plane. A color version of the results presented in tables 4.1, 4.3 and 4.4 can be obtained from <http://medi.uni-oldenburg.de/demo/ane/diss>.

The three source images are displayed in table 4.1 as the color image (converted to grey-levels) and as the corresponding RGB color planes. From visual inspection it is already obvious that the information contained in the three color planes is not mutually independent but highly correlated since the original images are recognizable in each of the color planes. This fact is also reflected in table 4.2 which displays the correlation coefficients  $c(s_i(f), s_j(f'))$ ,

$$c(s_i(f), s_j(f')) = \frac{E\{s_i(f) s_j(f')\} - E\{s_i(f)\} E\{s_j(f')\}}{\sqrt{E\{s_i^2(f)\} E\{s_j^2(f')\}}}, \quad (4.45)$$

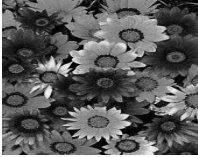











	color image	sources' spectral components		
		$f_{\text{red}}$ component	$f_{\text{green}}$ component	$f_{\text{blue}}$ component
$s_1$				
$s_2$				
$s_3$				

Table 4.1: Original images. Left column: Color images. Three columns to the right: Red, green and blue color planes of the color images.

between data in the different color planes of the different sources. Due to the use of the normalized correlation, the diagonal in table 4.2 is equal to one. Since the data in different color planes of the same source is correlated, high correlations are found for all elements of the 3 blocks on the diagonal of table 4.2. However, across different images, the correlations between any two color planes are close to zero, reflecting that the different source images are almost independent.

The sources were mixed by a different, randomly chosen mixing matrix  $\mathbf{A}(f)$  for each frequency  $f = f_{\text{red}}, f_{\text{green}}, f_{\text{blue}}$  which is shown in table 4.5. The resulting color planes of the mixed signals are displayed in table 4.3.

Separation was performed by jointly diagonalizing equation (4.25) for all possible frequency pairs, using the algorithm of Cardoso and Souloumiac (1996). Afterwards, remaining permutations were corrected for each pair of frequencies, as described in section 4.3.3. The obtained unmixing system  $\hat{\mathbf{A}}^{-1}(f)$  is shown in table 4.5. The unmixed image color planes, shown in table 4.4, and the total mixing-unmixing system  $(\hat{\mathbf{A}}^{-1}\mathbf{A})(f)$ , see table 4.5, demonstrate that the color planes have been successfully separated and that the separated components appear in the same order in every color plane.

		Source $s_1$			Source $s_2$			Source $s_3$		
		$f_{\text{red}}$	$f_{\text{green}}$	$f_{\text{blue}}$	$f_{\text{red}}$	$f_{\text{green}}$	$f_{\text{blue}}$	$f_{\text{red}}$	$f_{\text{green}}$	$f_{\text{blue}}$
$s_1$	$f_{\text{red}}$	1.00	0.80	0.42	-0.03	0.02	0.00	-0.03	-0.03	-0.02
$s_1$	$f_{\text{green}}$	0.80	1.00	0.61	-0.06	0.00	-0.01	0.01	0.00	0.02
$s_1$	$f_{\text{blue}}$	0.42	0.61	1.00	-0.03	-0.01	0.00	0.03	0.02	0.03
$s_2$	$f_{\text{red}}$	-0.03	-0.06	-0.03	1.00	0.65	0.48	0.04	0.03	0.03
$s_2$	$f_{\text{green}}$	0.02	0.00	-0.01	0.65	1.00	0.90	0.06	0.05	0.07
$s_2$	$f_{\text{blue}}$	0.00	-0.01	0.00	0.48	0.90	1.00	0.08	0.08	0.08
$s_3$	$f_{\text{red}}$	-0.03	0.01	0.03	0.04	0.06	0.08	1.00	0.99	0.99
$s_3$	$f_{\text{green}}$	-0.03	0.00	0.02	0.03	0.05	0.08	0.99	1.00	0.99
$s_3$	$f_{\text{blue}}$	-0.02	0.02	0.03	0.03	0.07	0.08	0.99	0.99	1.00

Table 4.2: Correlation coefficients computed between all frequency pairs of all sources. (Rounded to two significant digits.)

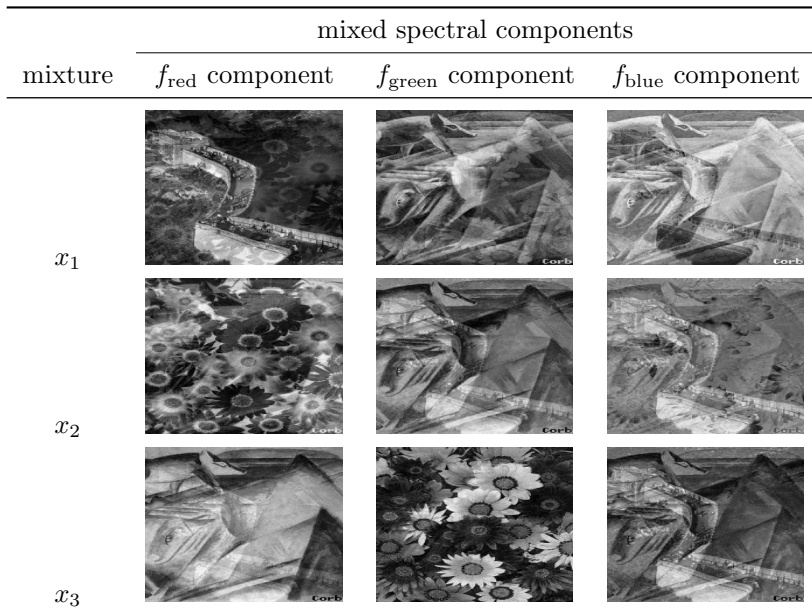


Table 4.3: Mixed color planes. A different, randomly chosen mixing matrix was used for each color plane to mix the three sources. (For better visual appearance, the contrast has been normalized for each image.)







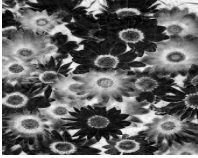

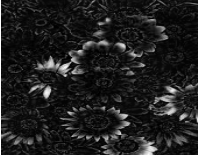
demix	separated spectral components		
	$f_{\text{red}}$ component	$f_{\text{green}}$ component	$f_{\text{blue}}$ component
$u_1$			
$u_2$			
$u_3$			

Table 4.4: Separated color planes obtained from the mixed images (cf. table 4.3) and the estimated separating matrices (cf. table 4.5).

	$f = f_{\text{red}}$	$f = f_{\text{green}}$	$f = f_{\text{blue}}$
$\mathbf{A}(f)$	$\begin{pmatrix} -0.43 & 0.29 & 1.19 \\ -1.67 & -1.15 & -0.04 \\ 0.13 & 1.19 & 0.33 \end{pmatrix}$	$\begin{pmatrix} 0.17 & -0.59 & 0.11 \\ -0.19 & 2.18 & 1.07 \\ 0.73 & -0.14 & 0.06 \end{pmatrix}$	$\begin{pmatrix} -0.10 & -1.34 & -0.69 \\ -0.83 & 0.71 & 0.86 \\ 0.29 & 1.62 & 1.25 \end{pmatrix}$
$\hat{\mathbf{A}}^{-1}(f)$	$\begin{pmatrix} 1.00 & -0.31 & -0.54 \\ -3.37 & 1.00 & 11.26 \\ -0.27 & 0.87 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.27 & -0.16 \\ -18.62 & 1.00 & 3.66 \\ -0.46 & -0.03 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.15 & 0.76 \\ 18.98 & 1.00 & 10.26 \\ 0.69 & -0.93 & 1.00 \end{pmatrix}$
$(\hat{\mathbf{A}}^{-1}\mathbf{A})(f)$	$\begin{pmatrix} 0.01 & 0.00 & \boxed{1.02} \\ 1.20 & \boxed{11.29} & -0.36 \\ \boxed{-1.21} & 0.12 & -0.03 \end{pmatrix}$	$\begin{pmatrix} 0.01 & 0.02 & \boxed{0.39} \\ -0.78 & \boxed{12.64} & -0.84 \\ \boxed{0.65} & 0.08 & -0.02 \end{pmatrix}$	$\begin{pmatrix} 0.00 & 0.01 & \boxed{0.39} \\ 0.37 & \boxed{-7.99} & 0.60 \\ \boxed{1.00} & 0.03 & -0.02 \end{pmatrix}$

Table 4.5: Mixing matrices  $\mathbf{A}(f)$ , separating matrices  $\hat{\mathbf{A}}^{-1}(f)$  computed by the algorithm, and the resulting total mixing-unmixing system  $(\hat{\mathbf{A}}^{-1}\mathbf{A})(f)$  for each frequency. (Rounded to two significant digits.)

#### 4.4.3 Speech signals

In the final experiment, convolutively mixed speech signals, recorded in a real room by Lee (1998b), were separated. The corresponding sound files can be obtained from <http://medi.uni-oldenburg.de/demo/ane/diss>.

The signals were transformed to the time-frequency spectrogram representation (e.g.



(Oppenheim and Schaefer, 1975) and the resulting spectrograms of the microphone signals were separated by the proposed algorithm. Spectrograms of mixed signals  $x_i(t)$  were computed using the short-time Fourier transformation (STFT)

$$x_i(t, f) = \sum_{\tau=0}^{2F-1} x_i(\tau + t) h(\tau) e^{-\pi i f \tau / F}, \quad (4.46)$$

where  $h(t)$  denotes the windowing function, and a window of length 64 ms, a window shift of 16 ms and a DFT length of 128 ms were used. Since the correlations across different frequencies are very low for spectrogram data, nine adjacent frequency channels were used to determine the separation matrix for each frequency  $f$  by the diagonalization method (4.25). After separation, the resulting spectrograms of the unmixed signals were transformed back into the time-domain using the overlap-add method (see also Oppenheim and Schaefer, 1975).

The original microphone signals and the unmixed signals obtained by the algorithm are shown in figure 4.1. As can be seen, and as can also be heard from the sound files of the unmixed signals, separation is quite good, but a clear crosstalk remains audible. For comparison, we have included unmixed signals obtained by the AMDecor algorithm (see Anemüller and Kollmeier, 2000) that almost perfectly separates the signals, cf. figure 4.1. Comparing the two algorithms' performance, the AMDecor algorithm is clearly superior.

It should be noted that further experiments with acoustic signals did not yield more promising results. Even in anechoic environment it was not possible to improve the signal-to-interference ratio by more than 5 dB using the proposed algorithm. This contrasts to an improvement of 15 dB using the AMDecor algorithm and 20 dB using the non-blind minimum-mean-squared-error method.

The reason for this comparably poor performance is seen in the property of the discrete Fourier transformation which results in almost decorrelated spectral components. If instead of the DFT of finite length, an infinitely long Fourier transformation was used, the correlations of different spectral components would be completely eliminated (Papoulis, 1991). In this sense, the remaining correlations can be regarded as a result of the finite window function  $h(t)$ . In contrast, correlations of frequency specific signal envelopes, as employed by the AMDecor algorithm, have a clear origin in the structure of human speech, and it may be argued that this is a more reliable criterion for source separation than the rather artificial effect of a windowing function. Therefore, it might not be surprising that the AMDecor algorithm outperforms the method presented in this paper.

From the point of view of signal statistics, the present algorithm is based on second-order statistics, whereas the AMDecor algorithm exploits signal properties that are related to the notion of fourth-order cross-cumulants from the field of higher-order statistics (e.g. Nikias and Petropulu, 1993). Therefore, the different performance of both algorithms can also be interpreted such that higher-order statistics is the more appropriate mathematical tool to capture across-frequency dependencies of Fourier transformed speech signals.

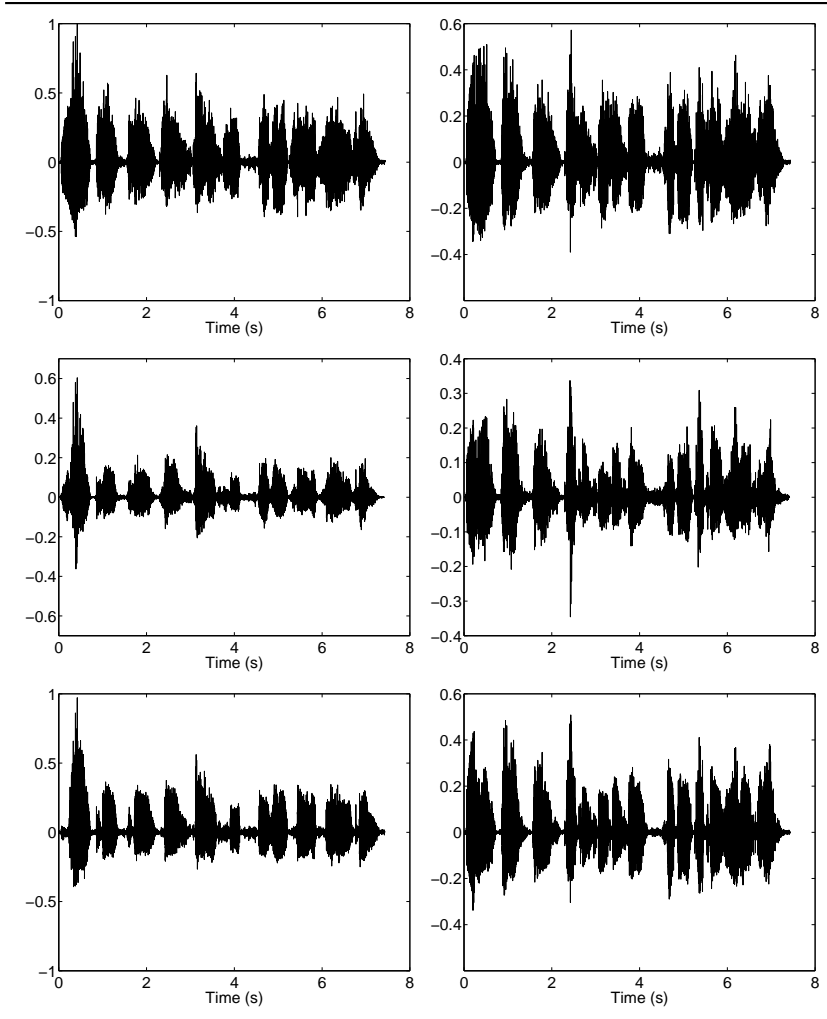


Figure 4.1: Separation of acoustics signals recorded in a real room. Top row: mixed signals recorded by Lee (1998b). Middle row: separated signals obtained by the proposed algorithm. Bottom row, included for comparison: separated speech signals obtained by the AMDecor algorithm (Anemüller and Kollmeier, 2000).

The advantage of the second-order method presented in this paper, however, is that the linear transformation properties of the covariance matrices allow for a fast solution by algebraic methods. In particular, the numerical optimization routine by which the eigenvalue and joint diagonalization equations are solved performs computations that involve the cross-covariance matrices of the mixed signals, only. In contrast, the non-linearity of the magnitude operation which enters the cost-function of the AMDecor algorithm, does not allow for an algebraic solution. Therefore, the separated signals have to be recomputed at each update step of the iterative optimization procedure, which results in a higher computational load.

## 4.5 Discussion

The present paper is based on the notion of multidimensional source signals with dimensions of stationary and non-stationary data and mixing systems. In dimensions of stationary data and mixing both the mixing system and the second order statistics of the source signals remain unchanged with respect to a parameter shift in this dimension. In contrast, the mixing system and second order statistics of the data may change under a parameter shift in those dimensions which are denoted as non-stationary. A typical example are color images which have been mixed with a different mixing matrix for each spectral band. While with respect to the spatial dimension the mixing system is constant and the data assumed to be stationary, both mixing and statistics of the data are distinct in different spectral bands. Motivated by the similar application of sound signal separation using time-frequency representations, the stationary dimensions are denoted as ‘time’ dimensions, and the non-stationary dimensions as ‘frequency’ dimensions.

We have proposed to compute cross-covariance matrices from mixed signals at different frequencies bands in order to reconstruct the sources. Depending on whether data at two or more frequencies is used for separation, the unmixing system is obtained by an eigenvalue decomposition or by solution of a simultaneous diagonalization problem, respectively. Permuted solutions at different frequencies are rearranged by a criterion based on the order of eigenvalues. Evaluation with Gaussian noise, image, and sound data has demonstrated the algorithms’ potential.

Some previous work in the literature has dealt with related problems. Gramss (1995) comments on across-frequency correlations for acoustic signal separation and suggests an iterative optimization scheme, however without addressing the permutation problem. Shamsunder and Giannakis (1997) describe an algorithm based on polyspectra and give a solution for the permutation invariance. Anemüller and Kollmeier (2000) propose an algorithm for the separation of speech signals based on across-frequency correlations of narrow-band signal envelopes, which is closely related to fourth-order cross-cumulants and also constitutes a solution to the permutation problem.

Diamantaras et al. (2000) propose an algorithm for the identification of two-input-two-output FIR channels which is similar to the method presented in this paper. By employing a whitening preprocessing step, a special case of the eigenvalue equations (4.13) and (4.15) is derived. As a consequence, the method proposed in (Diamantaras et al.,

2000) for correcting permutations is based on different principles than the method in the present paper, and is limited to the case of two sources. The unmixing system is estimated using correlations of pairs of two frequencies, only, and separation is performed for convolutively mixed one-dimensional signals.

In view of the existing literature, it is concluded that the present paper elaborates on some ideas previously mentioned in the literature on convolutive signal separation, and extends them to the field of multidimensional sources and to the use of an arbitrary number of sources and frequencies. Furthermore, a novel and appealingly simple solution to the problem of permutations in different frequency channels is given.

It is expected that the methods developed in the present paper can be useful in two applications. For the separation of data with multiple spectral bands, e.g., spectrogram sound data or spectral image data, correlations across different frequencies constitute a criterion for source separation that can be used on its own, or in addition to existing methods of decorrelation with respect to time- or spatial shifts, as outlined in section 4.3.4. By using this additional source of information, it should be possible to improve on the performance of source separation algorithms in a similar way as, e.g., decorrelation with multiple time-delays improves over decorrelation with only a single time-delay (e.g. Murata et al., 1998).

Furthermore, the separation of time-varying mixtures can be improved. Present approaches to time-varying mixtures average over short time segments to estimate the averaged unmixing system. The present method may improve the quality of separation since it allows to estimate the unmixing system for time  $t$  taking into account data from time  $t + \tau$  even though the unmixing system at both times is different, and without necessarily averaging over the entire time from  $t \dots t + \tau$ .

Generalization of the presented approach to take into account also information from higher-order statistics is an open issue. The ansatz presented by Anemüller and Kollmeier (2000) is regarded as a first step into this direction, which makes use of statistical quantities that are closely related to fourth-order cross-cumulants, however, at the expense of a higher computational load than the present algorithm. Further improvements in this direction, e.g., by incorporating ideas from the information maximization framework (e.g. Bell and Sejnowski, 1995), appear as a promising route.

## Chapter 5

# Summary and Conclusion

Three different algorithms for the problem of separating convolutively mixed acoustic signals have been proposed in the present thesis.

In the first approach (cf. chapter 2) the structure of the separating filters was limited to a signal delay and attenuation. Under this constraint, optimal separation can be achieved only in the free field, where the sound signals are superimposed with a finite propagation speed and attenuation, however, without echoes and reverberation. The free field assumption constitutes a first approximation to the true signal propagation and is expected to be appropriate only in rooms with little reverberation and a close distance between sources and microphones.

Adaptation of the filter coefficients is performed by an ICA algorithm for Fourier transformed speech signals which was derived from the principle of maximum likelihood. In the next step, the ICA algorithm was combined with the described filter structure. By devising an unwrapping algorithm for the phases of the complex valued filter coefficients, an algorithm was obtained which uses information from all frequencies to estimate the optimal separating filters.

By making efficient use of information contained in all frequency channels of the mixed signals, this algorithm achieves a very robust and fast convergence within approx. 0.2 s of signal time. Estimation of the separating filter is continuously adapted and computations can be performed in real-time. Therefore, the algorithm is also applicable for the separation of non-stationary signal mixing and represents the first published blind source separation algorithm which has been shown to separate moving speakers and track their position. Intermediate results were published in (Anemüller and Gramß, 1998) and (Anemüller and Gramß, 1999).

The ‘AMDecor’ algorithm presented in chapter 3 can be regarded as complementary to the first approach since no limiting assumptions were imposed with regard to the separating filters. Therefore, the algorithm can be employed to separate signals in rooms with echoes and reverberation. Rather, assumptions are made about the sources’ modulation structure which is assumed to bear similarities with modulations observed in speech signals.

The information transmitted in speech is coded in the frequency dependent change of the signal amplitude. This amplitude modulation of a single speech signal is not independent in different frequency channels, but highly correlated. Considering *two different* speech signals, however, the corresponding correlations vanish due to the assumption of independent sources. This correlation property of speech signals is employed for the task of blind source separation by requiring that the across-frequency correlations of signal amplitudes must vanish for the unmixed signals.

The advantage of this criterion for convolutive blind source separation is that it allows for the first time to achieve both separation of the source components and their consistent ordering across all frequency channels in a single processing step and without limitations for the separating filter. Furthermore, making use of correlations *across-frequency* exploits information for separation that is not considered in other algorithms. Thereby, quality of signal separation is improved, and it is possible to separate also signals which are inseparable for other algorithms. The application to standard data sets showed that the separation is improved compared to other state-of-the-art blind source separation methods. Evaluation using signals with strong reverberation proved that the quality of separation is close to the physical optimum also under very difficult conditions. The comparison with non-blind noise reduction schemes by means of an automatic speech recognition task (cf. appendix B) revealed that source separation exhibits the best improvement in recognition rate for strong interfering noise, whereas alternative methods are superior for soft noise. Intermediate results were presented in (Anemüller, 1999) and (Anemüller and Kollmeier, 2000).

The algorithm presented in chapter 4 also refrains from imposing constraints on the separating filters. Its spirit is very similar to the AMDecor algorithm (cf. chapter 3), however, while the latter has been shown to be based on higher-order statistics, this algorithm approaches the problem by using second-order correlations.

Since the approach involving second-order statistics results in a fully linear problem formulation — in contrast to the AMDecor approach which involves the non-linear magnitude operation — it admits an analytic solution which results in a system of eigenvalue equations and a system of diagonalization equations, respectively, both of which can be solved by efficient numerical techniques. The algorithm has been evaluated using synthetic data, image data and real room speech recordings.

Regarding the separation of acoustic signals, it is shown that the second-order correlations encountered in spectrogram data of speech signals result from the finite length of the discrete Fourier transformation. These second-order correlations are small compared to the envelope correlations used in the AMDecor algorithm. This fact is regarded as the reason why the resulting algorithm does not perform as good as the AMDecor algorithm on acoustic signals.

However, the proposed algorithm can still be of interest for the separation of sources like spectral image data, for which the notion of multidimensional sources has been introduced. Furthermore, the formalism of multidimensional sources applies to the separation of sources which are mixed with a time-varying mixing system, as well. The common approach in blind source separation algorithms for time-varying mixing systems is to average over short time-intervals during which the mixing system is re-

garded as constant. The proposed method appears to have the potential to obtain better signal separation with shorter averaging intervals since it allows to estimate the current mixing system from data at several time-points *without* assuming the stationarity of the mixing system over the corresponding time interval. This possibility could be of great interest for several applications, however, it remains to be demonstrated that the theoretical advantage can be achieved in practice.

In conclusion, the present thesis addressed the problem of convolutive blind source separation from different points of view.

By specifying a constrained model for the separating filter, rapid convergence and adaptation has been demonstrated.

By equipping the algorithms with appropriate models of statistical dependencies within each source, very good performance and correct ordering of unmixed spectral components have been achieved in difficult acoustical situations.

It has been wondered (e.g. Ikram and Morgan, 2000) why frequency domain based blind source separation algorithms are susceptible to local permutations of unmixed signal components at different frequencies, while time-domain algorithms do not appear to have this problem. The results presented in this thesis allow to draw the conclusion that local permutations occur only if the assumed model for the sources is not appropriate, e.g., if it is assumed that source components at different frequencies were independent. While frequency domain algorithms based on such assumptions are invariant with regard to local permutations, algorithms with a cost function defined in the time domain can detect permutations. Since applying the inverse Fourier transformation to permuted spectral components from different sources results, due to the central limit theorem, in a more Gaussian histogram of the time domain signals than would be the case if no permutations had been present, permutations of source components result in higher values of the time domain cost function and are therefore penalized by time domain blind source separation algorithms.

Finally, it has been shown that introducing the notion of multidimensional sources and modeling the dependencies by means of second-order statistics results in a novel approach which may be applicable in domains such as image processing and in the field of blind source separation involving time-varying mixing systems.





# Appendix A

## Technical Appendix

### A.1 Optimization under unitary matrix constraint

A preprocessing step is described which reduces the number of free parameters of the optimization problem (3.20). It is based on a standard pre-whitening method employed in several blind source separation algorithms (e.g. Comon, 1994; Cardoso and Souloumiac, 1996; Murata et al., 1998), which is slightly modified to fit to the application to speech signals. It is based on the fact that any separating matrix  $\mathbf{W}(f)$  can be written as the product of a ‘whitening matrix’  $\mathbf{V}(f)$  and a unitary matrix  $\mathbf{U}(f)$  (Comon, 1994),

$$\mathbf{W}(f) = \mathbf{U}(f) \mathbf{V}(f). \quad (\text{A.1})$$

Since the unmixed signals must be second order uncorrelated, the decorrelation is imposed on the signals in a pre-processing step. Hence, the recorded signals  $\mathbf{x}(T, f)$  are transformed to uncorrelated signals  $\tilde{\mathbf{x}}(T, f)$  by a matrix  $\mathbf{V}(f)$  such that

$$\tilde{\mathbf{x}}(T, f) = \mathbf{V}(f) \mathbf{x}(T, f) \quad (\text{A.2})$$

$$E \{ \tilde{\mathbf{x}}(T, f) \tilde{\mathbf{x}}^H(T, f) \} = \eta(f) \mathbf{I}, \quad (\text{A.3})$$

where  $\eta(f) \mathbf{I}$  is the rescaled identity matrix and  $\mathbf{V}(f)$  is chosen such that the total power of  $\tilde{\mathbf{x}}(T, f)$  at each frequency  $f$  equals the total power of  $\mathbf{x}(T, f)$  at the same frequency  $f$ .

The standard pre-whitening method sets the scaling of  $\mathbf{V}(f)$  such that  $\eta(f) = 1$  for all frequencies. However, this choice would result in the same signal power at all frequencies, which for speech signals amounts to an amplification of the high frequencies. Hence,  $\mathbf{V}(f)$  is chosen such that the signal power in each frequency channel is conserved. Note that the decorrelation is performed separately for each frequency and that it ensures second order decorrelation of the complex spectrograms  $\tilde{\mathbf{x}}(T, f)$ . The energy function (3.20), in contrast, is computed from the amplitude spectrograms and constitutes a more restrictive condition on the unmixed signals.

Any unitary matrix  $\mathbf{U}$  can be written as  $\mathbf{U} = \begin{pmatrix} y & z \\ z^* & y \end{pmatrix}$  with complex numbers  $y$  and  $z$  which fulfill  $|y|^2 + |z|^2 = 1$  (Cardoso and Souloumiac, 1996). Taking into account the invariance with respect to rescaling of the rows, the unitary separating matrix is parameterized as

$$\mathbf{U}(f) = \begin{pmatrix} \cos(\theta(f)) & \sin(\theta(f)) \exp(i\phi(f)) \\ -\sin(\theta(f)) \exp(-i\phi(f)) & \cos(\theta(f)) \end{pmatrix}. \quad (\text{A.4})$$

Hence, the number of parameters which need to be determined is reduced from two complex numbers without the preprocessing step to only two real angles after the preprocessing.

After preprocessing, the optimization scheme presented in section 3.4.4 is performed for the uncorrelated signals  $\tilde{\mathbf{x}}(T, f)$  and the unitary matrices  $\mathbf{U}(f)$  instead of the microphone signals  $\mathbf{x}(T, f)$  and the matrices  $\mathbf{W}(f)$ , respectively. Matrices  $\mathbf{U}(f)$  are parameterized by angles  $\theta(f)$  and  $\phi(f)$  and evaluation of the gradient of  $H$  with respect to  $\theta(f)$  and  $\phi(f)$  is performed numerically.

Since the preprocessing with the subsequent rotation results in separated signals with each output  $u_1(T, f), \dots, u_M(T, f)$  having equal mean power, it is necessary to rescale the output signals. This is done by first computing the total separating system as

$$\mathbf{W}(f) = \mathbf{U}(f) \mathbf{V}(f). \quad (\text{A.1})$$

and subsequently rescaling the rows of  $\mathbf{W}(f)$  such that each row has norm one and the diagonal of rescaled  $\mathbf{W}(f)$  is real. Afterwards, the output signals of the algorithm are computed from the matrix product (3.5) of rescaled matrix  $\mathbf{W}(f)$  with the microphone spectrograms  $\mathbf{x}(T, f)$ .

## A.2 Determination of the SIR

Signal-to-interference-ratios (SIRs) are computed based on knowledge of each source's energy transmission to the left and right microphone, respectively. Denote by  $\mathbf{s}^{(1)}(t) = [s_1^{(1)}(t), s_2^{(1)}(t)]^T$  the stereo signal of source one as recorded by the two microphones. Homologously, denote by  $\mathbf{s}^{(2)}(t) = [s_1^{(2)}(t), s_2^{(2)}(t)]^T$  the stereo signal corresponding to source two.

From the corresponding spectrograms,  $\mathbf{s}^{(1)}(T, f)$  and  $\mathbf{s}^{(2)}(T, f)$ , and the unmixing matrix  $\mathbf{W}(f)$  determined by the algorithm, the unmixed spectrograms

$$\mathbf{u}^{(1)}(T, f) = \mathbf{W}(f) \mathbf{s}^{(1)}(T, f) \quad (\text{A.5})$$

and

$$\mathbf{u}^{(2)}(T, f) = \mathbf{W}(f) \mathbf{s}^{(2)}(T, f) \quad (\text{A.6})$$

are computed.

In the case of perfect separation,  $\mathbf{u}^{(1)}(T, f)$  has non-zero signal components in only one component, e.g.,  $u_2^{(1)}(T, f) = 0$ . Conversely,  $\mathbf{u}^{(2)}(T, f)$  has non-zero signal-components in only the other component, i.e.,  $u_1^{(2)}(T, f) = 0$ .

For the more realistic case of cross talk energy between both components, the SIR is computed as follows. Denote by  $P_{ij}(f) = E\{|u_i^{(j)}(T, f)|^2\}$  the power in frequency  $f$  and component  $i$  of the  $j$ -th unmixed signal. The total power in component  $i$  is

$$P_{ij} = \sum_f P_{ij}(f) \quad (\text{A.7})$$

The signal to interference ratio (SIR) is defined as the ratio of direct-path energy to cross-talk energy where the assignment of direct-path and cross-talk is done such that the resulting SIR is maximized. For two sources, two assignments are possible, therefore the SIR is defined as

$$\text{SIR} = \max \left\{ \frac{P_{11} + P_{22}}{P_{12} + P_{21}}, \frac{P_{12} + P_{21}}{P_{11} + P_{22}} \right\}. \quad (\text{A.8})$$

Depending on which quotient in (A.8) is larger, the frequency specific SIR is computed as

$$\text{SIR}(f) = \begin{cases} \frac{P_{11}(f) + P_{22}(f)}{P_{12}(f) + P_{21}(f)} & \text{for SIR} = \frac{P_{11} + P_{22}}{P_{12} + P_{21}} \\ \frac{P_{12}(f) + P_{21}(f)}{P_{11}(f) + P_{22}(f)} & \text{for SIR} = \frac{P_{12} + P_{21}}{P_{11} + P_{22}} \end{cases} \quad (\text{A.9})$$

The SIR prior to separation is computed analogously from the stereo recordings  $\mathbf{s}^{(1)}(T, f)$  and  $\mathbf{s}^{(2)}(T, f)$  *without* the unmixing systems (A.5) and (A.6) being applied. The gain in SIR accomplished by the processing is computed as the difference between input and output SIR,

$$\text{SIR}_{\text{gain}} = \text{SIR}_{\text{out}} - \text{SIR}_{\text{in}} \quad (\text{A.10})$$

$$\text{SIR}_{\text{gain}}(f) = \text{SIR}_{\text{out}}(f) - \text{SIR}_{\text{in}}(f). \quad (\text{A.11})$$

### A.3 Non-blind correction of local permutations

Since for the purpose of evaluation the original source signals are known, this knowledge is exploited to correct for local permutations in the unmixed signals. The assignment of unmixed signals' components to each source is performed such that the components of the first unmixed signal correspond to the first source's components in each frequency channel. The correct assignment is determined by requiring that for each frequency the frequency specific SIR must be equal to

$$\text{SIR}(f) = \frac{P_{11}(f) + P_{22}(f)}{P_{12}(f) + P_{21}(f)}. \quad (\text{A.9})$$

After the unmixed components have been assigned accordingly, the total SIR is computed from (A.8).

## A.4 Non-blind MMSE separation

As a reference method, we perform non-blind separation based on the minimum-mean-squared-error method (MMSE) by determining the optimal linear reconstruction of the source signals' spectrograms from the mixed signals' spectrograms. The computation is performed in the frequency domain to obtain the optimal separation that can be achieved by a non-blind method, subject to the given filter-length.

The unmixing model for each frequency  $f$  is given by

$$w_{11}^{(ls)}(f) x_1(T, f) + w_{12}^{(ls)}(f) x_2(T, f) = s_1^{(1)}(T, f) + \epsilon_1(T, f) \quad (\text{A.12})$$

$$w_{21}^{(ls)}(f) x_1(T, f) + w_{22}^{(ls)}(f) x_2(T, f) = s_2^{(2)}(T, f) + \epsilon_2(T, f), \quad (\text{A.13})$$

where  $s_j^{(i)}(T, f)$  is defined as in appendix A.2. The parameters  $w_{ij}^{(ls)}$  are to be determined such that the reconstruction errors

$$E \{ |\epsilon_1(T, f)|^2 \} \quad \text{and} \quad E \{ |\epsilon_2(T, f)|^2 \} \quad (\text{A.14})$$

are minimized for all frequencies  $f$ .

The optimal solution of  $w_{ij}^{(ls)}(f)$  is determined from the linear equations

$$\begin{pmatrix} E\{|x_1(T, f)|^2\} & E\{x_1^*(T, f) x_2(T, f)\} \\ E\{x_1(T, f) x_2^*(T, f)\} & E\{|x_2(T, f)|^2\} \end{pmatrix} \begin{pmatrix} w_{11}^{(ls)}(f) \\ w_{12}^{(ls)}(f) \end{pmatrix} = \begin{pmatrix} E\{s_1^{(1)}(T, f) x_1^*(T, f)\} \\ E\{s_1^{(1)}(T, f) x_2^*(T, f)\} \end{pmatrix} \quad (\text{A.15})$$

$$\begin{pmatrix} E\{|x_1(T, f)|^2\} & E\{x_1^*(T, f) x_2(T, f)\} \\ E\{x_1(T, f) x_2^*(T, f)\} & E\{|x_2(T, f)|^2\} \end{pmatrix} \begin{pmatrix} w_{21}^{(ls)}(f) \\ w_{22}^{(ls)}(f) \end{pmatrix} = \begin{pmatrix} E\{s_2^{(2)}(T, f) x_1^*(T, f)\} \\ E\{s_2^{(2)}(T, f) x_2^*(T, f)\} \end{pmatrix}. \quad (\text{A.16})$$

## Appendix B

# Blinde Quellentrennung als Vorverarbeitung zur robusten Spracherkennung

### B.1 Einleitung

In diesem Beitrag evaluieren wir den Nutzen blinder Quellentrennung als Vorverarbeitungsstufe zum Zwecke robuster automatischer Spracherkennung. Blinde Quellentrennung (QT) ist eine Signalverarbeitungstechnik, die es ermöglicht, aus mehreren Aufnahmen akustischer Überlagerungen (etwa Sprache im Störgeräusch) die zugrunde liegenden Quellsignale (Sprache getrennt vom Störgeräusch) zu rekonstruieren. Ein spezieller Algorithmus für QT in verhallter Umgebung ist bereits vorgestellt worden (Anemüller, 1999). Eine potentielle Anwendung solcher Algorithmen besteht in der Störgeräuschbefreiung für die robuste automatische Spracherkennung. Das Perzeptionsmodell (PEMO) nach Dau et al. (Dau et al., 1996) wurde bereits zur Merkmalsextraktion in der automatischen Spracherkennung verwendet. Insbesondere in Kombination mit Neuronalen Netzen hat diese gehörgerechte Vorverarbeitung zu einer robusten Erkennungsleistung im Störgeräusch geführt (Tchorz and Kollmeier, 1999). Wir kombinierten den QT-Algorithmus mit einem Einzelworterkennungssystem auf Basis des PEMO, um eine weitere Verbesserung der Erkennungsleistung zu erreichen. Zur Evaluation vergleichen wir die Erkennungsraten bei QT-Vorverarbeitung mit denen ohne Vorverarbeitung und mit alternativen Störgeräuschunterdrückungssystemen. Berücksichtigt werden hierbei Aufnahmesituationen in verhallter und unverhallter Umgebung und bei unterschiedlichen Signal-Rausch Abständen.

---

<sup>1</sup>This appendix is a reprint of the publication (Anemüller et al., 2000).

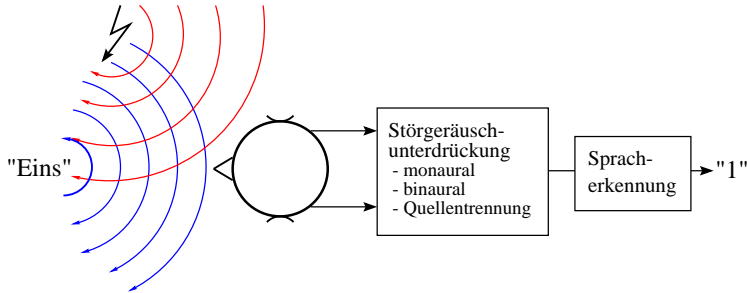


Abbildung B.1: Schematische Darstellung des Versuchsaufbaus.

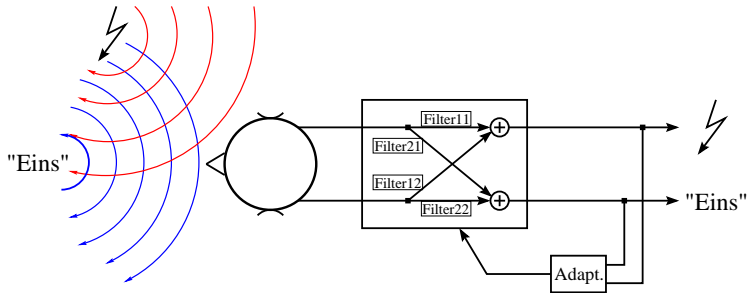


Abbildung B.2: Die Architektur des verwendeten Quellentrenners.

## B.2 Blinde Quellentrennung

Algorithmen zur blinden Quellentrennung zeichnen sich dadurch aus, dass sie sehr geringe Annahmen über die vorliegenden Signale machen. Es wird nur vorausgesetzt, dass die Signalquellen voneinander unabhängig sind, und dass die gleiche Anzahl Mikrofone wie Signalquellen vorhanden ist. Insbesondere sind die räumlichen Positionen von Quellen und Mikrofonen unbekannt — daher "blind" —, was blinde Quellentrennung für robuste Spracherkennung besonders interessant macht. Sind die Annahmen erfüllt, dann ist durch Filtern und Überlagern der Mikrofonsignale eine Rekonstruktion der getrennten Quellsignale, bis auf eine prinzipiell unbestimmbare Verzerrung, möglich.

Da die dazu benötigten Filter unbekannt sind, werden sie durch einen Optimierungsalgorithmus iterativ geschätzt, siehe Fig. B.2. Die Schlüsselfrage hierzu lautet, wie der Algorithmus bestimmt, ob die rekonstruierten Signale unabhängig oder noch vermischt sind. Kriterien hierfür können aufgrund verschiedener statistischer Maße definiert werden, siehe etwa (Nadal and Parga, 1997). Der von uns verwendete Algorithmus (Anemüller, 1999) benutzt — motiviert durch Eigenschaften von Sprache — die in verschiedenen Frequenzbändern korrelierte Amplitudenmodulation der Quell-

signale. Dazu werden zwischen den rekonstruierten Signalen die Korrelationen der frequenzspezifischen Einhüllenden *frequenzübergreifend* berechnet, also zwischen allen Frequenzbändern  $f_i$  des ersten rekonstruierten Signals und allen Frequenzbändern  $f_j$  des zweiten rekonstruierten Signals. Die Signale sind dann getrennt, wenn in diesem Sinn eine maximale Dekorrelation erreicht ist. Für eine genauere Beschreibung verweisen wir auf (Anemüller, 1999).

Der benutzte Algorithmus rekonstruiert jeweils die von einer Quelle an den Mikrofonen hervorgerufenen Signale; eine Entfaltung der Raumübertragungsfunktion wird also nicht vorgenommen. Tests mit verschiedenen Signalen zeigen, dass der Algorithmus eine gute Signaltrennung erreicht. Audio-Beispiele sind von der oben genannten WWW-Seite abrufbar.

## B.3 Robuste Spracherkennung

Das Perzeptionsmodell (PEMO) nach Dau et al. (Dau et al., 1996) ist ein funktionelles Modell der Signalverarbeitung im peripheren auditorischen System. Es ist in der Lage, das Antwortverhalten von Versuchspersonen in einer Vielzahl von psychoakustischen Experimenten quantitativ nachzubilden. Das PEMO extrahiert aus einem eintreffenden akustischen Signal die dazugehörige *interne Repräsentation*, welche sich bereits als ein robustes Merkmal für die automatische Spracherkennung bewährt hat (Tchorz and Kollmeier, 1999). Insbesondere in Kombination mit dem lokal-rekurrenten neuronalen Netz (LRNN) als Klassifikator übertrifft die PEMO Vorverarbeitung konventionelle Mel-Cepstralkoeffizienten deutlich an Robustheit gegenüber additiven Störgeräuschen (Kasper et al., 1997). Weiterhin wurde gezeigt, dass sich durch eine Filterung der eintreffenden Zeitsignale mittels monauraler (Kleinschmidt et al., 1998a) und binauraler (Kleinschmidt et al., 1998b) Algorithmen zu Störgeräuschreduktion die Erkennungsleistung des PEMO/LRNN Systems bei additiven Störgeräuschen beträchtlich steigern lässt. Voraussetzung ist dabei allerdings eine zuverlässige Sprachpausendetektion und Stationarität des Störgeräusches für die monaurale, bzw. die Kenntnis der Lage der Schallquellen im Raum für die binaurale Störgeräuschreduktion.

## B.4 Methoden

Es wurden Kunstkopfaufnahmen von Sprache und Störgeräusch aus reflexionsarmer und aus verhallter Umgebung benutzt. Die Aufnahme in verhallter Umgebung fand in einem Seminarraum mit einer Nachhallzeit  $T_{60}$  von ca. 0.5s statt. In allen Fällen betrug der Abstand zwischen Kunstkopf und Lautsprechern etwa 2.5m. Das Sprachsignal kam von vorn, das Störgeräusch von 30 Grad schräg rechts. Diese Signale wurden nachträglich abgemischt bei Signal-Rausch-Abständen (SNR) von -10dB, 0dB und 10dB.

Als Sprachsignale wurden die Wörter "Null" bis "Neun" aus dem ZIFKOM Datensatz verwendet. Insgesamt standen 2000 Artikulationen der Wörter, gesprochen von 200 verschiedenen Sprecherinnen und Sprechern, zur Verfügung. Diese wurden jeweils zur

Hälfte als Trainings- und als Testdatensatz für die sprecherunabhängige Spracherkennung benutzt. Als Störgeräusch diente ein sprachähnliches Rauschen ('babble-noise'), das aus der Überlagerung mehrerer Sprachsignale besteht. Zur Schätzung der optimalen Filter standen dem Quellentrenner für jede Versuchssituation nur die Wörter "Null" bis "Fünf" eines einzigen Sprechers, überlagert mit dem Störgeräusch, zur Verfügung, da die Benutzung des gesamten Testmaterials zu rechenaufwendig gewesen wäre. Die so gefundenen Filter dienten zur Trennung des gesamten Testmaterials in Sprache und Störgeräusch. Die Klassifikation in Sprach- bzw. Störsignal wurde anhand der erzielten Erkennungsrate vorgenommen. Die verwendeten Filter hatten eine Länge von 1536 taps bei einer Samplingrate von 16kHz. Diese große Filterlänge wurde gewählt, um sicherzustellen, dass die Trennung der Signale bei der gegebenen Raumakustik mit langer Nachhallzeit und großem Abstand zwischen den Kunstkopfmikrofonen und den Lautsprechern überhaupt möglich ist.

Zur Spracherkennung wurde die beschriebene Kombination aus PEMO-Vorverarbeitung und LRNN-Klassifikation benutzt. Hierbei wurden zwei neuronale Netzwerke benutzt, die sich darin unterscheiden, dass eines auf reflexionsarm aufgenommenes Trainingsmaterial und das zweite auf verhalltes Trainingsmaterial trainiert wurde.

Der beschriebene Aufbau, siehe Abb. 1, entspricht genau dem der Experimente von Kleinschmidt et al. (Kleinschmidt et al., 1999), so dass die durch den QT Algorithmus erreichte Verbesserung der Erkennungsleistung direkt mit den bereits vorliegenden Werten für den Ephraim-Malah Algorithmus und das binaurale Richtungsfilter nach Wittkop verglichen werden kann.

## B.5 Ergebnisse

Die Erkennungsraten in reflexionsarmer Umgebung für die drei verwendeten SNR-Werte sind in Fig. B.3 dargestellt. Für den SNR von  $-10\text{dB}$  liegt die Erkennungsrate ohne Vorverarbeitung nur unwesentlich über dem Zufallsniveau von 10%. Blinde Quellentrennung erreicht hier eine drastische Verbesserung bis hin zu fast 80% Erkennungsrate. Diese Verbesserung ist signifikant größer als die durch die alternativen Störgeräuschunterdrücker erreichten. Bei  $0\text{dB}$  SNR erzielt die Quellentrennung im Vergleich zu den anderen Algorithmen eine vergleichbare bzw. geringfügig niedrigere, jedoch signifikante Verbesserung. Bei einem Pegel von  $10\text{dB}$  SNR schließlich bricht die Erkennungsrate bei Quellentrennung ein und liegt sowohl unter der Erkennungsrate ohne Störgeräuschunterdrückung als auch unter dem für  $0\text{dB}$  SNR mit Quellentrennung erreichten Wert.

Es fällt auf, dass der auf verhallte Sprache trainierte LRNN-Klassifikator bei Quellentrennung in reflexionsarmer Umgebung besser klassifiziert als der auf reflexionsarm aufgenommene Sprache trainierte LRNN-Klassifikator. Dies ist vermutlich die Folge eines geringfügigen Kammfiltereffektes, der in diesem Fall bei der Quellentrennung als Artefakt auftrat und auch bei Hörtests wahrnehmbar war.

Die Erkennungsraten in verhallter Umgebung sind in Fig. B.4 dargestellt. Die Ergebnisse sind vergleichbar mit denen in reflexionsarmer Umgebung: bei  $-10\text{dB}$  SNR erreicht Quellentrennung die größte Verbesserung aller betrachteten Störgeräuschun-



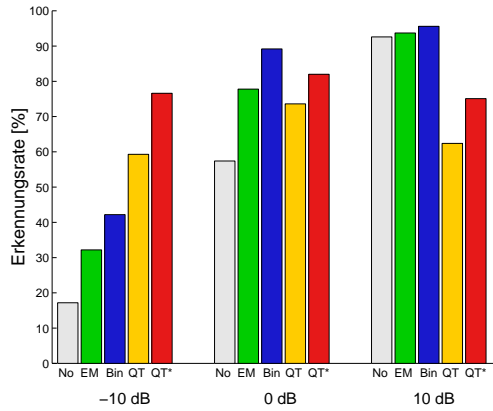


Abbildung B.3: Erkennungsleistung im reflexionsarmen Raum für drei SNR-Werte. *No*: Keine Störgeräuschunterdrückung, *EM*: monaural nach Ephraim–Malah, *Bin*: binaurales Richtungsfilter, *QT*: Quellentrenner und reflexionsarmum trainiertes LRNN, *QT\**: Quellentrenner und verhallt trainiertes LRNN

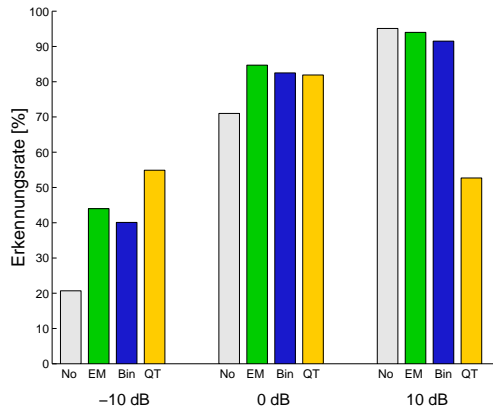


Abbildung B.4: Erkennungsleistung verhallter Umgebung. Bezeichnungen wie in Fig. B.3, außer *QT*: Quellentrenner und verhallt trainiertes LRNN

terdrücker; bei 0dB ist die Erkennungsrate für alle Störgeräuschunterdrücker ähnlich; bei 10dB bricht die Erkennungsrate bei Quellentrennung ein.

Der Grund für die schlechten Ergebnisse mit Quellentrennung bei 10dB SNR liegt vermutlich darin, dass bei diesem Pegel die Annahmen des Quellentrenners verletzt sind. Eine Schätzung des diffusen Aufnahmerauschens in den Sprachsignalen ergibt, dass dessen Pegel frequenzabhängig im Bereich von etwa  $-35\text{dB}$  bis  $-10\text{dB}$  relativ zum Sprachsignal liegt. Bei 10dB SNR erreicht damit das diffuse Aufnahmerauschen in einigen Frequenzbereichen vergleichbare Pegel wie das lokalisierte Störgeräusch. Es stellt damit effektiv eine dritte Signalquelle dar, was die Annahme von nur zwei Signalquellen verletzt, so dass der Quellentrenner keine Signaltrennung mehr erreichen kann.

## **B.6 Zusammenfassung**

Wegen ihrer minimalen Annahmen über Sprach- und Störsignal ist blinde Quellentrennung interessant als Störgeräuschunterdrückung für robuste Spracherkennung. Der verwendete Quellentrennungsalgorithmus erreicht erfahrungsgemäß eine gute Signaltrennung. Dies resultiert für SNR-Werte von  $-10\text{dB}$  in einer deutlichen Verbesserung der Erkennungsleistung des Spracherkenners. Bei 0dB SNR ist die Verbesserung durch den Quellentrenner vergleichbar mit den durch alternative Störgeräuschunterdrücker erreichten. Sind jedoch die Annahmen des Quellentrenners verletzt, in diesem Fall durch Aufnahmerauschen bei 10dB SNR, dann kann die Erkennungsleistung zusammenbrechen. Ein weiteres Problem für automatische Spracherkennung können durch die Quellentrennung erzeugte spektrale Veränderungen der Signale, wie etwa Nachhall, darstellen.

Bedanken möchten wir uns bei Klaus Kasper und Herbert Reininger von der Universität Frankfurt dafür, dass sie uns ihre LRNN Implementation zur Benutzung überlassen haben.

Diese Arbeit wurde von der Deutschen Forschungsgemeinschaft im Rahmen des Graduiertenkollegs Psychoakustik unterstützt.

# References

- Amari, S., Cichocki, A., and Yang, H. H., 1996.  
A new learning algorithm for blind signal separation.  
In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 757–763.  
7, 19
- Amari, S., Douglas, S. C., Cichocki, A., and Yang, H. H., 1997.  
Multichannel blind deconvolution and equalization using the natural gradient.  
In *Proceedings of the IEEE Workshop on Signal Processing Advances in Wireless Communications*, pages 101–104. Paris, France.  
8, 34
- Anemüller, J., 1999.  
Correlated modulation: A criterion for blind source separation.  
In *Joint meeting of the Acoustical Society of America and the European Acoustics Association*. Berlin, Germany.  
4 pages on CD-ROM proceedings.  
84, 91, 92, 93
- Anemüller, J. and Gramß, T., 1998.  
Blinde akustische Quellentrennung im Frequenzbereich.  
In A. Sill, editor, *Fortschritte der Akustik: DAGA 98*, pages 350–351. Deutsche Gesellschaft für Akustik (DEGA), Zürich, Switzerland.  
83
- Anemüller, J. and Gramß, T., 1999.  
On-line blind separation of moving sound sources.  
In J. F. Cardoso, C. Jutten, and P. Loubaton, editors, *Proceedings of the first international workshop on independent component analysis and blind signal separation*, pages 331–334. Aussois, France.  
29, 83
- Anemüller, J., Kleinschmidt, M., and Kollmeier, B., 2000.  
Blinde Quellentrennung als Vorverarbeitung zur robusten Spracherkennung.  
In V. Mellert, editor, *Fortschritte der Akustik: DAGA 2000*, pages 364–365. Deutsche Gesellschaft für Akustik (DEGA), Oldenburg, Germany.

- 12, 38, 91
- Anemüller, J. and Kollmeier, B., 2000.  
Amplitude modulation decorrelation for convolutive blind source separation.  
In P. Pajunen and J. Karhunen, editors, *Proceedings of the second international workshop on independent component analysis and blind signal separation*, pages 215–220. Helsinki, Finland.  
12, 27, 39, 66, 68, 79, 80, 81, 82, 84
- Attias, H. and Schreiner, C. E., 1998.  
Blind source separation and deconvolution: The dynamic component analysis algorithm.  
*Neural Computation*, 10:1373–1424.  
37, 62
- Back, A. D. and Weigend, A. S., 1997.  
A first application of independent component analysis to extracting structure from stock returns.  
*International Journal of Neural Systems*, 8:473–484.  
6
- Bell, A. J. and Sejnowski, T. J., 1995.  
An information maximization approach to blind separation and blind deconvolution.  
*Neural Computation*, 7:1129–1159.  
7, 16, 33, 65, 82
- Bell, A. J. and Sejnowski, T. J., 1996.  
Learning the higher-order structure of a natural sound.  
*Network: Computation in Neural Systems*, 7:261–266.  
6
- Bell, A. J. and Sejnowski, T. J., 1997.  
The ‘independent components’ of natural scenes are edge filters.  
*Vision Research*, 37:3327–3338.  
6, 66
- Belouchrani, A., Abed-Meraim, K., Cardoso, J. F., and Moulines, E., 1997.  
A blind source separation technique using second order statistics.  
*IEEE Transactions on Speech and Audio Processing*, 45:434–444.  
33, 65, 72
- Bienvenu, G. and Kopp, L., 1983.  
Optimality of high-resolution array processing using the eigensystem approach.  
*IEEE Transactions on Acoustics, Speech and Signal Processing*, 31:1235–1248.  
7
- Bishop, C. M., 1995.  
*Neural networks for pattern recognition*.

- Oxford University Press, Oxford.  
16
- Brehm, H. and Stammers, W., 1987.  
Description and generation of spherically invariant speech-model signals.  
*Signal Processing*, 12:119–141.  
16
- Bunse-Gerstner, A., Byers, R., and Mehrmann, V., 1993.  
Numerical methods for simultaneous diagonalization.  
*SIAM Journal on Matrix Analysis and Applications*, 14:927–949.  
71
- Capdevielle, V., Servière, C., and Lacoume, J. L., 1995.  
Blind separation of wide band sources in the frequency domain.  
In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2080–2083. Detroit, MI.  
7, 8, 34, 37, 66
- Cappé, O., 1994.  
Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor.  
*IEEE Transactions on Speech and Audio Processing*, 2:345–349.  
5
- Cardoso, J.-F. and Laheld, B. H., 1996.  
Equivariant adaptive source separation.  
*IEEE Transactions on Signal Processing*, 44:3017–3030.  
7, 16, 19, 33, 65
- Cardoso, J.-F. and Souloumiac, A., 1996.  
Jacobi angles for simultaneous diagonalization.  
*SIAM Journal on Matrix Analysis and Applications*, 17:161–164.  
47, 71, 76, 87, 88
- Chan, D. C. B., Rayner, P. J. W., and Godsill, S. J., 1996.  
Multi-channel signal separation.  
In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 649–652. Atlanta, GA.  
7
- Comon, P., 1994.  
Independent component analysis, a new concept?  
*Signal Processing*, 36:287–314.  
7, 16, 33, 47, 48, 65, 87
- Dau, T., Püschel, D., and Kohlrausch, A., 1996.  
A quantitative model of the “effective” signal processing in the auditory system i.

- Journal of the Acoustical Society of America*, 99:3615–3622.  
91, 93
- Diamantaras, K. I., Petropulu, A. P., and Chen, B., 2000.  
Blind two-input-two-output FIR channel identification based on frequency domain second-order statistics.  
*IEEE Transactions on Signal Processing*, 48:534–542.  
35, 81
- Ehlers, F. and Schuster, H. G., 1997.  
Blind separation of convolutive mixtures and an application in automatic speech recognition in noisy environment.  
*IEEE Transactions on Signal Processing*, 45:2608–2612.  
8, 34, 37, 66
- Ephraim, Y. and Malah, D., 1984.  
Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator.  
*IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-32:1109–1121.  
5
- Gerven, S. V. and Compernelle, D. V., 1995.  
Signal separation by symmetric adaptive decorrelation: Stability, convergence, and uniqueness.  
*IEEE Transactions on Signal Processing*, 43:1602–1612.  
7
- Gramss, T., 1995.  
A neural model for the separation of acoustic signals.  
In J. Bower, editor, *Computational Neuroscience: Trends in Research 1995*, pages 191–195. Monterey.  
35, 81
- Hall, J. W., Haggard, M. P., and Fernandes, M. A., 1984.  
Detection in noise by spectro-temporal pattern analysis.  
*Journal of the Acoustical Society of America*, 76:50–56.  
39
- Heckl, M. and Müller, H. A., editors, 1994.  
*Taschenbuch der technischen Akustik*.  
Springer, Berlin, 2nd edition.  
30, 53
- Hérault, J. and Jutten, C., 1986.  
Space or time adaptive signal processing by neural network models.  
In J. S. Denker, editor, *Neural networks for computing: AIP conference proceedings 151*. American Institute of Physics, New York.

7

Ikram, M. Z. and Morgan, D. R., 2000.

Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment.

In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

8, 85

Jung, T.-P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., and Sejnowski, T. J., 2000.

Removing electroencephalographic artifacts by blind source separation.

*Psychophysiology*, 37:163–178.

6

Jutten, C. and Héault, J., 1991.

Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture.

*Signal Processing*, 24:1–10.

7, 11, 33, 65

Jutten, C. and Taleb, A., 2000.

Source separation: From dusk till dawn.

In Pajunen and Karhunen (2000), pages 15–26.

7

Kablán, A. and Girolami, M., 2000.

Clustering of text documents by skewness maximization.

In Pajunen and Karhunen (2000), pages 435–440.

6

Kasper, K., Reininger, H., and Wolf, D., 1997.

Exploiting the potential of auditory preprocessing for robust speech recognition by LRNN.

In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1223–1227.

93

Kawamoto, M., Matsuoka, K., and Ohnishi, N., 1998.

A method of blind separation for convolved non-stationary.

*Neurocomputing*, 22:157–171.

7

Kleinschmidt, M., Marzinzik, M., and Kollmeier, B., 1998a.

Combining monaural noise reduction algorithms and perceptive preprocessing for robust speech recognition.

In T. Dau, V. Hohmann, and B. Kollmeier, editors, *Psychophysics, Physiology, and Models for Hearing*, pages 267–270. World Scientific, Singapore.

93

- Kleinschmidt, M., Tchorz, J., Wittkop, T., Hohmann, V., and Kollmeier, B., 1998b.  
Robuste Spracherkennung durch binaurale Richtungsfilterung und gehörgerechte Vorverarbeitung.  
In A. Sill, editor, *Fortschritte der Akustik: DAGA 98*, pages 396–397. Deutsche Gesellschaft für Akustik (DEGA), Zürich, Switzerland.  
93

- Kleinschmidt, M., Wittkop, T., and Kollmeier, B., 1999.  
Evaluation of monaural and binaural speech enhancement for robust auditory-based automatic speech recognition.  
In *Joint meeting of the Acoustical Society of America and the European Acoustics Association*. Berlin, Germany.  
94

- Kollmeier, B. and Koch, R., 1994.  
Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction.  
*Journal of the Acoustical Society of America*, 95:1593–1602.  
38

- Lambert, R. H., 1996.  
*Multichannel Blind Deconvolution: FIR Matrix Algebra and separation of Multipath Mixtures*.  
Ph.D. thesis, University of Southern California.  
34

- Lee, T.-W., 1998a.  
*Independent component analysis: Theory and applications*.  
Kluwer academic publishers, Boston.  
7, 18, 33, 58

- Lee, T.-W., 1998b.  
Sound recordings `rss_mA.wav` and `rss_mB.wav`.  
URL <http://tesla-e0.salk.edu/~tewon/Blind/Demos/>.  
58, 59, 60, 78, 80

- Lee, T.-W., Bell, A. J., and Lambert, R. H., 1997.  
Blind separation of delayed and convolved sources.  
In T. P. Michael Mozer, Michael Jordan, editor, *Advances in Neural Information Processing Systems*, volume 9, pages 758–764. MIT Press, Cambridge, MA.  
7, 34

- Lee, T.-W., Ziehe, A., Orglmeister, R., and 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*, volume 2, pages 1249–1252. Seattle, USA.  
8, 12, 37, 59
- Lindgren, U. A. and Broman, H., 1998.  
Source separation using a criterion based on second-order statistics.  
*IEEE Transactions on Signal Processing*, 46:1837–1850.  
7, 34
- MacKay, D. J. C., 1996.  
Maximum likelihood and covariant algorithms for independent component analysis.  
Technical report, Dept. of Physics, Cambridge University, England.  
URL <ftp://w01.ra.phy.cam.ac.uk/pub/mackay/ica.ps.gz>.  
17
- von der Malsburg, C. and Schneider, W., 1986.  
A neural cocktail-party processor.  
*Biological Cybernetics*, 54:29–40.  
5
- Matsuoka, K., Ohya, M., and Kawamoto, M., 1995.  
A neural net for blind separation of nonstationary signals.  
*Neural Networks*, 8:411–419.  
7, 33, 48, 49, 65
- McKeown, M. J., Makeig, S., Brown, G. G., Jung, T.-P., Kindermann, S. S., Bell, A. J., and Sejnowski, T. J., 1998.  
Analysis of fMRI data by blind separation into independent spatial components.  
*Human Brain Mapping*, 6:160–188.  
6
- Mejuto, C., Dapena, A., and Casteda, L., 2000.  
Frequency-domain infomax for blind separation of convolutive mixtures.  
In P. Pajunen and J. Karhunen, editors, *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Source Separation*, pages 315–320.  
37
- Michaelis, C., Gramss, T., and Strube, H. W., 1997.  
Glottal-to-noise excitation ratio — a new measure for describing pathological voices.  
*Acustica - acta acustica*, 83:700–706.  
39
- Molgedey, L. and Schuster, H. G., 1994.  
Separation of a mixture of independent signals using time delayed correlations.  
*Physical Review Letters*, 72:3634–3637.  
7, 33, 48, 65, 69, 72

- Murata, N., Ikeda, S., and Ziehe, A., 1998.  
An approach to blind source separation based on temporal structure of speech signals.  
Technical Report 98-2, BSIS, Riken Brain Science Institute, Tokyo, Japan.  
7, 8, 12, 20, 34, 37, 47, 66, 82, 87
- Nadal, J.-P. and Parga, N., 1997.  
Redundancy reduction and independent component analysis: Conditions on cumulants and adaptive approaches.  
*Neural Computation*, 9:1421–1456.  
92
- Nikias, C. L. and Petropulu, A. P., 1993.  
*Higher-order spectra analysis – A nonlinear signal processing framework*.  
Prentice Hall, Englewood Cliffs.  
63, 79
- Oppenheim, A. V. and Schaefer, R. W., 1975.  
*Digital signal processing*.  
Prentice-Hall, Englewood Cliffs.  
25, 34, 36, 79
- Pajunen, P. and Karhunen, J., editors, 2000.  
*Proceedings of the second international workshop on independent component analysis and blind signal separation*. Helsinki, Finland.  
101
- Papoulis, A., 1991.  
*Probability, random variables, and stochastic processes*.  
McGraw-Hill, New York, 3rd edition.  
18, 79
- Parra, L., 1998.  
Sound recordings `tvin1.wav` and `tvin2.wav`.  
URL [http://www.sarnoff.com/career\\_move/tech\\_papers/papers/](http://www.sarnoff.com/career_move/tech_papers/papers/).  
59
- Parra, L. and Spence, C., 2000a.  
Convolutional blind separation of non-stationary sources.  
*IEEE Transactions on Speech and Audio Processing*, 8:320–327.  
8, 12, 34, 37, 49, 58, 59, 65, 66
- Parra, L. and 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.  
12

- Parra, L., Spence, C., and Sajda, P., 2001.  
Statistical properties arising from the non-stationarity of natural signals.  
In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, Cambridge, MA.  
7, 23
- Parra, L., Spence, C., and Vries, B. D., 1998.  
Convolutional blind source separation based on multiple decorrelation.  
In *IEEE Neural Networks and Signal Processing Workshop*. Cambridge, UK.  
7, 8
- Paulus, E., 1998.  
*Sprachsignalverarbeitung: Analyse, Erkennung, Synthese*.  
Spektrum Akademischer Verlag, Heidelberg.  
39
- Pham, D. T., Garat, P., and Jutten, C., 1992.  
Separation of a mixture of independent sources through a maximum likelihood approach.  
In J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, editors, *Signal Processing VI: Theories and Applications*, pages 771–774.  
16, 17
- Platt, J. C. and Faggin, F., 1992.  
Networks for the separation of sources that are superimposed and delayed.  
In J. Moody, S. Hansen, and R. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 730–737. MIT Press, Cambridge, MA.  
12
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., 1992.  
*Numerical Recipes in C*.  
Cambridge University Press, Cambridge, second edition.  
47
- Priestley, M. B., 1981.  
*Spectral Analysis and Time Series*.  
Academic Press, London.  
66
- Sahlin, H. and Broman, H., 1998.  
Separation of real-world signals.  
*Signal Processing*, 64:103–104.  
12
- Servière, C., 1999.  
Blind source separation in presence of spatially correlated noises.

- In J. F. Cardoso, C. Jutten, and P. Loubaton, editors, *Proceedings of the first international workshop on independent component analysis and blind signal separation*, pages 497–502. Aussois, France.  
37
- Shamsunder, S. and Giannakis, G. B., 1997.  
Multichannel blind signal separation and reconstruction.  
*IEEE Transactions on Speech and Audio Processing*, 5:515–528.  
35, 81
- Smaragdis, P., 1998.  
Blind separation of convolved mixtures in the frequency domain.  
*Neurocomputing*, 22:21–34.  
19, 20
- Sompolinsky, H., Barkai, N., and Seung, H. S., 1995.  
On-line learning of dichotomies: Algorithms and learning curves.  
In J.-H. Oh, editor, *Neural networks: The statistical mechanics perspective*, pages 105–130.  
23
- Steeneken, H. J. M. and Houtgast, T., 1999.  
Mutual dependence of the octave-band weights in predicting speech intelligibility.  
*Speech Communication*, 28:109–123.  
39
- Strube, H. W., 1981.  
Separation of several speakers recorded by two microphones (cocktail-party processing).  
*Signal Processing*, 3:355–364.  
5
- Tchorz, J. and Kollmeier, B., 1999.  
A model of auditory perception as front end for automatic speech recognition.  
*Journal of the Acoustical Society of America*, 106:2040–2050.  
91, 93
- Tchorz, J. and Kollmeier, B., 2000.  
Noise suppression based on amplitude modulation analysis.  
*submitted to IEEE Transactions on Speech and Audio Processing*.  
39
- Tong, L., Liu, R.-w., Soon, V. C., and Huang, Y.-F., 1991.  
Indeterminacy and identifiability of blind identification.  
*IEEE Transactions on Circuits and Systems*, 38:499–509.  
14, 33, 36, 70

Torkkola, K., 1996a.

Blind separation of convolved sources based on information maximization.  
In *IEEE Workshop on Neural Networks for Signal Processing*. Kyoto, Japan.  
7

Torkkola, K., 1996b.

Blind separation of delayed sources based on information maximization.  
In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3509–3512. Atlanta, GA.  
12, 29

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.  
19

Verhey, J. L., Dau, T., and Kollmeier, B., 1999.

Within-channel cues in comodulation masking release (CMR): Experiments and model predictions using a modulation-filterbank model.  
*Journal of the Acoustical Society of America*, 106:2733–2745.  
39

Wachtler, T., Lee, T.-W., and Sejnowski, T. J., 2001.

The chromatic structure of natural scenes.  
*Journal of the Optical Society of America A-Optics, Image Science and Vision*, 18:65–77.  
66

Weinstein, E., Feder, M., and Oppenheim, A. V., 1993.

Multi-channel signal separation by decorrelation.  
*IEEE Transactions on Speech and Audio Processing*, 1:405–413.  
7, 34, 65

Wittkop, T., 2001.

*Two-channel noise reduction algorithms motivated by models of binaural interaction*.  
Ph.D. thesis, Fachbereich Physik, Universität Oldenburg.  
5, 30

Wittkop, T., Albani, S., Hohmann, V., Peissig, J., Woods, W., and Kollmeier, B., 1997.

Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction.  
*Acustica - acta acustica*, 83:684–699.  
5

- Yang, H. H. and Amari, S.-i., 1997.  
Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information.  
*Neural Computation*, 9:1457–1482.  
18
- Yellin, D. and Weinstein, E., 1994.  
Criteria for multichannel signal separation.  
*IEEE Transactions on Signal Processing*, 42:2158–2168.  
7
- Yellin, D. and Weinstein, E., 1996.  
Multichannel signal separation: Methods and analysis.  
*IEEE Transactions on Signal Processing*, 44:106–118.  
7, 8, 34
- Zelinski, R. and Noll, P., 1977.  
Adaptive transform coding of speech signals.  
*IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-25:299–309.  
16
- Ziehe, A., Müller, K.-R., Nolte, G., Mackert, B. M., and Curio, G., 2000.  
Artifact reduction in magnetoneurography based on time-delayed second-order correlations.  
*IEEE Transactions on Biomedical Engineering*, 47:75–87.  
6

# Danksagung

Mein Dank gilt Prof. Birger Kollmeier, in dessen Arbeitsgruppe ich die vorliegende Arbeit anfertigen durfte, und der sie mit kontinuierlicher Unterstützung und fortwährendem Interesse begleitete. Ich profitierte von dem großen Freiraum bei der Wahl des Forschungsthemas und dem großen Vertrauen, das er mir entgegenbrachte. Die Arbeit hat wesentlich durch seinen Einsatz für die hervorragenden Arbeitsbedingungen und vielfältige Anregungen gewonnen.

Prof. Volker Mellert danke ich für die freundliche Annahme des Korreferats.

Dankbar bin ich auch Tino Gramß, ohne den diese Arbeit nicht, und mit dem sie sicher anders geworden wäre. Ihm verdanke ich das Thema der blinden Quellentrennung, sowie viele Diskussionen, Anregungen und Weichenstellungen zu Beginn der Arbeit.

Den Mitgliedern der AG Medi möchte ich für die angenehme Atmosphäre danken, und für die vielen Einblicke, die ich in die verschiedenen Arbeitsgebiete, von Signalverarbeitung über Psychoakustik bis zu EEG-Signalen, erhalten konnte.

Besonders möchte ich Michael Kleinschmidt für die Zusammenarbeit bei der Spracherkennung und Störgeräuschunterdrückung danken.

Mein Dank gilt auch den mit-GNUs, die mit Lust und Frust unseren kleinen heliozentrischen Zoo gebaut und gepflegt haben, so daß bei Bedarf immer ausreichend flops und gigs verfügbar waren.

Weiterhin bin ich dankbar für das Privileg, daß ich die Arbeit mit zahlreichen anderen Forschern diskutieren konnte. Viele Fragen, Anregungen und Diskussionen haben wesentlich zum Gelingen der Arbeit beigetragen. Hierfür geht ein großer Dank an viele Wissenschaftler, von Berlin, Göttingen und Zürich bis Boston, Princeton und San Diego.

Jutta danke ich für ihre vielfache Unterstützung.





# Lebenslauf

Ich wurde am 21. Mai 1971 in Lippstadt (Westfalen) geboren. Nach dem Besuch der Grundschule Bad Sassendorf von 1977 bis 1981 wechselte ich auf das Aldegrevener Gymnasium in Soest, wo ich am 18. Mai 1990 das Abitur ablegte.

Im Anschluß an den Zivildienst in Lübeck und Soest schrieb ich mich zum Wintersemester 1992/93 an der Universität Oldenburg für das Physikstudium ein und legte dort am 8. Februar 1995 die Vordiplomsprüfung ab.

Nach den sechs Semestern des Grundkanons Physik ging ich nach England zum King's College, University of London, wo ich am Centre for Neural Networks, Dept. of Mathematics, im Studiengang *M.Sc. in Information Processing and Neural Networks* studierte. Meine Abschlußarbeit zum Thema *Coupled Synaptic and Neuronal Dynamics for Oscillator Networks* schrieb ich bei Prof. A. C. C. Coolen und schloß das Studium im September als *M.Sc. in Information Processing and Neural Networks* ab.

Seit 1997 forsche ich in der Arbeitsgruppe Medizinische Physik von Prof. B. Kollmeier an der Universität Oldenburg zum Thema Blinde Quellentrennung. Vom 1. März 1997 bis zum 29. Februar 2000 war ich Stipendiat im Graduiertenkolleg Psychoakustik und bin seitdem als wissenschaftlicher Angestellter in der Arbeitsgruppe Medizinische Physik tätig.