Source Separation for Hearing Aid Applications

Michael Syskind Pedersen

Kongens Lyngby 2006
IMM-PHD-2006-167

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

The main focuses in this thesis are on blind separation of acoustic signals and on a speech enhancement by time-frequency masking.

As a part of the thesis, an exhaustive review on existing techniques for blind separation of convolutive acoustic mixtures is provided.

A new algorithm is proposed for separation of acoustic signals, where the number of sources in the mixtures exceeds the number of sensors. In order to segregate the sources from the mixtures, this method iteratively combines two techniques: Blind source separation by independent component analysis (ICA) and time-frequency masking. The proposed algorithm has been applied for separation of speech signals as well as stereo music signals. The proposed method uses recordings from two closely-spaced microphones, similar to the microphones used in hearing aids.

Besides that, a source separation method known as *gradient flow beamforming* has been extended in order to cope with convolutive audio mixtures. This method also requires recordings from closely-spaced microphones.

Also a theoretical result concerning the convergence in gradient descent independent component analysis algorithms is provided in the thesis.

Resumé

I denne afhandling fokuseres hovedsagligt på blind kildeseparation af lydsignaler samt taleforbedring ved brug af tids-frekvensmaskering.

En grundig gennemgang af eksisterende teknikker til blind adskillelse af filtrerede akustiske signaler er præsenteret som en del af afhandlingen.

En ny algoritme til adskillelse af lydsignaler er foreslået, hvor antallet af kilder er større end antallet af mikrofoner. Til separation af kilder anvendes to teknikker: Blind kildeseparation ved hjælp af *independent component analysis* (ICA) og tids-frekvensmaskering. Metoden har været anvendt til adskillelse af talesignaler og stereo musiksignaler. Den foreslåede metode anvender optagelser fra to tætsiddende mikrofoner, magen til dem der anvendes i høreapparater.

Ud over dette, er en kildeseparationsmetode kendt som *gradient flow beamforming* udvidet, så metoden kan separere filtrerede lydsignaler. Denne metode kræver ligeledes tætsiddende mikrofoner.

Et teoretisk resultat, der omhandler konvergens af gradientnedstigning i ICA algoritmer, er ligeledes givet i denne afhandling.

Preface

This thesis was prepared at the Intelligent Signal Processing group at the Informatics Mathematical Modelling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering.

The thesis deals with techniques for blind separation of acoustic sources. The main focus is on separation of sources recorded at microphone arrays small enough to fit in a single hearing aid.

The thesis consists of a summary report and a collection of seven research papers written during the period June 2003 – May 2006, and published elsewhere. The contributions in this thesis are primarily in the research papers, while the main text for the most part can be regarded as background for the research papers.

This project was funded by the Oticon foundation.

Smørum, May 2006

Michael Syskind Pedersen

Papers Included in the Thesis

[A] Michael Syskind Pedersen and Chlinton Møller Nielsen. Gradient flow convolutive blind source separation. *Proceedings of the 2004 IEEE Signal Processing Society Workshop (MLSP)*, pp. 335–344, São Luís, Brazil, September 2004.

[B] Michael Syskind Pedersen, Jan Larsen, and Ulrik Kjems. On the Difference Between Updating The Mixing Matrix and Updating the Separation Matrix. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. vol. V pp. 297–300, Philadelphia, PA, USA. March 2005.

[C] Michael Syskind Pedersen, DeLiang Wang, Jan Larsen, and Ulrik Kjems. Overcomplete Blind Source Separation by Combining ICA and Binary Time-Frequency Masking. *Proceedings of IEEE Signal Processing Society Workshop (MLSP)*. pp. 15–20, Mystic, CT, USA. September 2005.

[D] Michael Syskind Pedersen, Tue Lehn-Schiøler, and Jan Larsen. BLUES from Music: BLind Underdetermined Extraction of Sources from Music. *Proceedings of Independent Component Analysis, and Blind Signal Separation Workshop (ICA)*. pp. 392–399, Charleston, SC, USA. March 2006.

[E] Michael Syskind Pedersen, DeLiang Wang, Jan Larsen, and Ulrik Kjems. Separating Underdetermined Convolutive Speech Mixtures. *Proceedings of Independent Component Analysis, and Blind Signal Separation Workshop (ICA)*. pp. 674–681, Charleston, SC, USA. March 2006.

[F] Michael Syskind Pedersen, DeLiang Wang, Jan Larsen, and Ulrik Kjems. Two-Microphone Separation of Speech Mixtures. *IEEE Transactions on Neural Networks*. April 2006. Submitted.

[G] Michael Syskind Pedersen, Jan Larsen, Ulrik Kjems, and Lucas Parra. A Survey of Convolutive Blind Source Separation Methods. To appear as *Chapter in Jacob Benesty, Yiteng (Arden) Huang, and M. Mohan Sondhi, editors, Springer Handbook on Speech Processing and Speech Communication.* 2006. Preliminary version.

Other Publications

The appendices contain the papers above, which have been written during the past three years. Three other publications written during the past three years are not included as a part of this thesis:

[[70]] Michael Syskind Pedersen, Lars Kai Hansen, Ulrik Kjems, and Karsten Bo Rasmussen. Semi-Blind Source Separation Using Head.Related Transfer Functions. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* vol. V pp. 713–716, Montreal, Canada. May 2004.

[[69]] Michael Syskind Pedersen. Matricks. *Technical Report.* IMM, DTU. 2005.

[[74]] Kaare Brandt Petersen and Michael Syskind Pedersen The Matrix Cookbook. Online Manual. 2006.

The work in [70] was mainly done during my Master's Thesis.

The work in [74] is an on-line collection of useful equations in matrix algebra called *The Matrix Cookbook*. This is joint work with Kaare Brandt Petersen, and we frequently update this paper with new equations and formulas. The most recent version of this manual can be found at `http://2302.dk/uni/matrixcookbook.html`.

The work in [69] also contains useful matrix algebra. This work was merged into *The Matrix Cookbook*.

Acknowledgements

I would like to thank my two supervisors Jan Larsen and Ulrik Kjems for excellent supervision. I would also like to thank the Oticon foundation for funding this project and Professor Lars Kai Hansen for suggesting me to do a Ph.D. I would also like to thank my colleagues at Oticon as well as my colleagues at the Intelligent Signal Processing (ISP) group at IMM, DTU for interesting conversations and discussions. It has been a pleasure to work with all these nice people.

A special thank goes to Professor DeLiang Wang whom I was visiting at The Ohio State University (OSU) during the first six months of 2005. I would also like to thank the people at Perception and Neurodynamics Laboratory at OSU for making my visit very pleasant.

Thanks to Malene Schlaikjer for reading my manuscript and for useful comments. I would also like to acknowledge all the other people who have assisted me through the project.

Contents

Chapter 1

Introduction

Many activities in human daily live involve processing of audio information. Much information about the surroundings is obtained through the perceived acoustic signal. Also much interaction between people occurs through audio communication, and the ability to listen and process sound is essential in order to take part of conversations with other people.

As humans become older, the ability to hear sounds degrades. Not only do weak sounds disappear, the time and frequency selectivity degrade too. Hereby, hearing impaired loose their ability to track sounds in noisy environments and thus the ability follow conversations.

One of the most challenging environments for human listeners to cope with is when multiple speakers are talking simultaneously. This problem is often referred to as the *cocktail-party problem* [29, 44], because in such a scenery, different conversations occur simultaneously and independent of each other. Humans with normal hearing actually perform remarkably well in such situations. Even in very noisy environments, they are able to track the sound of a single speaker among multiple speakers.

In order to cope with hearing impairment, hearing aids can assist people. One of the objectives of hearing aids is to improve the speech intelligibility and thereby help people to follow conversations better. One of the methods to improve the

intelligibility in difficult environments is to enhance the desired audio signal (often speech) and to suppress the background noise.

Today, different methods exist in order to enhance speech, and hereby increase the intelligibility in noisy environments [13]. Speech enhancement techniques can either be based on a single microphone recording or multi-microphone recordings. In speech enhancement methods, a desired speech signal is present in noise. The desired signal can be enhanced by either amplifying the speech signal or by suppressing the noise [13, 38, 24, 41].

In the following sections a more detailed discussion of the challenges in hearing and hearing aids will be given as well as a brief introduction to multi-microphone speech enhancement techniques which are considered in this thesis. This is presented in order to create the basis for the subsequent chapters.

1.1 Hearing and Hearing Aids

In order to understand hearing loss, it is important to have some basic knowledge about the human ear. In this section, the anatomy of the ear is introduced. Important concepts related to hearing is introduced and causes for hearing loss are reviewed. A simple introduction to the hearing aid is provided as well.

1.1.1 The Human Ear

The human ear can be divided into three parts: The outer ear, the middle ear, and the inner ear. An illustration of the ear is given in Figure 1.1. The outer ear is the visible part of the ear. It consists of the pinna and the auditory canal (meatus). Between the outer ear and middle ear is the eardrum (tympanic membrane) located. The eardrum is very sensitive to changes in air pressure. Sound waves cause the eardrum to vibrate. The middle ear is on the other side of the eardrum. The middle ear consists of a cavity (the tympanic cavity), and the three bones, the hammer, the anvil and the stirrup. The three bones transfer the sound waves from the eardrum to movements in the fluid inside the cochlea in the inner ear. In the cochlea, the sound waves are transformed into electrical impulses. The basilar membrane is located inside the cochlea. Inside the basilar membrane, hair cells are found. The hair cells can be divided into two groups: inner and outer hair cells. The inner hail cells mainly signal the movements of the cochlea to the brain. The outer hair cells mainly amplify the traveling wave in the cochlea. Depending on the frequency of the sound wave,

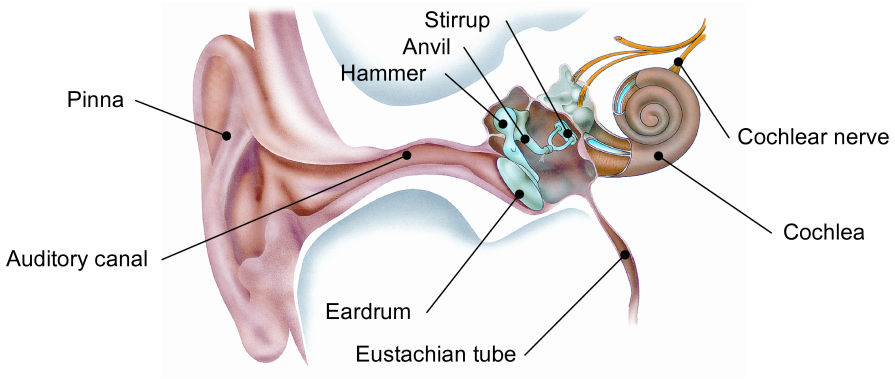

Figure 1.1: The ear can be divided into three parts, the outer ear, the middle ear, and the inner ear. Sound waves cause the eardrum to vibrate. In the middle ear, the hammer, the anvil, and the stirrup transfer the vibrations from the air into movements of the fluid inside the cochlea in the inner ear. In the cochlea, the movements are transferred into neural activity.

certain places in the basilar membrane are excited. This causes neural activity of certain hair cells. All together, there are about 12000 outer hair cells and 3500 hair cells [62].

1.1.2 Sound Level and Frequency Range

Sound waves occur due to changes in air pressure. The ear is very sensitive to changes in air pressure. Often the sound level is described in terms of intensity, which is the energy transmitted per second. The sound intensity is measured in terms of a reference intensity, I_0. The sound intensity ratio given in decibels (dB) is given as [62]

$$\text{number of dB} = 10 \log_{10}(I/I_0). \tag{1.1}$$

The reference intensity, with a sound pressure level (SPL) of 0 dB corresponds to a sound pressure of 20 μPa or 10^{-12} W/m^2. Humans can detect sound intensity ratios from about 0 dB SPL (with two ears and a sound stimuli of 1000 Hz) up to about 140 dB SPL. This corresponds to amplitudes with ratios that can vary by a factor of 10^7.

The minimum thresholds where sounds can be detected depend on the frequency and whether the sound is detected by use of one or two ears. This is illustrated

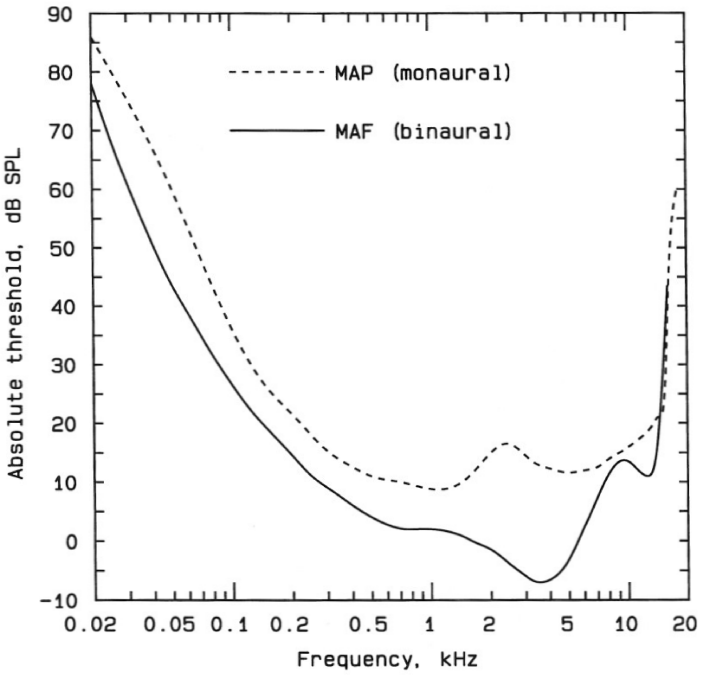

Figure 1.2: The minimum detectable sound as a function of the frequency. The figure shows both the minimum audible pressure (MAP) for monaural listening and the minimum audible field (MAF) for binaural listening. The MAP is the sound pressure measured by a small probe inside the ear canal. The MAF is the pressure measured at a point which was occupied by the listeners head. The figure is obtained from Moore (2003) [62, p. 56].

in Figure 1.2. As it can be seen, the frequency range for when sounds are audible goes from about 20 Hz up to about 20 kHz. It is important to notice that the minimum audible level also strongly varies with the frequency.

1.1.3 Hearing Impairment

Hearing loss can be divided into two types: Sensorineural loss and conductive loss. The sensorineural hearing loss is the most common type of hearing loss. The sensorineural loss is often caused by a defect in the cochlea (cochlea loss), but a sensorineural loss can also be caused by defects in higher levels in the auditory system such as the auditory nerve [62]. Defects in the cochlea is often

due to the loss of hair cells. The loss of hair cells reduces the neural activity. Hereby a hearing impaired experiences:

Reduced ability to hear sounds at low levels The absolute threshold, where sounds can be detected, is increased.

Reduced frequency selectivity The discrimination between sounds at different frequencies is decreased.

Reduced temporal processing The discrimination between successive sounds is decreased.

Reduced binaural processing The ability to combine information from the sounds received at the two ears is reduced.

Loudness recruitment Loudness recruitment means that the perceived loudness grows more rapidly than for a normal listener. This is illustrated in Figure 1.3.

All these different factors result in a reduced speech intelligibility for the person with a cochlear hearing loss, especially in noisy environments.

In a conductive hearing loss, the cochlea is typically not damaged. Here, the conduction in between the incoming sound and the cochlea is diminished. This decreased conduction can be caused by many factors:

Earwax If the auditory canal is closed by earwax, the sound is attenuated.

Disruptions in the middle ear If some of the three bones in the middle are disconnected, it may result in a conductive loss.

Otosclerosis Tissue growth on the stirrup may result in a conductive loss.

Otitis media Fluid in the middle ear causes a conductive loss.

1.1.4 Hearing Aids

An example of a (simplified) hearing aid is shown in Figure 1.4. The hearing loss is compensated by a frequency-dependent gain. Due to the loudness recruitment, the hearing aid has to amplify the sounds with a small amplitude more than the sounds with a higher amplitude. This reduction of the dynamic range is called compression. Depending on the type of hearing loss, many types of gain

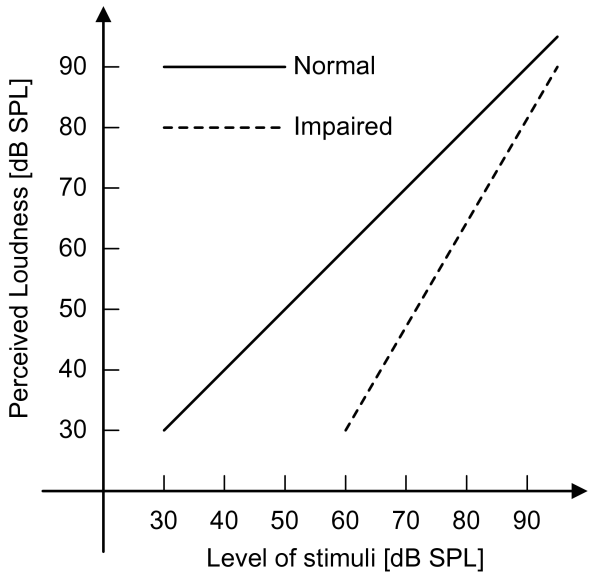

Figure 1.3: Loudness recruitment. For a normal listener, the perceived loudness level approximately corresponds to the stimuli level. For a hearing impaired with a cochlear hearing loss, the perceived loudness grows much more rapidly. The dynamic range of a hearing impaired is thus reduced.

strategies that compensate for the hearing loss exist. These different types are called *rationales*.

Before the compensation of the hearing loss, some audio pre-processing may be applied to the recorded acoustic signals. The purpose of this pre-processing step is to enhance the desired signal as much as possible before the compression algorithm compensates for the hearing loss. The audio pre-processing can be multi-microphone enhancement, that amplifies signals from certain directions. These techniques are known as beamforming. The pre-processing can also be based on a single microphone, here the enhancement/noise reduction is not based on the arrival direction of the sounds, but the enhancement relies more on the properties of the desired signal and the property of the unwanted noise.

In hearing aids, the signals have to be processed with as little delay as possible. If the audio signal is delayed too much compared to what the listener is seeing, the listener may not be able to fully combine the sound with vision, and the listener may loose the additional benefit from lip-reading. If the delay is e.g. more than 250 ms, most people find it difficult to carry on normal conversations

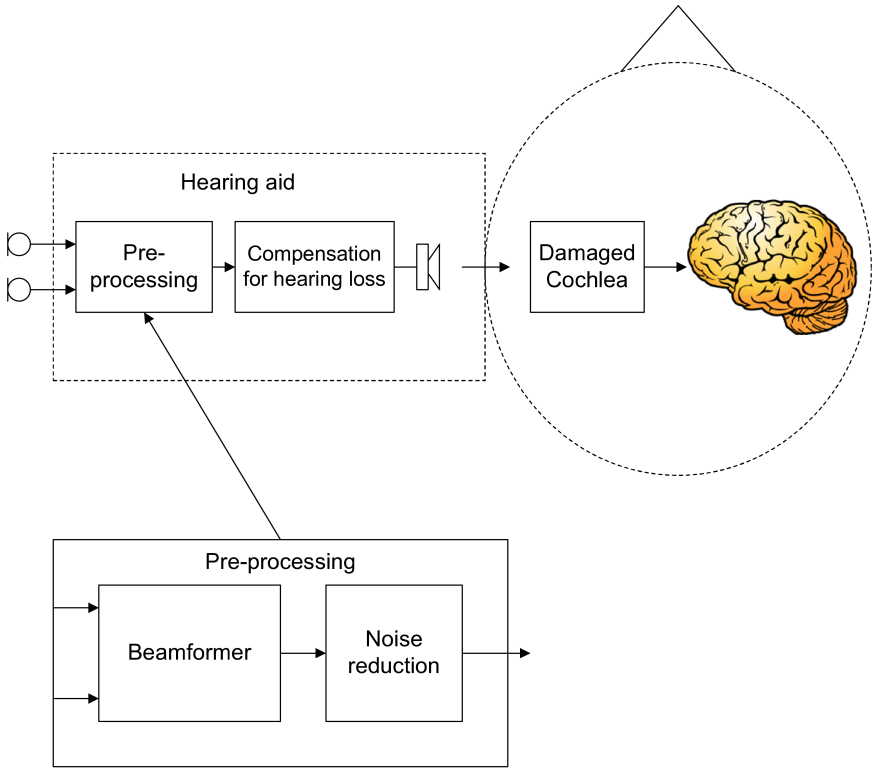

Figure 1.4: In a hearing aid, the damaged cochlea is compensated by a frequency-dependent gain and a compression algorithm. In order to enhance the desired audio signal, a pre-processing step is applied in the hearing aid. This enhancement may consist of a beamformer block that enhances a signal from a certain direction and a noise reduction block that reduces the noise based on the signal properties. The beamformer uses multiple microphone recordings, while the noise reduction is applied to a single audio signal.

[39]. Another problem is that often both the direct sound and the processed and hereby delayed sound reaches the eardrum. This is illustrated in Figure 1.5. Depending on the type of sound and the delay, the direct and the delayed sound may be perceived as a single sound or as two separate sounds. The perception of echoes and direct sound as a single sound is called the precedence effect. For example, a click is perceived as two separate clicks if the delay is more than as little as 5 milliseconds, while echoes from more complex sounds like speech are suppressed up to as much as 40 milliseconds [62, p. 253]. Even though the direct sound and the processed sound are perceived as a single sound, the

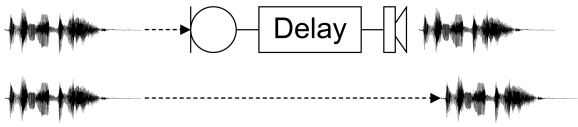

Figure 1.5: The sound obtained by the eardrum is often a combination of the direct sound and the sound, which has been processed through the hearing aid. The processed sound is delayed compared to the direct sound, and the resulting signal can therefore be regarded as a delay-and-sum filtered signal.

resulting signal is a delay and sum filtered signal (see Chapter 5). This comb filtering effect is undesired and one of the main reasons why the delay through the hearing aid should be kept as little as possible. For example: If a delay through a hearing aid is limited to e.g. 8 ms, and the sampling frequency is 16 kHz, the allowed delay corresponds to 128 samples.

1.2 Multi-microphone Speech Enhancement

When multiple microphones are available, spatial information can be utilized in order to enhance sources from a particular direction. Signals can be enhanced based on the geometry of the microphone array, or based on the statistics of the recorded signals alone. Many different solutions have been proposed to this problem and a brief review of some of the methods are given in the following. More detailed information on beamforming can be found in Chapter 5, and a much more detailed information on blind separation of sources can be found in Appendix G.

1.2.1 Beamforming

When spatial information is available, it is possible to create a direction dependent pattern, which enhances signals arriving from a desired direction while attenuating signals (noise) arriving from other directions. Such techniques are called *beamforming* [92, 20]. A beamformer can either be fixed, where the directional gain does not change or it can be adaptive, where the null gain direction adaptively is steered towards the noise source [35].

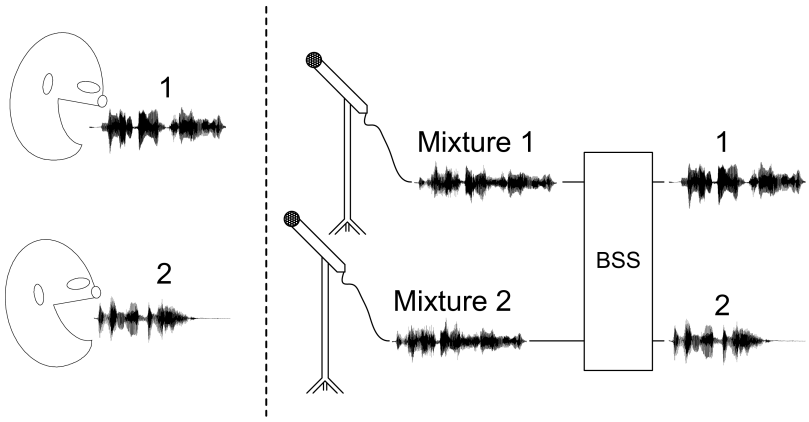

Figure 1.6: Illustration of the BSS problem. Mixtures of different audio signals are recorded by a number of microphones. From the mixtures, estimates of the source signals contained in the mixtures are found. Everything on the left side of the broken line cannot be seen from the blind separation box, hence the term *blind*.

1.2.2 Blind Source Separation and Independent Component Analysis

Often, the only available data are the mixtures of the different sources recorded at the available sensors. Not even the position of the different sensors is known. Still, it is sometimes possible to separate the mixtures and obtain estimates of the sources. The different techniques to obtain estimates of the different sources from the mixtures are termed *blind source separation* (BSS). The term *blind* refers to that only the mixtures are available. The BSS problem is illustrated in Figure 1.6. Here two people are talking simultaneously. Mixtures of the two voices are recorded by two microphones. From the recorded microphones, the separation filters are estimated. In order to separate sources, a model of the mixing system is required. Not only the direct path of the sources are recorded. Reflections from the surroundings as well as diffraction when a sound wave passes an object result in a filtering of the audio signals. Furthermore, different unknown characteristics from the microphones also contribute to the unknown filtering of the audio sources. Therefore the recorded audio signals are assumed to be convolutive mixtures. Given M microphones, the mth microphone signal

$x_m(t)$ is given by

$$x_m(t) = \sum_{n=1}^{N} \sum_{k=0}^{K-1} a_{mnk} s_n(t-k) + v_m(t) \tag{1.2}$$

Here each of the N source signals $s_n(t)$ is convolved with causal FIR filters of length K. a are the filter coefficients and $v(t)$ is the additional noise. In matrix form, the convolutive FIR mixture can be written as:

$$\mathbf{x}(t) \quad = \quad \sum_{k=0}^{K-1} \mathbf{A}_k \mathbf{s}(t-k) + \mathbf{v}(t) \tag{1.3}$$

Here, $\mathbf{A}_k$ is an $M \times N$ matrix which contains the kth filter coefficients. $\mathbf{v}(t)$ is the $M \times 1$ noise vector.

The objective in blind source separation is to estimate the original sources. An estimate of the sources can be found by finding separation filters, w_n, where the n'th filter ideally cancels all but the n'th source. The separation system can be written as

$$y_n(t) = \sum_{m=1}^{M} \sum_{l=0}^{L-1} w_{nml} x_m(t-l) \tag{1.4}$$

or in matrix form

$$\mathbf{y}(t) = \sum_{l=0}^{L-1} \mathbf{W}_l \mathbf{x}(t-l), \tag{1.5}$$

where $\mathbf{y}$(t) is the estimated sources.

A commonly used method to estimate the unknown parameters in the mixing/separation system is *independent component analysis* (ICA) [30, 50]. ICA relies on the assumption that the different sources are statistically independent from each other. If the sources are independent, methods based on higher order statistics (HOS) can be applied in order to separate the sources [26]. Alternatively, ICA methods based on the maximum likelihood (ML) principle have been applied [25]. Non-Gaussianity has as well been applied for source separation. Based on central limit theorem, each source in the mixture is further away from being Gaussian compared to the mixture.

Based on further assumptions on the sources, second order statistics (SOS) has shown to be sufficient for source separation. If the sources are uncorrelated and non-stationary, SOS alone can be utilized to segregate the sources [67]. Notice, when only SOS is used for source separation, the sources are not required to be independent, because no assumptions are made on statistics of an order higher than two.

A problem in many source separation algorithms is that the number of sources in the mixture is unknown. Furthermore, many source separation algorithms cannot separate more sources than the number of available microphones.

Not only the question concerning how many signals the mixture contains arises. In real-world systems, such as hearing aids, quite often only a single source in the pool of many sources is of interest. Which of the segregated signals is the target signal therefore have to be determined too. In order to determine the target signal among the segregated sources, additional information is required. Such information could e.g. be that the source of interest impinges the microphone array from a certain direction.

1.3 The Scope of This Thesis

The thesis has two main objectives:

1. **Source separation techniques** The first objective is to provide knowledge on existing methods within techniques for multi-microphone speech separation. These techniques include: blind source separation, beamforming, and computational auditory scene analysis (CASA).

2. **BSS for hearing aids** The second objective is to propose algorithms for separation of signals, especially signals recorded by a single hearing aid. Here, we limit ourself to the audio pre-processing step for hearing aids which was shown in Figure 1.4. We consider speech enhancement systems, where recordings from a microphone array are available. The size of a hearing aid limits the size of a microphone array in a hearing aid. The typical array dimension in a hearing aid is not greater than approximately 1.5 cm. Here, we mainly consider microphone arrays of such a size. We consider different techniques for separation/segregation of audio signals. The techniques are based on blind source separation by ICA and time-frequency masking.

As mentioned, the allowed latency and the processing power of a hearing aid are limited. The objective of this thesis is however not to build a functional hearing aid, but to reveal methods for separation of audio sources. Most of these methods have been developed as batch methods that require filters with filter lengths up to several thousand taps, which are much more than what can be allowed in a hearing aid.

We limit ourself to consider audio pre-processing algorithms that can be applied to listeners with normal hearing. Therefore, as a working assumption we assume that the compression (rationale) can compensate for the hearing impairment so that the pre-processing step can be evaluated by people without hearing impairment.

The main contributions of the thesis have been published elsewhere. This work is presented in the appendices. The main text of the thesis should be regarded as background for the papers in the appendices. The papers in the appendices can be organized into different groups:

Gradient flow beamforming In Appendix A the gradient flow beamforming model proposed by Cauwenberghs et al. [27] for instantaneous ICA is extended to convolutive mixtures. The actual source separation is performed in the frequency domain.

Difference between ICA parameterizations In Appendix B differences between parameterizations of maximum likelihood source separation based on the mixing matrix and the separation matrix are analyzed.

Combination of ICA and T-F masking In Appendix C–F it is demonstrated how two-by-two ICA and binary T-F masking can be applied iteratively in order to segregate underdetermined audio sources, having only two microphone recordings available.

Survey on convolutive BSS In Appendix G a survey on convolutive BSS methods is provided.

The background material in main text mostly serves as background for the publications in the Appendices A and Appendix C–F. Especially background material on the two source separation techniques known as *time-frequency masking* and *beamforming* is provided. Blind source separation is not considered in the main text, because the thorough survey on BSS of audio signal is given in Appendix G.

The main text of the thesis is organized as follows: In Chapter 2, different auditory models are described. This chapter provides background about how humans perceive sound. We present different time-frequency representations of acoustic signals. Basic knowledge about how sound is perceived like e.g. when a stronger sound masks a weaker sound is important in order to understand why the T-F masking technique that has been applied in some of the publications (Appendix C–F) works so surprisingly well. An accurate model of the auditory system is also a good foundation for a related topic: *auditory scene analysis*.

The following chapter (Chapter 3) provides a short description of cues in auditory scene analysis and how these cues can be mimicked by machines in computational auditory scene analysis (CASA) in order to segregate sounds. T-F masking and auditory scene analysis is closely connected. In both areas, the objective is to group units in time and in frequency in a way that only units belonging to the same source are grouped together.

Based on the establishment of auditory models and auditory scene analysis, Chapter 4 deals with the central subject on time-frequency masking.

Beamforming and small microphone array configurations are also central topics in this thesis and in hearing aid development. Limitations in linear source separation can be seen from the limitations in beamforming. A base knowledge about beamforming and on the limitations in microphone array processing is provided in Chapter 5 and it is a good starting point when reading the publications in Appendix A and Appendix C–F. In this chapter, we also consider simple beamforming-based source separation techniques.

In Chapter 6, we briefly summarize and discuss the results on source separation from the contributions in the appendices.

The conclusion goes in Chapter 7 along with a discussion of future work.

Chapter 2

Auditory Models

The objective of this chapter is to give the reader some basic knowledge about how the human perceives sound in the time-frequency domain. Some frequently used frequency scales that mimics the human frequency resolution are introduced; the Bark scale and the ERB scale. A frequently used auditory band-pass filterbank, *the Gammatone filterbank*, is also introduced in this chapter. A good model of the auditory system is important in order to understand why the T-F masking technique works so well in attenuating the noise while maintaining the target sound. Auditory models can also help understanding why some artifacts become audible while other modification to a signal is inaudible.

Depending on the frequency of the incoming sound, different areas of the basilar membrane are excited. Therefore we can say that the ear actually does an analysis of the sound signal, not only in time, but also in frequency. A time-frequency analysis can be described by a bank of band-pass filters as shown in Figure 2.1.

The different filters in the auditory filterbank can have different bandwidths and different delays. More information about an audio signal can be revealed, if the audio signal is presented simultaneous in time and in frequency, i.e. in the time-frequency (T-F) domain.

An example of a T-F distribution is the spectrogram, which is obtained by

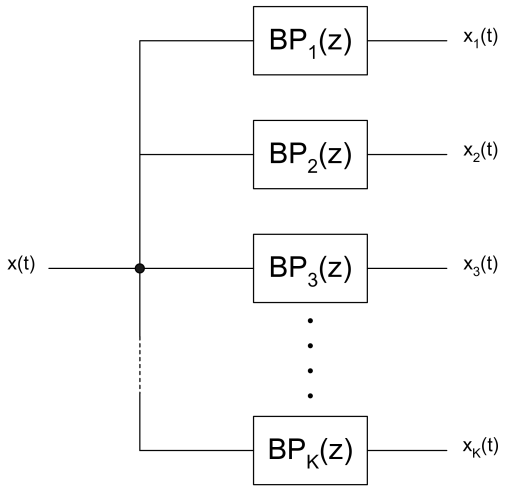

Figure 2.1: By a filterbank consisting of K band-pass filters, the signal $x(t)$ is transformed into the frequency domain. At time t the frequency domain signals $x_1(t), \ldots, x_K(t)$ are obtained.

the windowed short time Fourier transform (STFT), see e.g. [91]. Here, the frequency bands are equally distributed and the frequency resolution is the same for all frequencies.

The frequency resolution in the ear is however not linear. For the low frequencies, the frequency resolution is much higher than for the higher frequencies. In terms of perception, the width of the band-pass filters can be determined as a function of the center frequency of the band-pass filters.

When several sounds are present simultaneously, it is often experienced that a loud sound makes other weaker sounds inaudible. This effect is called *masking*. Whether one sound masks another sound depends on the level of the sounds, and how far the sounds are from each other in terms of frequency. In order to determine these masking thresholds, the *critical bandwidths* are introduced. The critical bandwidths are determined in terms of when the perception changes given a certain stimuli, e.g. whether a tone is masked by noise. Due to different ways of measuring the bandwidths, different sets of critical bandwidths have been proposed [43, 62]. Two well known critical bandwidth scales are the Bark critical bandwidth scale and the equivalent rectangular bandwidth (ERB) scale. Given the center frequency f_c (in Hz) of the band, the bandwidths can be

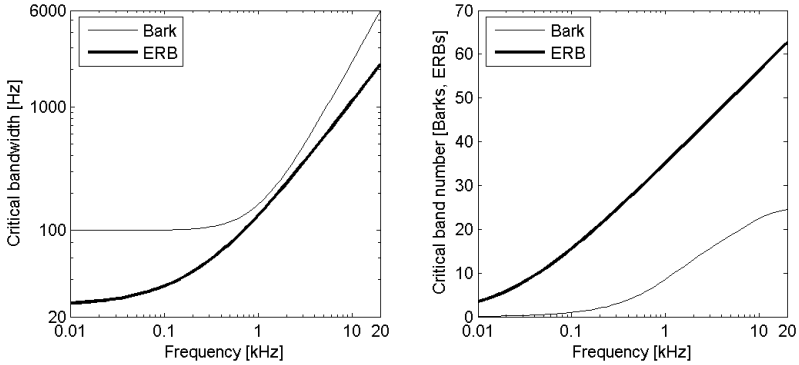

Figure 2.2: The left plot shows the width of the critical bands as function of frequency. The Bark critical bandwidth as well as the ERB critical bandwidth are shown. For frequencies above 10 kHz, the bandwidths are not well known. The right plot shows the critical band number as function of frequency. The critical band numbers are measured in Barks and in ERBs, respectively.

calculated as

$$\mathrm{BW}_{\mathrm{Bark}} = 25 + 75\left(1 + 1.4\left(\frac{f_c}{1000}\right)^{0.69}\right) \tag{2.1}$$

and

$$\mathrm{BW}_{\mathrm{ERB}} = 24.7(1 + 0.00437 f_c), \tag{2.2}$$

respectively [43]. The bandwidths as function of frequency are shown in Figure 2.2. The *critical band number* is found by stacking up critical bands until a certain frequency has been reached [43]. Because the critical bandwidth increases with increasing frequency, also the frequency distance between the critical band number grows with increasing frequency. The critical band numbers measured in Barks and in ERBs are calculated as function of the frequency f as [42]

$$\mathrm{Bark}(f) = 13 \arctan\left(0.76\frac{f}{1000}\right) + 3.5 \arctan\left(\left(\frac{f}{7500}\right)^2\right) \tag{2.3}$$

and [62]

$$\mathrm{ERB}(f) = 21.4 \log_{10}\left(4.37\frac{f}{1000} + 1\right), \tag{2.4}$$

respectively. The critical band numbers as function of frequency are also shown in Figure 2.2.

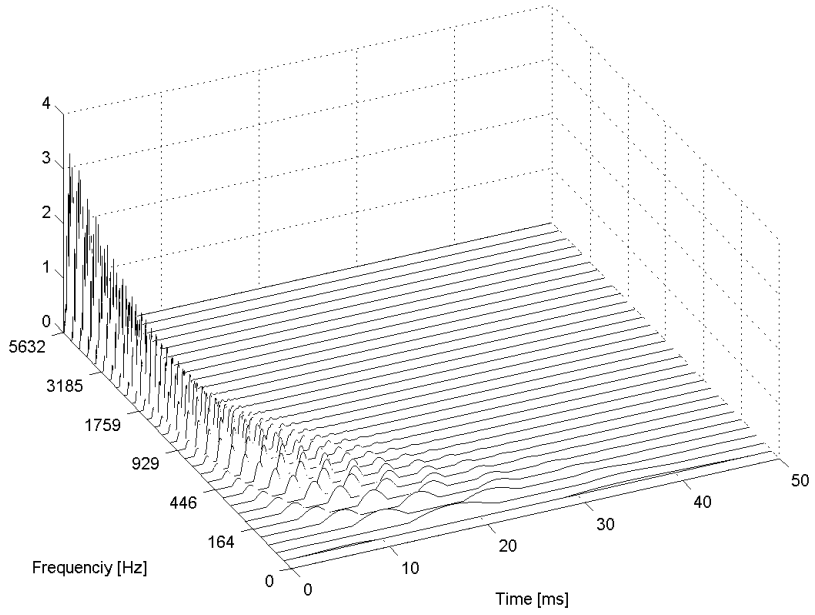

Figure 2.3: Gammatone auditory filters as function of the frequency and the time. It can be seen that the group delay of the low frequencies has longer impulse responses than the group delay of the high frequencies. In order to make the illustration clearer, the filter coefficients have been half-wave rectified. The filters with center frequencies corresponding to 1–20 ERBs are shown.

2.1 The Gammatone Filterbank

The impulse response of the Gammatone auditory filter of order n is given by the following formula [43, p. 254]:

$$g(t) = b^n t^{n-1} e^{-2\pi bt} \cos(2\pi f_c t + \varphi)$$

The envelope of the filter is thus given by $b^n t^{n-1} e^{-2\pi bt}$. This envelope is proportional to the Gamma distribution. In order to fit the response of the auditory nerve fibers of a human being with normal hearing well, $n = 4$ and depending on the center frequency, $b = 1.018$ ERBs. The impulse responses of a Gammatone filterbank are shown in Figure 2.3, and in Figure 2.4, the corresponding magnitude responses are shown. The cochlea is well modeled with a Gammatone

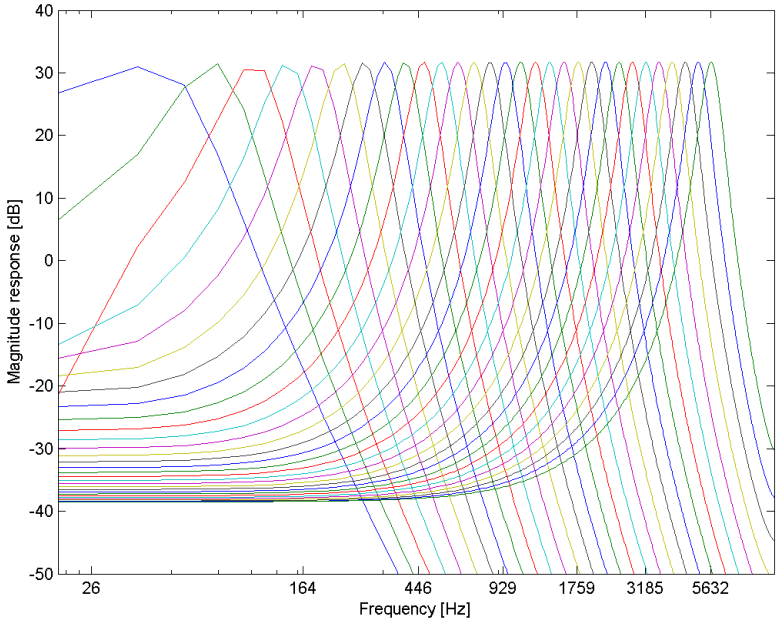

Figure 2.4: Magnitude responses of Gammatone auditory filters as function of the frequency on a logarithmic frequency scale. Magnitude responses of filters with center frequencies corresponding to 1–20 ERBs are shown.

filterbank. In the cochlear model, the potentials in the inner hair cells are modeled by half-wave rectifying and low-pass filtering the output of the filterbank (see e.g. [33]). A diagram of such a cochlear filtering is given in Figure 2.5.

2.2 Time-Frequency Distributions of Audio Signals

In this section different possible time-frequency distributions of audio signals are presented. As shown previously, the T-F processing of an audio signal can be regarded as the outputs of a bank of band-pass filters at different times. The spectrogram is obtained by the STFT. In Figure 2.6 three different time frequency distributions of the same speech signal are shown. The first T-F

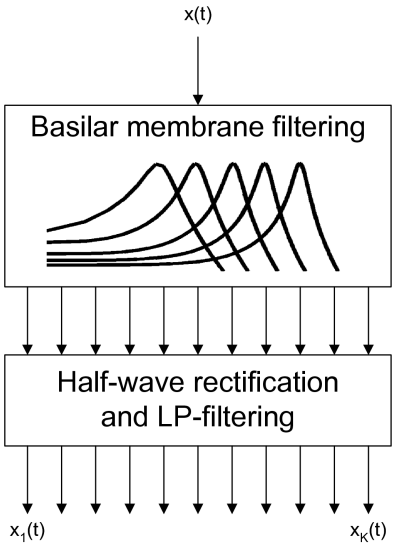

Figure 2.5: Cochlear filterbank. The signal is first band-pass filtered, e.g. by the Gammatone filterbank. Hereby, a non-linearity given by e.g a half-wave rectifier and a low-pass filter mimics the receptor potential in the inner hair cells.

distribution is a spectrogram with a linear frequency scale. We see that the frequency resolution of the Fourier transform is linear. The frequency resolution in the human ear is not linear. As it could be seen in Figure 2.2, at the lower frequencies, the human ear has a better frequency resolution than at the higher frequencies. The second T-F distribution in Figure 2.6 shows the spectrogram with a non-linear frequency distribution. By use of frequency warping [42], the frequency scale is chosen in order to follow the Bark frequency scale. With frequency warping, the Bark frequency scale can be approximated well by a delay line consisting of first-order all-pass filters [42]. Compared to the spectrogram with the linear frequency scale, the warped spectrogram has a better frequency resolution for the low frequencies on the expense of a worse frequency resolution for the high frequencies and different group delay across the frequencies.

The third T-F distribution in Figure 2.6 shows a so-called *cochleagram* [60, 86, 85]. The cochleagram uses a cochlear model to imitate the output response of the cochlea. Depending on the frequency of the stimuli, the neural activity has a maximum at a certain position on the basilar membrane.

In the shown cochleagram, the cochlea has been mimicked by the Gammatone

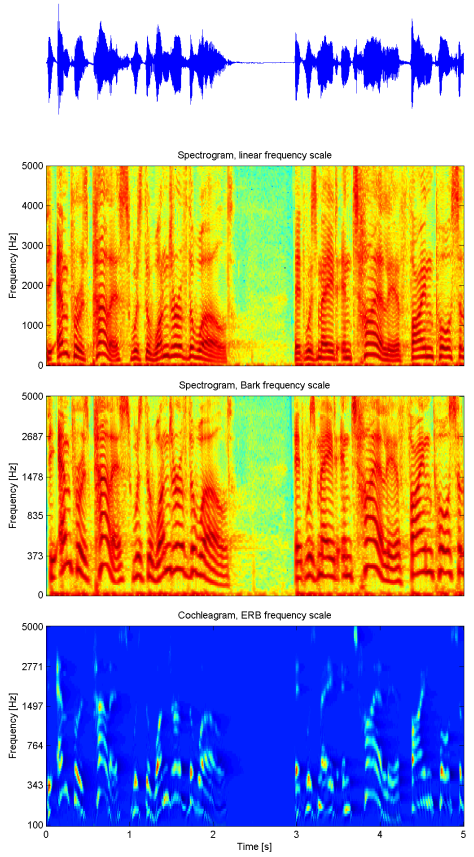

Figure 2.6: Three different time-frequency representations of a speech signal. The first T-F distribution is a spectrogram with a linear frequency distribution. The second T-F distribution shows the spectrogram, where the frequencies are weighted according to the Bark frequency scale. The frequency resolution is however higher than the resolution of the critical bands. The third T-F distribution is the so-called *cochleagram*. In the cochleagram, the frequency resolution corresponds to the frequency resolution in the human cochlea. Also here, the frequency scale is not linear, but follows the ERB frequency scale.

filterbank, followed by a hair cell model [61, 45] as it was illustrated in the diagram in Figure 2.5. The frequency scale in the shown cochleagram follows the ERB frequency scale. When the cochleagram is compared to the two spectrograms, we observe that the T-F distributions in the cochleagram is more sparse at the high frequencies compared to the lower frequencies. We thus have more spectral information in the high-frequency part of the two spectrograms than necessary.

2.2.1 Additional Auditory Models

Clearly more cues about a audio signal can be resolved, when the audio signal is decomposed into T-F components compared to when an audio signal is presented in either the time domain or in the frequency domain. However, not all perceptual properties can be resolved from an audio signal presented in the T-F domain. Other representations of an audio signal may resolve other perceptual cues. As an example, it is hard to resolve binaural cues from a single T-F distribution. On the other hand, the T-F distribution emphasizes other properties of an audio signal such as reverberations; even though they only have a minor influence on the perceived sound, the reverberations can clearly be seen in a spectrogram.

The slow varying modulations of a speech signal is not well resolved from the T-F distribution in the spectrogram. In order to better resolve this perceptual cue, a modulation spectrogram has been proposed [40]. Modulation filterbanks have also been applied into models for the auditory system [33]. Other modulation filterbanks have also been proposed. From some of the models, the audio signal can be reconstructed [84, 7, 83].

Chapter 3

Auditory Scene Analysis

Knowledge about the behavior of the human auditory system is important for several reasons. The auditory scene consists of different streams and the human auditory system is very good at paying attention to a single auditory stream at a time. In combination with auditory models, auditory scene analysis provides a good basis for understanding T-F masking, because the grouping in the brain and in the exclusive allocation in T-F masking are very similar.

An auditory stream may consist of several sounds [21]. Based on different *auditory cues*, these sounds are grouped together in order to create a single auditory stream. As it was illustrated in Figure 2.5, the basilar membrane in the cochlea performs a time-frequency analysis of the sound. This segmentation of an auditory signal into small components in time and frequency is followed by a grouping where each component is assigned a certain auditory stream. This segmentation and grouping of auditory components is termed auditory scene analysis [21]. A *principle of exclusive allocation* exists, i.e. when an auditory element has been assigned to a certain auditory stream, it cannot also exist in other auditory streams.

There are many similarities between auditory grouping and visual grouping. Like an auditory stream consists of several acoustic signals, also visual streams may consist of different objects which are grouped together, e.g. in vision many closely spaced trees are perceived as a forest while in the auditory domain, many

instruments playing simultaneously can be perceived as a single melody.

A speech signal is also perceived as a single stream even though it consists of different sounds. Some sounds originate from the vocal tract, other from the oral or nasal cavities. Still, a speech sound is perceived as a single stream, but two speakers are perceived as two different streams. Also music often consists of different instruments. Each instrument can be perceived as a single sound, but at same time, the instruments playing together are perceived as a single music piece.

Speech consists of voiced and unvoiced sounds. The voiced sounds can be divided into different groups such as vowels and sonorants. Vowels are produced from a turbulent airflow in the vocal tract. They can be distinguished from each other by the formant patterns. Sonorants are voiced speech sounds produced without a turbulent airflow in the vocal tract such as e.g. 'w' or the nasal sounds such as 'm' or 'n'. The unvoiced sounds are fricatives (noise) such as 'f' or 's' or affricates (stop sounds) such as 'p', 'd', 'g' or 't'.

Humans group sound signals into auditory streams based on different auditory cues. The auditory cues can be divided into two groups: *primitive* cues and *schema-based* cues [21, 31].

3.1 Primitive Auditory Cues

The primitive auditory cues are also called bottom-up cues. The cues are innate and they rely on physical facts which remain constant across different languages, music, etc. The primitive cues can be further divided into cues organized simultaneous, and cues which are organized sequentially. By simultaneous organization is meant acoustic components which all belong to the sound source at a particular time while sequential organization means that the acoustic components are grouped so that they belong to the same sound source across time. The following auditory cues are examples of primitive cues:

Spectral proximity Auditory components which are closely spaced in frequency tend to group together.

Common periodicity (Pitch) If the acoustic components have a common fundamental frequency (F_0), the sounds tend to group together. The cue becomes more strongly defined when many harmonics are present. Harmonics are frequencies which are multiples of the fundamental frequency.

Timbre If the two sounds have the same loudness and pitch, but still are dissimilar, they have different timbre. Timbre is what makes one instrument different from another. Timbre is multi-dimensional. One dimension of timbre is e.g. brightness.

Common fate Frequency components are grouped together when they change in a similar way. Common fate can be divided into different subgroups:

- Common onset: The auditory components tend to group, when a synchronous onset across frequency occurs.
- Common offset: The auditory components tend to group, when a synchronous offset across frequency occurs.
- Common modulation: The auditory components tend to group, if parallel changes in frequency occur (frequency modulation (FM)) or if the amplitudes change simultaneously across frequency (amplitude modulation (AM)).

Spatial cues When auditory components are localized at the same spatial position, they may group together, while components at different spatial positions may belong to different auditory streams. The human auditory system uses several cues to localize sounds [15]. Some cues are binaural, other cues are monaural:

- Interaural time difference (ITD): For low frequencies the time (or phase) difference between the ears is used to localize sounds. For frequencies above 800 Hz, the effect of the ITD begins to decrease and for frequencies above 1.6 kHz, the distance between the ears becomes greater than half a wavelength, and spatial aliasing occurs. Thus the ITD becomes ambiguous, and cannot be used for localization.
- Interaural Envelope Difference (IED): For signals with a slowly-varying envelope differences between the two ears, the envelope difference is used as a localization cue.
- Interaural level difference (ILD): For frequencies above approximately 1.6 kHz, the head attenuates the sound, when it passes the head (shadowing effect). The ILD is thus used for high frequencies to localize sounds.
- Diffraction from the head, reflections from the the shoulders and the pinna, are monaural cues which are used to localize sounds. The brain is able to use these special reflections for localization. These cues are most effective for high frequency sounds, and they are especially used to discriminate between whether a sound is arriving from the front or from the back.
- Head movements: Small head movements is another monaural cue used for sound localization.

- Visual cue: The spatial grouping becomes stronger if it is combined with a visual perception of the object.

Continuity If a sound is interrupted by e.g. a large noise burst so that a discontinuity in time occurs, the sound is often perceived as if it continues through the noise.

3.2 Schema-based Auditory Cues

The schema-based auditory cues are all based on stored knowledge. Here, the auditory system organizes acoustic components based on schemas. In schema-based scene analysis, the auditory system searches for familiar patterns in the acoustic environment. Therefore, the schema-based cues are also called *top-down* cues. Top-down means that on the basis of prior information, the brain makes a grouping decision at a higher level that influences the lower-level (primitive) grouping rules [36]. Contrary, the primitive cues are called *bottom-up* cues. Examples of schema-based cues are

Rhythm An expectation of a similar sound after a certain period is an example of an schema-based cue.

Attention In situations with several auditory streams, humans are able to voluntarily pay attention to a single stream. Whenever humans listen for something, it is part of a schema.

Knowledge of language Knowledge of a language makes it easier to follow such an auditory stream.

Phonemic restoration This cue is closely related to the continuity cue. Phonemes in words which are partly masked by noise bursts can sometimes be restored by the brain so that the partly incomplete word is perceived as a whole word.

3.3 Importance of Different Factors

Often different auditory cues may lead to different grouping of the acoustic elements in an auditory scene. Thus the cues compete against each other. Some auditory cues are stronger than others. For example, experiments have shown that frequency proximity is a stronger cue than the spatial origin of the sources

[21]. In listening experiments, variations are often seen across the listeners. Some of these variations can e.g. be explained by different schema-based auditory grouping across individuals. If a listener is exposed to a sentence several times, the words become easier to recognize.

3.4 Computational Auditory Scene Analysis

In computational auditory scene analysis (CASA), methods are developed in order to automatically organize the auditory scene according to the grouping cues. By use of the auditory grouping rules, each unit in time and frequency can be assigned to a certain auditory stream [96, 23, 31, 95]. When the T-F units have been assigned, it becomes easier to segregate the sources of interest from the remaining audio mixture.

Many computational models have been proposed. Some systems are based on a single auditory cue, while other systems are based on multiple cues. Some systems are based on single channel (monaural) recordings [96, 23, 94, 46, 48], whereas other systems are based on binaural recordings [65, 79, 78].

A commonly used cue for speech segregation is common periodicity. As an example, a CASA system based on pitch estimation has been proposed in [46]. When the system only uses pitch as a cue for segregation, it is limited to segregation of the voiced part of speech. Common onset and offset have also been used, together with the frequency proximity cue for speech segregation models [23, 47, 48]. By using onset and offset cues, both voiced and unvoiced speech can be segregated from a mixture [48]. Temporal continuity was used for segregation in [94].

The localization cues have also successfully been used to segregate sources from a mixture. The interaural time difference (ITD) and the interaural intensity difference (IID) have efficiently been used to segregate a single speaker from a mixture of several simultaneous speakers ITD/IID [65, 79, 78]. The IID has also been used in [28]. With strong models of the acoustic environment, also monaural localization cues have been used for monaural source separation [68].

Segregation of signals, where each sound is assumed to have different amplitude modulation, has also been performed. In [7], different music instruments have been segregated based on different amplitude modulation for each instrument.

Model-based methods have also been used for computational auditory grouping, segregation, and enhancement of speech [97, 65, 36, 12]. In [12], primitive cues

are used to divide the time-frequency representation of the auditory scene into fragments. Trained models are hereafter used to determine whether a fragment belongs to the speech signal or to the background.

Chapter 4

Time-Frequency Masking

To obtain segregation of sources from a mixture, the principle of exclusive allocation can be used together with the fact that speech is sparse. by sparseness is meat that speech signals from different sources only to some extent overlap in time and in frequency. Each unit in the T-F domain can thus be labeled so that it belongs to a certain source signal. Such a labeling can be implemented as a binary decision: The T-F unit is labeled with the value '1' if the unit belongs to the audio signal. Contrary, if the T-F unit does not belong to the signal of interest, it is labeled with the value '0'. This binary labeling of the T-F units results in a so-called *binary time-frequency mask*.

The separation is obtained by applying the T-F mask to the signals in the T-F domain, and the signals are reconstructed with a bank of synthesis filters. This is illustrated in Figure 4.1.

4.1 Sparseness in the Time-Frequency Domain

Speech is sparsely distributed in the T-F domain. Even in very challenging environments with some overlap between competing speakers, speech remains intelligible. In [22], experiments with binaural listening under anechoic conditions have shown that a speech signal is still intelligible even though there is up

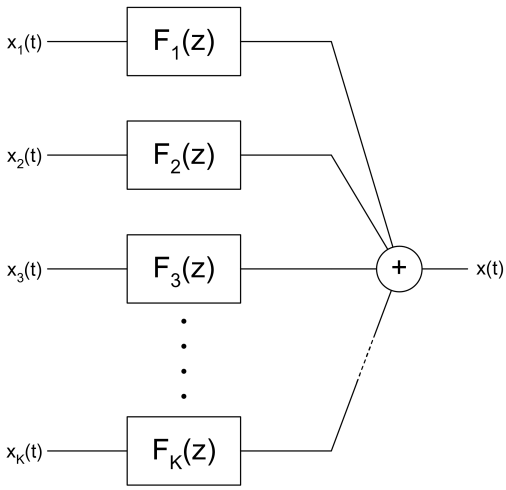

Figure 4.1: Like the T-F distrubution is obtained by a bank of bandpass filters (see Figure 2.1), the synthesis is also obtained by a bank of band-pass filters.

to six interfering speech-like signals, where all signals have the same loudness as the target signal. A speech signal is not active all the time. Thus speech is sparse in the time domain. Further, the speech energy is concentrated in isolated regions in time and frequency. Consequently, speech is even sparser in the T-F domain. This is illustrated in Figure 4.2. Here histogram values of speech amplitudes are shown in the case of one speech signal and a mixture of two speech signals. The amplitude values are shown both for the time domain and the time-frequency domain. Many low values indicate that the signal is sparse. As expected, one talker is sparser than two simultaneous talkers. It can also be seen that the T-F representation of speech is more sparse than the time domain representation.

Another way to show the validity of the sparseness in the T-F domain comes from the fact that the spectrogram of the mixture is almost equal to the maximum values of the individual spectrograms for each source in the logarithmic domain [82], i.e. for a mixture consisting of two sound sources

$$\log(e_1 + e_2) \approx \max(\log(e_1), \log(e_2)), \tag{4.1}$$

where e_1 and e_2 denotes the energy in a T-F unit of source 1 and source 2, respectively.

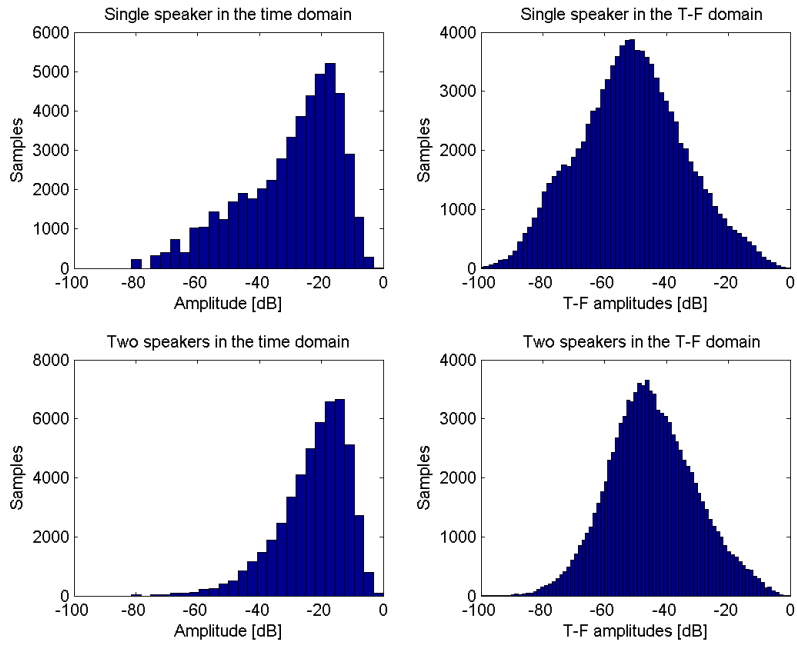

Figure 4.2: The histograms show the distribution of the amplitude of audio signals consisting of one and two speakers, respectively. The two left histograms show the amplitude distribution in the time domain, the right histograms show the amplitude distributions in the T-F domain obtained from the spectrograms in Figure 2.6 and Figure 4.4a. Many histogram values with small amplitudes indicate that the signal is sparse. It can be seen that the signals are sparser in the T-F domain compared to the time domain.

4.2 The Ideal Binary Mask

An optimal way to label whether a T-F unit belongs to the target signal or to the noise is for each T-F unit to consider the amplitude of the target signal and the amplitude of the interfering signals. For each T-F unit, if the target signal has more energy than all the interfering signals, the T-F unit is assumed to belong to the source signal. It is then labeled with the value '1'. Otherwise, the T-F unit is labeled with the value '0'. Given a mixture consisting of N audio

sources, the binary mask of the ith source in the mixture is thus given by

$$\mathrm{BM}_i(\omega,t) = \begin{cases} 1, & \text{if } |S_i(\omega,t)| > |X(\omega,t) - S_i(\omega,t)|; \\ 0, & \text{otherwise,} \end{cases} \tag{4.2}$$

where $S_i(\omega,t)$ is the ith source at the frequency unit ω and the time frame unit t and $X(\omega,t) - S_i(\omega,t)$ is the mixture in the T-F domain, where the ith source is absent. $|\cdot|$ denotes the absolute value. This mask has been termed the *ideal binary mask* [93] or the 0-dB mask [98]. Here 0 dB refers to that the decision boundary is when the local signal-to-noise ratio for a particular T-F unit is 0 dB. The ideal binary mask cannot be estimated in real-world applications, because it requires the knowledge of each individual source before mixing. With T-F masking techniques, the original source cannot be obtained, but due to the strong correlation between the signal obtained by the ideal binary mask and the original signal, the ideal binary mask has been suggested as a computational goal for binary T-F masking techniques [93, 37]. In theory, each original source in the mixture could be obtained from T-F masking, but it requires that the T-F mask is complex-valued. The quality and the sparsity of the ideal binary mask depends on the overall signal-to-noise ratio. If the noise is much stronger than the target signal, only few T-F units have a positive local SNR. Hereby the ideal binary mask becomes sparse, and quality of the estimated signal is poor.

To assign the T-F unit to the dominant sound, also corresponds well with auditory masking [62]. Within a certain frequency range where multiple sounds are present, the louder sound will mask the other sounds. The auditory masking phenomenon may also explain why T-F masking performs very well in segregating sources even though the sources overlap.

In Figure 4.3 and Figure 4.4, examples of ideal binary masks applied to speech mixtures are shown. Here, the mixture consists of a male speaker and a female speaker. The spectrogram of the mixture is shown in part a of the figures. The two ideal binary masks are calculated from equation (4.2) for all T-F units (ω,t) as

$$\mathrm{BM}_{\mathrm{male}}(\omega,t) = \begin{cases} 1, & \text{if } |S_{\mathrm{male}}(\omega,t)| > |S_{\mathrm{female}}(\omega,t)|; \\ 0, & \text{otherwise,} \end{cases} \tag{4.3}$$

and

$$\mathrm{BM}_{\mathrm{female}}(\omega,t) = \begin{cases} 1, & \text{if } |S_{\mathrm{female}}(\omega,t)| > |S_{\mathrm{male}}(\omega,t)|; \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

In order to obtain estimates of the two individual speakers in the frequency domain, the two binary masks are applied to the mixture by an element wise multiplication in the T-F domain, i.e.

$$\tilde{S}_i(\omega,t) = X(\omega,t) \circ \mathrm{BM}_i(\omega,t), \tag{4.5}$$

where $\circ$ denotes the element wise multiplication. The obtained spectrograms are shown in part c of Figure 4.3 and Figure 4.4. Like the spectrogram (analysis filter) is obtained by the STFT, the inversion of the spectrogram (synthesis) is obtained by the inverse STFT (ISTFT).

In Figure 4.3d and Figure 4.4d the spectrograms of the synthesized signals are shown. The spectrograms of the two original signals are shown in Figure 4.3e and Figure 4.4e, respectively.

Also in cases, where e.g. a Gammatone filterbank has been used for analysis, synthesis is possible. In order to recover an auditory model, especially the phase recovery is difficult [87]. The Gammatone filterbank has different group delay for different frequencies. This makes perfect synthesis difficult. Inversion of auditory filterbanks is discussed in [87, 54, 58, 57].

Consider again the spectrograms in part c. It is important to notice that even though a T-F unit in the binary mask is zero, its resulting synthesized signal contains energy in these T-F units, as it can be seen, when the spectrograms are compared to those in part d. This can be explained by considering the diagrams in Figure 4.5. When the signal representation is converted from the time domain into the T-F domain representation, the signal is represented in a higher-dimensional space. Because the dimension of the T-F domain is higher, different representations in the T-F domain may be synthesized into the same time domain signals. However, a time domain signal is only mapped into a single T-F representation. The T-F representation can also be viewed as a subband system with overlapping subbands [91]. Due to the overlapping bands, the gain in each band may also be adjusted in multiple ways, in order to obtain same synthesized signal.

When different sources overlap in the T-F domain, a binary mask may remove useful information from the target audio signal, because some areas in the T-F domain are missing. Recently, methods have been proposed in order to recover missing areas in the T-F domain [11, 80]. Based on the available signal, and training data, missing T-F units are estimated. The idea is that the training data which fits the missing T-F areas best are filled into these areas.

4.3 Distortions

When a T-F mask is applied to a signal, distortions may be introduced. These distortions are known as musical noise. Musical noise are distortions artificially introduced by the speech enhancement algorithm. Musical noise are short si-

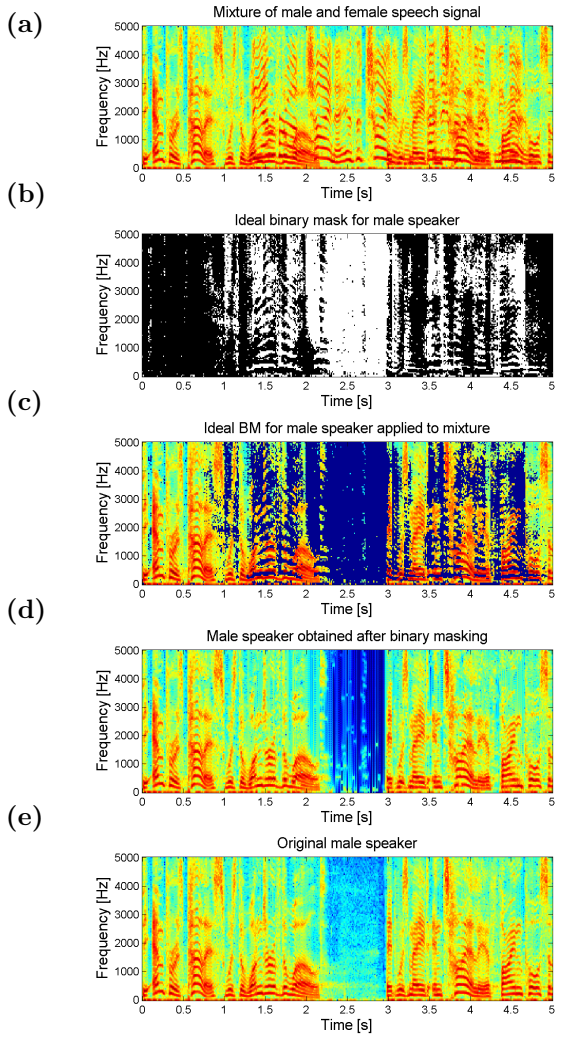

Figure 4.3: Segregation by binary masking. The male speaker (e) is segregated from the mixture (a) which consists of a male and a female speaker. The binary mask (b) is found such that T-F units where the male speaker has more energy than the female speaker has the value one, otherwise zero. The black T-F units have the value '1'; the white T-F units have the value '0'. The binary mask is applied to the mixture by an element wise multiplication and the spectrogram in (c) is thus obtained. The spectrogram of the estimated male speaker after synthesis is shown in (d).

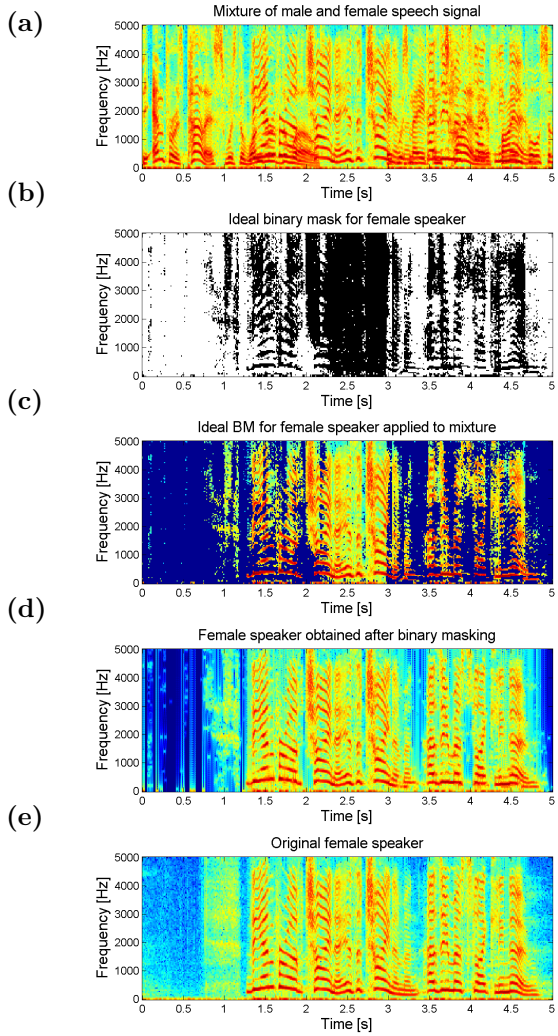

Figure 4.4: Segregation by binary masking like in Figure 4.3. Here the female speaker (e) is segregated from the mixture (a). The spectrogram of the estimated signal is shown in (d).

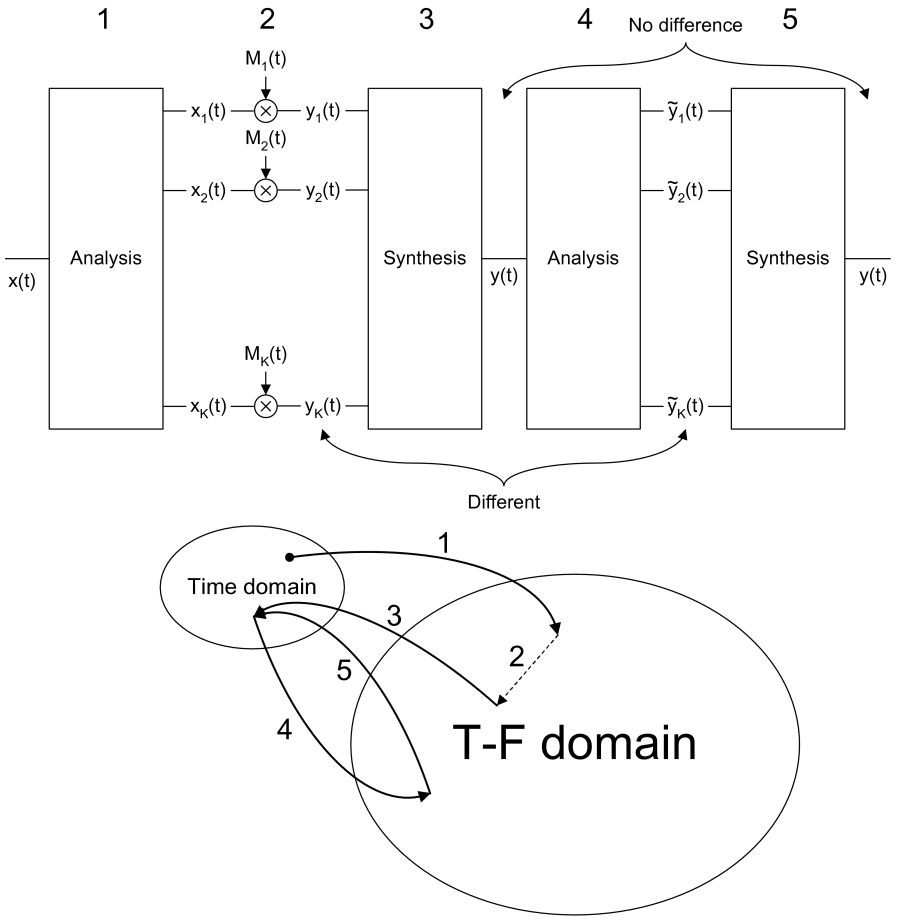

Figure 4.5: The time domain signal is mapped into the T-F domain by an analysis filterbank (step 1). The K bands in the T-F domain signal are modified by a T-F mask (step 2), where a gain is applied to each frequency band. The modified signal is transformed back into the time domain again by a synthesis filterbank (step 3). Because the dimension of the signal representation in the T-F domain is higher than the time domain representation, different T-F representations map into the same time-domain signal. Different time domain signals always map into different T-F representations.

nusoidal peaks at random frequencies and random times [24]. Distortion from musical noise deteriorates the quality of the speech signal. This deterioration of a sound signal can be explained by auditory scene analysis. Since the noise

occurs as random tones in time and frequency, it is very unlikely that the tones group with other acoustic components in the auditory scene. Because the tones occur at random times, they do not have any common onset/offset, or any rhythm. Since the tones occur at random frequencies, it is unlikely that there are common harmonics. Thus musical noise is perceived as many independent auditory streams that do not group together. Because these random tones are perceived as many unnatural sounds, a listener feels more annoyed by musical artifacts than by white noise [59]. The amount of musical noise also depends on how sparse the enhanced audio signal is. Peaks in the spectrogram far from the acoustic components are likely to be perceived as musical noise, while peaks closer to the audio signal are more likely to be below the masking threshold of the acoustic component. Two features were proposed in [59] in order to identify musical noise from speech: In the frequency axis of the spectrogram, the musical noise components are far from the speech components, so that they are not masked, and on the time axis in the spectrogram, the frequency magnitudes of the musical noise components vary faster than the speech components. Musical noise can be reduced e.g. by applying a more smooth mask or by smoothing the spectrogram. Smoothing in order to reduce the musical noise has been suggested in e.g. [1, 6, 3, 5, 4]. In [6], smoothing in time was proposed in order to reduce musical noise. The smoothing was applied by choosing a shift in the overlap-add ISTFT reconstruction which was much shorter than the window length.

Aliasing is another reason why energy is still present (and audible) within some of the removed T-F units. A binary mask can be regarded as a binary gain function multiplied to the mixture in the frequency domain. If the T-F analysis is viewed as a subband system, aliasing is introduced if the subband is decimated. The aliasing effects can however be avoided by a careful design of the analysis and synthesis filters. Effects from aliasing can also be reduced by using filterbanks without decimation. An example of a filterbank is the spectrogram. The spectrogram is usually decimated by a factor M/m, where M is the window length and m is the frame shift. In e.g. [28] filterbanks without decimation have been used.

4.4 Methods using T-F Masking

During the past two decades, numerous approaches for T-F masking based audio segregation have been proposed. To include some T-F areas and to omit other T-F areas in order to segregate speech signals was first proposed by Weintraub [96]. Physical cues from the acoustic signal was used to assign different T-F regions to the different sounds in the mixture. In [96, 97], the pitch was estimated and

used to allocate the different T-F regions. Most of the other CASA methods mentioned in Section 3.4 utilize T-F masking, e.g. [23, 94, 46, 48, 47, 12].

More recently, T-F masking has been used with techniques not directly inspired by the behavior of the human auditory system, but more based on the fact that audio signals, such as speech is sparse in the T-F domain. Clustering of sources based on histograms over time and amplitude differences between different microphone recordings was proposed in [51], and further extended in [9, 10, 75, 77, 98, 49, 76]. In [76], M-microphone beamforming and T-F masking was combined in order to cope with sources that overlap in the T-F domain. Hereby, $M-1$ simultaneous sources were allowed in each T-F unit. In [79, 78], ITD/IID histogram data were clustered in order to segregate speech signals. Direction of arrival (DOA) estimation based on clustering of delays between closely-spaced microphones has also been applied for binary mask estimation [89].

ICA and T-F masking have been combined in different ways. Audio signals can be segregated by finding the DOA of the signal and applying a binary mask which only allows arrival directions in a narrow band around the estimated DOA. This binary mask is usually very sparse and audible distortions are introduced from this very narrow binary mask [18]. In order to reduce the musical noise, it has been proposed to use DOA-based binary masks to remove $N-M$ signals from the mixture with N sources recorded at M microphones. Hereby the remaining problems consists of M sources recorded at M microphones. In this problem, the mixing matrix can be inverted, and the sources can be recovered with less musical artifacts [16, 17, 18, 2, 3]. In order to reduce musical artifacts even further, it has also been proposed to apply a continuous T-F mask in the first step contrary to the binary mask. The continuous mask is based on the DOA [5, 4].

T-F masks can also be applied to the output of an ICA algorithm. In order to cancel crosstalk and to enhance speech even further, the absolute value of two segregated signals from an ICA algorithm were compared, and a binary mask can be estimated based on comparison between the two output signals. The binary masks can then be applied to the two segregated signals in order to enhance them further [53]. This method has also been used as a pre-processing step for feature extraction and speech recognition [52]. A somewhat related, very computationally simple method for estimation of a binary mask is in the special case of the *better microphone* [63]. If recordings are available at multiple microphones and one of the microphones, *the better microphone*, is closer to the target speaker, the target speaker is dominant in this recording compared to the other recordings. Consequently a binary mask which primarily segregates the

primary speaker can simply be estimated as

$$\mathrm{BM}_{\mathrm{target}}(\omega,t) = \begin{cases} 1, & \text{if } |X_{\mathrm{better}}(\omega,t)| > |X_{\mathrm{other}}(\omega,t)|; \\ 0, & \text{otherwise,} \end{cases} \tag{4.6}$$

where $X_{\mathrm{better}}(\omega,t)$ denotes the recording at the better microphone and $X_{\mathrm{other}}(\omega,t)$ is the recording at the other microphone. In [63] a method similar to the one proposed in [53] was used. Here SIMO-ICA was used instead. In single-input-multiple-output (SIMO) ICA, each segregated output is estimated as if it was recorded at all the different microphones in the absence of other sources [88]. Hereby, if the separated source was recorded at two microphones, each separated source also exists as if it was recorded at the two microphones in absence of the other sources. In [63], the binary mask is used to remove the residual error of the SIMO output the error is found by comparison between the same different outputs of the same source.

ICA with T-F masking as a post-processing step have also been used iteratively in order to separate underdetermined mixtures [72, 73, 71]. The most complete explanation of this method is given in the in the paper in Appendix F.

Binary T-F masks have also been found based on trained recordings from single speakers [81]. Here, only a single microphone was used. Another one-microphone approach based on learned features and auditory cues can be found in [8].

4.5 Alternative Methods to Recover More Sources Than Sensors

Time-frequency masking is only one way to recover N sources from M recordings, where $N \geq M$. Statistical assumptions on the sources can be used to compute the maximum a posteriori estimates of the sources [55, 66], i.e.

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} P(\mathbf{s}|\mathbf{x},\mathbf{A}) \tag{4.7}$$

$$= \arg\max_{\mathbf{s}} P(\mathbf{x}|\mathbf{A},\mathbf{s})P(\mathbf{s}), \tag{4.8}$$

where $\mathbf{x}$ is the mixtures, $\mathbf{A}$ is the estimated mixing parameters, $\hat{\mathbf{s}}$ is the estimate of the sources $\mathbf{s}$ and $P(\cdot)$ denotes the probability density function.

A third method the recover the sources is to assume that the number of sources active at the same time always equals the number of sensors. Hereby each mixing process can be inverted and the obtained sources are then combined afterwards. This assumption has been used in e.g. [19].

Chapter 5

Small Microphone Arrays

In many applications, it is desirable to amplify, sounds arriving from a certain direction and to attenuate sounds arriving from other directions. Such direction-dependent gains can be obtained by processing sounds recorded by a microphone array. For hearing aid applications, a microphone array typically consists of two or three microphones placed close to each other near the ear canal. A typical microphone placement is shown in Figure 5.1. The purpose of this chapter is to review methods for signal enhancement by beamforming. Two types of array configurations are considered in this, a linear microphone array and a circular four-microphone array. These two array types have also been used in the proposed algorithms, which can be found in the appendices.

5.1 Definitions of Commonly Used Terms

In this section we define some commonly used terms.

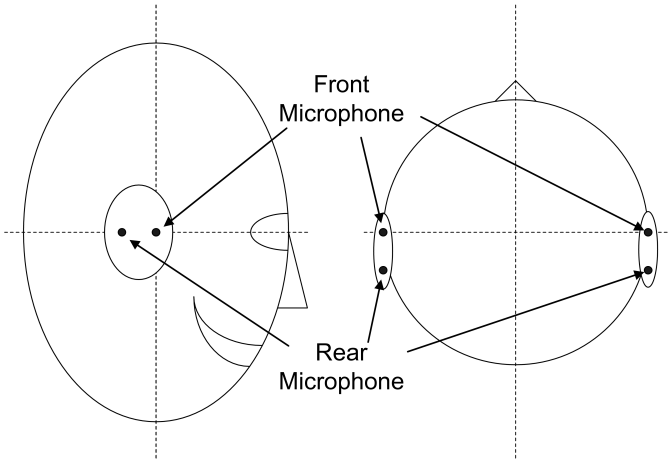

Figure 5.1: Typical placement for a two-microphone array for hearing aid applications. The microphone placements are shown form the side as well as from above.

5.1.1 Free Field Assumption

When a sound propagates in a free field, the sound wave does not pass any obstacles on its way. Thus if the sound source is a point source, the sound propagates spherically from its origin. Because the surface area of a sphere depends on the radius squared, every time the distance r to the source is doubled, the sound intensity I is decreased by a factor of four. The intensity is thus proportional to $1/r^2$. In a free field, the sound level is decreased by 6 dB every time the source distance is doubled [43]. In a room, the sound signal is reflected from the walls and these effects are recorded at the microphones too. Therefore, due to room reverberations and under the assumption that the sound is not attenuated by obstacles on its path, the sound level will be above that of a free field.

When a microphone array is located in a free field, and the distance to the sound source is much larger than the distance d between the microphones ($r \gg d$), the impinging sound wave can be assumed to be a plane wave. Under this far-field assumption, the delay between the array elements and thus the source direction does not depend on r (see Figure 5.2).

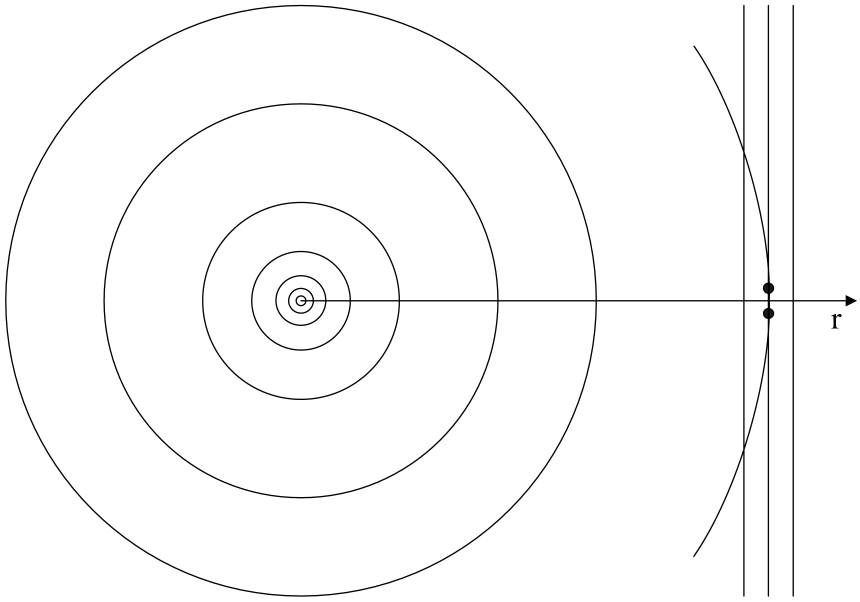

Figure 5.2: In a free field, the sound level is decreased by 6 dB every time the distance r from the source to the microphone array is doubled. It is assumed that the sound wave is planar if the distance d between the elements in the array is much smaller than r.

5.1.2 Spherically Isotopic Noise Field

In a spherically isotropic noise field, all arrival directions are equally likely. We also denote such a field, a diffuse noise field. The coherence between the different microphone signals depends on the frequency f and the microphone distance d and it is given by [20]

$$\Gamma(f,d) = \frac{\sin(\frac{2\pi f}{c}d)}{\frac{2\pi f}{c}d} \tag{5.1}$$

Samples drawn from a spherically isotropic distribution is shown in Figure 5.3.

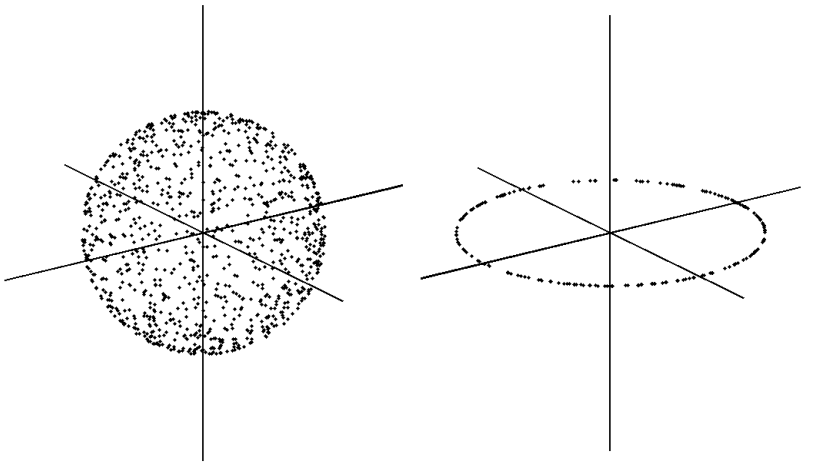

Figure 5.3: The left plot shows samples drawn from a spherical isotropic field. The samples are equally distributed on a sphere. The right plot shows samples drawn from a cylindrical isotropic field. Here the samples are equally distributed on a circle.

5.1.3 Cylindrically Isotopic Noise Field

In a cylindrically isotropic noise field, all arrival directions in the $x-y$-plane are equally likely and sounds only arrive from directions in the $x-y$-plane. Samples drawn from a cylindrically isotropic distribution is shown in Figure 5.3.

5.2 Directivity Index

A beamformer has a response that varies with the direction. The direction is described as a function of the spherical coordinates ϕ and θ. The angles are shown in Figure 5.4. The beamformer can be evaluated with respect to how

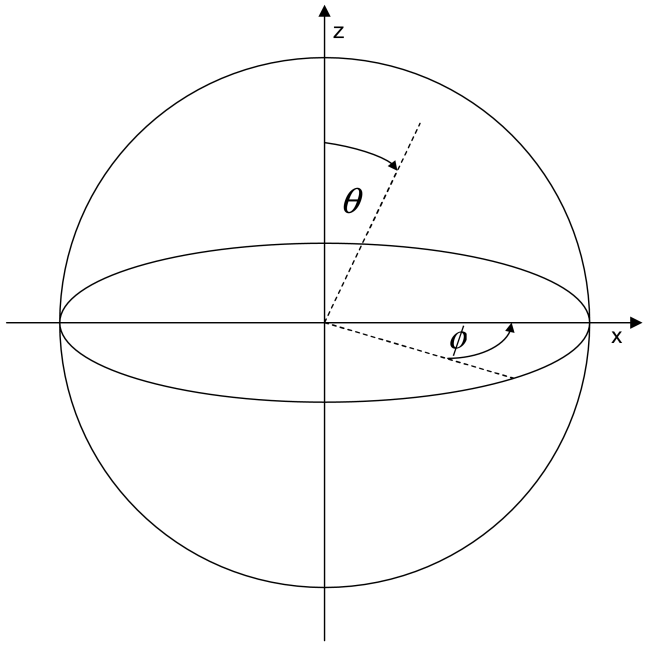

Figure 5.4: The beamformer response depends on the arrival direction of the source signal. In the far-field, the arrival direction can be described as function of ϕ and θ.

well it suppresses a diffuse noise field. In a diffuse noise field, the noise arrives from all directions with the same probability. We also call this type of noise for omnidirectional. The directivity factor (DF) is defined as the ratio between the response from the target direction $R(\phi_0, \theta_0)$ and the response from all directions integrated over the sphere [64].

$$\mathrm{DF} = \frac{4\pi[R(\phi_0, \theta_0)]^2}{\int_0^{\pi} \int_0^{2\pi} [R(\theta, \phi)]^2 \sin(\theta) d\phi d\theta}. \tag{5.2}$$

More frequently, the directivity index (DI) is used. The DI is the DF measured in dB, i.e.,

$$\mathrm{DI} = 10 \log \left(\frac{4\pi[R(\phi_0, \theta_0)]^2}{\int_0^{\pi} \int_0^{2\pi} [R(\theta, \phi)]^2 \sin(\theta) d\phi d\theta} \right) \tag{5.3}$$

If the directivity gain is independent of ϕ, the integral reduces to [90]

$$\mathrm{DI} = 10 \log \left(\frac{2[R(\theta_0)]^2}{\int_0^{\pi} [R(\theta)]^2 \sin(\theta) d\theta} \right) \tag{5.4}$$

As we can see, the directivity factor describes how well a beamformer suppresses sounds which arrive from all directions compared to the sound from the desired direction.

In some situations, the noise is assumed mainly to arrive from the back while the desired signals mainly is assume to arrive from the front direction. Therefore, as an alternative directional signal to noise ratio, also the ratio between the (desired) signals arriving from the front and (unwanted) signals arriving from the back can be found. This ratio is called the *front-to-back* ratio (FBR) [14] and is calculated as

$$\mathrm{FBR} = 10\log\left(\frac{\int_{\theta_0-\pi/2}^{\theta_0+\pi/2}\int_{\phi_0-\pi/2}^{\phi_0+\pi/2}[R(\theta,\phi)]^2\sin(\theta)d\phi d\theta}{\int_{\theta_0+\pi/2}^{\theta_0+3\pi/2}\int_{\phi_0+\pi/2}^{\phi_0+3\pi/2}[R(\theta,\phi)]^2\sin(\theta)d\phi d\theta}\right). \tag{5.5}$$

If the target direction is $(\theta_0, \phi_0) = (0, 0)$, we can change the boundaries and the FBR can be written as [34]

$$\mathrm{FBR} = 10\log\left(\frac{\int_{0}^{\pi/2}\int_{0}^{2\pi}[R(\theta,\phi)]^2\sin(\theta)d\phi d\theta}{\int_{\pi/2}^{\pi}\int_{0}^{2\pi}[R(\theta,\phi)]^2\sin(\theta)d\phi d\theta}\right). \tag{5.6}$$

Further, if the directivity gain is independent of ϕ, the integral reduces to

$$\mathrm{FBR} = 10\log\left(\frac{\int_{0}^{\pi/2}[R(\theta)]^2\sin(\theta)d\theta}{\int_{\pi/2}^{\pi}[R(\theta)]^2\sin(\theta)d\theta}\right). \tag{5.7}$$

5.3 Microphone Arrays

Directivity can either be obtained by adding or by subtracting microphone signals. When the sources are added and possibly delayed, the beamformer is called a *delay-sum beamformer*. When the sources instead are subtracted, we term microphone array a *superdirective beamformer*[34]. The two types of beamformers are illustrated in Figure 5.5. The delay T between the two microphone signals depends on the arrival angle θ, the microphone distance d, and the sound velocity c as

$$T = \frac{d}{c}\cos(\theta). \tag{5.8}$$

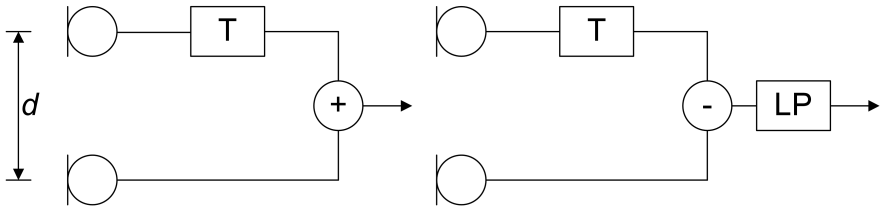

Figure 5.5: In order to obtain directivity, signals recorded at two microphones can either be summed together, or subtracted from each other. By varying the delay in the delay element, The placement of the null can be steered by adjusting the delay T in the delay element. When the difference between the microphone signals are used, the DC component of the signal is removed. In order to compensate for this high-pass effect, the differential delay is followed by a low-pass filter.

5.3.1 The Delay-Sum Beamformer

When the microphones signals are added, we have a null gain at the frequency where one microphone signal is delayed by half a wavelength compared to the other microphone signal, i.e. for $d = \lambda/2$. A null direction is a direction of which there is no directive gain. The direction of the null can be determined by varying the delay element. With a two-microphone array in a spherically isotopic noise field a DI of 3 dB can be obtained at the frequency corresponding to

$$f = \frac{c}{2d}. \tag{5.9}$$

This is illustrated in Figure 5.6, where the two microphones are placed along the x-axis. The distance between the microphones is half a wavelength. Since the microphone signals are added without any delay, the maximum directivity is obtained in the $y - z$ plane. In Figure 5.7 directivity patterns are shown for different wavelengths. The two crosses indicate that the microphones are placed along the x-axis. When $\lambda = 2d$, signals arriving at the end direction of the array are completely canceled out. We also see that the delay-sum beamformer is inefficient for $\lambda \ll 2d$. When $\lambda > 2d$, spatial aliasing occurs and multiple null directions and sidelobes occur in the beampattern.

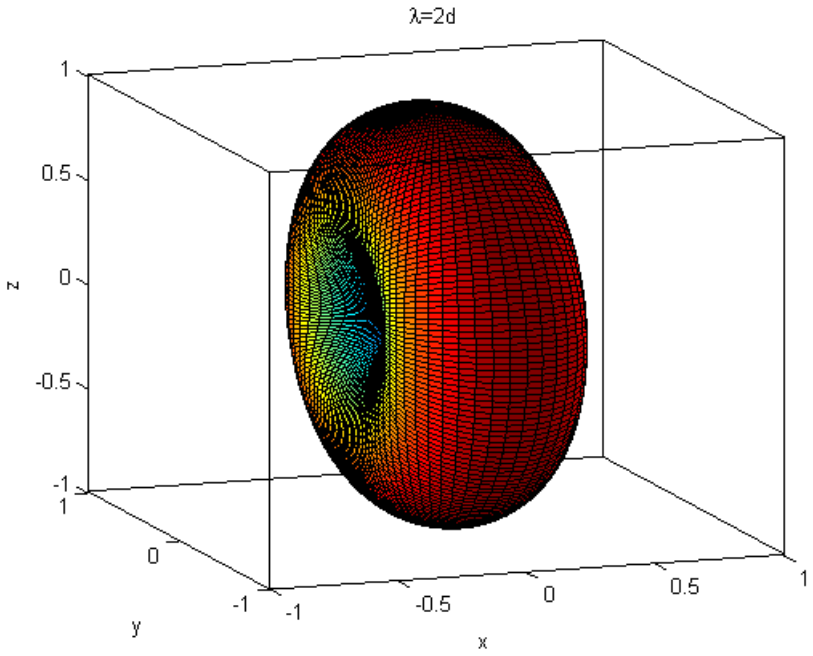

Figure 5.6: Normalized 3D directivity pattern obtained for a delay-sum beamformer. A two-microphone array is placed along the x-axis. The pattern is shown for a wavelength equal to twice the microphone distance. A signal arriving from the from the direction along the x-axis is canceled out because the microphone signals are out of phase, while a signal arriving from a direction perpendicular to the x-axis is in phase and therefore has maximum directivity.

5.3.2 Superdirective Beamformers

Higher directivity can be obtained if instead the difference between the microphone signals is used. For a two-microphone array, the microphone response can be written as function of the arrival angle θ and the frequency as

$$R(\theta, f) = s_1 g_1(\theta) e^{j\frac{kd}{2}\cos(\theta)} + s_2 g_2(\theta) e^{j\frac{kd}{2}\cos(\theta)}, \tag{5.10}$$

where s_1 and s_2 are the sensitivities of the microphones, and $g_1(\theta)$ and $g_2(\theta)$ are the angular sensitivities of the two microphones. $k = 2\pi/\lambda = 2\pi f/c$ is the wave number. If the two microphones have omnidirectional responses, $g_1 = g_2 = 1$, (5.10) reduces to

$$R(\theta) = s_1 e^{-j\frac{kd}{2}\cos(\theta)} + s_2 e^{j\frac{kd}{2}\cos(\theta)}. \tag{5.11}$$

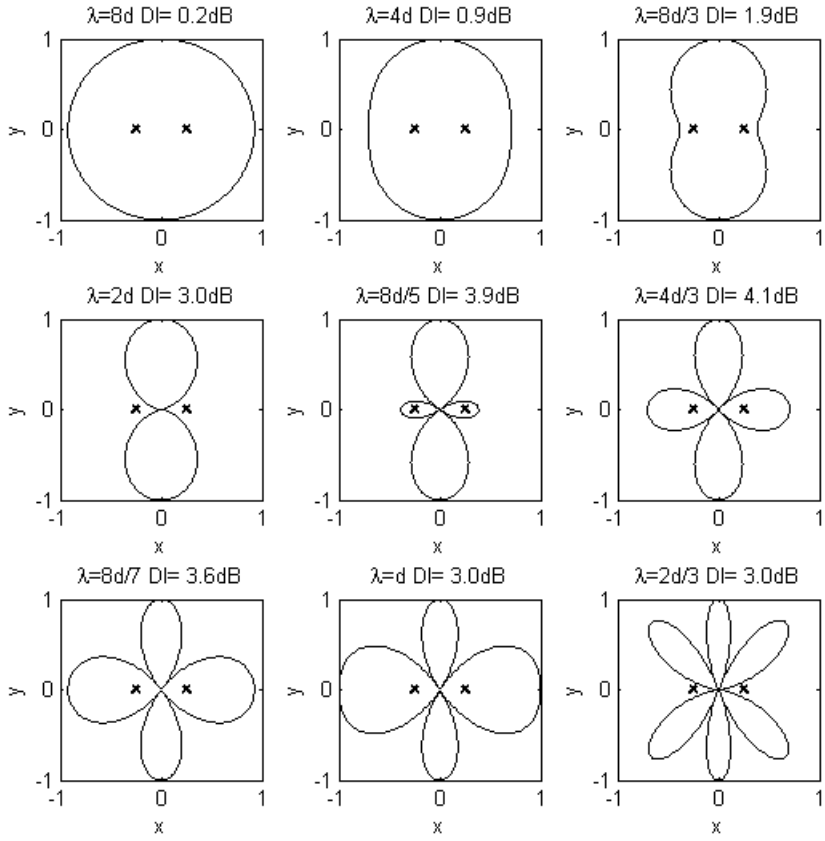

Figure 5.7: Directivity patterns obtained for a delay-sum beamformer for different wavelengths. The crosses indicate that the microphone array is located along the x-axis. For hearing aid applications, the spatial under-sampling is usually avoided by choosing a small microphone distance. Thus the main benefit is that the additional microphone noise is reduced.

By assuming that the microphone distance is much smaller than the wavelength, i.e. $kd \ll \pi$ [90, 34], we can approximate the response by a first order Taylor

polynomial, i.e.

$$\begin{aligned} R(\theta) &\approx s_1(1 - j\frac{kd}{2}\cos(\theta)) + s_2(1 + j\frac{kd}{2}\cos(\theta)) && (5.12)\\ &= (s_1 + s_2) + j\frac{kd}{2}(s_2 - s_1)\cos(\theta) && (5.13)\\ &= A + B\cos(\theta) && (5.14) \end{aligned}$$

Polar curves of the form $R(\theta) = A + B\cos(\theta)$ are called limaçon patterns. By assuming that the target source comes from the $\theta = 0$ direction, the directivity index can be found by inserting (5.14) into (5.4) [90]:

$$\begin{aligned} \text{DI} &= 10\log\left(\frac{2[R(\theta_0)]^2}{\int_0^\pi [R(\theta)]^2\sin(\theta)d\theta}\right) && (5.15)\\ &= 10\log\left(\frac{2[A+B]^2}{\int_0^\pi [A + B\cos(\theta)]^2\sin(\theta)d\theta}\right) && (5.16)\\ &= 10\log\left(\frac{2[A+B]^2}{A^2 + \frac{1}{3}B^2}\right) && (5.17) \end{aligned}$$

The maximum directivity can be found by maximizing (5.17) with respect to A and B. The highest DI (6 dB) is obtained for $A = 1/4$ and $B = 3/4$. The directivity pattern with the highest directivity index is called a *hypercardioid*.

Also the maximum FBR can be found. Here, we insert (5.14) into (5.7):

$$\begin{aligned} \text{FBR} &= 10\log\left(\frac{\int_0^{\pi/2}[R(\theta)]^2\sin(\theta)d\theta}{\int_{\pi/2}^{\pi}[R(\theta)]^2\sin(\theta)d\theta}\right) && (5.18)\\ &= 10\log\left(\frac{\int_0^{\pi/2}[A + B\cos(\theta)]^2\sin(\theta)d\theta}{\int_{\pi/2}^{\pi}[A + B\cos(\theta)]^2\sin(\theta)d\theta}\right) && (5.19)\\ &= 10\log\left(\frac{3A^2 + 3AB + B^2}{3A^2 - 3AB + B^2}\right) && (5.20) \end{aligned}$$

We maximize (5.20) with respect to A and B, and the highest FBR (11.4 dB) is obtained for $A = (\sqrt{3}-1)/2$ and $A = (3-\sqrt{3})/2$. The pattern with the highest FBR is called the *supercardioid*.

In Table 5.1 different first order directional patterns are listed. The omnidirectional pattern which has the same gain for all directions, is actually a delay-sum beamformer. Examples of first order directional patterns are shown in Figure 5.8, and in Figure 5.9, a 3D hypercardioid directional pattern is shown. As mentioned, we assumed that $kd \ll \pi$. In Figure 5.10, the hypercardioid is

Table 5.1: Characterization of different limaçon patterns [34, 90].

Pattern	A	B	DI (dB)	FBR (dB)	Null direction
Omnidirectional	1	0	0	0	–
Dipole	0	1	4.8	0	90°
Cardioid	$\frac{1}{2}$	$\frac{1}{2}$	4.8	8.5	180°
Hypercardioid	$\frac{1}{4}$	$\frac{3}{4}$	6.0	8.5	109°
Supercardioid	$\frac{\sqrt{3}-1}{2}$	$\frac{3-\sqrt{3}}{2}$	5.7	11.4	125°

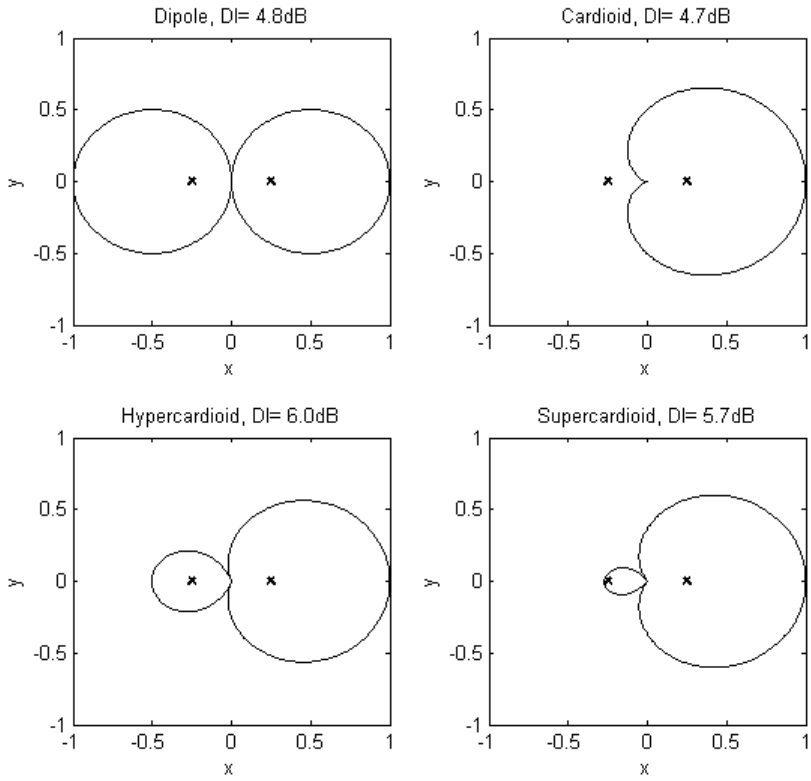

Figure 5.8: Normalized directivity patterns obtained for different values of A and B in (5.14). The wavelength is $\lambda = 20d$. The crosses indicate that the microphone array is located on the x-axis.

plotted as function of wavelength (or frequency). As we see, as kd increases, the pattern begins to change shape, and the directivity index decreases. When

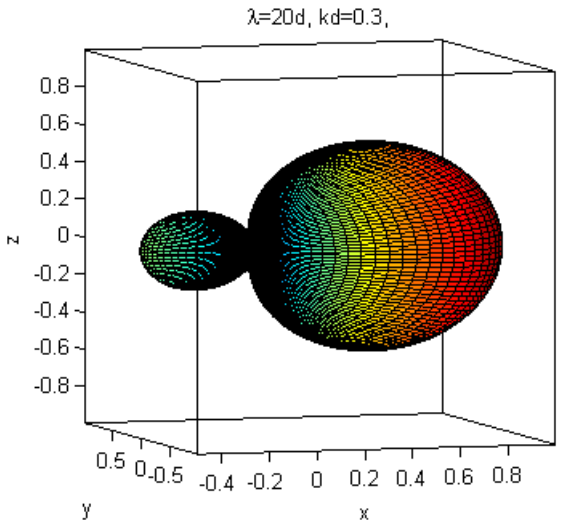

Figure 5.9: 3D directivity pattern obtained for a superdirective the-microphone beamformer. The shown directional pattern is a hypercardioid. The hypercardioid has the maximum directivity index. The two-microphone array is placed on the x-axis.

$\lambda > d/2$, sidelobes begins to appear. Compared to the delay-sum beamformer, where a null occurred at a certain frequency, the null direction of the superdirective beamformer is independent of the frequency. However, frequency dependent nulls appear if there is spatial aliasing. In Figure 5.11 we compare the response of the delay-sum beamformer and the superdirective beamformer as function of frequency. It can be seen that while the delay-sum beamformer have the null placement at wavelengths of $\lambda/2 + n\lambda$, where $n = 0, 1, 2, \ldots$, the superdirective beamformer have the null placement at DC. In order to obtain a flat gain over a wide frequency range, the high-pass effect in the superdirective beamformer is compensated by a low-pass filter. Such a low-pass filtered response is also shown in the figure, with a first-order LP-filter given by

$$H_{\mathrm{LP}}(z) = \frac{1}{1 - \xi z^{-1}}. \tag{5.21}$$

In order to ensure stability, $\xi < 1$. If noise is present in the system, a low-pass

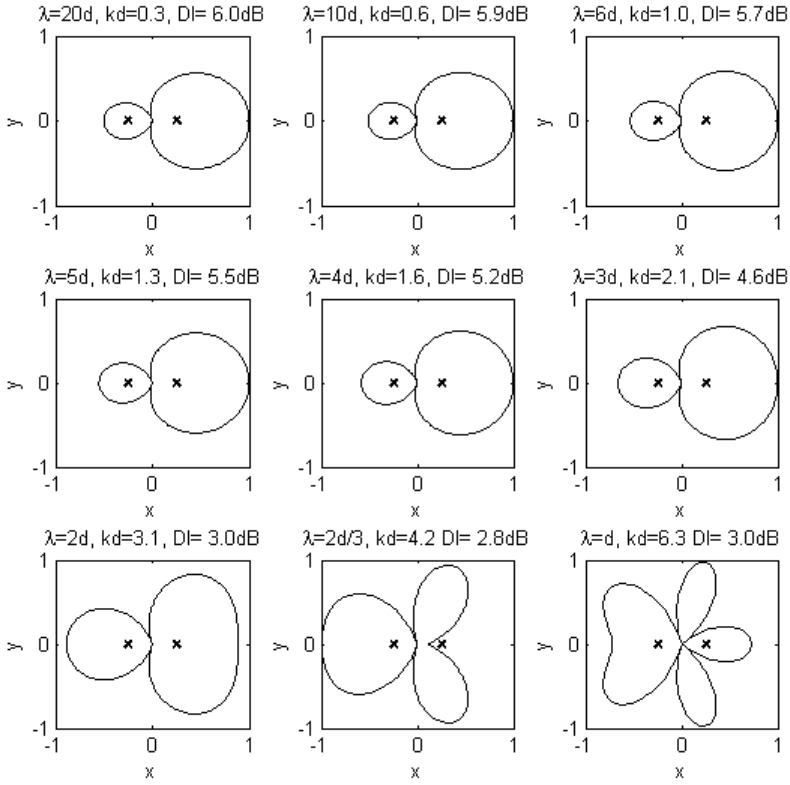

Figure 5.10: Normalized directivity patterns obtained for a delay-sum beamformer for different wavelengths. The crosses indicate that the microphone array is located on the x-axis. The DI is found as the ratio between the direction with maximum gain and all other directions. As kd increases, the pattern begins to change shape and the DI decreases.

filter would amplify the low-frequency noise too much, therefore ξ cannot be too close to one. In Figure 5.11, $\xi = 0.8$.

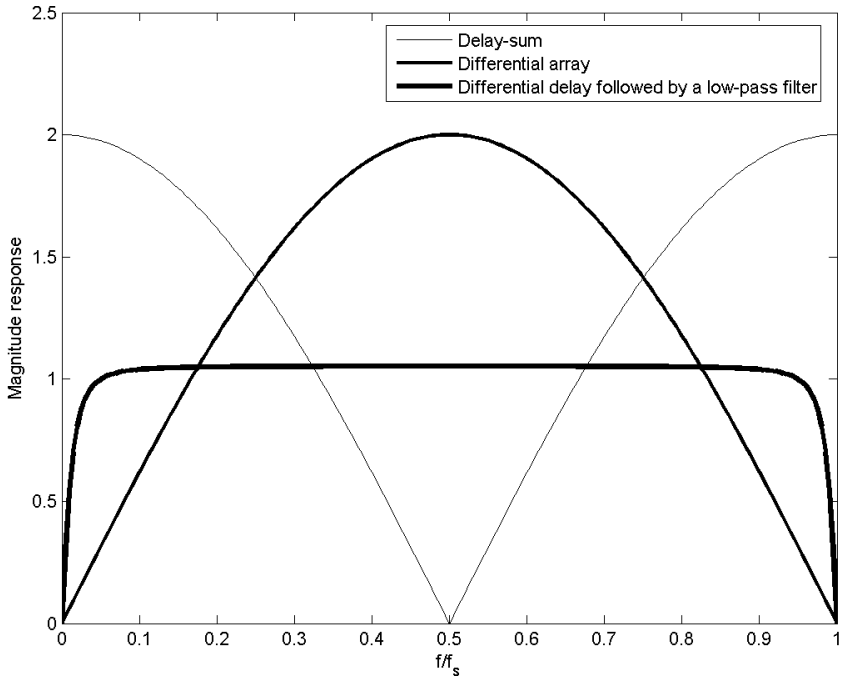

Figure 5.11: Magnitude responses as function of the frequency for a delay-sum array and a differential array. The delay between the two microphones corresponds to one sample. As it can be seen, the differential microphone array has a zero gain for DC signals. To compensate for this high-pass effect and obtaining a flat response over a wider frequency range, the differential beamformer is followed by a low-pass filter. This resulting gain is also shown.

5.3.3 Linear Microphone Arrays

In the previous sections, we considered microphone arrays consisting of two microphones. Some of the results can also be extended to linear arrays consisting of N microphones. Figure 5.12 shows linear arrays containing N omnidirectional microphones. Again, directivity can be obtained by either summing or finding the difference between all the microphone signals. As it could be seen, the delay-sum beamformer and the differential beamformer require different array dimensions in order to provide the desired responses. This is illustrated in Figure 5.12 too. If the delay-sum beamformer has to work efficiently, the

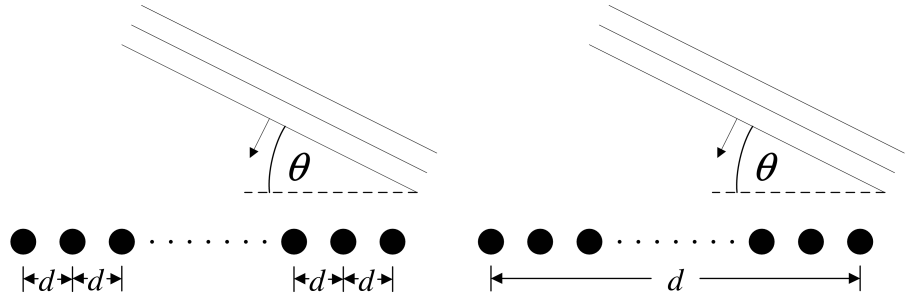

Figure 5.12: A linear microphone array. The microphone array consists of N equally spaced microphones. For a delay-sum beamformer, the distance between two adjacent microphones is d, where $d = \lambda/2$. For a superdirectional microphone array, the size of the whole array is d, where $d \ll \lambda/2$.

distance between microphones next to each other in the array should be $d = 2\lambda$. Contrary, for the differential beamformer, the size of the whole array should be d, where $\lambda \gg d$. Thus, the size of a superdirectional microphone array has to be much smaller than the size of a delay-sum array. It can be shown that for an N-microphone delay-sum beamformer, the maximum directivity for the frequency corresponding to $\lambda = d/2$ can be written as function of N as [64, 32]

$$\mathrm{DI}_{\mathrm{max,DS}} = 10\log(N). \tag{5.22}$$

Also the maximum DI for a differential beamformer can be found as function of the number of microphones. In [34], it is shown that the maximum directivity of a superdirective beamformer is given by

$$\begin{aligned} \mathrm{DI}_{\mathrm{max,SD}} &= 10\log(N^2) && (5.23)\\ &= 20\log(N). && (5.24) \end{aligned}$$

Even though there are some advantages of the superdirectional beamformer, there are also some disadvantages that limits the feasibility. Because the array size of the superdirective beamformer should be much smaller than the wavelength, there are some physical limits. An acoustic high-fidelity signal is in the frequency range of about 30–16000 Hz [62]. At a frequency of 16 kHz, $\lambda/2 \approx 11$ mm. Thus, the array size should be much smaller than 11 mm.

Another problem with differential beamformers is mismatch between microphones. Most likely each microphone in the array has different frequency dependent amplitude and phase response, and the microphone placement may also be uncertain. Microphone mismatch deteriorates the directional gain.

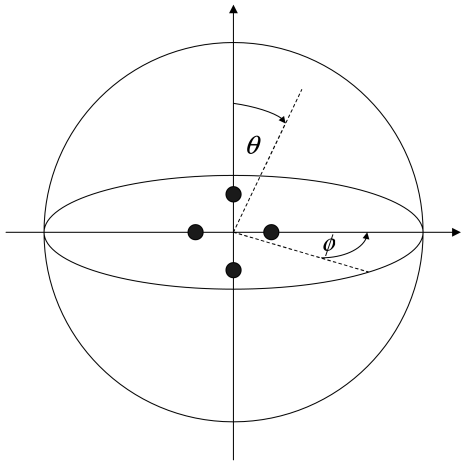

Figure 5.13: A circular four-microphone array placed in a coordinate system at the locations (1,0,0), (-1,0,0), (0,0,1) and (0,0,-1)

The superdirectional microphone array is also sensitive to microphone noise. As it was shown in Figure 5.11, the low frequencies of a superdirectional beamformer has to be amplified. Hereby also the possible microphone noise is amplified. The sensitivity of the microphone array increases as function of the array order $(N-1)$. Further, as kd becomes smaller, the proportionally with $1/kd$ [34]. Therefore, currently superdirectional arrays that consists of more than about 3–4 microphones are only of theoretical interest.

5.3.4 Circular Four-microphone Array

Consider the microphone array in Figure 5.13. The four microphones placed in the $x-z$-plane at the locations (1,0,0), (-1,0,0), (0,0,1), and (0,0,-1) have the sensitivities s_{100}, s_{-100}, s_{001} and s_{00-1}, respectively. The response $r(\theta,\phi)$ is given by

$$r(\theta,\phi) = s_{100}e^{j\tau_x} + s_{-100}e^{-j\tau_x} + s_{001}e^{j\tau_z} + s_{00-1}e^{-j\tau_z}, \tag{5.25}$$

where $\tau_x = \frac{kd}{2}\sin(\theta)\cos(\phi)$ and $\tau_z = \frac{kd}{2}\cos(\theta)$. Thus the response can be rewritten as

$$\begin{aligned} r(\theta,\phi) &= s_{100}e^{j\frac{kd}{2}\sin(\theta)\cos(\phi)} + s_{-100}e^{-j\frac{kd}{2}\sin(\theta)\cos(\phi)} \\ &+ s_{001}e^{j\frac{kd}{2}\cos(\theta)} + s_{00-1}e^{-j\frac{kd}{2}\cos(\theta)}. \end{aligned} \tag{5.26}$$

Again, we assume that $kd \ll \pi$. Thus the exponential can be rewritten using the second order Taylor expansion, i.e., $e^x \approx 1 + x + \frac{x^2}{2}$. Hereby

$$\begin{aligned}
r(\theta,\phi) &= s_{100}(1 + j\frac{kd}{2}\sin(\theta)\cos(\phi) - \frac{(kd)^2}{8}\sin^2(\theta)\cos^2(\phi)) \\
&\quad + s_{-100}(1 - j\frac{kd}{2}\sin(\theta)\cos(\phi) - \frac{(kd)^2}{8}\sin^2(\theta)\cos^2(\phi)) \\
&\quad + s_{001}(1 + j\frac{kd}{2}\cos(\theta) - \frac{(kd)^2}{8}\cos^2(\theta)) \\
&\quad + s_{00-1}(1 - j\frac{kd}{2}\cos(\theta) - \frac{(kd)^2}{8}\cos^2(\theta)) \\
&= (s_{100} + s_{-100} + s_{001} + s_{00-1}) \\
&\quad + j\frac{kd}{2}\sin(\theta)\cos(\phi)(s_{100} - s_{-100}) \\
&\quad - \frac{(kd)^2}{8}\sin^2(\theta)\cos^2(\phi))(s_{100} + s_{-100}) \\
&\quad + j\frac{kd}{2}\cos(\theta)(s_{001} - s_{00-1}) \\
&\quad - \frac{(kd)^2}{8}\cos^2(\theta)(s_{001} + s_{00-1}) \\
&= A + B\sin(\theta)\cos(\phi) + C\sin^2(\theta)\cos^2(\phi) \\
&\quad + D\cos(\theta) + E\cos^2(\theta),
\end{aligned} \tag{5.27}$$

where $A = (s_{100}+s_{-100}+s_{001}+s_{00-1})$, $B = j\frac{kd}{2}(s_{100}-s_{-100})$, $C = -\frac{(kd)^2}{8}(s_{100}+s_{-100})$, $D = j\frac{kd}{2}(s_{001} - s_{00-1})$, and $E = -\frac{(kd)^2}{8}(s_{001} + s_{00-1})$. This expression can be used to find the directivity index. With the desired direction given by $(\theta, \phi) = (\pi, 0)$, we insert (5.27) into (5.3):

$$\mathrm{DI} = 10\log\Big(\frac{4\pi[A + B + C]^2}{\int_0^{2\pi}\int_0^{\pi}[A + B\sin(\theta)\cos(\phi) + C\sin^2(\theta)\cos^2(\phi) + D\cos(\theta) + E\cos^2(\theta)]^2\sin(\theta)d\theta d\phi}\Big),$$

which reduces to

$$\mathrm{DI} = 10\log\left(\frac{[A + B + C]^2}{A^2 + \frac{1}{3}B^2 + \frac{1}{5}C^2 + \frac{2}{3}A(C + E) + \frac{1}{5}D^2 + \frac{1}{5}E^2 + \frac{2}{15}CE}\right). \tag{5.28}$$

Notice, $A = (C + E)/\frac{-(kd)^2}{8}$. Therefore (5.28) can be rewritten as

$$\mathrm{DI} = 10\log\Big(\frac{[(C + E)/\frac{-(kd)^2}{8} + B + C]^2}{[(C + E)/\frac{-(kd)^2}{8}]^2 + \frac{1}{3}B^2 + \frac{1}{5}C^2 + \frac{2}{3}((C + E)/\frac{-(kd)^2}{8})(C + E) + \frac{1}{5}D^2 + \frac{1}{5}E^2 + \frac{2}{15}CE}\Big).$$

Hereby, it can be seen that the DI is dependent on the wave number k and hereby is dependent on the frequency f. For different frequencies, the DI is

Table 5.2: The DI is maximized with respect to different frequencies, with $d = 10$ mm

f [Hz]	B	C	D	E	DI [dB]
0	-0.8000	-1.0000	0	1.0000	8.2930
500	-0.8001	-0.9999	0	1.0005	8.8926
1000	-0.8005	-0.9995	0	1.0018	8.8914
2000	-0.8027	-0.9988	0	1.0080	8.8866
10000	-0.8025	-0.8792	0	1.1271	8.6759

maximized with respect to B, C, D and E. The results[1] are given in Table 5.2. The solution for $f = 0$ is independent of the frequency because $A = 0$, and the DI will be constant for all frequencies. Notice, all these directivity indices are smaller than the similar maximum DI of 9.5 dB which can be obtained with a linear three-microphone array [90]. 3D-plots of the directivity are shown in Figure 5.14 and Figure 5.15. The four microphone summing coefficients are given as

$$s_{100} = -j\frac{B}{kd} - \frac{4C}{(kd)^2} \tag{5.29}$$

$$s_{-100} = j\frac{B}{kd} - \frac{4C}{(kd)^2} \tag{5.30}$$

$$s_{001} = -j\frac{D}{kd} - \frac{4E}{(kd)^2} \tag{5.31}$$

$$s_{00-1} = j\frac{D}{kd} - \frac{4E}{(kd)^2} \tag{5.32}$$

$$\tag{5.33}$$

Notice, by choosing another direction of the source signal than $(\theta, \phi) = (\pi, 0)$ in (5.3), another maximum value of the DI could be found.

5.4 Considerations on the Average Delay between the Microphones

Consider two microphones placed in a free field as illustrated in Figure 5.16. We denote the delay between the microphones by τ_z. The small index z indicates that the microphones are placed along the z-axis.

[1] The values of B, C, D, and E have been found by use of the Matlab function *fminsearch*.

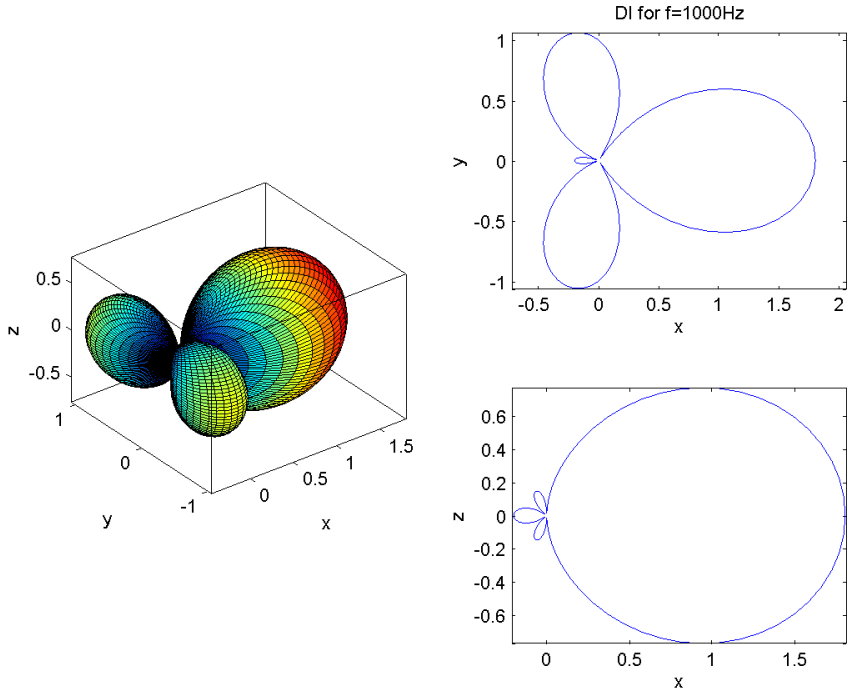

Figure 5.14: 3D directivity plot. The directivity is found for $B = -0.8$, $C = -1$, $D = 0$ and $E = 0$. The frequency used in the calculations is $f = 1000$ Hz and $d = 10$ mm. The DI is also shown for the $z = 0$ and $y = 0$. As it can be seen the directivity is not rotation symmetric as in the case where the microphone array is linear.

5.4.1 Average delay in Spherically Diffuse Noise Field

We can find the probability distribution for τ_z given all arrival directions are equally likely. In a spherically diffuse noise field, all directions are equally likely. If all directions are equally likely, the spherical coordinates are random variables Θ, Φ. Φ has a uniform distribution with the following probability density function (pdf):

$$f_\Phi(\phi) = \begin{cases} \frac{1}{2\pi}, & 0 < \theta \leq 2\pi; \\ 0, & \text{otherwise.} \end{cases}, \tag{5.34}$$

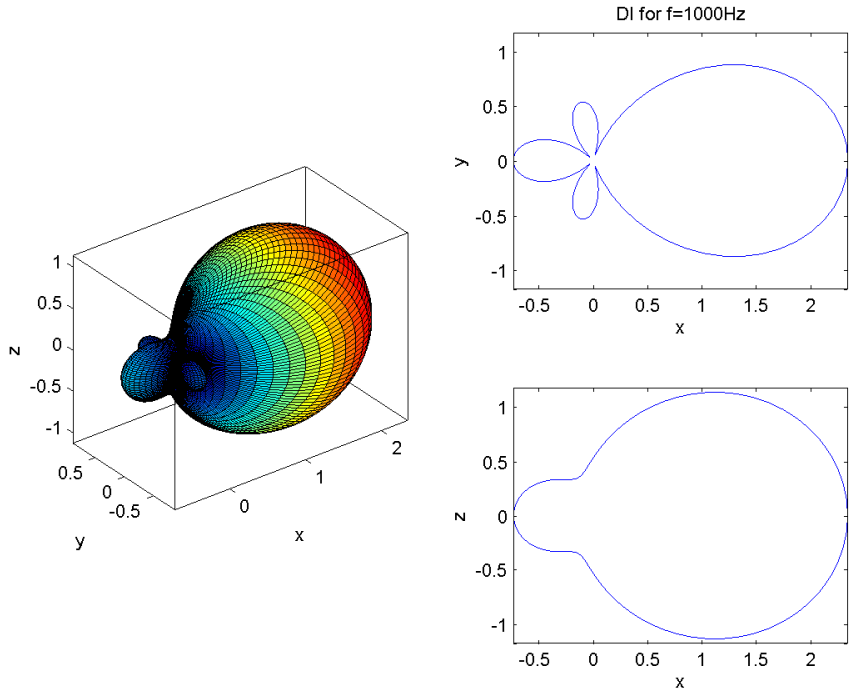

Figure 5.15: 3D directivity plot with the values in Table 5.2 optimized for $f = 1000$ Hz and $d = 10$ mm.

If Φ (the longitude) has a uniform distribution, Θ (the latitude) does not have a uniform distribution. This can be seen by e.g. considering a globe. On a globe, both the latitude and the longitude angles are uniformly distributed. At the poles of a globe, the areas formed between the latitude and the longitude lines are smaller than at the areas at the equatorial region. Uniformly distributed longitude and latitude thus results in a non-uniform distribution of points on the sphere (with a relatively higher probability of being near the poles). In order to determine the distribution for θ, consider the unit sphere in Figure 5.17. The area between the two circles $d\Omega$ is given by

$$d\Omega = 2\pi r d\theta, \tag{5.35}$$

where $r = \sin(\theta)$. Hereby

$$\begin{aligned} d\Omega &= 2\pi \sin(\theta) d\theta & (5.36)\\ &= -2\pi d(\cos(\theta)). & (5.37) \end{aligned}$$

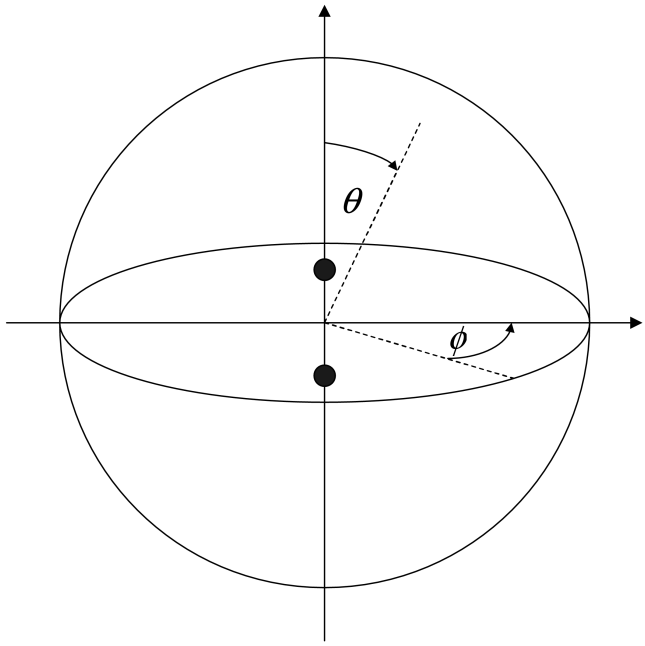

Figure 5.16: Two-microphone array placed in a free-field.

The probability distribution of $d\Omega$, $P(d\Omega)$ is found by dividing by the whole area of the unit sphere, i.e. 4π. Hereby

$$P(d\Omega) = \frac{d\Omega}{4\pi} = -\frac{d(\cos(\theta))}{2}. \tag{5.38}$$

If each $d\Omega$ has the same size and is equally likely, $P(d\Omega)$ follows an uniform distribution. Hereby, $\cos(\theta)$ follows an uniform distribution too. Therefore, Θ is a function of a random variable Ψ:

$$\Theta = \arccos(\Psi), \tag{5.39}$$

where Ψ is uniformly distributed with

$$f_\Psi(\psi) \quad = \quad \begin{cases} \frac{1}{2}, & -1 < \psi < 1; \\ 0, & \text{otherwise.} \end{cases} \tag{5.40}$$

To find the pdf of Θ, the following equation is used [56, p. 125]

$$f_\Theta(\theta) = \sum_k \frac{f_\Psi(\psi)}{|d\theta/d\psi|}\bigg|_{\psi=\psi_k} \tag{5.41}$$

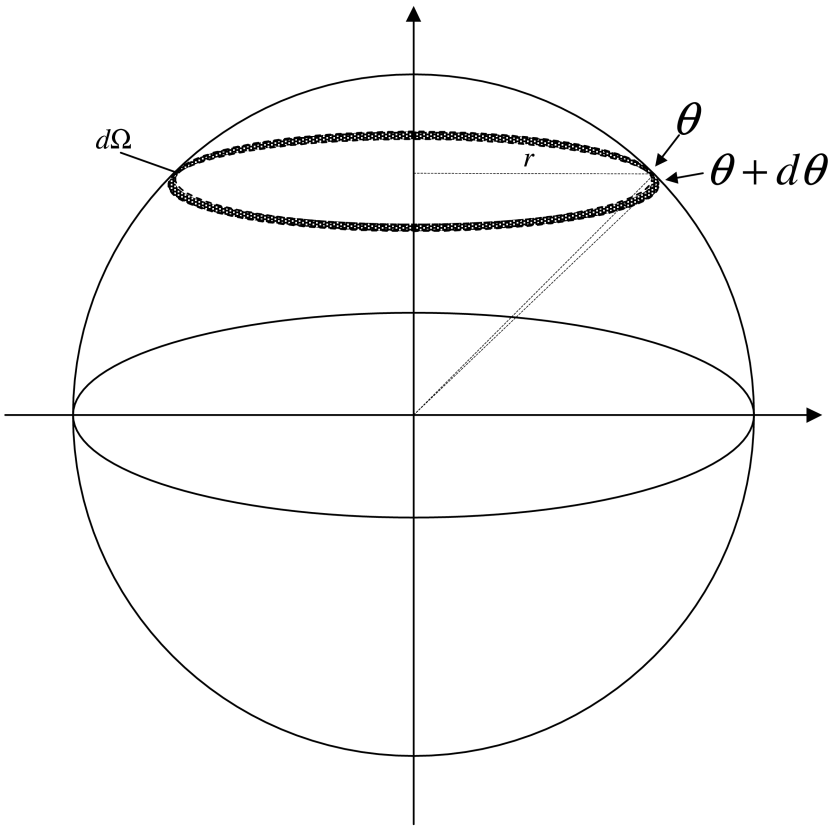

Figure 5.17: The area between the two broken circles is given by $d\Omega = 2\pi r d\theta$.

Here, $d\theta/d\psi$ is

$$\frac{d(\arccos(\psi))}{d\psi} = \frac{-1}{\sqrt{1-\psi^2}} \tag{5.42}$$

Inserting (5.42) into (5.41) with $\psi = \cos(\theta)$ yields

$$f_\Theta(\theta) = \begin{cases} \frac{\sqrt{1-\cos^2(\theta)}}{2} = \frac{\sin(\theta)}{2}, & 0 \le \theta \le \pi; \\ 0, & \text{otherwise}, \end{cases} \tag{5.43}$$

The pdf's for Θ and Φ are shown in Figure 5.18.

The delay τ_z is described by the random variable T_z, where

$$T_z \quad = \quad \cos(\Theta). \tag{5.44}$$

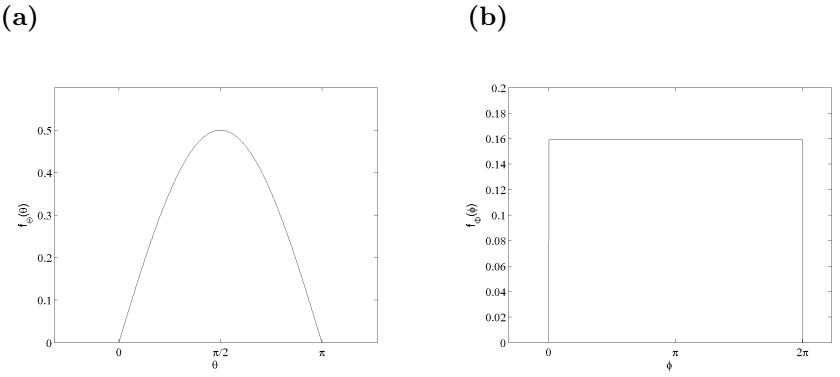

Figure 5.18: When all arrival directions for a source signal are equally likely, the distributions for the spherical coordinates θ and ϕ are $f_\Theta(\theta) = \frac{\sin(\theta)}{2}$ and $f_\Phi(\phi) = \frac{1}{2\pi}$, respectively.

The pdf for T_z is found by inserting (5.39) into (5.44)

$$T_z = \cos(\arccos(\Psi)) \tag{5.45}$$
$$= \Psi. \tag{5.46}$$

Hereby we observe that average delay between the microphones T_z is uniformly distributed with

$$f_{T_z}(\tau_z) = \begin{cases} \frac{1}{2}, & -1 < \tau_z < 1; \\ 0, & \text{otherwise.} \end{cases} \tag{5.47}$$

5.4.2 Average delay in Cylindrical Diffuse Noise Field

If, instead the sound was impinging from directions equally distributed on a circle in a plane, the probability function of the delay would different. Now the delay is described by the random variable T, where

$$T = \cos(\Psi), \tag{5.48}$$

with Ψ uniformly distributed as in equation (5.34). The pdf for T is given by [56, p. 126]

$$f_T(\tau) = \begin{cases} \frac{1}{\pi\sqrt{1-\tau^2}}, & -1 < \tau < 1; \\ 0, & \text{otherwise,} \end{cases} \tag{5.49}$$

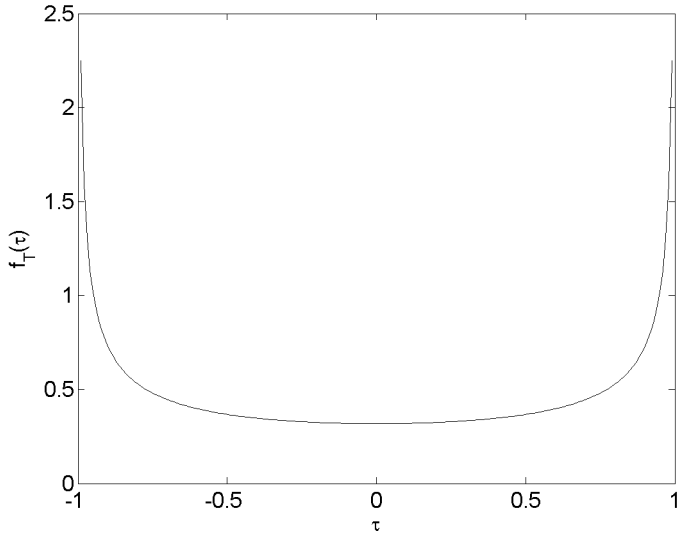

Figure 5.19: Probability density function for the delay between two microphones in a cylindrical diffuse noise field.

The pdf is shown in Figure 5.19

To conclude, if a sound impinges an array, and all arrival directions are equally likely, all possible delays between the two microphones are equally likely too. We also get some side results. In Figure 5.16, the microphone was placed symmetrically on the z-axis. We could as well have chosen to place the microphone array along the x or the y axis. If the array is placed on the x-axis , the delay would have been given as

$$\begin{aligned} T_x &= \sin(\Theta)\cos(\Phi) && (5.50)\\ &= \sin(\arccos(\Psi))\cos(\Phi) && (5.51)\\ &= \sqrt{1-\Psi^2}\cos(\Phi) && (5.52) \end{aligned}$$

Due to symmetry, we know that $\sqrt{1-\Psi^2}\cos(\Phi)$ simply reduces to a uniform distribution with the same distribution as Ψ. Likewise, if the microphones were placed along the y-axis, the delay would be

$$\begin{aligned} T_y &= \sin(\Theta)\sin(\Phi) && (5.53)\\ &= \sqrt{1-\Psi^2}\sin(\Phi) && (5.54)\\ &= \Psi', && (5.55) \end{aligned}$$

where Ψ' is uniformly distributed like Ψ.

We also see that the delay between the two microphone signals is not uniformly distributed if the noise field is cylindrically distributed.

5.4.3 Average Delay in Spherically Diffuse Noise Field for a Circular Four-microphone Array.

The probability densities for the average delays can also be found for the delays between the microphones in the circular array. We consider the two delays τ_x and τ_z. τ_x is the delay between the two microphones placed on the x-axis and τ_z is the delay between the two microphones placed on the z-axis in Figure 5.13, respectively.

Again, we assume a source signal arrives at the microphone array from a random direction. Given the spherical coordinates (θ, ϕ), the two (normalized) delays τ_x and τ_z are given by

$$\tau_x = \sin(\theta)\cos(\phi) \tag{5.56}$$
$$\tau_z = \cos(\theta). \tag{5.57}$$

From Section 5.4.1, we know that the delay between two microphones in a spherically diffuse noise field are given by uniform distributions. Thus the marginal distributions are given by

$$f_{T_x}(\tau_x) = \begin{cases} \frac{1}{2}, & -1 < \tau_x < 1; \\ 0, & \text{otherwise}, \end{cases} \tag{5.58}$$

and

$$f_{T_z}(\tau_z) = \begin{cases} \frac{1}{2}, & -1 < \tau_z < 1; \\ 0, & \text{otherwise}. \end{cases} \tag{5.47}$$

The two marginal probability density functions $f_{T_x}(\tau_x)$ and $f_{T_z}(\tau_z)$ are not independent of each other. If one of the two delays are given, the conditional distributions can be found, i.e. $f_{T_x}(\tau_x|\tau_z)$ and $f_{T_z}(\tau_z|\tau_x)$. In order to find $f_{T_x}(\tau_x|\tau_z)$, $\theta = \arccos(\tau_z)$ is inserted into (5.56). Hereby

$$\tau_x = \cos(\phi)\sin(\arccos(\tau_z)) \tag{5.59}$$
$$= \cos(\phi)\sqrt{1-\tau_z^2} \tag{5.60}$$
$$= \cos(\phi)a. \tag{5.61}$$

Here, it can be seen that the conditional distribution is given by a constant a multiplied by $\cos(\Phi)$. The pdf for the random variable, which we denote

$K = \cos(\Phi)$, have the density function obtained from equation (5.49), i.e.

$$f_K(k) = \begin{cases} \frac{1}{\pi\sqrt{1-k^2}}, & -1 < k < 1; \\ 0, & \text{otherwise.} \end{cases} \tag{5.62}$$

The conditional pdf for $T_x = aK$ is given by

$$f_{T_x}(\tau_x|\tau_z) = \frac{1}{a} f_K\Big(\frac{\tau_x}{a}\Big), a > 0 \tag{5.63}$$

$$= \frac{1}{\sqrt{1-\tau_z^2}} \frac{1}{\pi\sqrt{1-(\frac{\tau_x}{\sqrt{1-\tau_z^2}})^2}} \tag{5.64}$$

$$= \begin{cases} \frac{1}{\pi\sqrt{1-\tau_x^2-\tau_z^2}}, & -\sqrt{1-\tau_z^2} < \tau_x < \sqrt{1-\tau_z^2}; \\ 0, & \text{otherwise.} \end{cases} \tag{5.65}$$

Here, the constraint $\tau_x^2 + \tau_z^2 \leq 1$ has been applied. Also, the joint distribution can be found by

$$f_{T_x,T_z}(\tau_x, \tau_z) = f_{T_x}(\tau_x|\tau_z) f_{T_z}(\tau_z) \tag{5.66}$$

$$= \frac{1}{\pi\sqrt{1-\tau_x^2-\tau_z^2}} \frac{1}{2} \tag{5.67}$$

$$= \begin{cases} \frac{1}{2\pi\sqrt{1-\tau_x^2-\tau_z^2}}, & \tau_x^2 + \tau_z^2 \leq 1; \\ 0, & \text{otherwise.} \end{cases} \tag{5.68}$$

Finally, $f_{T_z}(\tau_z|\tau_x)$ is found by

$$f_{T_z}(\tau_z|\tau_x) = \frac{f_{T_x,T_z}(\tau_x, \tau_z)}{f_{T_x}(\tau_x)} \tag{5.69}$$

$$= \begin{cases} \frac{1}{\pi\sqrt{1-\tau_x^2-\tau_z^2}}, & -\sqrt{1-\tau_x^2} < \tau_z < \sqrt{1-\tau_x^2}; \\ 0, & \text{otherwise.} \end{cases} \tag{5.70}$$

The joint distribution is shown in Figure 5.20, and the two conditional distributions are shown in Figure 5.21.

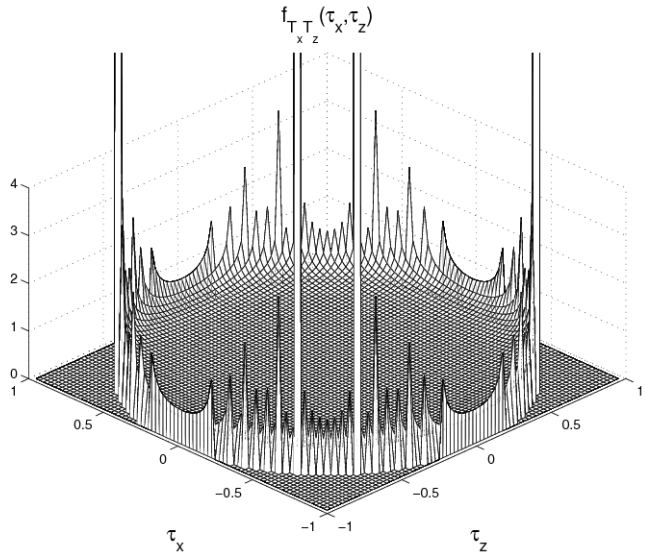

Figure 5.20: The joint distribution is given by $f_{T_x,T_z}(\tau_x, \tau_z) = \frac{1}{2\pi\sqrt{1-\tau_x^2-\tau_z^2}}$.

$f_{T_x}(\tau_x|\tau_z)$ $f_{T_z}(\tau_z|\tau_x)$

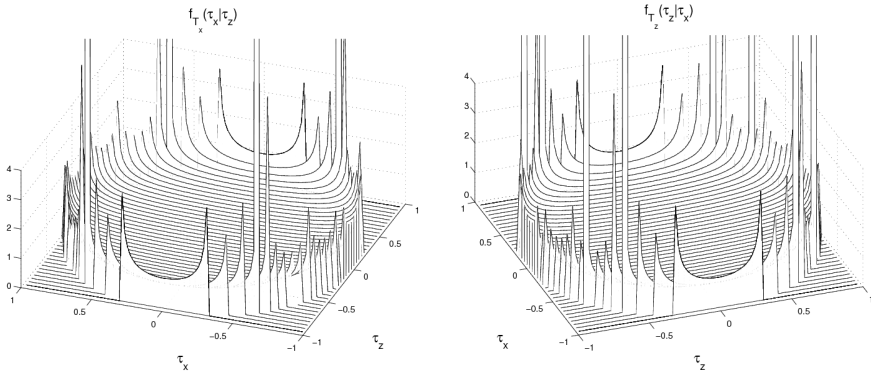

Figure 5.21: The two conditional distributions, $f_{T_x}(\tau_x|\tau_z)$ and $f_{T_z}(\tau_z|\tau_x)$.

Chapter 6

Source Separation

In this chapter, the source separation methods considered in this thesis are summarized. The more detailed descriptions of the proposed algorithms are provided in the appendices. Also a theoretical result concerning ICA is summarized in this chapter.

An important contribution in this thesis is provided in Appendix G. Here, we provide an exhaustive survey on blind separation of convolutive mixtures. In this survey, we provide a taxonomy wherein most of the proposed convolutive blind source separation methods can be classified. We cite most of the published work about separation of convolutive mixtures. The survey is a pre-print of a version which is going to be published as a book chapter.

The objective in this thesis was to develop source separation algorithms for microphone arrays small enough to fit in a hearing aid. Two source separation algorithms are presented in this thesis. Both methods take advantage of differential microphone signals, where directional microphone gains are obtained from closely spaced microphones.

Gradient flow beamforming is a method, where separation of delayed sources recorded at a small microphone array can be obtained by instantaneous ICA. The used microphone array is a circular microphone array consisting of four microphones similar to the one shown in Figure 5.13. Such an array has a size

that could be applicable for a hearing aid device. For real-world applications, we have to take convolutive sources into account. The principles of gradient flow beamforming proposed by Cauwenberghs [27] is described in Appendix A. Here, we also propose how we can extend the gradient flow framework in order to cope with separation of convolutive mixtures. The proposed extension is verified by experiments on simulated convolutive mixtures.

The papers in Appendix C–F all concern the same topic and this work is another important contribution in this thesis. As mentioned, a problem in many source separation algorithms is that the number of sources has to be known in advance and the number of sources cannot exceed the number of mixtures. In order to cope with these problems, we propose a method based on independent component analysis and time-frequency masking in order to iteratively segregate an unknown number of sources with only two microphones available. First, we consider what happens when instantaneous ICA is applied to mixtures, where the number of sources in the mixtures exceed the number of sensors. We find that the mixtures are separated into different components, where each component is as independent as possible from the other components. In the T-F domain the two ICA outputs are compared in order to estimate binary T-F masks. These masks are then applied to the original two microphone signals, and hereby some of the signals can be removed by the T-F mask. The T-F mask is applied to each of the original signals, and therefore the signals are maintained as stereo signals. ICA can then be applied again to the binary masked stereo signals, and the procedure is continued iteratively, until all but one signal is removed from the mixture. This iterative procedure is illustrated in Figure 6.1 and in Figure 6.2. Experiments on simulated instantaneous mixtures show that our method is able to segregate mixtures that consist of up to seven simultaneous speech signals. Another problem, where this method can be applied is to separation of stereo music into the individual instruments and vocals. When each individual source and vocalist are available, tasks such as music transcription, identification of instruments or identification of the vocalist become easier. In Appendix D, we apply the method for separation of stereo music. Here, we demonstrate that instruments that are located at spatially different positions can be segregated from each other.

As mentioned, for real world signals, it is important that the method can take reverberations into account. Motivated by that, we propose to change the instantaneous ICA method in our iterative method by a convolutive ICA algorithm. This is described in the paper in Appendix E. Furthermore, in the paper in Appendix F we show that the method, with some extensions, is able to segregate convolutive mixtures, even though an instantaneous ICA algorithm is used. This is an advantage because instantaneous ICA is computationally less expensive than convolutive ICA. In this paper we also provide a more thorough evaluation of our proposed method. We demonstrate that the method is ca-

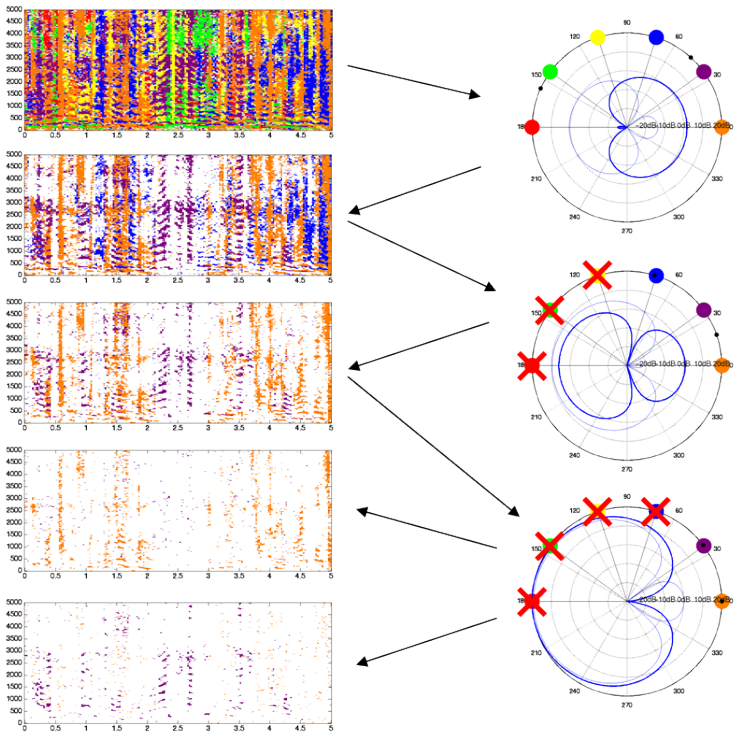

Figure 6.1: The principles of the proposed method. The T-F plot in the upper left corner shows the T-F distribution of a mixture consisting of six sources. Each color denote time-frequency areas, where one of the six sources have more energy then the other five sources. ICA is applied to the two mixtures. The two outputs of the ICA method can be regarded as directional gains, where each of the two ICA outputs amplify certain sources and attenuate other sources. This is shown in the directional plot in the right side of the figure. The six colored dots illustrate the spatial location of each of the six sources. By comparing the two ICA outputs, which corresponds to comparing the directional patterns, a binary mask can be created that remove the signals from directions, where one directional pattern is greater than the other directional pattern. The binary mask is applied to the two original signals. The white T-F areas show the areas which have been removed from the mixture. As it can be seen, some of colors in the T-F distribution are removed, while other colors remain in the T-F distribution. ICA is then applied again, and from the new directional patterns, yet another signal from the original mixture is removed. Finally, all but one signal is extracted from the mixture by the binary mask.

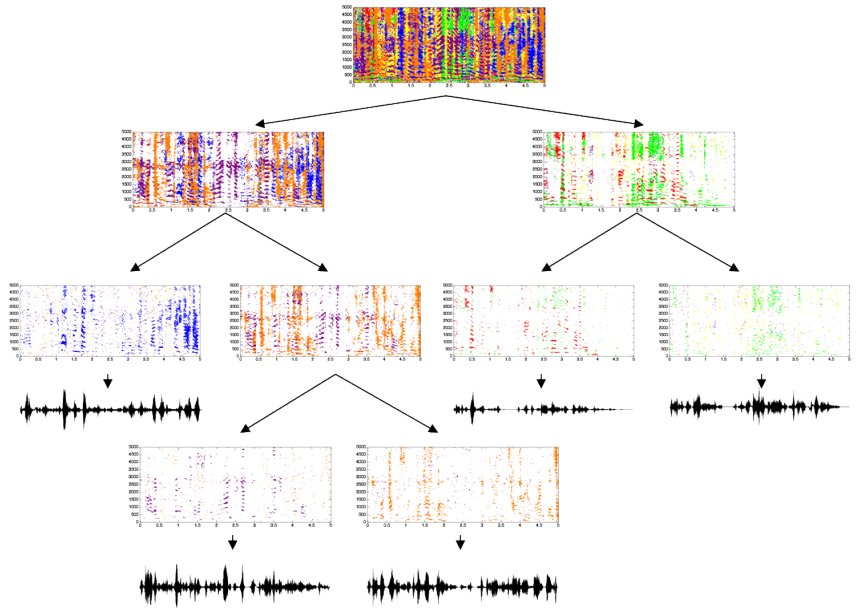

Figure 6.2: The different colors in the T-F distributions indicate areas in time and frequency, where one source has more energy than all the other sources (ideal binary masks) together. We see that each source is dominating in certain T-F areas. For each iteration, binary masks are found that removes some of the speakers from the mixture. The white areas are the areas which have been removed. This iterative procedure is continued until only one speaker (color) is dominant in the remaining mixture.

pable of segregating four speakers convolved with real recorded room impulse responses obtained from a room with a reverberation time of $T_{60} = 400$ ms.

Appendix B contains a theoretical paper concerning ICA. In ICA, either the mixing matrix $\mathbf{A}$ or the separation matrix $\mathbf{W}$ can be found. Here, we argue that it is easier to apply a gradient descent search in order to minimize the cost function, when the cost function is expressed as function of the separation matrix compared to when the cost function is expressed as function of mixing matrix. Examples on typical cost functions are shown in Figure 6.3. This is because the points where the mixing matrix is singular are mapped into infinity when the mixing matrix is inverted. The cost function have an infinite value in the singular points. In the separation domain, these points are far away from the area, where the cost function is minimized. Therefore, the gradient search converges faster in the separation domain than in the mixing domain. This is

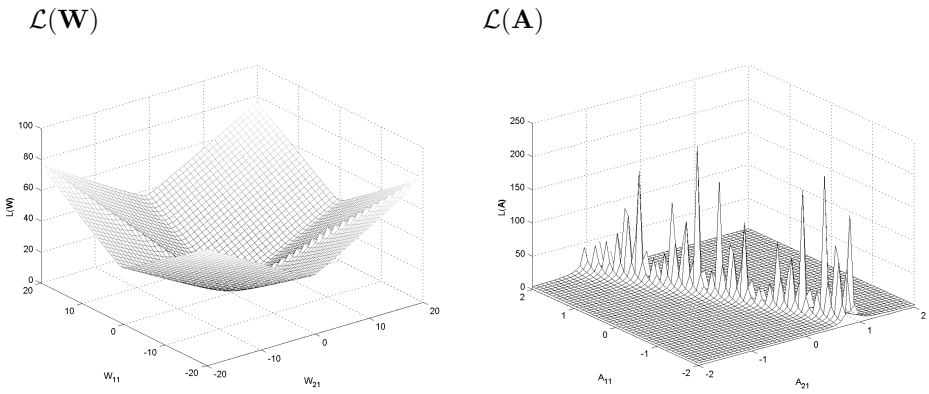

Figure 6.3: The negative log likelihood cost functions given as function of the parameters in the separation space $\mathcal{L}(\mathbf{W})$ and in the mixing matrix space $\mathcal{L}(\mathbf{A})$.

validated by experiments. However, if instead the natural gradient is used, there is no difference between the convergence rate, when the gradient search in the mixing matrix domain and the separation matrix domain are compared. These results are illustrated in Figure 6.4.

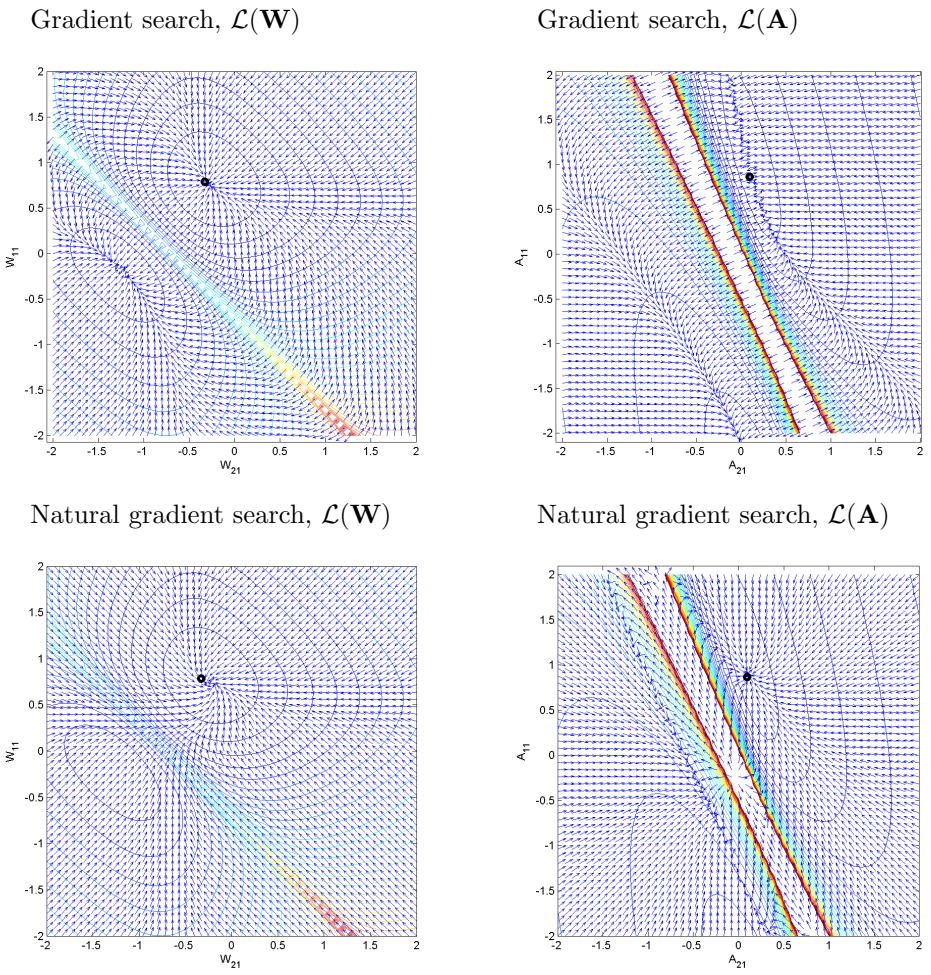

Figure 6.4: The gradient descent directions shown on the two negative log likelihood cost functions from Figure 6.3. The circle shows the desired minima. Both gradient directions and the natural gradient directions are shown.

Chapter 7

Conclusion

In this thesis, the focus has been on aspects within enhancement of audio signals. Especially applications concerning enhancement of audio signals as a front-end for hearing aids have been considered. In particular, acoustic signals recorded by small microphone arrays have been considered. Issues concerning beamforming and small microphone arrays were described in Chapter 5. In particular, two-microphone arrays and a circular four-microphone array were considered. In that context, we also proposed an extension of gradient flow beamforming in order to cope with convolutive mixtures. This was presented in Appendix A.

Two different topics within enhancement of acoustic signals have been considered in this thesis, i.e. blind source separation and time-frequency masking.

One of the main objectives in this work was to provide a survey on the work done within the topic of blind separation of convolutive mixtures. The objective of this survey was to provide a taxonomy wherein the different source separation methods can be classified, and we have classified most of the proposed methods within blind separation of convolutive mixtures. However, in this survey we do not evaluate and compare the different methods. The reader can use the survey in order to achieve an overview of convolutive blind source separation, or the reader can use the survey to obtain knowledge of the work done within a more specific area of blind source separation. Furthermore, also other source separation have been reviewed, i.e. CASA and beamforming techniques.

To develop source separation methods for small microphone arrays is the other main topic of this thesis. Especially source separation techniques by the combination of blind source separation and time-frequency masking. Time-frequency masking methods were reviewed in Chapter 4. T-F masking is a speech enhancement method, where different gains are applied to different areas in time and in frequency.

We have presented a novel method for blind separation of underdetermined mixtures. Here, traditional methods for blind source separation by independent component analysis were combined with the binary time-frequency masking techniques. The advantage of T-F masking is that it can be applied to a single microphone recording, where other methods such as ICA (based on a linear separation model) and beamforming require multiple microphone recordings. We therefore apply the blind source separation techniques on mixtures recorded at two microphones in order to iteratively estimate the binary mask, but we apply the binary T-F masking technique to do the actual segregation of the signals. Our method was evaluated, and it successfully separated instantaneous speech mixtures consisting of up to seven simultaneous speakers, with only two sensor signals available. The method was also evaluated on convolutive mixtures from mixtures mixed with real room impulse responses. We were able to segregate sources from mixtures consisting of four sources under reverberant conditions. Furthermore, we have also shown that the proposed method is applicable for segregation of single instruments or vocal sounds from stereo music. The proposed algorithm has several advantages. The method does not require that the number of sources is known in advance, and the method can segregate the sources from the mixture even though the number of sources exceeds the number of microphones. Furthermore, the segregated sources are maintained as stereo signals.

In this thesis a theoretical result regarding ICA was presented too. This result is not related to the other work done. We show why gradient descent update is faster, when the log likelihood cost function is given as function of the inverse of the mixing matrix compared to when it is given as function of the mixing matrix. When the natural gradient was applied the difference between the two parameterizations disappeared and fast convergence was obtained in both cases.

Outlook and Future Work

A question that rises is whether the proposed methods are applicable for hearing aids. As mentioned in the introduction, the processing delay through a hearing aid should be kept as small as possible. In this thesis, it was chosen to disregard

these delay constraints, because the source separation problem still is not completely solved, even without these hard constraints. In this thesis, we therefore used frequency resolutions much higher than what can be obtained in a hearing aid. Future work could be to constrain these methods in order to make them more feasible for hearing aids.

In the traditional blind source separation, very long separation filters have to be estimated. The requirement for these very long filters means that these methods are hard to apply in environments, where the mixing filters changes rapidly. Segregation by T-F masking does not require such long filters, and hence this technique may be more applicable for separation of sources in acoustic environments, where a fast adaptation is required.

When T-F masking techniques are applied, the acoustic signal is not perfectly reconstructed. This is not a problem as long as the perceptual quality is high. However, it is likely that artifacts are audible in the signals to which the T-F mask has been applied. The use of non-binary masks as well as possible reconstruction of the missing spectral information may reduce such artifacts. Perceptual information such as auditory masking has been applied to speech enhancement algorithms in order to reduce the musical artifacts [41]. Such information could also be used to constrain the gain in T-F masking methods.

As it was demonstrated in Chapter 4, the actual gain in time and in frequency does not correspond to the gain applied by the time-frequency mask. The influence from the signal analysis and synthesis on the actually applied gain is an area which only have been considered by few (if any) within the T-F masking community.

Publications

The papers that have been produced during the past three years are presented in the following appendices.

Appendix A

Gradient Flow Convolutive Blind Source Separation

GRADIENT FLOW CONVOLUTIVE BLIND SOURCE SEPARATION

Michael Syskind Pedersen*
Informatics and Mathematical Modelling, Building 321
Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
Phone: +45 4525 3904
E-mail: msp@imm.dtu.dk
Web: imm.dtu.dk/~msp

Chlinton Møller Nielsen
Technology & Innovation, Bang & Olufsen
Peter Bangs Vej 15, DK-7600 Struer, Denmark
Phone: +45 9684 4058.
Email: chn@bang-olufsen.dk

Abstract. Experiments have shown that the performance of instantaneous gradient flow beamforming by Cauwenberghs et al. is reduced significantly in reverberant conditions. By expanding the gradient flow principle to convolutive mixtures, separation in a reverberant environment is possible. By use of a circular four-microphone array with a radius of 5 mm, and applying convolutive gradient flow instead of just applying instantaneous gradient flow, experimental results show an improvement of up to around 14 dB can be achieved for simulated impulse responses and up to around 10 dB for a hearing aid application with real impulse responses.

INTRODUCTION

The *gradient flow* blind source separation technique proposed by Cauwenberghs et al. [5] uses a four microphone array to separate 3 sound signals. The gradient flow can be regarded as a preprocessing step in order to enhance the difference between the signals before a blind separation algorithm is applied. The gradient flow technique requires small array sizes. Small array sizes occur in some source separation applications such as hearing aids. Here the physical dimensions of the microphone array may limit the separation performance due to the very small difference between the recorded signals. In the literature, some attempts exist to separate sound signals by use of microphone arrays with a dimension of about 1 cm [2, 6, 7]. These techniques

*This work was supported by the Oticon Foundation

are either based on beamforming, blind source separation [3], or a combination of these techniques. The gradient flow method is able to estimate delayed versions of the source signals, as well as the source arrival angles. As shown in the simulations, the model may fail in reverberant environments, i.e. when each of the source signals is convolved in time. Here, a model is proposed that extends the instantaneous gradient flow model to a convolutive gradient flow model. Simulations show that the convolutive model is able to cope with reverberant situations, in which the instantaneous model fails.

INSTANTANEOUS GRADIENT FLOW MODEL

The gradient flow model is described into details in [5, 8, 10, 11]. Each signal x_{pq} is received by a sensor placed at location (p, q), which is shown in Figure 1. At a point in the coordinate system $\mathbf{r}$, there is a delay, $\tau(\mathbf{r})$, between an incoming wavefront and the origin. The delay with respect to the n'th source signal, s_n is denoted as $\tau^n(\mathbf{r})$. It is assumed that the sources are located in the far-field. Hence the wavefront of the incoming waves is linear. Using that assumption the delay can be described the following way [5]:

$$\tau(\mathbf{r}) \approx \frac{1}{c}\mathbf{r} \cdot \mathbf{u}, \tag{1}$$

where $\mathbf{u}$ is a unit vector pointing in the direction of the source and c is the velocity of the wave.
Now consider a sensor placed at the coordinates (p, q) as in Figure 1. The time delay from the source can be expressed as

$$\tau^n_{pq} = p\tau^n_1 + q\tau^n_2, \tag{2}$$

where $\tau^n_1 = \mathbf{r}_1 \cdot \mathbf{u}_n/c$ and $\tau^n_2 = \mathbf{r}_2 \cdot \mathbf{u}_n/c$. τ^n_1 and τ^n_2 are the time differences in the directions of the two orthogonal vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ as shown in Figure 1. The point $\mathbf{r}_{pq}$ can be described as $\mathbf{r}_{pq} = p\mathbf{r}_1 + q\mathbf{r}_2$.

Description of field

The field is described by the incoming waves. At the center of the coordinate system, the contribution to the field from the n'th source is given by $s_n(t)$. By using the Taylor series expansion, the field from the n'th source at the point $\mathbf{r}$ in the coordinate system is given by $s_n(t + \tau^n(\mathbf{r}))$, where [11]

$$s_n(t + \tau^n(\mathbf{r})) = s_n(t) + \frac{1}{1!}\tau^n(\mathbf{r})\dot{s}_n(t) + \frac{1}{2!}(\tau^n(\mathbf{r}))^2\ddot{s}_n(t) + \dots. \tag{3}$$

Here, $\dot{}$ and $\ddot{}$ denote the 1'st and 2'nd order derivative, respectively. Hence, the received signal at $\mathbf{r}_{pq}$ can be written as

$$x_{pq}(t) = \sum_{n=1}^{N} s_n(t+\tau^n(\mathbf{r})) = \sum_{n=1}^{N} s_n(t) + \frac{1}{1!}\tau^n(\mathbf{r})\dot{s}_n(t) + \frac{1}{2!}(\tau^n(\mathbf{r}))^2\ddot{s}_n(t) + \dots. \tag{4}$$

Additionally, a noise term $\varepsilon_{pq}(t) \propto N(0, \sigma)$ can be added [5]. The received signal can be approximated by using only the first two terms of (4):

$$x_{pq}(t) = \sum_{n=1}^{N} s_n(t + \tau^n(\mathbf{r})) \approx \sum_{n=1}^{N} s_n(t) + \tau^n(\mathbf{r})\dot{s}_n(t). \tag{5}$$

Notice, the Taylor approximation only holds, if the dimension of the array is not too large (see [5] for details).

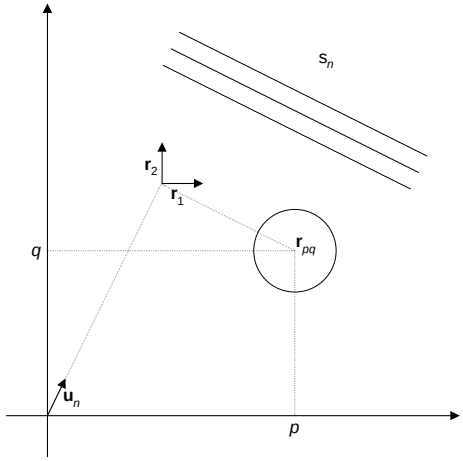

Figure 1: Sensor placed at the point $\mathbf{r}$ with the position coordinates (p, q) so that the point is described the following way: $\mathbf{r}_{pq} = p\mathbf{r}_1 + q\mathbf{r}_2$, where $\mathbf{r}_1$ and $\mathbf{r}_2$ are orthogonal vectors. The time delay between (p, q) and the origin with respect to the n'th source signal is denoted as τ^n_{pq}.

Gradient Flow

The spatial derivatives along the position coordinates (p, q) around the origin in the coordinate system are found of various orders (i, j) [11].

$$\xi_{ij}(t) \equiv \frac{\partial^{i+j}}{\partial^i p \partial^j q} x_{pq}(t) \mid_{p=q=0} \tag{6}$$

$$= \sum_{n=1}^{N} (\tau_1^n)(\tau_2^n) \frac{d^{i+j}}{d^{i+j}t} s_n(t) \tag{7}$$

Additionally, the derivative of the sensor noise $\nu_{ij}(t)$ may be added. Corresponding to (5), the 0'th and 1'st order terms yield:

$$\xi_{00}(t) = \sum_n s_n(t) \tag{8}$$

$$\xi_{10}(t) = \sum_n \tau_1^n \frac{ds^n(t)}{dt} = \sum_n \tau_1^n \dot{s}_n(t) \tag{9}$$

$$\xi_{01}(t) = \sum_n \tau_2^n \frac{ds^n(t)}{dt} = \sum_n \tau_2^n \dot{s}_n(t) \tag{10}$$

The estimates of the 0'th order term $\xi_{00}(t)$, i.e. the estimate of the field in the origin, can be obtained from the sensors as the average of the signals since the sensors are symmetrically distributed around the origin at the four coordinates (0,1), (1,0), (0,-1) and (-1,0):

$$\xi_{00}(t) \approx \frac{1}{4}(x_{-1,0} + x_{1,0} + x_{0,-1} + x_{0,1}). \tag{11}$$

The estimates of the two 1'st order derivatives can as well be estimated from the sensors:

$$\xi_{10}(t) = \frac{\partial x}{\partial p} \approx \frac{\Delta x}{\Delta p} = \frac{x_{1,0} - x_{-1,0}}{1 - (-1)} = \frac{1}{2}(x_{1,0} - x_{-1,0}) \tag{12}$$

$$\xi_{01}(t) = \frac{\partial x}{\partial q} \approx \frac{\Delta x}{\Delta q} = \frac{x_{0,1} - x_{0,-1}}{1 - (-1)} = \frac{1}{2}(x_{0,1} - x_{0,-1}) \tag{13}$$

By taking the time derivative of $\xi_{00}(t)$, the following equation can be obtained.

$$\frac{d}{dt}\xi_{00}(t) = \sum_{n=1}^{N} \frac{d}{dt} s_n(t) \tag{14}$$

Thus, the following instantaneous linear mixture can be obtained.

$$\begin{bmatrix} \dot{\xi}_{00}(t) \\ \xi_{10}(t) \\ \xi_{01}(t) \end{bmatrix} \approx \begin{bmatrix} 1 & \cdots & 1 \\ \tau_1^1 & \cdots & \tau_1^N \\ \tau_2^1 & \cdots & \tau_2^N \end{bmatrix} \begin{bmatrix} \dot{s}_1(t) \\ \vdots \\ \dot{s}_N(t) \end{bmatrix} \tag{15}$$

This equation is of the type $\mathbf{x} = \mathbf{A}\mathbf{s}$, where only $\mathbf{x}$ is known. Assuming that the source signals $\mathbf{s}$ are independent, (15) can be solved by independent component analysis (see e.g. [3]).

EXTENSION TO CONVOLUTIVE MIXTURES

As mentioned in [5], the instantaneous model (15), may be extended to convolutive mixtures. In Figure 2, a situation is shown in which each source signal does not only arrive from a single direction. Here, reflections of each

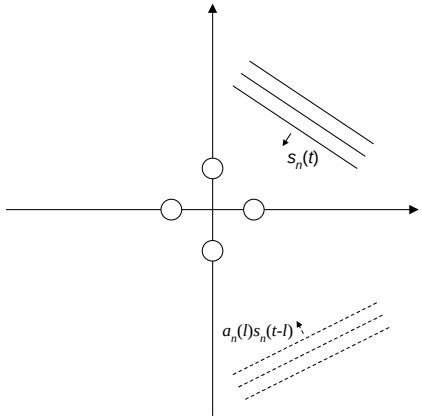

Figure 2: At the time t, a signal $s_n(t)$ originating from source n is arriving at the sensor array. At the same time, reflections from the same source arrive from other directions. These reflections are attenuated by the factor a_n and delayed by the time lag l. Each signal received at the sensor array are therefore convolved mixtures of the original source signals. For simplification, only a single source and a single reflection is shown.

source signal may be present too. Each reflection is delayed by a factor l and attenuated by an attenuated by a factor $a_n(l)$. Now, similarly to (4) the received signal x_{pq} at the sensor at position (p, q) is described as

$$x_{pq}(t) = \sum_{n=1}^{N} \sum_{l=0}^{L} a_n(l) s_n(t + \tau^n(\mathbf{r}, l) - l), \tag{16}$$

where L is the assumed maximum time delay. Using the Taylor expansion, each received mixture can be written as

$$x_{pq}(t) = \sum_{n=1}^{N} \sum_{l=0}^{L} a_n(l) \big[s_n(t-l) + \tau^n(\mathbf{r}, l) \dot{s}_n(t-l) + \frac{\tau^n(\mathbf{r}, l)}{2} \ddot{s}_n(t-l) + \dots \big] \tag{17}$$

Using only the first two terms of the Taylor expansion and inserting $\tau^n(\mathbf{r}_{pq}, l) = p\tau_1^n(l) + q\tau_2^n(l)$, (17) can be written as

$$x_{pq}(t) \approx \sum_{n=1}^{N} \sum_{l=0}^{L} a_n(l) \big[s_n(t-l) + (p\tau_1^n + q\tau_2^n) \dot{s}_n(t-l) \big]. \tag{18}$$

Similar to the instantaneous mixture case, the spatial derivatives of the convolutive mixture can be found from (6). The 0'th order and the 1'st order derivatives are then similarly to (8)–(10):

$$\xi_{00}(t) \quad = \quad \sum_{n} \sum_{l} a_n(l) s_n(t-l) \tag{19}$$

$$\xi_{10}(t) = \sum_n \sum_l a_n(l)\tau_1^n(l)\frac{ds^n(t-l)}{dt} = \sum_n \sum_l a_n(l)\tau_1^n(l)\dot{s}_n(t-l) \quad (20)$$

$$\xi_{01}(t) = \sum_n \sum_l a_n(l)\tau_2^n(l)\frac{ds^n(t-l)}{dt} = \sum_n \sum_l a_n(l)\tau_2^n(l)\dot{s}_n(t-l) \quad (21)$$

The time derivative of $\xi_{00}(t)$ is expressed as

$$\dot{\xi}_{00}(t) = \sum_{n=1}^{N} \sum_{l=0}^{L} a_n(l)\dot{s}_n(t-l). \quad (22)$$

By expressing (22), (20) and (21) with matrix notation, the following expression can be obtained:

$$\begin{bmatrix} \dot{\xi}_{00}(t) \\ \xi_{10}(t) \\ \xi_{01}(t) \end{bmatrix} \approx \sum_{l=0}^{L} \begin{bmatrix} a_1(l) & \cdots & a_N(l) \\ a_1(l)\tau_1^1(l) & \cdots & a_N(l)\tau_1^N(l) \\ a_1(l)\tau_2^1(l) & \cdots & a_N(l)\tau_2^N(l) \end{bmatrix} \begin{bmatrix} \dot{s}_1(t-l) \\ \vdots \\ \dot{s}_N(t-l) \end{bmatrix} \quad (23)$$

$$= \begin{bmatrix} a_1(l) & \cdots & a_N(l) \\ a_1(l)\tau_1^1(l) & \cdots & a_N(l)\tau_1^N(l) \\ a_1(l)\tau_2^1(l) & \cdots & a_N(l)\tau_2^N(l) \end{bmatrix} * \begin{bmatrix} \dot{s}_1(t) \\ \vdots \\ \dot{s}_N(t) \end{bmatrix}, \quad (24)$$

where $*$ is the convolution operator. This is a convolutive mixture problem of the well-known type $\mathbf{x} = \mathbf{A} * \mathbf{s}$, where only an estimate of $\mathbf{x}$ is known. These estimates are found similarly to the instantaneous case from (11)–(13).

FREQUENCY DOMAIN SEPARATION

In [8], the *Jade* algorithm [4] was successfully applied to solve the instantaneous mixing ICA problem (15). The *Jade* algorithm is based on joint diagonalization of 4'th order cumulants. In order to solve the convolutive mixing problem (23), the problem is transformed into the frequency domain [9]. Hereby, the convolution in the time domain can be approximated by multiplications in the frequency domain, i.e. for each frequency bin,

$$\xi(f,m) \approx \mathbf{A}(f)\dot{\mathbf{s}}(f,m), \quad (25)$$

where m denotes the index of the frame of which the short-time Fourier transform STFT is calculated. f denotes the frequency. When solving the ICA problem in the frequency domain, different permutations for each frequency band may occur. In order to solve the frequency permutations, the method suggested in [1] has been used. It is assumed that the mixing matrices in the frequency domain will be smooth. Therefore, the mixing matrix at frequency band k, $\mathbf{A}(f_k)$ is compared to the mixing matrix at band $k-1$, $\mathbf{A}(f_{k-1})$. This is done by calculating the distance between any possible permutations of $\mathbf{A}(f_k)$ and $\mathbf{A}(f_{k-1})$, i.e.

$$D(p) = \sum_{i,j} |a_{ij}^{(p)}(f_k) - a_{ij}(f_{k-1})|, \quad (26)$$

where p represents the p'th permutation. The permutation which yields the smallest distance is assumed to be the correct permutation. Notice, for an $N \times N$ mixing matrix, there are $N!$ different permutations. Therefore this method becomes slow for large N. For a 3×3 mixing matrix there are only six possible permutations.

EXPERIMENTS

Signals with synthetic impulse responses

Three speech sentences have been artificially mixed – two female speakers and one male speaker. The duration of each speech signal is 10 seconds, and the speech signals have a sampling frequency of 20 kHz. A demonstration of separated sounds is available at `wwm.imm.dtu.dk/~msp`. The microphone array consists of four microphones. These are placed in a horizontal plane. An application for such a microphone array is shown in Figure 3, where the four microphones are placed in a single hearing aid. Here, the distance between the microphones and the center of the array is 5 mm. By use of the gradient flow method, it is possible to separate up to three sources [8]. If there are more than three sources, an enhancement of the signals may be achieved even though full separation of all sources isn't possible. In the first experiment, a convolutive mixture of the three sources is simulated. The arrival angles as well as the attenuation factor of the reverberations have been chosen randomly. The maximum delay in this experiment has been chosen to 25 samples. No sensor noise is added. The differentiator has been chosen to be a 1000 order FIR differentiator estimated with a least squares approach (even though a smaller order could be sufficient). The integrator is implemented as a first order Alaoui IIR filter as in [8]. Here, all 200000 samples have been used to estimate the separated sounds. In order to achieve on-line separation, the separated sounds may be estimated using blocks of shorter duration [8]. The instantaneous Jade performs well if only the direct sounds are present, but if reverberations exist too, the separation performance is significantly reduced. The signal to interference ratio improvement is calculated as

$$\Delta\mathrm{SIR}(i) = 10\log\Big(\frac{\langle (y_{i,s_i})^2\rangle}{\langle (\sum_{j\neq i} y_{i,s_j})^2\rangle}\Big) - 10\log\Big(\frac{\langle (x_{10,s_i})^2\rangle}{\langle (\sum_{j\neq i} x_{10,s_j})^2\rangle}\Big), \qquad (27)$$

Here, y_{i,s_j} is the i'th separated signal, where only the j'th of the original signals has been sent through the mixing and unmixing system. x_{10,s_i} is the recorded signal at the microphone at position (1,0) with only the i'th source signal active. $\langle\cdot\rangle$ denotes the expectation over all samples.
The ΔSIR has been found for different DFT lengths as well as the case, where the instantaneous Jade has been applied to the convolutive mixture. Hamming windows of the same length as the DFT has been used. An STFT overlap of 75% has been used. Table 5.1 shows the separation results of the convolutive mixture. As it can be seen, the length of the DFT should be at

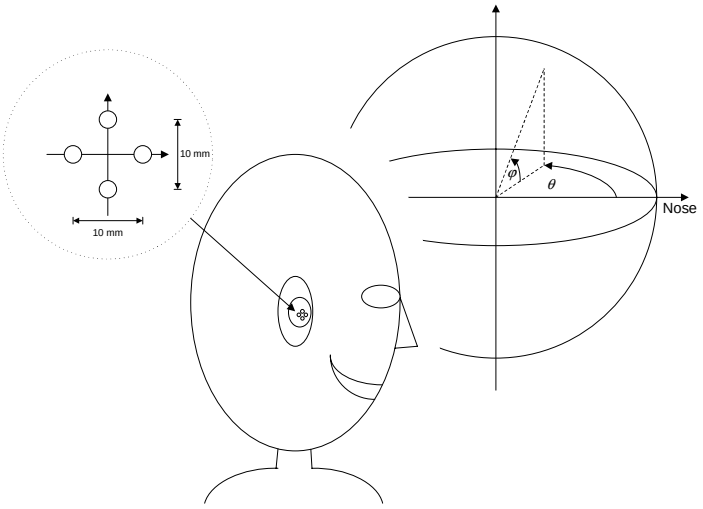

Figure 3: Four microphones are placed in a hearing aid. The distance between the microphones and the center of the array is 5 mm. By using such a configuration, it is possible to separate up to three independent sound sources. The azimuth angle, θ is defined according to the figure so that 0° is the direction of the nose. Likewise, the elevation angle φ is defined according to the figure so that 0° corresponds to the horizontal plane. Both angles increase in the counterclockwise direction.

least 256, in order to separate all three sources. By keeping the DFT length constant at 512, the length of the mixing filters were increased. Here the sources could be separated when the maximum delay of the mixing filters were up to 200 samples. By increasing the maximum delay to 400 samples, the separation failed. It can be seen that, the FIR separating filters have to be significantly longer than the mixing filters in order to ensure separation.

Real impulse responses

A four-microphone array has been placed in a dummy ear on the right side of a head and torso simulator. In an anechoic room, impulse responses have been estimated from different directions. No sensor noise has been added. Due to the recordings in an anechoic room, the only reflections existing are those from the head and torso simulator. The separation results are shown in Table 5.1. The performance is not as good as in the case of the synthetic impulse responses. In contrast to the synthetic impulse responses, the microphones may have different amplitude and phase responses. This may reduce the performance. The "UK female" seems to be the hardest sound to separate, but from the listening tests, it is easy to determine the separated sound from the two other speech signals.

Table 1: Three synthetic, artificially mixed speech signals have been separated. The maximum delay of each convolutive mixture is 25 samples. The arrival elevation angles (φ) and the azimuth (θ) angles of the direct sounds are given. The ΔSIR have been found for the instantaneous case and for different DFT lengths. The best separation is achieved with a DFT length of 256 or 512.

	UK Male	UK female	DK female
θ	0°	-112.5°	-157.5°
φ	0°	-21°	14°
Instantaneous JADE	9.5 dB	2.4 dB	2.5 dB
DFT length=64	10.2 dB	2.4 dB	14.2 dB
DFT length=128	11.0 dB	0.5 dB	11.5 dB
DFT length=256	9.0 dB	9.2 dB	14.6 dB
DFT length=512	8.9 dB	8.5 dB	16.5 dB
DFT length=1024	6.5 dB	8.7 dB	16.2 dB

Table 2: Signals generated from real impulse responses recorded by a four-microphone array placed in the right ear of a head and torso simulator inside an anechoic room. No noise has been added. Here, the "UK female" is the hardest sound to separate. When listening to the sounds, all of them seems to be separated. When the DFT becomes too long, the separation decreases. One explanation could be that the attempt to solve the permutation ambiguity fails.

	UK Male	UK female	DK female
θ	0°	-112.5°	-157.5°
φ	0°	-21°	14°
Instantaneous JADE	2.6 dB	2.2 dB	9.6 dB
DFT length=64	10.6 dB	1.6 dB	8.3 dB
DFT length=128	11.7 dB	-0.4 dB	5.8 dB
DFT length=256	13.1 dB	0.5 dB	6.1 dB
DFT length=512	13.9 dB	-0.2 dB	3.6 dB
DFT length=1024	9.8 dB	0.0 dB	2.6 dB

CONCLUSION AND FUTURE WORK

The performance by the instantaneous gradient flow beamforming is reduced significantly in reverberant mixtures. By expanding the gradient flow principle to convolutive mixtures, it is possible to separate convolutive mixtures in cases where the instantaneous gradient flow beamforming fails. It has been shown that the extension to convolutive mixtures can be achieved by solving a convolutive ICA problem (23) instead of solving an instantaneous ICA problem (15). A frequency domain Jade algorithm has been used to solve the convolutive mixing problem. In order to cope with a more difficult reverberant environment, other convolutive separation algorithms should be investigated. The mixing coefficients (23) are expected to have certain values. E.g. the first row in the mixing matrices is significantly larger than the

two other rows. Prior information on the coefficients of the mixing filters could as well be used in order to improve the separation. The knowledge of the delays in the mixing filters may as well be used in order to determine the arrival angles of the mixed sounds.

ACKNOWLEDGEMENT

The authors would like to thank Jan Larsen and Ulrik Kjems for useful comments and valuable discussions. We also acknowledge the financial support by the Oticon Foundation.

REFERENCES

[1] W. Baumann, B.-U. Köhler, D. Kolossa and R. Orglmeister, "Real Time Separation of Convolutive Mixtures," in **ICA2001**, San Diego, California, USA, December 9–12 2001, pp. 65–69.

[2] W. Baumann, D. Kolossa and R. Orglmeister, "Beamforming-Based Convolutive Source Separation," in **ICASSP2003**, Hong Kong, April 2003, vol. V, pp. 357–360.

[3] J.-F. Cardoso, "Blind Signal Separation: Statistical Principles," **Proceedings of the IEEE**, vol. 9, no. 10, pp. 2009–2025, October 1998.

[4] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," **IEE Proceedings-F**, vol. 140, no. 6, pp. 362–370, December 1993.

[5] G. Cauwenberghs, M. Stanacevic and G. Zweig, "Blind Broadband Source Localization and Separation in Miniature Sensor Arrays," in **IEEE Int. Symp. Circuits and Systems (ISCAS'2001)**, May 6–9 2001, vol. 3, pp. 193–196.

[6] G. W. Elko and A.-T. N. Pong, "A simple adaptive first-order differential microphone," in **Proceedings of 1995 Workshop on Applications of Single Processing to Audio and Accoustics**, October 15–18 1995, pp. 169–172.

[7] T. J. Ngo and N. Bhadkamkar, "Adaptive blind separation of audio sources by a physically compact device using second order statistics," in **First International Workshop on ICA and BSS**, Aussois, France, January 1999, pp. 257–260.

[8] C. M. Nielsen, **Gradient Flow Beamforming utilizing Independent Component Analysis**, Master's thesis, Aalborg University, Institute of Electronic Systems, January 5 2004.

[9] V. C. Soon, L. Tong, Y. F. Huang and R. Liu, "A wideband blind identification approach to speech acquisition using a microphone array," in **IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)**, San Francisco, California USA: IEEE, March 23–26 1992, vol. 1, pp. 293–296.

[10] M. Stanacevic, G. Cauwenberghs and G. Zweig, "Gradient Flow Broadband Beamforming and Source Separation," in **ICA'2001**, December 2001.

[11] M. Stanacevic, G. Cauwenberghs and G. Zweig, "Gradient Flow Adaptive Beamforming and Signal Separation in a Miniature Microphone Array," in **ICASSP2002**, Florida, USA, May 13–17 2002, vol. IV, pp. 4016–4019.

Appendix B

On the Difference Between Updating The Mixing Matrix and Updating the Separation Matrix

ON THE DIFFERENCE BETWEEN UPDATING THE MIXING MATRIX AND UPDATING THE SEPARATION MATRIX

Michael Syskind Pedersen, Ulrik Kjems

Oticon A/S,
Strandvejen 58
DK-2900 Hellerup, Denmark
{msp,uk}@oticon.dk

Jan Larsen

Technical University of Denmark
Informatics and Mathematical Modelling
Richard Petersens Plads, Building 321
DK-2800 Kongens Lyngby, Denmark
jl@imm.dtu.dk

ABSTRACT

When the ICA source separation problem is solved by maximum likelihood, a proper choice of the parameters is important. A comparison has been performed between the use of a mixing matrix and the use of the separation matrix as parameters in the likelihood. By looking at a general behavior of the cost function as function of the mixing matrix or as function of the separation matrix, it is explained and illustrated why it is better to select the separation matrix as a parameter than to use the mixing matrix as a parameter. The behavior of the natural gradient in the two cases has been considered as well as the influence of pre-whitening.

1. INTRODUCTION

Consider the independent component analysis (ICA) problem, where n sources $\mathbf{s} = [s_1, \ldots, s_n]^T$ are transmitted through a linear mixing system and observed by n sensors. The mixing system is described by the mixing matrix $\mathbf{A}$, and the observations are denoted by $\mathbf{x} = [x_1, \ldots, x_n]^T$. This leads to the following equation

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{1}$$

where only the observations $\mathbf{x}$ are known. The objective is to find an estimate $\mathbf{y}$ of the original sources. This can be done by estimating the separation mixing matrix $\mathbf{W} = \mathbf{A}^{-1}$, so that

$$\mathbf{y} = \mathbf{W}\mathbf{x}, \tag{2}$$

Notice, the source estimates may be arbitrarily permuted or scaled. The separation matrix can either be found directly or it can be found by finding an estimate of the mixing matrix and afterwards inverting the mixing matrix, provided that $\mathbf{A}$ is invertible. Here, the classical likelihood source separation is considered.

2. LIKELIHOOD SOURCE SEPARATION

A possible method for solving the ICA problem is the maximum likelihood principle [1]. The ML is closely related to other ICA methods [2] such as the infomax method [3], or maximum a posteriori MAP methods [4], [5]. In maximum likelihood source separation, the probability of a dataset given the parameters $\boldsymbol{\theta}$ of the model should be maximized. In this particular case, for the data $\mathbf{x}$, the parameters are given by either the separation matrix $\mathbf{W}$ or by the mixing matrix $\mathbf{A}$. Thus, the likelihood can be expressed by either

$$p(\mathbf{x}|\mathbf{W}) = |\det \mathbf{W}| \prod_m p_m(\sum_n W_{mn} x_n). \tag{3}$$

or

$$p(\mathbf{x}|\mathbf{A}) = \frac{1}{|\det \mathbf{A}|} \prod_m p_m(\sum_n A^{-1}_{mn} x_n) \tag{4}$$

Here, $p_m(\sum_n A^{-1}_{mn} x_n) = p_m(\sum_n W_{mn} x_n) = p(s_m)$ is the probability density function of the m'th source signal. For source signals such as speech, a heavy tailed source distribution is chosen. One way of maximizing the likelihood, is to minimize the negative logarithm of the likelihood. Given the likelihood functions in (3) and (4), the negative log likelihood functions are given in terms of either $\mathbf{W}$ or in terms of $\mathbf{A}$ as:

$$\mathcal{L}(\mathbf{W}) = -\ln|\det(\mathbf{W})| - \sum_m \ln p_m(\sum_n W_{mn} x_n) \tag{5}$$

$$\mathcal{L}(\mathbf{A}) = \ln|\det(\mathbf{A})| - \sum_m \ln p_m(\sum_n A^{-1}_{mn} x_n). \tag{6}$$

The respective gradients of (5) and (6) are given by [6]

$$\frac{\partial \mathcal{L}(\mathbf{W})}{\partial \mathbf{W}} = -(\mathbf{I} + \mathbf{z}\mathbf{y}^T)\mathbf{A}^T \tag{7}$$

$$\frac{\partial \mathcal{L}(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{W}^T(\mathbf{I} + \mathbf{z}\mathbf{y}^T). \tag{8}$$

Here $\mathbf{z} = \frac{\partial \ln p_m(\mathbf{y})}{\partial \mathbf{y}}$ is a nonlinear mapping of $\mathbf{y}$. Choosing $\mathbf{z} = -\tanh(\mathbf{y})$ corresponds to a probability density function for $\mathbf{y}$ proportional to $\frac{1}{\cosh(\mathbf{y})}$. This pdf is heavier tailed than e.g. a Gaussian distribution. $\mathbf{I}$ is the identity matrix. The gradient descent update steps are then

$$\mathbf{W} := \mathbf{W} + \mu_W(\mathbf{I} + \mathbf{z}\mathbf{y}^T)\mathbf{A}^T \tag{9}$$

$$\mathbf{A} := \mathbf{A} - \mu_A \mathbf{W}^T(\mathbf{I} + \mathbf{z}\mathbf{y}^T), \tag{10}$$

where μ_W and μ_A are learning rates. The learning rates can be constant or they can vary as a function of the update step. These algorithms may as well be made into iterative batch versions [7] by averaging over the samples:

$$\mathbf{W} := \mathbf{W} + \mu_W(\mathbf{I} + E[\mathbf{z}\mathbf{y}^T])\mathbf{A}^T \tag{11}$$

$$\mathbf{A} := \mathbf{A} - \mu_A \mathbf{W}^T(\mathbf{I} + E[\mathbf{y}\mathbf{z}^T]). \tag{12}$$

Here, $E[\cdot]$ denotes the expectation and each sample is assumed to be independent of the other samples.

Thanks to the Oticon Foundation for funding.

3. COMPARISON BETWEEN THE LIKELIHOOD FUNCTIONS

First consider the cost function $\mathcal{L}(\mathbf{A})$. In many source separation problems, the values of the mixing matrix will be relatively close to zero. Large values of $\mathbf{A}$ are not very likely. If it is assumed that there is a limit on how large $|A_{ij}|$ can be, the n^2-dimensional space occupied by the cost function $\mathcal{L}(\mathbf{A})$ is limited by this maximum value and it is possible to "view" the whole cost function because it only occupies a finite part of the $\mathbf{A}$-space. Because the whole cost function is within a finite space, the points where $\mathbf{A}$ is singular exist in this space too. At a singular point, the cost function $\mathcal{L}(\mathbf{A})$ becomes infinitely large. This makes it hard for gradient descent algorithms to find a minima in a limited space with the existence of infinite values. Now consider $\mathcal{L}(\mathbf{W})$. The space spanned by the $n \times n$ elements in $\mathbf{W}$ is infinitely large because a limit in the $\mathbf{A}$-space doesn't limit the $\mathbf{W}$-space. Now consider the behavior of the singular points in the $\mathbf{A}$-space when they are mapped into the $\mathbf{W}$-space. Recall that the $\{i, j\}$'th element of an inverse matrix can be written as

$$W_{ij} = (\mathbf{A}^{-1})_{ij} = \frac{\text{adj}(A_{ij})}{\det \mathbf{A}}, \tag{13}$$

where adj is the adjoint matrix. The adjoint matrix, can be found by the following steps:

1. Remove the jth row and the ith column of $(\mathbf{A})_{ij}$.
2. Find the determinant of the remaining part and
3. multiply by $(-1)^{i+j}$.

This means that the $\mathbf{A}^{-1}$ is proportional to $\frac{1}{\det \mathbf{A}}$. At the points where $\mathbf{A}$ is singular, its determinant is 0. Thus, when $\mathbf{A}$ becomes singular, $\mathbf{W}$ becomes infinitely large so all the points in the $\mathbf{A}$-space, where $\mathcal{L}(\mathbf{A}) = \infty$ are mapped into the $\mathbf{W}$-space far away from the origin and will therefore not disturb the gradient. Because large values of $\mathbf{A}$ are unlikely, $|\det \mathbf{A}|$ is prevented from becoming too large and hereby, the determinant of $\mathbf{W}$ is prevented from being close to 0. Hence, it is unlikely that $\mathbf{W}$ becomes singular.

4. SIMULATION EXAMPLE

The elements of a 3×3 mixing matrix have been drawn from a Gaussian distribution with zero mean and a standard deviation equal to one:

$$\mathbf{A} = \begin{bmatrix} 0.8644 & 0.8735 & -1.1027 \\ 0.0942 & -0.4380 & 0.3962 \\ -0.8519 & -0.4297 & -0.9649 \end{bmatrix} \tag{14}$$

Hereby,

$$\mathbf{W} = \mathbf{A}^{-1} = \begin{bmatrix} 0.7872 & 1.7481 & -0.1818 \\ -0.3275 & -2.3546 & -0.5927 \\ -0.5492 & -0.4949 & -0.6120 \end{bmatrix} \tag{15}$$

In order to find $E[\mathbf{z}\mathbf{y}^T]$, 3×1000 samples have been drawn from the $1/\cosh$-distribution[1]. The behavior of the two cost functions $\mathcal{L}(\mathbf{A})$ and $\mathcal{L}(\mathbf{W})$ are considered as function of two parameters in $\mathbf{A}$ and $\mathbf{W}$, respectively while the other parameters are kept constant. Figure 1 and figure 2 show the cost function $\mathcal{L}(\mathbf{W})$ as function of W_{11} and W_{21}. Figure 1 shows the negative direction of the gradients. The circle ($\circ$) is placed at the correct values of W_{11} and W_{12}, which also can be seen in (15). It can be seen that the negative gradient directions are pointing toward the global minimum, where the circle is located, or toward a local minimum. When considering figure 2, it can be seen that as $\mathcal{L}(\mathbf{W})$ is increased, when $|W_{11}|$ or $|W_{21}|$ is increased. Now consider figure 3. Here $\mathcal{L}(\mathbf{A})$ is shown as function of A_{11} and A_{21}. Here too, the negative gradient directions point toward the minima. In figure 4 the shape of $\mathcal{L}(\mathbf{A})$ can be seen. The values, where $\mathbf{A}$ is close to singular can clearly be seen and not far from these singular values, the global minimum exists. Due to these huge differences in the cost function within a quite small range, it can be hard to find the correct solution. This is also illustrated in figure 5. Here the value of the two cost functions $\mathcal{L}(\mathbf{A})$ and $\mathcal{L}(\mathbf{W})$ are shown as function of the number of iterations. The two learning rates are kept constant. They have been chosen such that the cost functions are minimized as fast as possible. The two learning rates has been found to be $\mu_A = 0.03$ and $\mu_W = 0.3$. It can be seen that more iterations are needed in order to find $\mathbf{A}$ than to find $\mathbf{W}$. Further, it can be seen that $\mathcal{L}(\mathbf{A})$ hasn't reached the minimum after 200 iterations. Actually, after 200 iterations the sources are not separated at all when $\mathcal{L}(\mathbf{A})$ is minimized. This can be explained by considering the cost function in figure 4. At the areas around the minimum of $\mathcal{L}(\mathbf{A})$, the cost function has almost the same value as at the minimum. This makes it very hard to minimize, since the gradient decent steps are very small. Even after 500 iterations, the separation quality [8] of the three sources is only between 13 and 41 dB while the separation quality of the sources, where the $\mathcal{L}(\mathbf{W})$ is minimized is between 36 and 88 dB. Even though only $\mathcal{L}(\mathbf{A})$ and $\mathcal{L}(\mathbf{W})$ have been investigated as function of two parameters in each matrix, the shown behavior of $\mathcal{L}(\mathbf{A})$ and $\mathcal{L}(\mathbf{W})$ is believed to be a general behavior for any of the parameters.

4.1. Natural Gradient learning

By using natural gradient descent [9] instead of gradient descent, the cost functions may be minimized with a smaller number of iterations. The natural gradients are obtained by multiplying the gradient in (7) by $\mathbf{W}^T\mathbf{W}$ on the right side and the gradient in (8) by $\mathbf{A}\mathbf{A}^T$ on the left side. Hereby, the natural gradient steps are given by [10]

$$\Delta \mathbf{W}_{NG} = -(\mathbf{I} + E[\mathbf{z}\mathbf{y}^T])\mathbf{W} \tag{16}$$

$$\Delta \mathbf{A}_{NG} = \mathbf{A}(\mathbf{I} + E[\mathbf{z}\mathbf{y}^T]). \tag{17}$$

The natural gradient update steps have been used in the separation problem. As it can be seen in figure 5, the separation performance works equally well whether the natural gradient is used in the A-domain or in the W-domain. Hereby, it seems that the natural gradient is able to erase the convergence difference between updating the algorithm in the A-domain and in the W-domain.

4.2. Pre-whitening

Pre-whitening of the data may simplify the separation problem. After pre-whitening the data $\mathbf{x}$ is uncorrelated and

$$E[\mathbf{x}\mathbf{x}^T] = \mathbf{I} \tag{18}$$

The update equations (11), (12), (16) and (17) have been applied in the case, where the data has been pre-whitened. Figure 6 shows

[1] Artificial data y which is $\frac{1}{\cosh}$-distributed can be generated from uniformly distributed data as $Y = \ln|\tan(X)|$, where X is a uniformly distributed random variable over the interval $0 < x < \pi$.

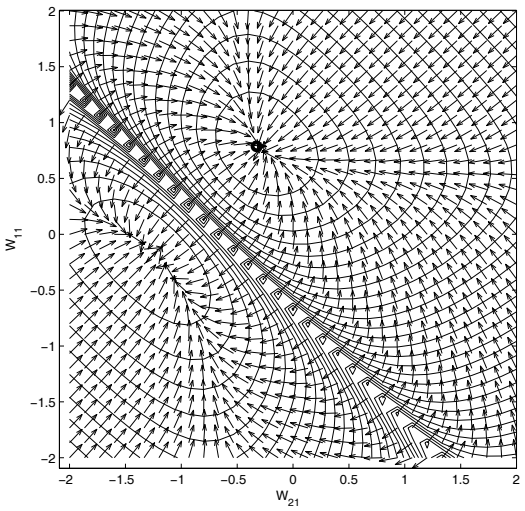

Fig. 1. The cost function $\mathcal{L}(\mathbf{W})$ as function of W_{11} and W_{21}. The other elements in $\mathbf{W}$ are held constant by their true value. The direction of the gradients are shown as well.

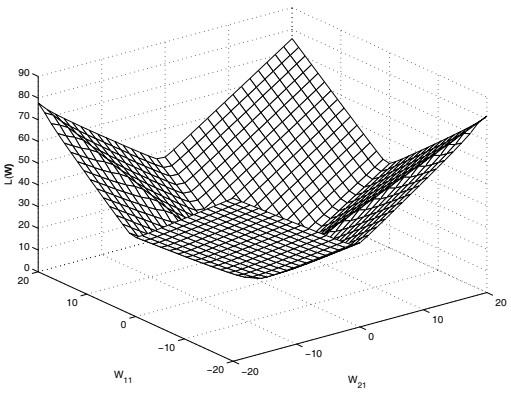

Fig. 2. The cost function $\mathcal{L}(\mathbf{W})$ as function of W_{11} and W_{21}. As it can be seen, as the parameters in $\mathbf{W}$ are increased, $\mathcal{L}(\mathbf{W})$ is increased too.

how the cost functions are minimized as function of the number of iterations. As it can be seen, the convergence time is significantly improved. Still, when $\mathbf{A}$ is updated in the $\mathbf{A}$-domain without the natural gradient, convergence is slow compared to updating in the $\mathbf{W}$-domain.

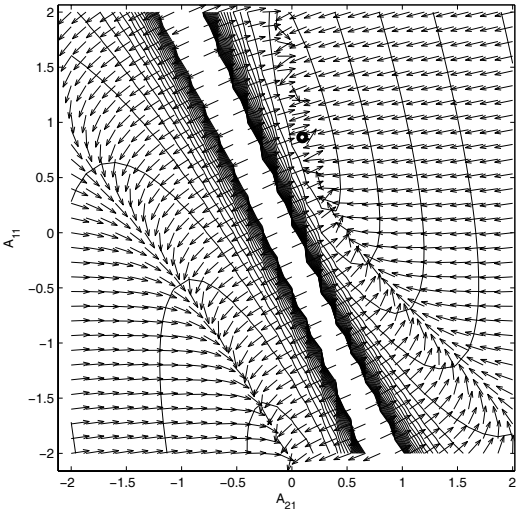

Fig. 3. The cost function $\mathcal{L}(\mathbf{A})$ as function of A_{11} and A_{21}. The other elements in $\mathbf{A}$ are held constant by their true value. The direction of the gradients are shown as well.

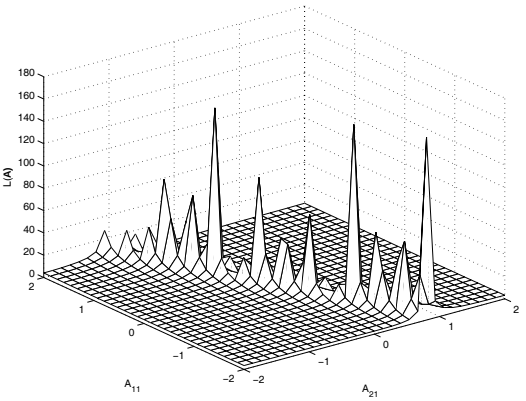

Fig. 4. The cost function $\mathcal{L}(\mathbf{A})$ as function of A_{11} and A_{21}. As it can be seen, the cost function is dominated by a high ridge, where the mixing matrix is close to singular, and some flat areas. Compared to the cost function in figure 2, it is much harder to find the global minima.

5. CONCLUSION

When performing source separation based on minimizing a cost function by gradient descent, the shape of the cost function is important. By comparing the negative log likelihood cost function as either function of the mixing matrix or as function of the separation matrix, the contours of the cost functions are very different. Due

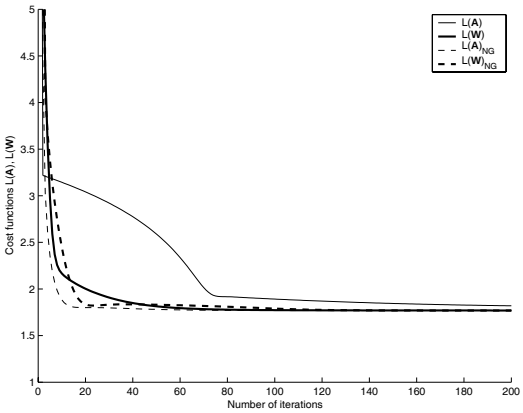

Fig. 5. The cost function $\mathcal{L}(\mathbf{A})$ and $\mathcal{L}(\mathbf{W})$ as function of the number of iterations. The constant learning rates are selected in order to minimize the number of iterations in order to ensure convergence. After 200 iterations, only $\mathcal{L}(\mathbf{W})$ has been minimized. Even though $\mathcal{L}(\mathbf{A})$ seems to have been minimized as well, $\mathbf{A}$ has not been correctly estimated. Only a value of $\mathbf{A}$ somewhere at the flat areas in figure 4 has been found, and much more iterations are needed in order to find the correct value of $\mathbf{A}$. Also, the minimization as function of the iterative update by use of the natural gradient is shown. Here, the cost functions are minimized by use of a smaller number of iterations and fast convergence is achieved for the update of $\mathbf{A}$ as well as $\mathbf{W}$. By using the natural gradient, the difference between updating the algorithm in the two domains seems to have disappeared.

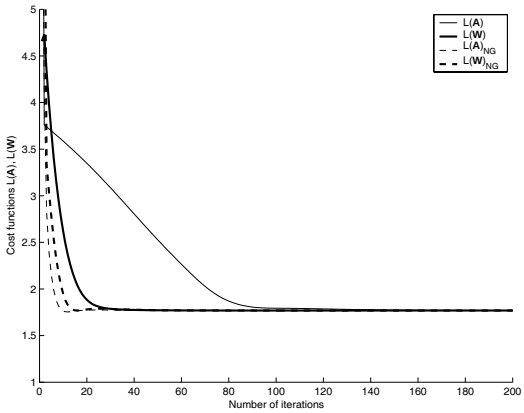

Fig. 6. The cost function $\mathcal{L}(\mathbf{A})$ and $\mathcal{L}(\mathbf{W})$ as function of the number of iterations as in figure 5. Contrary to figure 5, the data has been pre-whitened before the four update equations have been applied to the data. As it can be seen, pre-whitening increases the convergence speed significantly. Still, the update in the $\mathbf{A}$-domain without the natural gradient has the slowest convergence speed.

to these different behaviors of the cost functions, it has been found that it is much easier to minimize the negative log likelihood, when it is a function of the separation matrix than as function of the mixing matrix. If the natural gradient is applied in the mixing domain, it is able to cope with the difficult contour of the cost function. But in problems, where the natural gradient is hard to find, a proper choice of the parameters may be crucial. Also pre-whitening has been considered. By pre-whitening the data before applying the ICA algorithm, the convergence is significantly increased. The results may be generalized to more difficult problems such as e.g. convolutive ICA.

6. REFERENCES

[1] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, *Independent Component Analysis*, Wiley, 2001.

[2] Jean-François Cardoso, "Blind signal separation: Statistical principles," *Proceedings of the IEEE*, vol. 9, no. 10, pp. 2009–2025, October 1998.

[3] Anthony J. Bell and Terrence J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[4] Pedro A.d.F.R. Højen-Sørensen, Ole Winther, and Lars Kai Hansen, "Mean field approaches to independent component analysis," *Neural Computation*, vol. 14, pp. 889–918, 2002.

[5] Kevin H. Knuth, "Bayesian source separation and localization," in *Proceedings of the SPIE Conference on Bayesian Infernce on Inverse problems*, San Diego, California, July 1998, pp. 147–158.

[6] David J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 1st edition, 2003.

[7] Nikos Vlassis and Yoichi Motomura, "Efficient source adaptivity in independent component analysis," *IEEE Transactions on Neural Networks*, vol. 12, no. 3, pp. 559–565, May 2001.

[8] D.W.E. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Int. Workshop Independent Component Analysis and Blind Signal Separation*, Aussois, France, January 11–15 1999, pp. 261–266.

[9] Shun-ichi Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251–276, 1998.

[10] Andrezej Cichocki and Shun-ichi Amari, *Adaptive Blind Signal and Image Processing*, Wiley, 2002.

Appendix C

Overcomplete Blind Source Separation by Combining ICA and Binary Time-Frequency Masking

OVERCOMPLETE BLIND SOURCE SEPARATION BY COMBINING ICA AND BINARY TIME-FREQUENCY MASKING

Michael Syskind Pedersen[1,2]*, DeLiang Wang*[3]*, Jan Larsen*[1] *and Ulrik Kjems*[2]

[1] Informatics and Mathematical Modelling, Technical University of Denmark
Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark
[2]Oticon A/S, Strandvejen 58, DK-2900 Hellerup, Denmark
[3] Department of Computer Science and Engineering & Center for Cognitive Science,
The Ohio State University, Columbus, OH 43210-1277, USA

ABSTRACT

A limitation in many source separation tasks is that the number of source signals has to be known in advance. Further, in order to achieve good performance, the number of sources cannot exceed the number of sensors. In many real-world applications these limitations are too strict. We propose a novel method for overcomplete blind source separation. Two powerful source separation techniques have been combined, *independent component analysis* and *binary time-frequency masking*. Hereby, it is possible to iteratively extract each speech signal from the mixture. By using merely two microphones we can separate up to six mixed speech signals under anechoic conditions. The number of source signals is not assumed to be known in advance. It is also possible to maintain the extracted signals as stereo signals.

1. INTRODUCTION

Blind source separation (BSS) addresses the problem of recovering N unknown source signals $\mathbf{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ from M recorded mixtures $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^T$ of the source signals. The term blind refers to that only the recorded mixtures are known. An important application for BSS is separation of speech signals. The recorded mixtures are assumed to be linear superpositions of the source signals, i.e.

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \boldsymbol{\nu}(n), \tag{1}$$

where $\mathbf{A}$ is an $M \times N$ mixing matrix and n denotes the discrete time index. $\boldsymbol{\nu}(n)$ is additional noise. A method to retrieve the original signals up to an arbitrary permutation and scaling is independent component analysis (ICA) [1]. In ICA, the main assumption is that the source signals are independent. By applying ICA, an estimate $\mathbf{y}(n)$ of the source signals can be obtained by finding a (pseudo)inverse $\mathbf{W}$ of the mixing matrix so that

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n). \tag{2}$$

Many methods require that the number of source signals is known in advance. Another drawback of most of these methods is that the number of source signals is assumed not to exceed the number of microphones, i.e. $M \geq N$. Even if the mixing process $\mathbf{A}$ is known, it is not invertible, and in general, the independent components cannot be recovered exactly [1]. In the case of more sources than sensors, the *overcomplete/underdetermined* case, successful separation often relies on the assumption that the source signals are sparsely distributed - either in the time domain, in the frequency domain or in the time-frequency (T-F) domain [2], [3], [4], [5]. If the source signals do not overlap in the time-frequency domain, high-quality reconstruction could be obtained [4].

However, there is overlap between the source signals. In this case, good separation can still be obtained by applying a binary time-frequency mask to the mixture [3], [4]. In *computational auditory scene analysis*, the technique of T-F masking has been commonly used for years (see e.g. [6]). Here, source separation is based on organizational cues from auditory scene analysis [7]. More recently the technique has also become popular in blind source separation, where separation is based on non-overlapping sources in the T-F domain [8]. T-F masking is applicable to source separation/ segregation using one microphone [6], [9] or more than one microphone [3], [4]. T-F masking can be applied as a binary mask. For a binary mask, each T-F unit is either weighted by one or by zero. In order to reduce musical noise, more smooth masks may also be applied [10]. An advantage of using a binary mask is that only a binary decision has to be made [11]. Such a decision can be based on, e.g., clustering [3], [4], [8], or direction-of-arrival [12]. ICA has been used in different combinations with the binary mask. In [12], separation is performed by removing signals by masking $N - M$ signals and afterwards applying ICA in order to separate the remaining M signals. ICA has also been used the other way around. In [13], it has been applied to separate two signals by using two microphones. Based on the ICA outputs, T-F masks are estimated and a mask is applied to each of the ICA outputs in order to improve the signal to noise ratio.

In this paper, a novel method for separating an arbitrary number of speech signals is proposed. Based on the output of a square (2×2) ICA algorithm and binary T-F masks, this method iteratively segregates signals from a mixture until an estimate of each signal is obtained.

2. GEOMETRICAL INTERPRETATION OF INSTANTANEOUS ICA

We assume that there is an unknown number of acoustical source signals but only two microphones. It is assumed that each source signal arrives from a certain direction and no reflections occur, i.e. an anechoic environment. In order to keep the problem simple, the source signals are mixed by an instantaneous mixing matrix as in eq. (1). Due to delays between the microphones, instantaneous ICA with a real-valued mixing matrix usually is not applica-

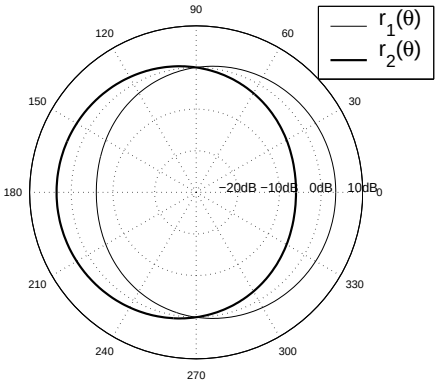

Fig. 1. The two directional microphone responses are shown as function of the direction θ.

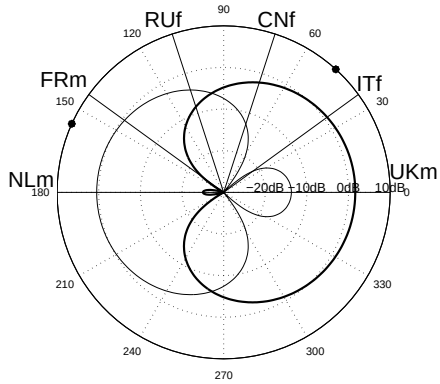

Fig. 2. The polar plots show the gain for different directions. ICA is applied with two sensors and six sources. The two dots at the periphery show the null directions. The lines pointing out from the origin denote the true direction of the speech sources. The three-letter abbreviations (see table 1) identifies the different speech signals which have been used. As it can be seen from the figure, the ICA solution tends to place the null towards sources spatially close to each other. Therefore, each of the two outputs is a group of signals spatially close to each other.

Table 1. The six speech signals. All speakers use raised voice as if they were speaking in a noisy environment.

Abbreviation	Description
CNf	Female speech in Chinese
NLm	Male speech in Dutch
FRm	Male speech in French
ITf	Female speech in Italian
UKm	Male speech in English
RUf	Female speech in Russian

ble to signals recorded at an array of microphones, but if the microphones are placed at exact same location and the microphones have different responses for different directions, the separation of delayed sources can be approximated by the instantaneous model [14]. Hereby, a combination of microphone gains correspond to a certain directional pattern. Therefore, two directional microphone responses are used. The two microphone responses are chosen as functions of the direction θ as $r_1(\theta) = 1 + 0.5\cos(\theta)$ and $r_2(\theta) = 1 - 0.5\cos(\theta)$, respectively. The two microphone responses are shown in figure 1. It is possible to make two such directional patterns by adding and subtracting omnidirectional signals from two microphones placed closely together. Hence, the mixing system is given by

$$\mathbf{A}(\boldsymbol{\theta}) = \begin{bmatrix} r_1(\theta_1) & \cdots & r_1(\theta_N) \\ r_2(\theta_1) & \cdots & r_2(\theta_N) \end{bmatrix}. \quad (3)$$

Different speech signals are used as source signals. The used signals are sampled with a sampling frequency of 10 kHz and the duration of each signal is 5 s. The speech signals are shown in table 1.

2.1. More sources than sensors

Now consider the case where $N \geq (M = 2)$. When there are only two mixed signals, a standard ICA algorithm only has two output signals $\mathbf{y}(n) = [y_1(n), y_2(n)]^T$. Since the number of separated signals obtained by (2) is smaller than the number of source signals, $\mathbf{y}$ does not contain the separated signals. Instead $\mathbf{y}$ is another linear superposition of each of the source signals, where the weights are given by $\mathbf{G} = \mathbf{WA}$ instead of just $\mathbf{A}$ as in (1). Hereby, $\mathbf{G}$ just corresponds to another weighting depending on θ. These weights make $y_1(n)$ and $y_2(n)$ as independent as possible. This is illustrated in figure 2. An implementation of the infomax ICA algorithm [15] has been used. The BGFS method has been used for optimization [16][1]. The figure shows the two estimated spatial responses from $\mathbf{G}(\boldsymbol{\theta})$ in the overdetermined case. The response of the m'th output is given by $|\mathbf{w}_m^T\mathbf{a}(\theta)|$, where $\mathbf{w}_m$ is the separation vector from the m'th output and $\mathbf{a}(\theta)$ is the mixing vector for the arrival direction θ [17]. By varying θ over all possible directions, directivity patterns can be created as shown in figure 2. The estimated null placement is illustrated by the two round dots placed at the periphery of the polar plot. The lines pointing out from the origin illustrate the correct direction of the source signals. Here, the sources are uniformly distributed in the interval $[0^\circ \leq \theta \leq 180^\circ]$. As it can be seen, the nulls do not cancel single sources out. Rather, a null is placed at a direction pointing towards a *group* of sources which are spatially close to each other. Here, it can be seen that the first output, $y_1(n)$, the signals NLm and FRm are dominating and in the second output, $y_2(n)$, the signals UKm, ITf and CNf are dominating. The sixth signal, RUf exists in both outputs. This new weighting of the signals can be used to estimate binary masks.

[1]Matlab toolbox available from `http://mole.imm.dtu.dk/toolbox/ica/`

3. BLIND SOURCE EXTRACTION WITH ICA AND BINARY MASKING

A flowchart for the algorithm is given in figure 3. As described in the previous section, a two-input-two-output ICA algorithm is applied to the input mixtures, disregarding the number of source signals that actually exist in the mixture. The two output signals are arbitrarily scaled. The scaling is fixed by using knowledge about the microphone responses. Hereby, the two null directions can be found. The two output signals are scaled such that where one directional response has a null, the other response has a unit gain. The two re-scaled output signals, $\hat{y}_1(n)$ and $\hat{y}_2(n)$ are transformed into the frequency domain e.g. by use of the Short-Time Fourier Transform STFT so that two spectrograms are obtained:

$$\hat{y}_1 \quad \rightarrow \quad Y_1(\omega, t) \tag{4}$$
$$\hat{y}_2 \quad \rightarrow \quad Y_2(\omega, t), \tag{5}$$

where ω denotes the frequency and t is the time index. The binary masks are then determined by for each T-F unit comparing the amplitudes of the two spectrograms:

$$\mathrm{BM1}(\omega, t) \quad = \quad \tau|Y_1(\omega, t)| > |Y_2(\omega, t)| \tag{6}$$
$$\mathrm{BM2}(\omega, t) \quad = \quad \tau|Y_2(\omega, t)| > |Y_1(\omega, t)|, \tag{7}$$

where τ is a threshold. Next, each of the the two binary masks is applied to the original mixtures in the T-F domain, and by this non-linear processing, some of the speech signals are *removed* by one of the masks while other speakers are removed by the other mask. After the masks have been applied to the signals, they are reconstructed in the time domain by the inverse STFT. If there is only a single signal left in the masked output, defined by the selection criteria in section 3.1, i.e. all but one speech signal have been masked, this signal has been extracted from the mixture and it is saved. If there are more than one signal left in the masked outputs, ICA is applied to the two masked signals again and a new set of masks are created based on (6), (7) and the previous masks. The use of the previous mask ensures that T-F units that have been removed from the mixture are not reintroduced by the next mask. This is done by an element-wise multiplication between the previous mask and the new mask. This iterative procedure is followed until all masked outputs consist of only a single speech signal. Notice, the output signals are maintained as two signals. Stereo signals created with directional microphones placed at the same location with an angle between the directional patterns of 90° (here 180°) are termed XY-stereo.

3.1. Selection criterion

Further processing on a pair of masked signals should be avoided in two cases. If all but one signal have been removed or if too much has been removed so that there is no signal left after applying the mask. The decisions are based on the eigenvalues of the covariance matrix between the masked sensor signals. The covariance matrix is calculated as

$$\mathbf{R} = \langle \hat{\mathbf{x}} \hat{\mathbf{x}}^T \rangle, \tag{8}$$

where $\langle \cdot \rangle$ denotes the expectation with respect to the whole signal, and $\hat{\mathbf{x}}$ is the two time domain signals of which the binary mask has been applied. If $\hat{\mathbf{x}}$ only contains one signal, the covariance matrix is singular, and the smallest eigenvalue $\lambda_{\min}$ is approximately equal to zero [18]. Since parts of the other signals may remain after masking, the smallest eigenvalue is equal to the noise variance

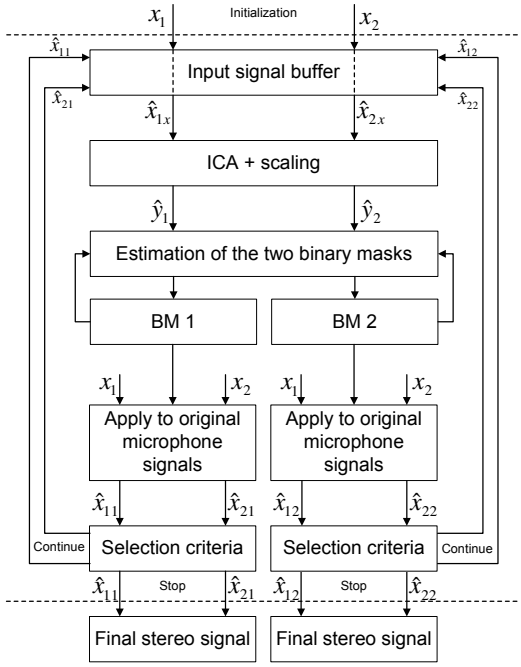

Fig. 3. Flowchart showing the main steps of the proposed algorithm. From the output of the ICA algorithm, binary masks are estimated. The binary masks are applied to the original signals which again are processed through the ICA step. Every time the output from one of the binary masks is detected as a single signal, the signal is stored. The iterative procedure stops when all outputs only consist of a single signal.

of these remaining signals. Therefore, if $\lambda_{\min}$ is smaller than a certain noise threshold $\tau_{\lambda_{\min}}$, it is assumed that there is less than two signals and no further processing is necessary. In order to discriminate between zero or one signal, the largest eigenvalue $\lambda_{\max}$ is considered. If $\lambda_{\max}$ is smaller than a certain threshold $\tau_{\lambda_{\max}}$, the output is considered of such a bad quality that the signal should be thrown away.

3.2. Finding the remaining signals

Since some signals may have been removed by both masks, all T-F units that have not been assigned the value '1' are used to create a *remaining mask*, and the procedure is applied to the mixture signal of which the remaining mask is applied, to ensure that all signals are estimated. Notice, this step has been omitted from figure 3.

4. EVALUATION

The algorithm described above has been implemented and evaluated with mixtures of the six signals from table 1. For the STFT,

an FFT length of 2048 has been used. This gives a frequency resolution of 1025 frequency units. A Hanning window with a length of 512 samples has been applied to the FFT signal and the frame shift is 256 samples. A high frequency resolution is found to be necessary in order to obtain good performance. The sampling frequency of the speech signals is 10 kHz. The three thresholds τ, $\tau_{\lambda_{\min}}$ and $\tau_{\lambda_{\max}}$ have been found from initial experiments. In the ICA step, the separation matrix is initialized by the identity matrix, i.e. $\mathbf{W} = \mathbf{I}$. In order to test robustness, $\mathbf{W}$ was also initialized with a random matrix with values uniformly distributed over the interval [0,1]. The different initialization did not affect the result. When using a binary mask, it is not possible to reconstruct the speech signal as if it was recorded in the absence of the interfering signals, because the signals partly overlap. Therefore, as a computational goal for source separation, the *ideal binary mask* has been suggested [11]. The ideal binary mask for a signal is found for each T-F unit by comparing the energy of the desired signal to the energy of all the interfering signals. Whenever the signal energy is highest, the T-F unit is assigned the value '1' and whenever the interfering signals have more energy, the T-F unit is assigned the value '0'. As in [9], for each of the separated signals, the percentage of energy loss P_{EL} and the percentage of noise residue P_{NR} are calculated:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \quad (9)$$

$$P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)}, \quad (10)$$

where $O(n)$ is the estimated signal, and $I(n)$ is the recorded mixture resynthesized after applying the ideal binary mask. $e_1(n)$ denotes the signal present in $I(n)$ but absent in $O(n)$ and $e_2(n)$ denotes the signal present in $O(n)$ but absent in $I(n)$. Also the signal to noise ratio (SNR) is found. Here the SNR is defined using the resynthesized speech from the ideal binary mask as the ground truth

$$\text{SNR} = 10\log_{10}\left[\frac{\sum_n I^2(n)}{\sum_n (I(n)-O(n))^2}\right]. \quad (11)$$

The algorithm has been applied to mixtures consisting of up to six signals. In all mixing situations, the signals have been uniformly distributed in the interval $[0^\circ \leq \theta \leq 180^\circ]$. The separation results are shown in figure 4 and in table 2.

Two ideal binary masks have been found – one for each microphone signal. In all cases, all the signals have been segregated from the mixture. In most cases also the correct number of signals is estimated. Only in the case of three mixtures, one of the source signals is estimated twice. The double extraction is caused by the selection criteria. Based on the chosen thresholds, the selection criteria in some cases allows a signal to be extracted more than once. In the case of the six mixtures from figure 2, the six estimated binary masks are shown in figure 5 along with the estimated ideal binary masks from each of the two microphone signals. The input SNR (SNR_i) is shown in figure 4 too. The SNR_i is the ratio between the desired signal and the noise in the recorded mixtures. The separation quality decreases when the number of signals is increased. This is expected because when the number of mixed signals is increased, the mixtures become less sparse. Random distributions of the source directions as well as more than six signals have also been examined. Here, in general, not all the sources are separated from each other. If the arrival angles between signals are too narrow, these signals may be detected as a single signal, and they are not separated. Listening tests validate the separation results. This method differs from previous methods which use a binary mask and two microphones [3], [4]. In [3], binaural cues have been applied for separation, i.e interaural time and intensity differences. In [4], the separation is likewise based on amplitude and time difference of each source. Here separation is based on clustering of T-F units that have similar amplitude and phase properties. In our approach too, separation can only be achieved if the source signals have different spatial positions, but the separation criterion is based on independence between the source signals.

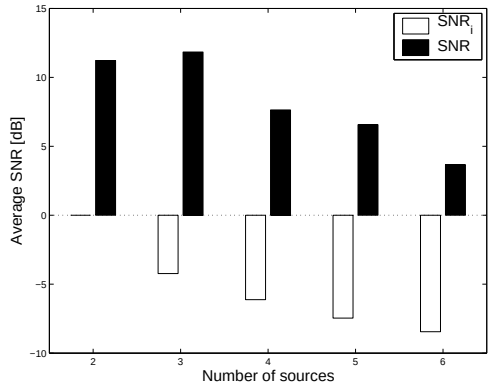

Fig. 4. The signal to noise ratio as function of the number of source signals. The average SNR for the mixtures before separation (SNR_i) is shown as well as the average SNR after separation calculated by eq. (11). In the case of three signals, the incorrectly estimated signal is ignored (see table 2).

5. CONCLUDING REMARKS

A novel method of blind source separation of has been described. Based on sparseness and independence, the method iteratively extracts all the speech signals without knowing the signals in advance. An advantage of this method is that stereo signals are maintained through the processing. So far, the method has been applied to successful separation of up to six speech signals under anechoic conditions by use of two microphones. Future work will include separation of mixtures in reverberant environment, a more blind solution of the scaling problem, and improved techniques for the stopping criteria based on detection of a single signal. Alternative to using a linear frequency scale, a frequency scale that models the auditory system more accurately could be used, because an auditory-based front-end is reported to be more robust than a Fourier-based analysis in the presence of background interference [9]. The use of more than two sensors could also be investigated. By using more than two sensors, a better resolution can be obtained

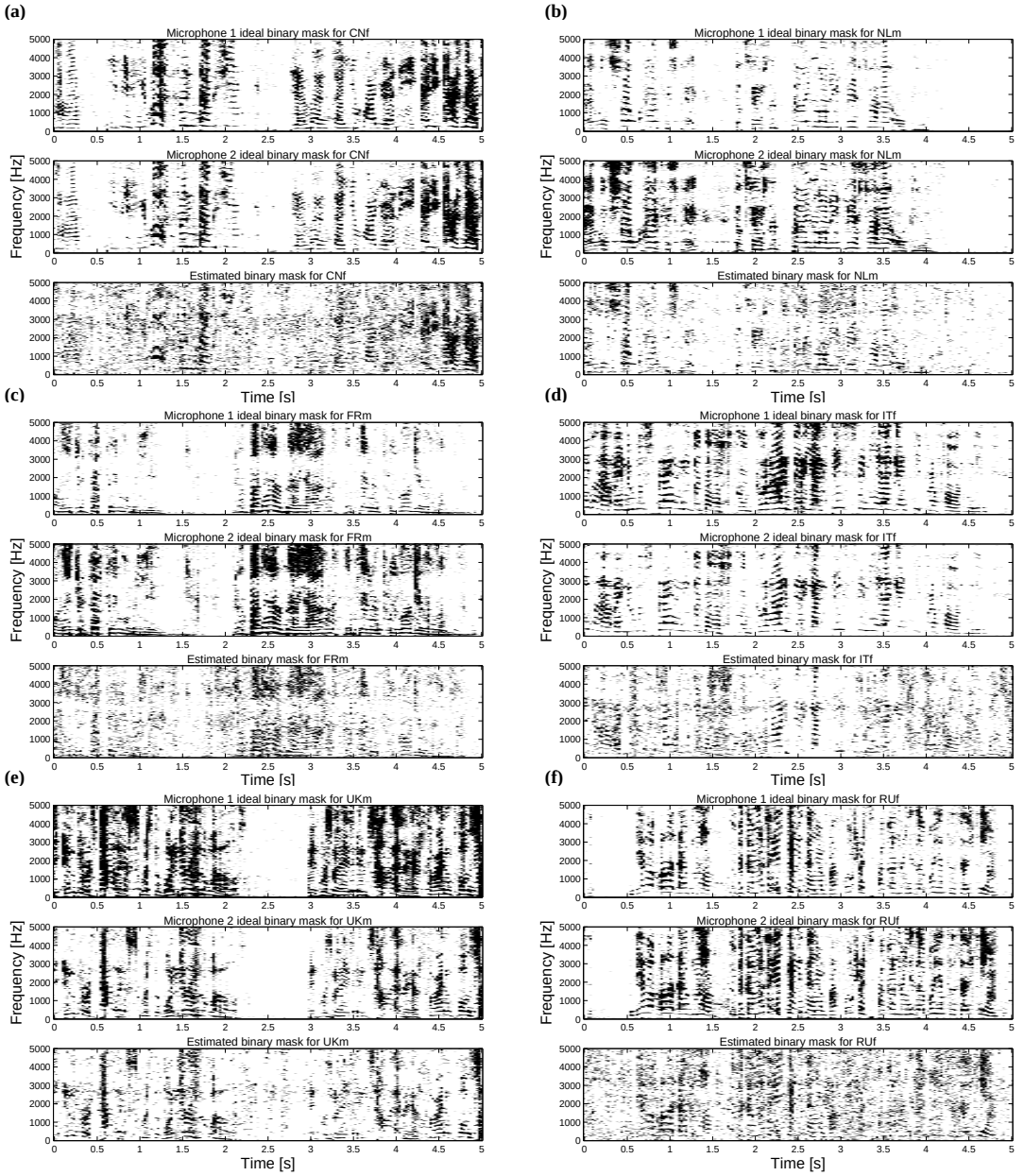

Fig. 5. For a mixture of 6 mixed speech signals, binary masks have been estimated for each of the 6 speech signals. The black areas correspond to the mask value '1' and the white areas correspond to the mask value '0'. The results are shown together with the calculated *ideal binary masks* of each of the two microphone signals. The signals (**a**)–(**f**) appear in the order which they were extracted from the mixture. The first three signals (**a**)–(**c**) were extracted after two iterations, the next two signals (**d**), (**e**) were extracted after three iterations. The last signal (**f**) was extracted from the remaining mask as described in section 3.2.

Table 2. Separation results. Mixtures consisting from two up to six signals have been separated from each other successfully. In most cases, the correct number of sources has been extracted. Only in the case of three source signals, one of the signals has been estimated twice. Here the average performance has been calculated with(†) and without the extra signal. The signals appear in the order which they were extracted from the mixture.

Separated	Microphone 1		Microphone 2	
Signal	$P_{EL}(\%)$	$P_{NR}(\%)$	$P_{EL}(\%)$	$P_{NR}(\%)$
UKm	0.01	8.42	6.83	0.00
FRm	7.13	0.00	0.00	6.11
Average	3.57	4.21	3.41	3.06
NLm	0.11	2.46	3.84	0.06
CNf	5.28	0.16	0.26	2.81
CNf†	86.39	13.12	88.97	63.95
RUf	6.74	11.55	6.17	17.26
Average†	24.63	6.82	24.81	21.02
Average	4.04	4.72	3.43	6.71
CNf	1.27	13.25	3.78	13.79
RUf	2.14	17.64	17.26	3.24
FRm	5.37	2.77	1.01	10.79
UKm	19.60	8.00	14.67	4.60
Average	7.09	10.41	9.18	8.11
RUf	10.65	20.00	24.17	17.70
NLm	8.11	4.13	13.58	1.84
FRm	9.81	17.68	1.32	22.37
ITf	19.20	4.37	4.87	6.92
CNf	4.74	15.55	5.13	16.93
Average	10.50	12.35	9.81	13.15
CNf	8.72	28.20	6.77	21.92
NLm	11.96	15.45	16.32	11.47
FRm	16.05	34.95	29.05	28.72
ITf	29.69	26.87	20.36	23.08
UKm	35.56	6.14	23.26	8.38
RUf	19.58	46.57	28.14	35.33
Average	20.26	26.36	20.65	21.48

and ambiguous arrival angles may be avoided. Also applications for other types of sparse signals could be examined.

6. ACKNOWLEDGEMENTS

The work was performed while M.S.P. was a visiting scholar at The Ohio State University Department of Computer Science and Engineering. M.S.P was supported by the Oticon Foundation. M.S.P and J.L are partly also supported by the European Commission through the sixth framework IST Network of Excellence: Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL), contract no. 506778. D.L.W was supported in part by an AFOSR grant (FA9550-04-1-0117) and an AFRL grant (FA8750-04-0093).

7. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley, 2001.

[2] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time fourier transform," in *Proc. ICA'2000*, 2000, pp. 87–92.

[3] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, October 2003.

[4] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[5] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete bss for convolutive mixtures based on hierarchical clustering," in *Proc. ICA'2004*, Granada, Spain, September 22–24 2004, pp. 652–660.

[6] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 684–697, May 1999.

[7] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, 2 edition, 1990.

[8] A. Jourjine, S. Richard, and Ö. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proc. ICASSP'2000*, Istanbul, Turkey, June 2000, vol. V, pp. 2985–2988.

[9] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1135–1150, September 2004.

[10] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proc. ICASSP2005*, March 18–23 2005, vol. III, pp. 81–84.

[11] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, Pierre Divenyi, Ed., pp. 181–197. Kluwer, Norwell, MA, 2005.

[12] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ica," in *Proc. ICA'2004*, September 22–24 2004, pp. 898–905.

[13] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," in *Proc. ICA'2004*, Granada, Spain, September 22–24 2004, pp. 832–839.

[14] M. Ito, Y. Takeuchi, T. Matsumoto, H. Kudo, M. Kawamoto, T. Mukai, and N. Ohnishi, "Moving-source separation using directional microphones," in *Proc. ISSPIT'2002*, December 2002, pp. 523–526.

[15] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[16] H.B. Nielsen, "Ucminf - an algorithm for unconstrained, nonlinear optimization," Tech. Rep. IMM-TEC-0019, IMM, Technical University of Denmark, 2001.

[17] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays*, Digital Signal Processing. Springer, 2001.

[18] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acous., Speech and Signal Processing*, vol. ASSP-33, no. 2, pp. 387–392, April 1985.

Appendix D

BLUES from Music: BLind Underdetermined Extraction of Sources from Music

BLUES from Music: BLind Underdetermined Extraction of Sources from Music

Michael Syskind Pedersen, Tue Lehn-Schiøler, and Jan Larsen

ISP, IMM, DTU*
{msp,tls,jl}@imm.dtu.dk

Abstract. In this paper we propose to use an instantaneous ICA method (BLUES) to separate the instruments in a real music stereo recording. We combine two strong separation techniques to segregate instruments from a mixture: ICA and binary time-frequency masking. By combining the methods, we are able to make use of the fact that the sources are differently distributed in both space, time and frequency. Our method is able to segregate an arbitrary number of instruments and the segregated sources are maintained as stereo signals. We have evaluated our method on real stereo recordings, and we can segregate instruments which are spatially different from other instruments.

1 Introduction

Finding and separating the individual instruments from a song is of interest to the music community. Among the possible applications is a system where e.g. the guitar is removed from a song. The guitar can then be heard by a person trying to learn how to play. At a later stage the student can play the guitar track with the original recording. Also when transcribing music to get the written note sheets it is a great benefit to have the individual instruments in separate channels. Transcription can be of value both for musicians and for people wishing to compare (search in) music. On a less ambitious level identifying the instruments and finding the identity of the vocalist may aid in classifying the music and again make search in music possible. For all these applications, separation of music into its basic components is interesting. We find that the most important application of music separation is as a preprocessing step.

Examples can be found where music consists of a single instrument only, and much of the literature on signal processing of music deals with these examples. However, in the vast majority of music several instruments are played together, each instrument has its own unique sound and it is these sounds in unison that produce the final piece. Some of the instruments are playing at a high pitch and

* This work is supported by the Danish Technical Research Council (STVF), through the framework project "Intelligent Sound", STVF no. 26-04-0092 and the Oticon Foundation.

some at a low, some with many overtones some with few, some with sharp onset and so on. The individual instruments furthermore each play their own part in the final piece. Sometimes the instruments are played together and sometimes they are played alone. Common for all music is that the instruments are not all playing at the same time. This means that the instruments to some extent are separated in time and frequency. In most modern productions the instruments are recorded separately in a controlled studio environment. Afterwards the different sources are mixed into a stereo signal. The mixing typically puts the most important signal in the center of the sound picture hence often the vocal part is located here perhaps along with some of the drums. The other instruments are placed spatially away from the center. The information gained from the fact that the instruments are distributed in both space, frequency and time can be used to separate them.

Independent component analysis (ICA) is a well-known technique to separate mixtures consisting of several signals into independent components [1]. The most simple ICA model is the instantaneous ICA model. Here the vector $\mathbf{x}(n)$ of recorded signals at the discrete time index n is assumed to be a linear superposition of each of the sources $\mathbf{s}(n)$ as

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \boldsymbol{\nu}(n), \tag{1}$$

where $\mathbf{A}$ is the mixing matrix and $\boldsymbol{\nu}(n)$ is additional noise. If reverberations and delays between the microphones are taken into account, each recording is a mixture of different filtered versions of the source signals. This model is termed the *convolutive* mixing model.

The separation of music pieces by ICA and similar methods has so far not received much attention. In the first attempts ICA was applied to separation of mixed audio sources [2]. A standard (non-convolutive) ICA algorithm is applied to the time-frequency distribution (spectrogram) of different music pieces. The resulting model has a large number of basis functions and corresponding source signals. Many of these arise from the same signal and thus a postprocessing step tries to cluster the components. The system is evaluated by listening tests by the author and by displaying the separated waveforms. Plumbley et al. [3] presents a range of methods for music separation, among these are an ICA approach. Their objective is to transcribe a polyphonic single instrument piece. The convolutive ICA model is trained on a midi synthesized piece of piano music. Mostly, only a single note is played making it possible for the model to identify the notes as a basis. The evaluation by comparing the transcription to the original note sheets showed good although not perfect performance. Smaragdis et al. has presented both an ICA approach [4] and a Non-negative Matrix Factorization (NMF) approach [5] to music separation. The NMF works on the power spectrogram assuming that the sources are additive. In [6] the idea is extended to use convolutive NMF. The NMF approach is also pursued in [7] where an artificial mixture of a flute and piano is separated and in [8] where the drums are separated from polyphonic music. In [9] ICA/NMF is used along with a vocal discriminant to extract the vocal.

Time-Frequency (T-F) masking is another method used to segregate sounds from a mixture (see e.g. [10]). In *computational auditory scene analysis*, the technique of T-F masking has been commonly used for years. Here, source separation is based on organizational cues from auditory scene analysis [11]. When the source signals do not overlap in the time-frequency domain, high-quality reconstruction can be obtained [12]. However, when there are overlaps between the source signals good separation can still be obtained by applying a binary time-frequency mask to the mixture [12, 13]. Binary masking is also consistent with constraints from auditory scene analysis such as people's ability to hear and segregate sounds [14]. More recently the technique has also become popular in blind source separation, where separation is based on non-overlapping sources in the T-F domain [15]. T-F masking is applicable to source separation/ segregation using one microphone [10, 16] or more than one microphone [12, 13]. In order to segregate stereo music into independent components, we propose a method to combine ICA with T-F masking in order to iterative separate music into spatially independent components. ICA and T-F masking has previously been combined. In [**?**], ICA has been applied to separate two signals from two mixtures. Based on the ICA outputs, T-F masks are estimated and a mask is applied to each of the ICA outputs in order to improve the signal to noise ratio.

Section 2 provides a review of ICA on stereo signals. In section 3 it is described how to combine ICA with masking in the time frequency domain. In section 4 the algorithm is tested on real music. The result is evaluated by comparing the separated signals to the true recordings given by the master tape containing the individual instruments.

2 ICA on stereo signals

In stereo music, different music sources (song and instruments) are mixed so that the sources are located at spatially different positions. Often the sounds are recorded separately and mixed afterwards. A simple way to create a stereo mixture is to select different amplitudes for the two signals in the mixture. Therefore, we assume that the stereo mixture $\mathbf{x}$ at the discrete time index n can be modeled as an instantaneous mixture as in eqn. (1), i.e.

$$\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1N} \\ a_{21} \cdots a_{2N} \end{bmatrix} \begin{bmatrix} s_1(n) \\ \vdots \\ s_N(n) \end{bmatrix} + \begin{bmatrix} \nu_1(n) \\ \nu_2(n) \end{bmatrix}. \tag{2}$$

Each row in the mixing matrix $[a_{1i}\ a_{2i}]^T$ contains the gain of the i'th source in the stereo channels. The additional noise could e.g. be music signals which do not origin from a certain direction. If the gain ratio a_{1i}/a_{2i} of the i'th source is different from the gain ratio from any other source, we can segregate this source from the mixture. A piece of music often consists of several instruments as well as singing voice. Therefore, it is likely that the number of sources is greater than two. Hereby we have an *underdetermined* mixture. In [17] it was shown

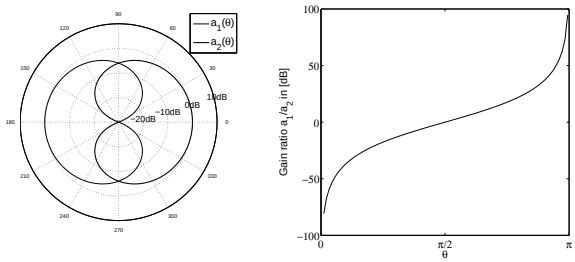

Fig. 1. The two stereo responses $a_1(\theta)$ and $a_2(\theta)$ are shown as function of the direction θ. The monotonic gain ratio is shown as function of the direction θ.

how to extract speech signals iteratively from an underdetermined instantaneous mixture of speech signals. In [17] it was assumed that a particular gain ratio a_{1i}/a_{2i} corresponded to a particular spatial source location. An example of such a location-dependant gain ratio is shown in Fig 1. This gain ratio is obtained by selecting the two gains as $a_1(\theta) = 0.5\big(1 - \cos(\theta)\big)$ and $a_2(\theta) = 0.5\big(1 + \cos(\theta)\big)$.

2.1 ICA solution as an adaptive beamformer

When there are no more sources than sensors, an estimate $\tilde{\boldsymbol{s}}(n)$ of the original sources can be found by applying a (pseudo) inverse linear system, to eqn. (1).

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n) = \mathbf{W}\mathbf{A}\mathbf{s}(n) \tag{3}$$

where $\mathbf{W}$ is a 2×2 separation matrix. From eqn. (3) we see that the output $\mathbf{y}$ is a function of $\mathbf{s}$ multiplied by $\mathbf{WA}$. Hereby we see that $\mathbf{y}$ is just a different weighting of $\mathbf{s}$ than $\mathbf{x}$ is. If the number of sources is greater than the number of mixtures, not all the sources can be segregated. Instead, an ICA algorithm will estimate $\mathbf{y}$ as two subsets of the mixtures which are as independent as possible, and these subsets are weighted functions of $\mathbf{s}$. The ICA solution can be regarded as an adaptive beamformer which in the case of underdetermined mixtures places the zero gain directions towards different groups of sources. By comparing the two outputs, two binary masks can be found in the T-F domain. Each mask is able to remove the group of sources towards which one of the ICA solutions places a zero gain direction.

3 Extraction with ICA and binary masking

A flowchart of the algorithm is presented in Fig. 2. As described in the previous section, a two-input-two-output ICA algorithm is applied to the input mixtures, disregarding the number of source signals that actually exist in the mixture. As

shown below the binary mask is estimated by comparing the amplitudes of the two ICA outputs and hence it is necessary to deal with the arbitrary scaling of the ICA algorithm. As proposed in [1], we assume that all source signals have the same variance and the outputs are therefore scaled to have the same variance. From the two re-scaled output signals, $\hat{y}_1(n)$ and $\hat{y}_2(n)$, spectrograms are obtained by use of the Short-Time Fourier Transform (STFT):

$$y_1 \rightarrow \hat{y}_1 \rightarrow Y_1(\omega, t) \tag{4}$$

$$y_2 \rightarrow \hat{y}_2 \rightarrow Y_2(\omega, t), \tag{5}$$

where ω is the frequency and t is the time index. The binary masks are then found by a bitwise amplitude comparison between the two spectrograms:

$$\mathrm{BM1}(\omega, t) = \tau|Y_1(\omega, t)| > |Y_2(\omega, t)| \tag{6}$$

$$\mathrm{BM2}(\omega, t) = \tau|Y_2(\omega, t)| > |Y_1(\omega, t)|, \tag{7}$$

where τ is a threshold that determines the sparseness of the mask. As τ is increased, the mask is sparser. We have chosen $\tau = 1.5$. Next, each of the two binary masks is applied to the original mixtures x_1 and x_2 in the T-F domain, and by this non-linear processing, some of the music signal are *removed* by one of the masks while other parts of music are removed by the other mask. After the masks have been applied to the signals, they are reconstructed in the time domain by the inverse STFT and and two sets of masked output signals $(\hat{x}_{11}, \hat{x}_{21})$ and $(\hat{x}_{12}, \hat{x}_{22})$ are obtained.

In the next step, it is considered whether the masked output signals consists of more than one signal. The masked output signals are divided into three group defined by the selection criterion in section 3.1. It is decided whether there is one signal in the segregated output signal, more than one signal in the segregated output, or if the segregated signal contains too little energy, so that the signal is expected to be of too poor quality.

There is no guarantee that two different outputs are not different parts of the same separated source signal. By considering the correlation between the segregated signals in the time domain, it is decided whether two outputs contains the same signal. If so, their corresponding two masks are merged. Also the correlation between the segregated signals and the signals with too poor quality is considered. From the correlation coefficient, it is decided whether the mask of the segregated signal is extended by merging the mask of the signal of poor quality. Hereby the overall quality of the new mask is higher.

When no more signal consist of more than one signal, the separation procedure stops. After the correlation between the output signals have been found, some masks still have not been assigned to any of the source signal estimates. All these masks are then combined in order to create a *background mask*. The background mask is then applied to the original two mixtures, and possible sounds that remain in the background mask are found. The separation procedure is then applied to the remaining signal to ensure that there is no further signal hidden. This procedure is continued until the remaining mask does not change any more. Note that the final output signals are maintained as stereo signals.

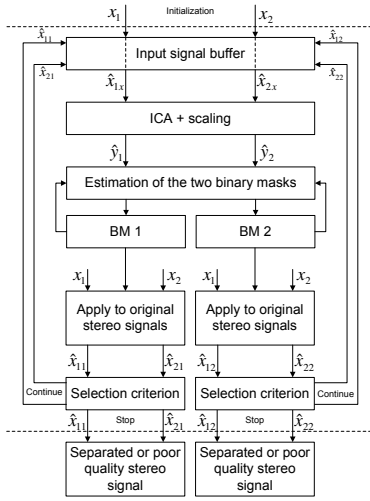

Fig. 2. Flowchart showing the main steps of the algorithm. From the output of the ICA algorithm, binary masks are estimated. The binary masks are applied to the original signals which again are processed through the ICA step. Every time the output from one of the binary masks is detected as a single signal, the signal is stored. The iterative procedure stops when all outputs only consist of a single signal. The flowchart has been adopted from [17].

3.1 Selection criterion

It is important to decide whether the algorithm should stop or whether the processing should proceed. The algorithm should stop separating when the signal consists of only one source or when the mask is too sparse so that the quality of the resulting signal is unsatisfactory. Otherwise, the separation procedure should proceed. We consider the covariance matrix between the output signals to which the binary mask has been applied, i.e. $\mathbf{R}_{xx} = \langle \mathbf{x}\mathbf{x}^H \rangle$. If the covariance matrix is close to singular, it indicates that there is only one source signal. To measure the singularity, we find the condition number of $\mathbf{R}_{xx}$. If the condition number is below a threshold, it is decided that $\mathbf{x}$ contains more than one signal and the separation procedure continues. Otherwise, it is assumed that $\mathbf{x}$ consists of a single source and the separation procedure stops.

4 Results

The method has been applied to different pieces of music. The used window length was 512, the FFT length was 2048. The overlap between time frames

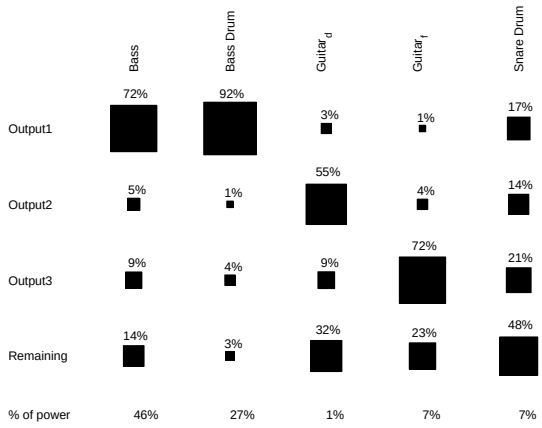

Fig. 3. Correlation coefficients between the extracted channels and the original stereo channels. The coefficients has been normalized such that the columns sum to one. The last row shows the percentage of power of the tracks in the mixture.

was 75%. The sampling frequency is 10 kHz. Listening tests confirm that the method is able to segregate individual instruments from the stereo mixture. We do not observe that correlations can be heard. However, musical artifacts are audible. Examples are available on-line for subjective evaluation [18]. In order to evaluate the method objectively, the method has been applied to 5 seconds of stereo music, where each of the different instruments has been recorded separately, processed from a mono signal into a stereo signal, and then mixed. We evaluate the performance by calculating the correlation between the segregated channels and the original tracks. The results are shown in Fig. 3 As it can be seen from the figure, the correlation between the estimated channels and the original channels is quite high. The best segregation has been obtained for those channels, where the two channels are made different by a gain difference. Among those channels is the guitars, which are well segregated from the mixture. The more omnidirectional (same gain from all directions) stereo channels cannot be segregated by our method. However, those channels are mainly captured in the remaining signal, which contains what is left when the other sources has been segregated. Some of the tracks have the same gain difference. Therefore, it is hard to segregate the 'bass' from the 'bass drum'.

5 Conclusion

We have presented an approach to segregate single sound tracks from a stereo mixture of different tracks while keeping the extracted signals as stereo signals.

The method utilizes that music is sparse in the time, space and frequency domain by combining ICA and binary time-frequency masking. It is designed to separate tracks from mixtures where the stereo effect is based on a gain difference. Experiments verify that real music can be separated by this algorithm and results on an artificial mixture reveals that the separated channel is highly correlated with the original recordings.

We believe that this algorithm can be a useful preprocessing tool for annotation of music or for detecting instrumentation.

References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley (2001)
2. Casey, M., Westner, A.: Separation of mixed audio sources by independent subspace analysis. In: Proc. ICMC. (2000)
3. Plumbley, M.D., Abdallah, S.A., Bello, J.P., Davies, M.E., Monti, G., Sandler, M.B.: Automatic music transcription and audio source separation. Cybernetics and Systems **33** (2002) 603–627
4. Smaragdis, P., Casey, M.: Audio/visual independent components. Proc. ICA'2003 (2003) 709–712
5. Smaragdis, P., Brown, J.C.: Non-negative matrix factorization for polyphonic music transcription. Proc. WASPAA 2003 (2003) 177–180
6. Smaragdis, P.: Non-negative matrix factor deconvolution; extraction of multiple sound sourses from monophonic inputs. Proc. ICA'2004 (2004) 494–499
7. Wang, B., Plumbley, M.D.: Musical audio stream separation by non-negative matrix factorization. In: Proc. DMRN Summer Conf. (2005)
8. Helén, M., Virtanen, T.: Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In: Proc. EUSIPCO'2005. (2005)
9. Vembu, S., Baumann, S.: Separation of vocals from polyphonic audio recordings. In: Proc. ISMIR2005. (2005) 337–344
10. Wang, D.L., Brown, G.J.: Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Networks **10** (1999) 684–697
11. Bregman, A.S.: Auditory Scene Analysis. 2 edn. MIT Press (1990)
12. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Processing **52** (2004) 1830–1847
13. Roman, N., Wang, D.L., Brown, G.J.: Speech segregation based on sound localization. J. Acoust. Soc. Amer. **114** (2003) 2236–2252
14. Wang, D.L.: On ideal binary mask as the computational goal of auditory scene analysis. In Divenyi, P., ed.: Speech Separation by Humans and Machines. Kluwer, Norwell, MA (2005) 181–197
15. Jourjine, A., Rickard, S., Yilmaz, O.: Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In: Proc. ICASSP. (2000) 2985–2988
16. Hu, G., Wang, D.L.: Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Networks **15** (2004) 1135–1150
17. Pedersen, M.S., Wang, D., Larsen, J., Kjems, U.: Overcomplete blind source separation by combining ICA and binary time-frequency masking. In: Proc. MLSP. (2005)
18. http://www.intelligentsound.org/demos/demos.htm.

Appendix E

Separating Underdetermined Convolutive Speech Mixtures

Separating Underdetermined Convolutive Speech Mixtures

Michael Syskind Pedersen[1,2], DeLiang Wang[3], Jan Larsen[1], and Ulrik Kjems[2]

[1] Informatics and Mathematical Modelling, Technical University of Denmark
Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark
[2]Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark
[3] Department of Computer Science and Engineering & Center for Cognitive Science,
The Ohio State University, Columbus, OH 43210-1277, USA
{msp,jl}@imm.dtu.dk, uk@oticon.dk, dwang@cse.ohio-state.edu

Abstract. A limitation in many source separation tasks is that the number of source signals has to be known in advance. Further, in order to achieve good performance, the number of sources cannot exceed the number of sensors. In many real-world applications these limitations are too restrictive. We propose a method for underdetermined blind source separation of convolutive mixtures. The proposed framework is applicable for separation of instantaneous as well as convolutive speech mixtures. It is possible to iteratively extract each speech signal from the mixture by combining *blind source separation* techniques with *binary time-frequency masking*. In the proposed method, the number of source signals is not assumed to be known in advance and the number of sources is not limited to the number of microphones. Our approach needs only two microphones and the separated sounds are maintained as stereo signals.

1 Introduction

Blind source separation (BSS) addresses the problem of recovering N unknown source signals $\mathbf{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ from M recorded mixtures $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^T$ of the source signals. The term 'blind' refers to that only the recorded mixtures are known. An important application for BSS is separation of speech signals. The recorded mixtures are assumed to be linear superpositions of the source signals. Such a linear mixture can either be instantaneous or convolutive. The instantaneous mixture is given as

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \boldsymbol{\nu}(n), \tag{1}$$

where $\mathbf{A}$ is an $M \times N$ mixing matrix and n denotes the discrete time index. $\boldsymbol{\nu}(n)$ is additional noise. A method to retrieve the original signals up to an arbitrary permutation and scaling is independent component analysis (ICA) [1]. In ICA, the main assumption is that the source signals are independent. By applying ICA, an estimate $\mathbf{y}(n)$ of the source signals can be obtained by finding a (pseudo)inverse $\mathbf{W}$ of the mixing matrix so that

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n). \tag{2}$$

Notice, this inversion is not exact when noise is included in the mixing model. When noise is included as in (1), $\mathbf{x}(n)$ is a nonlinear function of $\mathbf{s}(n)$. Still, the inverse system is assumed to be approximated by a linear system.

The convolutive mixture is given as

$$\mathbf{x}(n) = \sum_{k=0}^{K-1} \mathbf{A}_k \mathbf{s}(n-k) + \boldsymbol{\nu}(n) \tag{3}$$

Here, the source signals are mixtures of filtered versions of the original source signals. The filters are assumed to be causal and of finite length K. The convolutive mixture is more applicable for separation of speech signals because the convolutive model takes reverberations into account. The separation of convolutive mixtures can either be performed in the time or in the frequency domain. The separation system for each discrete frequency ω is given by

$$\mathbf{Y}(\omega, t) = \mathbf{W}(\omega)\mathbf{X}(\omega, t), \tag{4}$$

where t is the time frame index. Most methods, both instantaneous and convolutive, require that the number of source signals is known in advance. Another drawback of most of these methods is that the number of source signals is assumed not to exceed the number of microphones, i.e. $M \geq N$.

If $N > M$, even if the mixing process is known, it may not be invertible, and the independent components cannot be recovered exactly [1]. In the case of more sources than sensors, the *underdetermined/overcomplete* case, successful separation often relies on the assumption that the source signals are sparsely distributed in the time-frequency domain [2], [3]. If the source signals do not overlap in the time-frequency domain, high-quality reconstruction could be obtained [3].

However, there is overlap between the source signals. In this case, good separation can still be obtained by applying a binary time-frequency (T-F) mask to the mixture [2], [3]. In *computational auditory scene analysis*, the technique of T-F masking has been commonly used for years (see e.g. [4]). Here, source separation is based on organizational cues from auditory scene analysis [5]. More recently the technique has also become popular in blind source separation, where separation is based on non-overlapping sources in the T-F domain [6]. T-F masking is applicable to source separation/ segregation using one microphone [4],[7],[8] or more than one microphone [2], [3]. T-F masking is typically applied as a binary mask. For a binary mask, each T-F unit is either weighted by one or zero. An advantage of using a binary mask is that only a binary decision has to be made [9]. Such a decision can be based on, e.g., clustering [2], [3], [6], or direction-of-arrival [10]. ICA has been used in different combinations with the binary mask. In [10], separation is performed by first removing $N-M$ signals via masking and afterwards applying ICA in order to separate the remaining M signals. ICA has also been used in the other way around. In [11], it has been applied to separate two signals by using two microphones. Based on the ICA outputs, T-F masks are estimated and a mask is applied to each of the ICA outputs in order to improve the signal to noise ratio (SNR).

In this paper, we propose a method to segregate an arbitrary number of speech signals in a reverberant environment. We extend a previously proposed method for separation of instantaneous mixtures [12] to separation of convolutive mixtures. Based on the output of a square (2×2) blind source separation algorithm and binary T-F masks, our method segregates speech signals iteratively from the mixtures until an estimate of each signal is obtained.

2 Blind Extraction by combining BSS and binary masking

With only two microphones, it is not possible to separate more than two signals from each other because only one null direction can be placed for each output. This fact does not mean that the blind source separation solution is useless in the case of $N > M$. In [12] we examined what happened if an ICA algorithm was applied to an underdetermined 2-by-N mixture. When the two outputs were considered, we found that the ICA algorithm separates the mixtures into subspaces, which are as independent as possible. Some of the source signals are mainly in one output while other sources mainly are present in the other output.

A flowchart for the algorithm is given in Fig. 1. As described in the previous section, a two-input-two-output blind source separation algorithm has been applied to the input mixtures, regardless the number of source signals that actually exist in the mixture. The two output signals are arbitrarily scaled. Different methods have been proposed in order to solve the scaling ambiguity. Here, we assume that all source signals have the same variance as proposed in [1] and the outputs are therefore scaled to have the same variance.

The two re-scaled output signals, $\hat{y}_1(n)$ and $\hat{y}_2(n)$, are transformed into the frequency domain e.g. using the Short-Time Fourier Transform STFT so that two spectrograms are obtained:

$$\hat{y}_1 \rightarrow Y_1(\omega, t) \tag{5}$$

$$\hat{y}_2 \rightarrow Y_2(\omega, t), \tag{6}$$

where ω denotes the frequency and t is the time frame index. The binary masks are then determined for each T-F unit by comparing the amplitudes of the two spectrograms:

$$\mathrm{BM1}(\omega, t) = \tau |Y_1(\omega, t)| > |Y_2(\omega, t)| \tag{7}$$

$$\mathrm{BM2}(\omega, t) = \tau |Y_2(\omega, t)| > |Y_1(\omega, t)|, \tag{8}$$

where τ is a threshold. Next, each of the two binary masks is applied to the original mixtures in the T-F domain, and by this non-linear processing, some of the speech signals are *removed* by one of the masks while other speakers are removed by the other mask. After the masks have been applied to the signals, they are reconstructed in the time domain by the inverse STFT. If there is only a single signal left in the masked output, defined by the selection criteria

in Section 2.3, i.e. all but one speech signal have been masked, this signal is considered extracted from the mixture and it is saved. If there are more than one signal left in the masked outputs, the procedure is applied to the two masked signals again and a new set of masks are created based on (7), (8) and the previous masks. The use of the previous mask ensures that T-F units that have been removed from the mixture are not reintroduced by the next mask. This is done by an element-wise multiplication between the previous mask and the new mask. This iterative procedure is followed until all masked outputs consist of only a single speech signal. When the procedure stops, the correlation between the segregated sources are found in order to determine whether a source signal has been segregated more than once. If so, the source is re-estimated by merging the two correlated masks. It is important to notice that the iteratively updated mask always is applied to the original mixtures and not to the previously masked signal. Hereby a deterioration of the signal due to multiple iterations is avoided.

2.1 Finding the background signals

Since some signals may have been removed by both masks, all T-F units that have not been assigned the value '1' are used to create a *background mask*, and the procedure is applied to the mixture signal after the remaining mask is applied, to ensure that all signals are estimated. Notice that this step has been omitted from Fig. 1.

2.2 Extension to convolutive mixtures

Each convolutive mixture is given by a linear superposition of filtered versions of each of the source signals. The filters are given by the impulse responses from each of the sources to each of the microphones. An algorithm capable of separating convolutive mixtures is used in the BSS step. Separation still relies on the fact that the source signals can be grouped such that one output mainly contains one part of the source signals and the other output mainly contains the other part of the signals. In order to avoid arbitrary filtering, only the cross channels of the separation filters have been estimated. The direct channel is constrained to be an impulse. Specifically, we employ the frequency domain convolutive BSS algorithm by Parra and Spence [13][1].

2.3 Selection criterion

In order to decide if all but one signal have been removed, we consider the envelope statistics of the signal. By considering the envelope histogram, it can be determined whether one or more than one signal is present in the mixture. If only one speech signal is present, many of the amplitude values are close to zero.

[1] Matlab code is available from `http://ida.first.gmd.de/~harmeli/download/download_convbss.html`

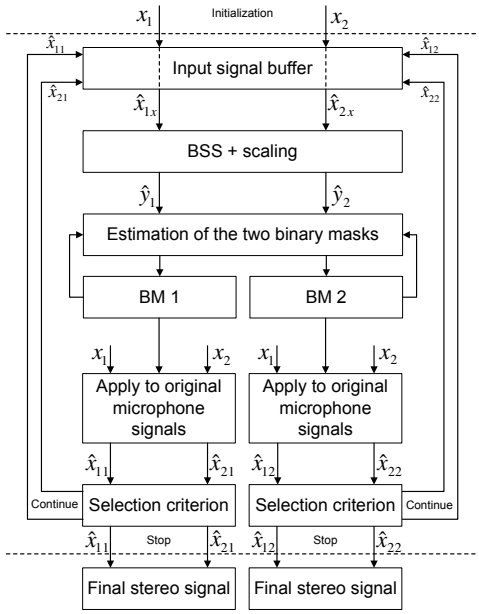

Fig. 1. Flowchart showing the main steps of the proposed algorithm. From the output of the BSS algorithm, binary masks are estimated. The binary masks are applied to the original signals which again are processed through the BSS step. Every time the output from one of the binary masks is detected as a single signal, the signal is stored. The iterative procedure stops when all outputs only consist of a single signal. The flowchart has been adopted from [12].

If more speech signals are present, less amplitude values are close to zero. In order to discriminate between one and more than one speech signals in the mixture, we measure the width of the histogram as proposed in [14] as the distance between the 90% and the 10% percentile normalized to the 50% percentile, i.e.

$$\text{width} = \frac{P_{90} - P_{10}}{P_{50}}. \tag{9}$$

Further processing on a pair of masked signals should be avoided if there is one or zero speech signals in the mixture. If the calculated width is smaller than two, we assume that the masked signal consists of more than one speech signal. We discriminate between zero and one signal by considering the energy of the segregated signal. This selection criterion is more robust to reverberations than the correlation-based criterion used in [12].

3 Evaluation

The algorithm described above has been implemented and evaluated with instantaneous and convolutive mixtures. For the STFT, an FFT length of 2048 has been used. A Hanning window with a length of 512 samples has been applied to the FFT signal and the frame shift is 256 samples. A high frequency resolution is found to be necessary in order to obtain good performance. The sampling frequency of the speech signals is 10 kHz, and the duration of each signal is 5 s. The thresholds have been found from initial experiments. In the ICA step, the separation matrix is initialized by the identity matrix, i.e. $\mathbf{W} = \mathbf{I}$. When using a binary mask, it is not possible to reconstruct the speech signal as if it was recorded in the absence of the interfering signals, because the signals partly overlap. Therefore, as a computational goal for source separation, we employ the *ideal binary mask*[9]. The ideal binary mask for a signal is found for each T-F unit by comparing the energy of the desired signal to the energy of all the interfering signals. Whenever the signal energy is higher, the T-F unit is assigned the value '1' and whenever the interfering signals have more energy, the T-F unit is assigned the value '0'. As in [8], for each of the separated signals, the percentage of energy loss P_{EL} and the percentage of noise residue P_{NR} are calculated as well as the signal to noise ratio (SNR) using the resynthesized speech from the ideal binary mask as the ground truth:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)}, \quad P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)}, \quad \text{SNR} = 10 \log_{10} \left[\frac{\sum_n I^2(n)}{\sum_n (I(n) - O(n))^2} \right],$$

where $O(n)$ is the estimated signal, and $I(n)$ is the recorded mixture resynthesized after applying the ideal binary mask. $e_1(n)$ denotes the signal present in $I(n)$ but absent in $O(n)$ and $e_2(n)$ denotes the signal present in $O(n)$ but absent in $I(n)$. The input signal to noise ratio, SNR_i, is found too, which is the ratio between the desired signal and the noise in the recorded mixtures.

Table 1. Separation results for four convolutively mixed speech mixtures. A manual selection criterion was used.

Signal No.	P_{EL}(%)	P_{NR}(%)	SNR_i (dB)	SNR (dB)
1	66.78	20.41	-4.50	1.35
2	32.29	41.20	-4.50	1.24
3	52.86	19.08	-3.97	2.12
4	15.78	30.39	-6.67	2.91
Average	41.93	27.77	-4.91	1.91

Convolutive mixtures consisting of four speech signals have also been separated. The signals are uniformly distributed in the interval $0° \leq \theta \leq 180°$. The mixtures have been obtained with room impulse responses synthesized using the image model [15]. The estimated room reverberation time is $T_{60} \approx 160$ ms. The

Table 2. Separation results for four convolutively mixed speech mixtures. The selection criterion as proposed in Section 2.3 was used.

Signal No.	P_{EL}(%)	P_{NR}(%)	SNR$_i$ (dB)	SNR (dB)
1	39.12	46.70	-4.50	0.63
2	64.18	18.62	-4.50	1.45
3	26.88	33.73	-3.97	2.40
4	45.27	32.49	-6.67	1.69
Average	43.86	32.88	-4.91	1.54

distance between the microphones is 20 cm. The method has been evaluated with and without the proposed selection criterion described in Section 2.3. When the selection criterion was not used, it has been decided when a source signal has been separated by listening to the signals. The separation results are shown in Table 1 and Table 2. The average input SNR is -4.91 dB. When the selection criterion was applied manually, the average SNR after separation is 1.91 dB with an average SNR gain of 6.8 dB. When selection criterion was applied as proposed, the average SNR after separation is 1.45 dB with an average SNR gain of 6.4 dB, which is about half a dB worse than selecting the segregated signals manually. It is not always that all the sources are extracted from the mixture. Therefore the selection criterion could be further improved. For separation of instantaneous mixtures an SNR gain of 14 dB can be obtained, which is significantly higher than that for the reverberant case. This may be explained by several factors. Errors such as misaligned permutations are introduced from the BSS algorithm. Also, convolutive mixtures are not as sparse in the T-F domain as instantaneous mixtures. Further, the assumption that the same signals group into the same groups for all frequencies may not hold. Some artifacts (musical noise) exist in the segregated signals. Especially in the cases, where the values of P_{EL} and P_{NR} are high. Separation results are available for listening at `www.imm.dtu.dk/~msp`.

As mentioned earlier, several approaches have been recently proposed to separate more than two sources using two microphones by employing binary T-F masking [2], [3], [10]. These methods use clustering of amplitude and time differences between the microphones. In contrast, our method separates speech mixtures by iteratively extracting individual source signals. Our results are quite competitive although rigorous statements about comparison are difficult because the test conditions are different.

4 Concluding remarks

A novel method of blind source separation of underdetermined mixtures has been described. Based on sparseness and independence, the method iteratively extracts all the speech signals. The linear processing from BSS methods alone cannot separate more sources than the number of recordings, but with the additional nonlinear processing introduced by the binary mask, it is possible to separate more sources than the number of sensors. Our method is applicable

to separation of instantaneous as well as convolutive mixtures and the output signals are maintained as stereo signals. An important part of the method is the detection of when a single signal exists at the output. Future work will include better selection criteria to detect a single speech signal, especially in a reverberant environment. More systematic evaluation and comparison will also be given in the future. The assumption of two microphones may be relaxed and the method may also be applicable to other signals than speech which also have significant redundancy.

Acknowledgements The work was performed while M.S.P. was a visiting scholar at The Ohio State University Department of Computer Science and Engineering. M.S.P was supported by the Oticon Foundation. M.S.P and J.L are partly also supported by the European Commission through the sixth framework IST Network of Excellence: PASCAL. D.L.W was supported in part by an AFOSR grant and an AFRL grant.

References

1. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley (2001)
2. Roman, N., Wang, D.L., Brown, G.J.: Speech segregation based on sound localization. J. Acoust. Soc. Amer. **114** (2003) 2236–2252
3. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Processing **52** (2004) 1830–1847
4. Wang, D.L., Brown, G.J.: Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Networks **10** (1999) 684–697
5. Bregman, A.S.: Auditory Scene Analysis. 2 edn. MIT Press (1990)
6. Jourjine, A., Rickard, S., Yilmaz, O.: Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In: Proc. ICASSP. (2000) 2985–2988
7. Roweis, S.: One microphone source separation. In: NIPS'00. (2000) 793–799
8. Hu, G., Wang, D.L.: Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Networks **15** (2004) 1135–1150
9. Wang, D.L.: On ideal binary mask as the computational goal of auditory scene analysis. In Divenyi, P., ed.: Speech Separation by Humans and Machines. Kluwer, Norwell, MA (2005) 181–197
10. Araki, S., Makino, S., Sawada, H., Mukai, R.: Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA. In: Proc. ICA'2004. (2004) 898–905
11. Kolossa, D., Orglmeister, R.: Nonlinear postprocessing for blind speech separation. In: Proc. ICA'2004, Granada, Spain (2004) 832–839
12. Pedersen, M.S., Wang, D.L., Larsen, J., Kjems, U.: Overcomplete blind source separation by combining ICA and binary time-frequency masking. In: Proceedings of the MLSP workshop, Mystic, CT, USA (2005)
13. Parra, L., Spence, C.: Convolutive blind separation of non-stationary sources. IEEE Trans. Speech and Audio Processing **8** (2000) 320–327
14. Büchler, M.C.: Algorithms for Sound Classification in Hearing Instruments. PhD thesis, Swiss Federal Institute of Technology, Zurich (2002)
15. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Amer. **65** (1979) 943–950

Appendix F

Two-Microphone Separation of Speech Mixtures

Two-microphone Separation of Speech Mixtures

Michael Syskind Pedersen, DeLiang Wang, Jan Larsen, and Ulrik Kjems

***Abstract*— Separation of speech mixtures, often referred to as the cocktail party problem, has been studied for decades. In many source separation tasks, the separation method is limited by the assumption of at least as many sensors as sources. Further, many methods require that the number of signals within the recorded mixtures be known in advance. In many real-world applications these limitations are too restrictive. We propose a novel method for underdetermined blind source separation using an instantaneous mixing model which assumes closely spaced microphones. Two source separation techniques have been combined, *independent component analysis* (ICA) and *binary time-frequency masking*. By estimating binary masks from the outputs of an ICA algorithm, it is possible in an *iterative* way to extract basis speech signals from a convolutive mixture. The basis signals are afterwards improved by grouping similar signals. Using two microphones we can separate in principle an arbitrary number of mixed speech signals. We show separation results for mixtures with as many as seven speech signals under instantaneous conditions. We also show that the proposed method is applicable to segregate speech signals under reverberant conditions, and we compare our proposed method to another state-of-the-art algorithm. The number of source signals is not assumed to be known in advance and it is possible to maintain the extracted signals as stereo signals.**

***Index Terms*— Underdetermined speech separation, ICA, time-frequency masking, ideal binary mask.**

I. Introduction

THE problem of extracting a single speaker from a mixture of many speakers is often referred to as the cocktail party problem [1], [2]. Human listeners cope remarkably well in adverse environments, but when the noise level is exceedingly high, human speech intelligibility also suffers. By extracting speech sources from a mixture of many speakers, we can potentially increase the intelligibility of each source by listening to the separated sources.

Blind source separation addresses the problem of recovering N unknown source signals $\mathbf{s}(n) = [s_1(n), \ldots, s_N(n)]^T$ from M recorded mixtures $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^T$ of the source signals. n denotes the discrete time index. Each of the recorded mixtures $x_i = x_i(n)$ consists of $N_s = f_s T$ samples, where f_s is the sampling frequency and T denotes the duration in seconds. The term 'blind' refers to that only the recorded mixtures are known. The mixture is assumed to be a linear superposition of the source signals, sometimes with additional noise, i.e.,

$$\mathbf{x}(n) = \mathbf{A}\mathbf{s}(n) + \boldsymbol{\nu}(n), \tag{1}$$

Michael Syskind Pedersen and Ulrik Kjems are with Oticon A/S, Kongebakken 9, DK-2765, Denmark. Email: {msp, uk}@oticon.dk, DeLiang Wang is with the Department of Computer Science & Engineering and the Center for Cognitive Science, The Ohio State University, Columbus, OH 43210-1277, U.S.A. Email: dwang@cse.ohio-state.edu. Jan Larsen is with the Intelligent Signal Processing Group at the Department of Informatics and Mathematical Modelling, Technical University of Denmark, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark. Email: jl@imm.dtu.dk.

where $\mathbf{A}$ is an $M \times N$ mixing matrix. $\boldsymbol{\nu}(n)$ is additional noise. Also, $\mathbf{A}$ is assumed not to vary as function of time. Often, the objective is to estimate one or all of the source signals. An estimate $\mathbf{y}(n)$ of the original source signals can be found by applying an (pseudo) inverse linear operation, i.e.,

$$\mathbf{y}(n) = \mathbf{W}\mathbf{x}(n), \tag{2}$$

where $\mathbf{W}$ is an $N \times M$ separation matrix. Notice that this inversion is not exact when noise is included in the mixing model. When noise is included as in (1), $\mathbf{y}(n)$ is a nonlinear function of $\mathbf{x}(n)$ [3]. In this paper, the inverse is approximated by a linear system.

In real environments, a speech signal does not only arrive from a single direction. Rather, multiple reflections from the surroundings occur as delayed and filtered versions of the source signal. In this situation, the mixing model is better approximated by a *convolutive* mixing model. The convolutive FIR mixture is given as

$$\mathbf{x}(n) \quad = \quad \sum_{k=0}^{K-1} \mathbf{A}_k \mathbf{s}(n-k) + \boldsymbol{\nu}(n) \tag{3}$$

Here, the source signals are mixtures of filtered versions of the anechoic source signals. The filters are assumed to be causal and of finite length K. Numerous algorithms have been proposed to solve the convolutive problem [4], but few are able to cope with underdetermined as well as reverberant conditions [5]–[9].

Independent Component Analysis (ICA) describes a class of methods that retrieve the original signals up to an arbitrary permutation and scaling [10]. Successful separation relies on assumptions on the statistical properties of the source signals. To obtain separation, many ICA methods require that at most one source be Gaussian. Many algorithms assume that the source signals are independent or the source signals are non-Gaussian [11]–[14]. Other methods are able to separate the source signals using only second order statistics. Here, it is typically assumed that the sources have different correlation [15]–[17] or the source signals are non-stationary [18], [19]. Blind source separation algorithms have been applied in many areas such as feature extraction, brain imaging, telecommunications, and audio separation [10].

ICA methods have several drawbacks. Often, it is required that the number of source signals is known in advance and only few have addressed the problem of determining the number of sources in a mixture [20], [21]. Further, standard formulation requires that the number of source signals does not exceed the number of microphones. If the number of sources is greater than the number of mixtures, the mixture is called *underdetermined* (or *overcomplete*). In this case, the independent components cannot be recovered exactly without incorporating

additional assumptions, even if the mixing process **A** is known [10]. Additional assumptions include knowledge about the geometry, or detailed knowledge about the source distributions [22]. For example, the source signals are assumed to be sparsely distributed - either in the time domain, in the frequency domain or in the time-frequency (T-F) domain [8], [23]–[26]. Sparse sources have a limited overlap in the T-F domain. The validity of non-overlapping sources in the T-F domain comes from the observation that the spectrogram of a mixture is approximately equal to the maximum of the individual spectrograms in the logarithmic domain [27]. When the source signals do not overlap in the time-frequency domain, high-quality reconstruction can be obtained [8]. The property of non-overlapping sources in the T-F domain has been denoted as the W-disjoint orthogonality [28]. Given the short-time Fourier transform (STFT) of two speech signals $S_i(\omega, t)$ and $S_j(\omega, t)$, the W-disjoint orthogonality property can be expressed as

$$S_i(\omega, t)S_j(\omega, t) = 0, \forall i \neq j, \forall \omega, t, \tag{4}$$

where t is the time frame index and ω is the discrete frequency index. This property holds, for example, when tones are disjoint in frequency.

However, there is overlap between the source signals but good separation can still be obtained by applying a binary time-frequency mask to the mixture [24], [8]. In *computational auditory scene analysis* [29], the technique of T-F masking has been commonly used for many years (see e.g. [30]). Here, source separation is based on organizational cues from auditory scene analysis [31]. Binary masking is consistent with perceptual constraints regarding human ability to hear and segregate sounds [32]. Especially, time-frequency masking is closely related to the prominent phenomenon of auditory masking [33]. More recently the technique has also become popular in the ICA community to deal with non-overlapping sources in the T-F domain [28]. T-F masking is applicable to source separation/segregation using one microphone [30], [34], [35] or more than one microphone [8], [24]. T-F masking is typically applied as a binary mask. For a binary mask, each T-F unit (the signal element at a particular time and frequency) is either weighted by one or by zero. In order to reduce artifacts, soft masks may also be applied [36]. Also by decreasing the downsampling factor in the signal analysis and synthesis, a reduction of musical noise is obtained [37].

An advantage of using a T-F binary mask is that only a binary decision has to be made [32]. Such a decision can be based on clustering from different ways of direction-of-arrival estimation [8], [24], [28], [38]. ICA has been used in different combinations with the binary mask. In [38], separation is performed by removing $N - M$ signals by masking and then applying ICA in order to separate the remaining M signals. In [39], ICA has been used the other way around. Here, ICA is applied to separate two signals by using two microphones. Based on the ICA outputs, T-F masks are estimated and a mask is applied to each of the ICA outputs in order to improve the signal to noise ratio (SNR).

In this paper, we propose a novel approach to separating an arbitrary number of speech signals. Based on the output of a square (2×2) ICA algorithm and binary T-F masks, our approach iteratively segregates signals from a mixture until an estimate of each signal is obtained. Our method is applicable to both instantaneous and convolutive mixtures. A preliminary version of our work has been presented in [40], where we demonstrated the ability of our proposed framework to separate up to six speech mixtures from two instantaneous mixtures. In [41] it has been demonstrated that the approach can be used to segregate stereo music recordings into single instruments or singing voice. In [42] we described an extension to separate convolutive speech mixtures.

The paper is organized as follows. In Section II, we show how instantaneous real-valued ICA can be interpreted geometrically and how the ICA solution can be applied to underdetermined mixtures. In Sections III and IV we develop a novel algorithm that combines ICA and binary T-F masking in order to separate instantaneous as well as convolutive underdetermined speech mixtures. In Section V, we systematically evaluate the proposed method and compare it to existing methods. Further discussion is given in Section VI, and Section VII concludes the paper.

II. Geometrical interpretation of instantaneous ICA

We assume that there is an unknown number of acoustical source signals but only two microphones. It is assumed that each source signal arrives from a distinct direction and no reflections occur, i.e., we assume an anechoic environment in our mixing model. We assume that the source signals are mixed by an instantaneous time-invariant mixing matrix as in Eq. (1). Due to delays between the microphones, instantaneous ICA with a real-valued mixing matrix usually is not applicable to signals recorded at an array of microphones. Nevertheless, if the microphones are placed at the exact same location and have different gains for different directions, the separation of delayed sources can be approximated by the instantaneous mixing model [43]. Hereby, a combination of microphone gains corresponds to a certain directional pattern. The assumption that the microphones are placed at the exact same location can be relaxed. A similar approximation of delayed mixtures to instantaneous mixtures is provided in [44]. There, the differences between closely spaced omnidirectional microphones are used to create directional patterns, where instantaneous ICA can be applied. In the Appendix, we show how the recordings from two closely spaced omnidirectional microphones can be used to make two directional microphone gains.

Therefore, a realistic assumption is that two directional microphone responses recorded at the same location are available. For evaluation purposes, we have chosen appropriate microphone responses; the frequency independent gain responses are chosen as functions of the direction θ as $r_1(\theta) = 1 + 0.5\cos(\theta)$ and $r_2(\theta) = 1 - 0.5\cos(\theta)$, respectively. The two microphone responses are shown in Fig. 1. Hence, instead of having a mixing system where a given microphone delay corresponds to a given direction, a given set of microphone gains corresponds to a certain direction, and the mixing system

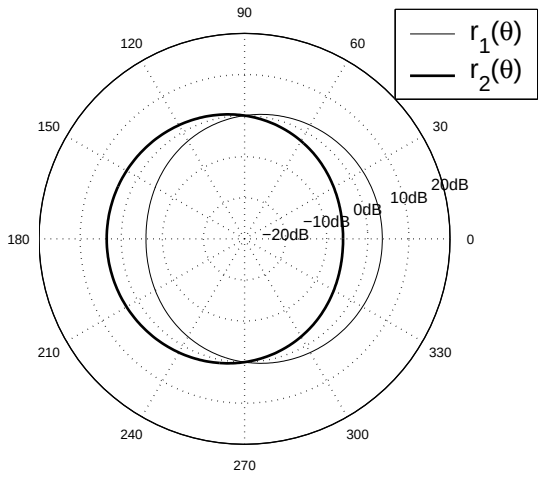

Fig. 1. The two directional microphone responses are shown as function of the direction θ.

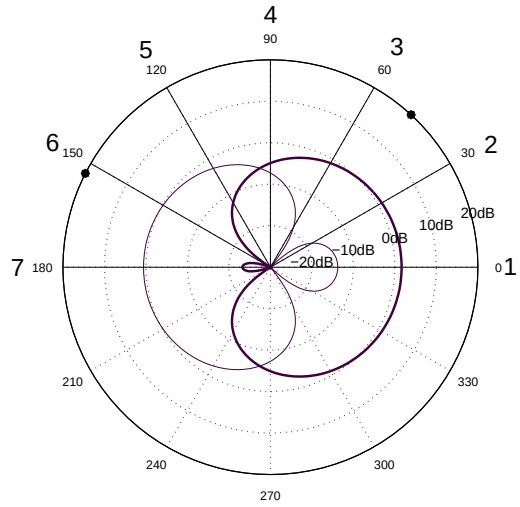

Fig. 2. The polar plots show the gain for different directions. ICA is applied with two sensors and seven sources. The two dots at the outer perimeter show the null directions. We see that each row of the 2×2 ICA solution can make just one null direction in the interval $0° \leq \theta \leq 180°$. Symmetric nulls exist in the interval $180° \leq \theta \leq 360°$. The lines pointing out from the origin denote the true direction of the seven numbered speech sources. The ICA solution tends to place the null towards sources spatially close to each other, and each of the two outputs represents a group of spatially close signals.

is given by

$$\mathbf{A}(\boldsymbol{\theta}) = \begin{bmatrix} r_1(\theta_1) & \cdots & r_1(\theta_N) \\ r_2(\theta_1) & \cdots & r_2(\theta_N) \end{bmatrix}. \quad (5)$$

For the instantaneous case, the separation matrix $\mathbf{W}$ can be regarded as direction-dependent gains. For an $M \times M$ separation matrix, it is possible to have at most $M-1$ null directions, i.e., directions from which the interference signal is canceled out, see e.g. [45], [46]. Signals arriving from other directions are not completely canceled out, and they thus have a gain greater than $-\infty$ dB.

Now consider the case where $N \geq M = 2$. When there are only two mixed signals, a standard ICA algorithm only has two output signals $\mathbf{y}(n) = [y_1(n), y_2(n)]^T$. Since the number of separated signals obtained by (2) is smaller than the number of source signals, $\mathbf{y}$ does not contain the separated signals. Instead, if the noise term is disregarded, $\mathbf{y}$ is another linear superposition of the source signals, i.e.

$$\mathbf{y} = \mathbf{G}\mathbf{s}, \quad (6)$$

where the weights are given by $\mathbf{G} = \mathbf{W}\mathbf{A}$ instead of just $\mathbf{A}$ as in (1). Thus, $\mathbf{G}$ just corresponds to another weighting of each of the source signals depending on θ. These weights make $y_1(n)$ and $y_2(n)$ as independent as possible even though $y_1(n)$ and $y_2(n)$ themselves are not single source signals. This is illustrated in Fig. 2. The figure shows the two estimated spatial responses from $\mathbf{G}(\boldsymbol{\theta})$ in the underdetermined case. The response of the m'th output is given by $\mathbf{g}_m(\theta) = |\mathbf{w}_m^T \mathbf{a}(\theta)|$, where $\mathbf{w}_m$ is the separation vector from the m'th output and $\mathbf{a}(\theta) = [r_1(\theta), r_2(\theta)]^T$ is the mixing vector for the arrival direction θ [45]. By varying θ over all possible directions, directivity patterns can be created as shown in Fig. 2. The estimated null placement is illustrated by the two round dots placed at the perimeter of the outer polar plot. The lines pointing out from the origin illustrate the direction of the seven source signals. Here, the sources are equally distributed in the interval $0° \leq \theta \leq 180°$. As shown in the figure, typically the nulls do not cancel single sources out. Rather, a null is placed at a direction pointing towards a *group* of sources which are spatially close to each other. Here, it can be seen that in the first output, $y_1(n)$, the signals 5, 6 and 7 dominate and in the second output, $y_2(n)$, the signals 1, 2 and 3 dominate. The last signal, 4 exists with almost equal weight in both outputs. As we show in Section III, this new weighting of the signals can be used to estimate binary masks reliably. Similar equivalence has been shown between ICA in the frequency domain and adaptive beamforming [46]. In that case, for each frequency, $\mathbf{Y}(\omega) = \mathbf{G}(\omega)\mathbf{S}(\omega)$.

III. Blind Source Extraction with ICA and Binary Masking

A. Algorithm for instantaneous mixtures

The input to our algorithm is the two mixtures x_1 and x_2 of duration N_s. The algorithm can be divided into three main parts: a *core procedure*, a *separation stage* and a *merging stage*. The three parts are presented in Fig. 3, Fig. 4 and Fig. 5, respectively.

1) Core procedure: Fig. 3 shows the *core procedure*. The core procedure is performed iteratively for a number of cycles in the algorithm. The inputs to the core procedure are two input mixtures x_a and x_b and a binary mask (step A), which has been applied to the original signals x_1 and x_2 in order to obtain

x_a and x_b. In the initial application of the core procedure, $x_a = x_1$ and $x_b = x_2$, and BM is all ones.

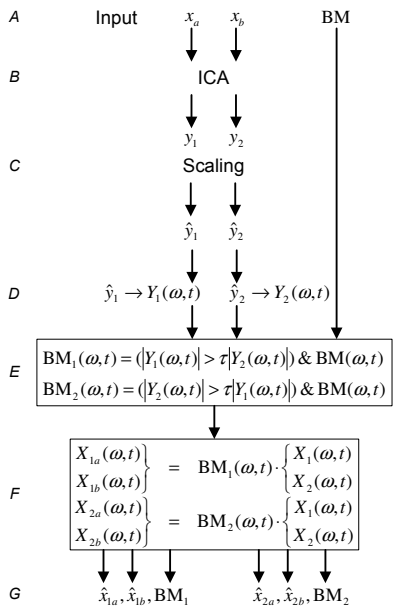

Fig. 3. Flowchart showing the *core procedure* of the algorithm. The algorithm has three input signals: The two input mixtures $x_a = [x_a(0), x_a(1), \cdots, x_a(N_s)]$ and $x_b = [x_b(0), x_b(1), \cdots, x_b(N_s)]$, and a binary mask which has been applied to the two original mixtures in order to obtain x_a and x_a. Source separation by ICA is applied to the two original signals in order to obtain y_1 and y_2. $\hat{y}_1$ and $\hat{y}_2$ are obtained by normalizing the two signals with respect to the variance. The re-scaled signals are transformed into the T-F domain, where the two binary masks are obtained by comparing the corresponding T-F units of the two T-F signals and multiplying by the input binary mask to prevent re-introduction of already masked T-F units. The two estimated masks are then applied in the T-F domain to the original signals $x_1 \rightarrow X_1(\omega,t)$ and $x_2 \rightarrow X_2(\omega,t)$. The output consists of the two estimated binary masks and the four masked signals.

As described in the previous section, a two-input two-output ICA algorithm is applied to the input mixtures, regardless of the number of source signals that actually exist in the mixture (step *B*). The two outputs y_1 and y_2 from the ICA algorithm are arbitrarily scaled (step *C*). Since the binary mask is estimated by comparing the amplitudes of the two ICA outputs, it is necessary to solve the scaling problem. In [40], we solved the scaling problem by using the knowledge about the microphone responses. Here we use a more 'blind' method to solve the scaling ambiguity. As proposed in [10], we assume that all source signals have the same variance and the outputs are therefore scaled to have the same variance. The two re-scaled output signals, $\hat{y}_1$ and $\hat{y}_2$ are transformed into the frequency domain (step *D*), e.g. by use of the STFT so that two spectrograms are obtained:

$$\hat{y}_1 \rightarrow Y_1(\omega,t) \quad (7)$$

$$\hat{y}_2 \rightarrow Y_2(\omega,t), \quad (8)$$

where ω denotes the frequency and t the time window index. From the two time-frequency signals, two binary masks are

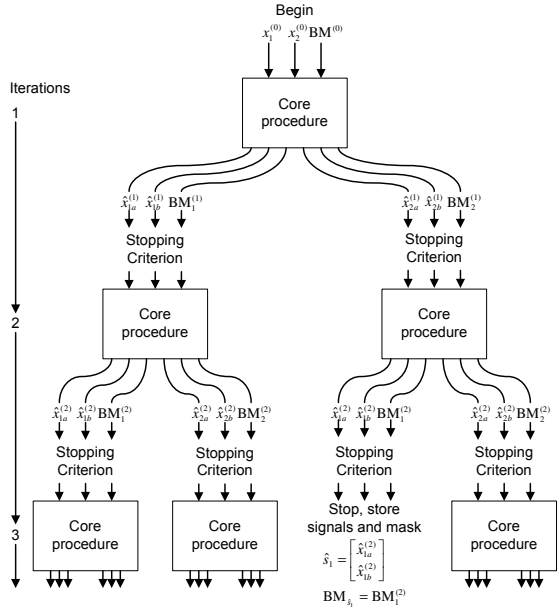

Fig. 4. The *separation stage*. Separation is performed iteratively by the core procedure as described in Fig. 3. The stopping criterion is applied to each set of outputs from the core procedure. If the output consists of more than one speech signal, the core procedure is applied again. If the output consists of only a single source signal, the output and its corresponding mask are stored. The core procedure is applied to the outputs iteratively until all outputs consist of only a single signal. The outputs are stored either as a candidate for a separated stereo sound signal $\hat{s}$ or a separated stereo signal of poor quality $\hat{p}$.

estimated. The binary masks are determined for each T-F unit by comparing the amplitudes of the two spectrograms (step *E*):

$$\mathrm{BM}_1(\omega,t) = \begin{cases} 1, & \text{if } |Y_1(\omega,t)| > \tau|Y_2(\omega,t)|; \\ 0, & \text{otherwise.} \end{cases} \quad \forall \omega, t \quad (9)$$

$$\mathrm{BM}_2(\omega,t) = \begin{cases} 1, & \text{if } |Y_2(\omega,t)| > \tau|Y_1(\omega,t)|; \\ 0, & \text{otherwise.} \end{cases} \quad \forall \omega, t \quad (10)$$

where τ is a parameter. The parameter τ in (9) and (10) controls how sparse the mask should be, i.e., how much of the interfering signals should be removed at each iteration. If $\tau = 1$, the two estimated masks together contain the same number of retained T-F units (i.e. equal to 1) as the previous mask. If $\tau > 1$, the combination of the two estimated masks is more sparse, i.e. having fewer retained units, than the previous binary mask. This is illustrated in Fig. 6. In general, when $\tau > 1$, the convergence is faster at the expense of a sparser resulting mask. When the mask is sparser, musical noise becomes more audible. The performance of the algorithm is considered for $\tau = 1$ and $\tau = 2$. We do not consider the case where $0 < \tau < 1$ as some T-F units would be assigned the value '1' in both estimated masks. In order to ensure that the binary mask becomes sparser for every iteration, a simple logical AND operation between the previous mask and the estimated mask is applied.

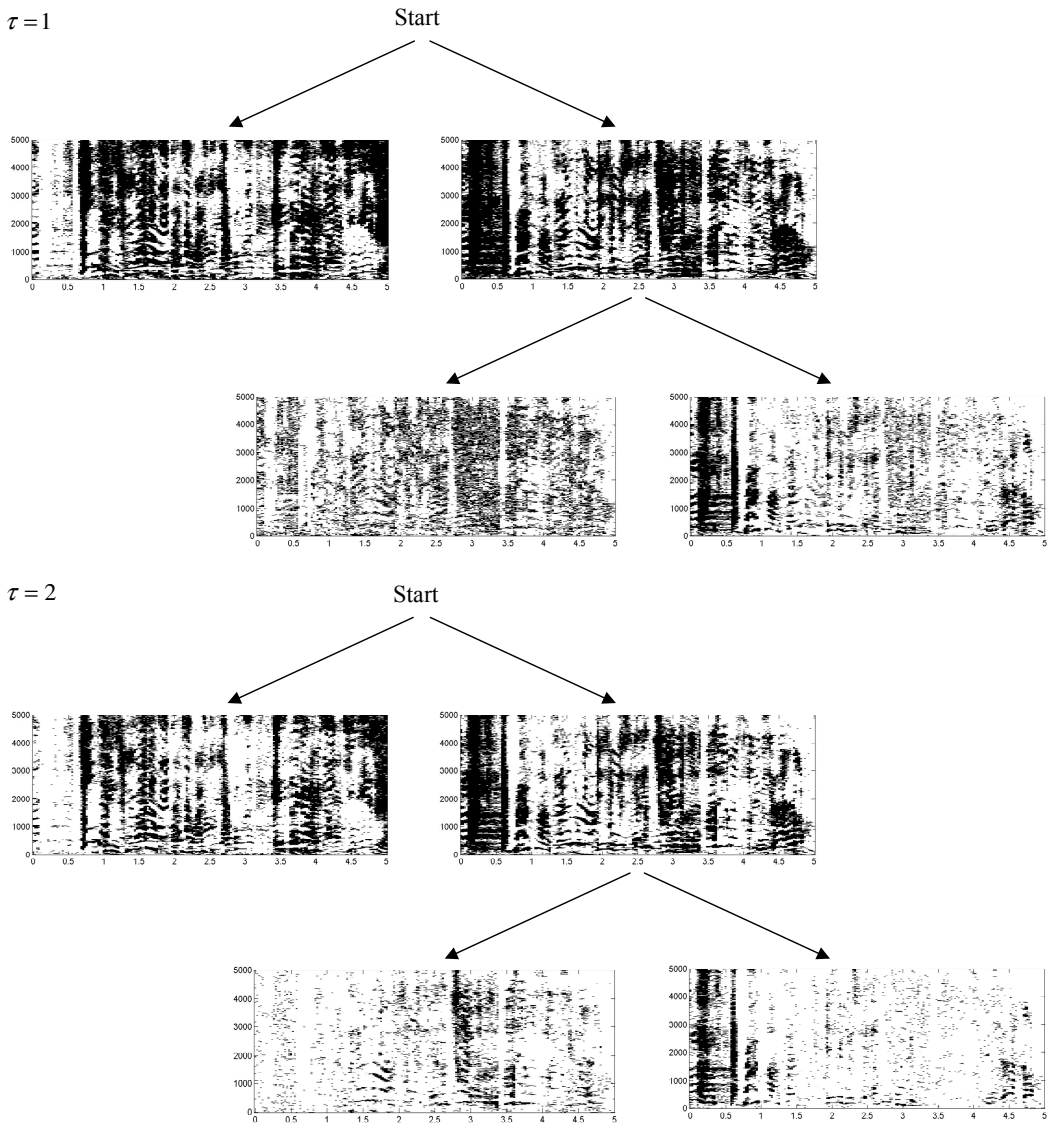

Fig. 6. The first two iterations for the estimations of the binary masks. Black indicates '1', and white '0'. For each iteration, two new masks are estimated by comparison of the ICA output as shown in equations (9) and (10). The previous mask ensures that no T-F units are re-introduced. The plot above shows the case of $\tau = 1$. When $\tau = 1$, the estimated masks contain the same T-F units as the mask in the previous iteration. The plot below shows the case of $\tau = 2$. Here the two estimated masks together contain less T-F units than the binary mask at the previous iteration. Therefore τ can be used to control the convergence speed. The separation performance with the $\tau = 1$ and $\tau = 2$ is presented in Table V and VI, respectively.

Next, each of the two binary masks is applied to the original mixtures in the T-F domain (step *F*), and by this non-linear processing, some of the speech signals are *attenuated* by one of the masks while other speakers are attenuated by the other mask. After the masks have been applied to the signals, they are reconstructed in the time domain by the inverse STFT (step *G*).

Time-frequency decomposition can be obtained in many ways, of which the STFT is only one way. The STFT has a linear frequency scale. A linear frequency scale does not accord well with human perception of sounds. The frequency representation in the human ear is closer to a logarithmic scale. The frequency resolution at the low frequencies is much higher than that at the high frequencies [33]. Therefore, T-F decomposition, where the frequency spacing is logarithmic may be a better choice than a linear scale. T-F decomposition

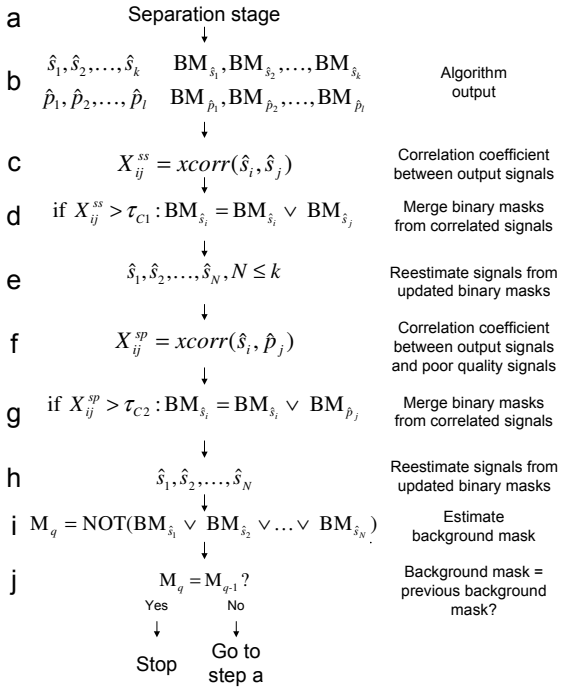

Fig. 5. Flowchart showing the steps of the *merging stage*. The details of the separation stage in step 'a' are shown in Fig. 3 and in Fig. 4. From the separation stage, the outputs shown in step 'b' are available. $\hat{s}_1, \ldots, \hat{s}_k$ denote the k separated signals, and $\hat{p}_1, \ldots, \hat{p}_l$ denotes the l separated signals of poor quality. BM denotes the corresponding binary mask of the estimated signal. The outputs from the main algorithm are further processed in order to improve the separated signals. Masks of output signals which are correlated are merged. Also masks output signals which are correlated with signals of poor quality are merged with these masks. A *background mask* is estimated from T-F units that have not been used so far. This mask is used to execute the main algorithm again. If the background mask has not changed, the segregated signals are not changed any further and the algorithm stops.

based on models of the cochlea are termed *cochleagrams* [29]. Different filterbanks can be used in order to mimic the cochlea, including the Gammatone filterbank [47]. Frequency warping of a spectrogram is another option, e.g. to fit the Bark frequency scale [48].

2) Separation stage: Fig. 4 shows the *separation stage*, i.e. how the core procedure is applied iteratively in order to segregate all the source signals from the mixture. At the beginning, the two recorded mixtures are used as input to the core procedure. The initial binary mask, $\mathrm{BM}^{(0)}$ has the value '1' for all T-F units. A stopping criterion is applied to the two sets of masked output signals. The masked output signals are divided into three categories defined by the stopping criterion in Section IV:

1) The masked signal is of poor quality.
2) The masked signal consists of mainly one source signal.
3) The masked signal consists of more than one source signal.

In the first case, the poor quality signal is stored for later use and marked as a poor quality signal. We denote these signals as $\hat{p}$. When we refer to a signal of poor quality, we mean a signal whose mask only contains few T-F units. Such a signal is distorted with many artifacts. In the second case, the signal is stored as a candidate for a separated source signal. We denote those signals as $\hat{s}$. In the third case, the masked signal consists of more than one source. Further separation is thus necessary, and the core procedure is applied to the signals. T-F units that have been removed by a previous mask cannot be re-introduced in a later mask. Thus, for each iteration, the estimated binary masks become sparser. This iterative procedure is followed until no more signals consist of more than one source signal.

3) Merging stage: The objective of our proposed method is to segregate all the source signals from the mixture. Because a signal may be present in both ICA outputs, there is no guarantee that two different estimated masks do not lead to the same separated source signal. In order to increase the probability that all the sources are segregated and no source has been segregated more than once, a *merging stage* is applied. Further, the merging stage can also improve the quality of the estimated signals. The merging steps are shown in Fig. 5. The output of the separation stage (step a) is shown in step b. The output of the algorithm consists of the k segregated sources, $\hat{s}_1, \ldots, \hat{s}_k$, the l segregated signals of poor quality, $\hat{p}_1, \ldots, \hat{p}_l$, and their corresponding binary masks. In the merging stage, we identify binary masks that mainly contain the same source signal. A simple way to decide whether two masks contain the same signal is to consider the correlation between the masked signals in the time domain. Notice that we cannot find the correlation between the binary masks. The binary masks are disjoint with little correlation. Because we have overlap between consecutive time frames, segregated signals that originate from the same source are correlated in the time domain.

In step c, the correlation coefficients between all the separated signals are found. If the normalized correlation coefficient between two signals is greater than a threshold τ_{C1}, a new signal is created from a new binary mask as shown in step d and e. The new mask is created by applying the logical OR operation to the two masks associated with the two correlated signals. Here, we just find the correlation coefficients from one of the two microphone signals and assume that the correlation coefficient from the other channel is similar.

Even though a segregated signal is of poor quality, it might still contribute to improve the quality of the extracted signals. Thus, the correlation between the signals with low quality (energy) and the signals that contain only one source signal is found (step f). If the correlation is greater than a threshold τ_{C2}, the mask of the segregated signal is expanded by merging the mask of the signal of poor quality (step g and h). Hereby the overall quality of the new mask should be higher, because the new mask is less sparse. After the correlations between the output signals have been found, some T-F units still have not been assigned to any of the source signal estimates. As illustrated in Fig. 7, there is a possibility that some of the sources in the mixture have not been segregated. In the direction where the gains from the two ICA outputs are almost equal, there is a higher uncertainty on the binary decision,

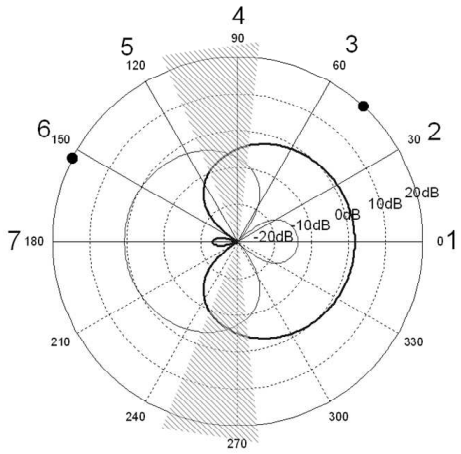

Fig. 7. As in Fig. 2 the polar plots show the gain for different directions. Comparison between the gains determines the binary masks. Within the shaded areas, the gain is almost equal. Source signals that arrive from a direction close to where the gains are almost equal will (depending on the parameter τ) either exist in both masked signals or in none of the masked signals. Therefore, the algorithm may fail to segregate such source signals from the mixture.

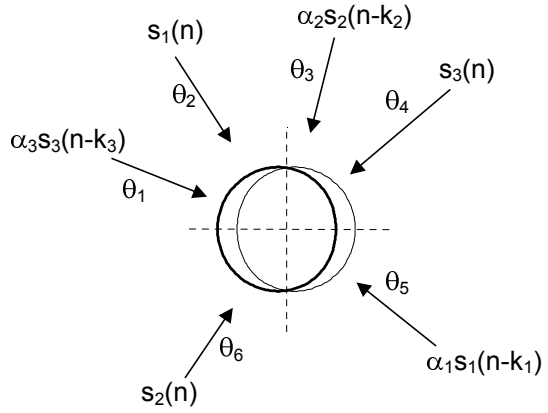

Fig. 8. A simple reverberant environment with three sources each having one reflection. As in Fig. 1 the impinging signals are recorded by a two-microphone array with directional responses, so that each direction corresponds to a certain set of directional microphone responses. Here, each reflection can be regarded as a single source impinging the microphone array.

which means that a source in that area may appear in both outputs. Furthermore, if $\tau > 1$ some T-F units in the shaded area of Fig. 7 are assigned the value '0' in both binary masks. Therefore, sources are assumed to exist in the T-F units which have not been assigned to a particular source yet. Thus, a *background mask* is created from all the T-F units which have not been assigned to a source (step i). The background mask is then applied to the original two mixtures, and possible sounds that remain in the background mask are hereby extracted. The separation algorithm is then applied to the remaining signal to ensure that there is no further signal to extract. This process continues until the remaining mask does not change any more (step j). Notice that the final output signals are maintained as two signals.

B. Modified algorithm for convolutive mixtures

In a reverberant environment, reflections from the signals generally arrive from different directions. In this situation, the mixing model is given by (3). Again, we assume that the sounds are recorded by a two-microphone array with directional responses given in Fig. 1. A simple reverberant environment is illustrated in Fig. 8. Here three sources $s_1(n)$, $s_2(n)$ and $s_3(n)$ are impinging the two-microphone array and direction-dependent gains are obtained. Also one reflection from each of the sources is recorded by the directional microphones: $\alpha_1 s_1(n-k_1)$, $\alpha_2 s_2(n-k_2)$ and $\alpha_3 s_3(n-k_3)$. In this environment, we can write the mixture with an instantaneous mixing model $\mathbf{x} = \mathbf{As}$ with $\mathbf{s} = [\alpha_3 s_3(n-k_3), s_1(n), \alpha_2 s_2(n-k_2), s_3(n), \alpha_1 s_1(n-k_1), s_2(n)]^T$ and

$$\mathbf{A}(\boldsymbol{\theta}) = \begin{bmatrix} r_1(\theta_1) & \cdots & r_1(\theta_6) \\ r_2(\theta_1) & \cdots & r_2(\theta_6) \end{bmatrix}. \qquad (11)$$

We can therefore apply the iterative instantaneous ICA algorithm to the mixture, and we can segregate the convolutive mixture into numerous components, as independent as possible, where each component is a source or a reflection impinging from a certain direction. Similarly, a merging stage can determine if two segregated components originate from the same source.

When the method is applied to reverberant mixtures, we observe that the estimated binary masks becomes more frequency dependent so that the binary mask for some frequencies mainly contains zeroes and for other frequency bands mainly contains ones. This results in band-pass filtered versions of the segregated signals. For example, one binary mask mainly contains the high-frequency part of a speech signal, while another mask mainly contains a low-frequency part of the same speech signal. This high-pass and low-pass filtered versions are poorly correlated in the time-domain. In order to merge these band-pass filtered speech signals that originate from the same source, we compute the correlation between the envelopes of the signals instead. This approach has successfully been applied in frequency domain ICA in order to align permuted frequencies [49], [50]. The following example shows that the envelope correlation is a better merging criterion than just finding the correlation between the signals, when the signals are bandpass-filtered.

Two speech signals A and B with a sampling rate of 10 kHz are each convolved with a room impulse response having $T_{60} = 400$ ms. Both signals are divided into a high-frequency (HF) part, and a low frequency (LF) part. Hereby four signals A_{LF}, A_{HF}, B_{LF}, and B_{HF} are obtained. The two LF signals are obtained from binary masks which contain ones for frequencies below 2500 Hz and zeros otherwise, and the two HF signals are obtained from binary masks which contain ones for frequencies above 2500 Hz and zeros otherwise. We

TABLE I
CORRELATION BETWEEN HIGH- AND LOW-PASS FILTERED SPEECH SIGNALS, THE ENVELOPE OF THE SIGNALS AND THE SMOOTHED ENVELOPE OF THE SIGNALS.

	A_{LF}	A_{HF}	B_{LF}	B_{HF}
A_{LF}	1	0.0006	0.0185	0.0001
A_{HF}		1	0.0001	0.0203
B_{LF}			1	0.0006
B_{HF}				1

	$\mathcal{E}(A_{\mathrm{LF}})$	$\mathcal{E}(A_{\mathrm{HF}})$	$\mathcal{E}(B_{\mathrm{LF}})$	$\mathcal{E}(B_{\mathrm{HF}})$
$\mathcal{E}(A_{\mathrm{LF}})$	1	0.0176	0.0118	0.0131
$\mathcal{E}(A_{\mathrm{HF}})$		1	0.0106	0.0202
$\mathcal{E}(B_{\mathrm{LF}})$			1	0.0406
$\mathcal{E}(B_{\mathrm{HF}})$				1

	$\hat{\mathcal{E}}(A_{\mathrm{LF}})$	$\hat{\mathcal{E}}(A_{\mathrm{HF}})$	$\hat{\mathcal{E}}(B_{\mathrm{LF}})$	$\hat{\mathcal{E}}(B_{\mathrm{HF}})$
$\hat{\mathcal{E}}(A_{\mathrm{LF}})$	1	0.0844	0.0286	0.0137
$\hat{\mathcal{E}}(A_{\mathrm{HF}})$		1	0.0202	0.0223
$\hat{\mathcal{E}}(B_{\mathrm{LF}})$			1	0.0892
$\hat{\mathcal{E}}(B_{\mathrm{HF}})$				1

now find the correlation coefficients between the four signals and the envelopes. The envelope can be obtained in different ways. The envelope $\mathcal{E}$ of the signal $x(n)$ can be calculated as [51]

$$\mathcal{E}(x(n)) = |x(n) + j\mathcal{H}(x(n))|, \tag{12}$$

where $\mathcal{H}(x(t))$ denotes the Hilbert transform, and j denotes the imaginary unit. Alternatively, we can obtain a smoother estimate $\hat{\mathcal{E}}$ as

$$\hat{\mathcal{E}}(x(n)) = \hat{\mathcal{E}}(x(n-1)) + \alpha(n)\big(|x(n)| - \hat{\mathcal{E}}(x(n-1))\big), \tag{13}$$

where

$$\alpha = \begin{cases} 0.04, & \text{if } |x(n)| - \hat{\mathcal{E}}(x(n-1)) > 0; \\ 0.01, & \text{if } |x(n)| - \hat{\mathcal{E}}(x(n-1)) < 0. \end{cases} \tag{14}$$

The above values of α have been found experimentally. The attack time and release time of the low-pass filter have been chosen differently in order to track the onsets easily. We initialize (13) by setting $\hat{\mathcal{E}}(x(0)) = 0$.

To prevent the DC component of the envelope from contributing to the correlation, the DC components are removed from the envelopes by a high-pass filter, before the correlation coefficient between the envelopes is computed. In Table I, the correlation coefficients between the four signals have been found, as well as the correlations between the envelopes and the smoothed envelopes. It is desirable that the correlation between signals that originate from the same source be high while the correlation between different signals be low. As it can be seen, the correlation coefficients between the signals do not indicate that A_{LF} and A_{HF} (or B_{LF} and B_{HF}) belong to the same source signal. When the correlation coefficients between the envelopes are considered, the correlations between A_{LF} and A_{HF} (or B_{LF} and B_{HF}) are a little higher than the cross-correlation between the source signals. The best result is obtained for the correlation between the smoothed envelopes. Here the correlations between A_{LF} and A_{HF} (or B_{LF} and B_{HF}) are significantly higher than the correlations between the different sources. In the reverberant case, we thus merge masks based on correlation between the smoothed envelope. We have also tried to apply the envelope-based merging criterion in the instantaneous case, but found that the simple correlation-based criterion gives better results. The reason, we suspect, is that the temporal fine structure of a signal that is present in the instantaneous case but weakened by reverberation is more effective than the signal envelope for revealing correlation.

IV. STOPPING CRITERION

As already mentioned, it is important to decide whether the algorithm should stop or the processing should repeat. The algorithm should stop when the signal consists of only one source or when the mask is too sparse (hence the quality of the resulting signal will be poor). Otherwise, the separation procedure should continue. When there is only one source in the mixture, the signal is expected to arrive only from one direction and thus the rank of the mixing matrix is one. We propose a stopping criterion based on the covariance matrix of the masked sensor signals. An estimate of the covariance matrix is found as

$$\mathbf{R}_{xx} = \langle \mathbf{x}\mathbf{x}^T \rangle = \frac{1}{N_s}\mathbf{x}\mathbf{x}^T, \tag{15}$$

where N_s is the number of samples in $\mathbf{x}$. By inserting (1), and assuming that the noise is independent with variance σ^2, the covariance can be written as function of the mixing matrix and the source signals:

$$\begin{aligned} \mathbf{R}_{xx} &= \langle (\mathbf{A}\mathbf{s} + \boldsymbol{\nu})(\mathbf{A}\mathbf{s} + \boldsymbol{\nu})^T \rangle & (16) \\ &= \mathbf{A}\langle \mathbf{s}\mathbf{s}^T \rangle \mathbf{A}^T + \langle \boldsymbol{\nu}\boldsymbol{\nu}^T \rangle & (17) \\ &= \mathbf{A}\langle \mathbf{s}\mathbf{s}^T \rangle \mathbf{A}^T + \sigma^2 \mathbf{I} & (18) \\ &= \boldsymbol{\Psi} + \sigma^2 \mathbf{I}, & (19) \end{aligned}$$

where $\boldsymbol{\Psi} = \mathbf{A}\mathbf{R}_{ss}\mathbf{A}^T$ of size $M \times M$. We assume that the masked sensor signal consists of a single source if the condition number (based on the 2-norm) [52] is greater than a threshold τ_c, i.e.

$$\mathrm{cond}(\mathbf{R}_{xx}) > \tau_c. \tag{20}$$

A high condition number indicates that the matrix is close to being singular. Since $\mathbf{R}_{xx}$ is symmetric and positive definite, $\mathrm{cond}(\mathbf{R}_{xx}) = \max \mathrm{eig}(\mathbf{R}_{xx}) / \min \mathrm{eig}(\mathbf{R}_{xx})$, where $\mathrm{eig}(\mathbf{R}_{xx})$ is the vector of eigenvalues of $\mathbf{R}_{xx}$. Because the desired signals are speech signals, we bandpass filter the masked mixed signals before we calculate the covariance matrix, so that only frequencies where speech dominates are considered. The cutoff frequencies of the bandpass filter are chosen to be 500 and 3500 Hz.

In order to discriminate between zero and one source signal, we consider the power of the masked signal. If the power of the masked signal has decreased by a certain amount compared to the power of the original mixture, the signal is considered to be of poor quality. We define this amount by the parameter τ_E, which is measured in dB.

This stopping criterion is applied for instantaneous as well as convolutive mixtures. In the case of convolutive mixtures, the stopping criterion aims at stopping when the energy of the segregated signal mainly comes from a single direction, i.e. the iterative procedure should stop when only a single

reflection from a source remains in the mixture. Note that, as illustrated in Fig. 8, our algorithm for convolutive mixtures treats each reflection as a distinct sound source. Because many reflections have low energy compared to the direct path, a high number of segregated signals of poor quality are expected in the reverberant case.

V. Evaluation

A. Evaluation Metrics

When using a binary mask, it is not possible to reconstruct the speech signal perfectly, because the signals partly overlap. An evaluation method that takes this into account is therefore used [53]. As a computational goal for source separation, the *ideal binary mask* has been suggested [32]. The ideal binary mask for a signal is found for each T-F unit by comparing the energy of the signal to the energy of all the interfering signals. Whenever the signal energy is higher within a T-F unit, the T-F unit is assigned the value '1' and whenever the combined interfering signals have more energy, the T-F unit is assigned the value '0'. The ideal binary mask produces the optimal SNR gain of all binary masks in terms of comparing with the entire signal [34].

As in [34], for each of the separated signals, the percentage of energy loss P_{EL} and the percentage of noise residue P_{NR} are calculated:

$$P_{EL} = \frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \quad (21)$$

$$P_{NR} = \frac{\sum_n e_2^2(n)}{\sum_n O^2(n)}, \quad (22)$$

where $O(n)$ is the estimated signal, and $I(n)$ is the signal re-synthesized after applying the ideal binary mask. $e_1(n)$ denotes the signal present in $I(n)$ but absent in $O(n)$ and $e_2(n)$ denotes the signal present in $O(n)$ but absent in $I(n)$. The performance measure P_{EL} can be regarded as a weighted sum of the T-F unit power present in the ideal binary mask, but absent in the estimated mask, while the performance measure P_{NR} can be regarded as a weighted sum of the T-F unit power present in the estimated binary mask, but absent in the ideal binary mask.

Also the output signal-to-noise ratio ($\mathrm{SNR_o}$) can be measured. Here the $\mathrm{SNR_o}$ is defined using the re-synthesized speech from the ideal binary mask as the ground truth

$$\mathrm{SNR_o} = 10\log_{10}\left[\frac{\sum_n I^2(n)}{\sum_n (I(n)-O(n))^2}\right]. \quad (23)$$

If instead the original signal is used as the ground truth in the numerator in (23), the relatively low target energy from the T-F units that have been assigned the value '0' will also contribute. Because there is good perceptual correlation between the true speech signal and the resynthesized speech signal from the ideal mask [32], we should not let the inaudible values of the true signal contribute disproportionately to the SNR estimation. Therefore, it is better to use the ideal mask as the ground truth. Also the signal-to-noise ratio before separation, the input SNR ($\mathrm{SNR_i}$), is calculated. The $\mathrm{SNR_i}$ is the ratio between the desired signal and the interfering signals in the recorded masked mixtures. The SNR gain is measured in dB by

$$\Delta\mathrm{SNR} = \mathrm{SNR_o} - \mathrm{SNR_i}. \quad (24)$$

If we instead were using the original signals as ground truth, the SNR gain would be about 1-2 dB lower (see also [34]).

B. Setup and parameter choice

For evaluation, twelve different speech signals - six male and six female - from eleven different languages have been used. All speakers raised voice as if they were speaking in a noisy environment. The duration of each of the signals is five seconds and the sampling frequency is $f_s = 10$ kHz. All the source signals have approximately the same loudness. Separation examples and Matlab source code are available online [54], [55]. The signal positions are chosen to be seven positions equally spaced in the interval $0^\circ \leq \theta \leq 180^\circ$ as shown in Fig. 2. Hereby, the minimum angle between two signals is 30°. During the experiments, each mixture is chosen randomly and each source is randomly assigned to one of the seven positions.

We have experimented with several different random mixtures. Sometimes the method fails in separating all the mixtures. In those cases, typically two segregated signals are merged because they are too correlated, resulting in $N-1$ segregated signals, where one of the segregated signals consists of two source signals which are spatially close to each other. Alternatively, one source signal may occur twice resulting in $N+1$ separated signals. Therefore, as another success criterion we also count the number of times where all N sources in the mixture have been segregated into exactly N signals and each of the N sources are dominating in exactly one of the segregated signals. We call the ratio "correctness of detected source number" or "Correct #" in the result tables. We then calculate the average performance from those where the number of sources has been correctly detected when the algorithm stops. Although not all signals are correctly separated, it is still useful for some applications to recover some of the signals. Subjective listening could determine which of the source signals in the mixture the segregated signal is closest to. Here we use an automatic method to determine the pairing between the segregated signal and a source signal by comparing the corresponding estimated mask of the segregated signal and the ideal masks of different source signals. The source signal whose corresponding ideal mask is closest (in terms of most number of ones in common) to the estimated mask is determined to correspond to the segregated source. This method correlates well with subjective listening.

Different instantaneous ICA algorithms can be applied to the method. For evaluation we use an implementation of the INFOMAX ICA algorithm [13] which uses the BFGS (Broyden-Fletcher-Goldfarb-Shanno) optimization method [56], [57].

TABLE II

ROBUSTNESS OF τ_C AND τ_E FOR INSTANTANEOUS MIXTURES OF $N = 4$ AND $N = 6$ SIGNALS.

	τ_C					
	$N = 4$			$N = 6$		
τ_E	2000	3000	4000	2000	3000	4000
	15.45	15.34	15.24	13.85	14.04	13.87
15	10/10	10/10	9/10	6/10	5/10	5/10
	15.34	15.23	15.18	13.91	13.94	14.06
20	10/10	10/10	10/10	8/10	9/10	6/10
	15.64	15.19	14.36	14.39	13.86	14.06
25	4/10	4/10	5/10	1/10	4/10	6/10

ΔSNR and the number of times
(out of the ten cases) where all signals have been segregated

Unless otherwise stated, the parameter τ in Equations (9) and (10) is set to $\tau = 1$.

1) Choice of thresholds: Different thresholds have to be chosen. The thresholds have been determined from initial experiments as described below.

Regarding the two correlation thresholds, τ_{C1} and τ_{C2} shown in Fig. 5, our experiments show that most correlations between the time signals are very close to zero. Two candidates for separated signals are merged if the correlation coefficient is greater than 0.1. If τ_{C1} is increased, some signals may not be merged even though they mainly contain the same source. If τ_{C2} is decreased, the probability of merging different source signals is increased. The low energy signals are even less correlated with the candidates for separated signals. Therefore, we have chosen $\tau_{C2} = 0.03$. If τ_{C2} is increased, the masks become sparser, and more artifacts occur. If τ_{C2} becomes smaller, noise from other sources becomes more audible.

The thresholds in the stopping criterion are estimated from the initial experiments too. The condition number related threshold is chosen to be $\tau_C = 3000$. The signal is considered to contain too little energy when the energy of the segregated signal has decreased to $\tau_E = -20$ dB, when the power of a recorded mixture is normalized to 0 dB.

The robustness of the two thresholds τ_C and τ_E has been evaluated. τ_C has been evaluated for the values 2000, 3000 and 4000. Likewise, τ_E has been evaluated for the values 15, 20 and 25 dB. For each pair of τ_C and τ_E ten different random speech mixtures drawn from the pool of twelve speech signals are segregated. The experiment has been performed for mixtures consisting of four or six speech signals. In each case, ΔSNR is measured. Also the number of times (out of ten) where exactly all the sources in the mixture are been segregated is found. The results are reported in Table II. As it can be seen, the ΔSNR does not vary much as function of the two thresholds. The number of times where the method fails to segregate exactly N speech signals from the mixture is minimized for $\tau_C = 3000$ and $\tau_E = 20$ dB, which will be used in the evaluation.

The algorithm could be applied to a mixture several times, each time with different thresholds. Such a procedure could increase the chance of extracting all the sources from the mixture.

TABLE III

PERFORMANCE FOR DIFFERENT WINDOW LENGTHS

Window length	$P_{EL}(\%)$	$P_{NR}(\%)$	ΔSNR	Correct #
256 (25.6 ms)	9.17	11.38	13.56	44/50
512 (51.2 ms)	6.07	8.62	15.23	46/50
1024 (102.4 ms)	6.86	9.92	14.75	46/50

The window length is given in samples and in milliseconds.
The DFT length is four times the window length.
The number of signals in each instantaneous mixture is $N = 4$.

2) Window function: In [8], the Hamming window is found to perform slightly better than other window functions. In the following, the Hamming window will be used.

3) Window length: Different window lengths have been tried. The overlap factor is selected to be 75%. An overlap factor of 50% has also been considered, but better performance is obtained with 75% overlap.

With an overlap of 75% the separation has been evaluated for window lengths of 256, 512 and 1024 samples, which with $f_s = 10$ kHz give window shifts of 12.8, 25.6 and 51.2 ms, respectively. For a Hamming window the 3 dB bandwidth of the main lobe is 1.30 samples [58]. The frequency (spectral) resolution is thus 50.8, 25.4 and 12.7 Hz, respectively. The DFT length is four times the window length. Hence, the spectrogram resolution is 513, 1025 and 2049, respectively. By selecting a DFT length longer than the window length, the spectrogram becomes smoother, and when listening to the segregated signals, the quality becomes much better too. When the DFT size is longer than the window size, there is more overlap between the different frequency bands. Furthermore, artifacts from aliasing are reduced by zero-padding the window function.

The results are shown in Table III. The average performance is given for fifty random mixtures, each consisting of four speech sources. The highest SNR improvement is achieved for a window length of 512. A similar performance is achieved for the window length of 1024, while the window length of 256 performs a little worse. In the following experiments, we use a window length of 512.

4) ICA algorithm: We have chosen to use the INFOMAX algorithm [13] for evaluation, but other ICA algorithms could be used also. To examine how much the performance of our method depends on the chosen ICA algorithm, we have compared the INFOMAX and the JADE algorithm [59] in the ICA step. In both cases, the code is available online [56], [60]. The two algorithms have been applied to the same fifty mixtures each consisting of four signals drawn from the pool of twelve signals. The results are given in Table IV. As it can be seen, the performance of our method does not depend much on whether the chosen ICA algorithm is the INFOMAX or the JADE algorithm.

C. Separation results for instantaneous mixtures

Tables V and VI show the average separation performance for mixtures of N signals for $\tau = 1$ and $\tau = 2$. For each N, the algorithm has been applied fifty times to different speaker mixtures from the pool of twelve speakers at N of the seven random positions.

TABLE IV
COMPARISON BETWEEN JADE AND INFOMAX ICA ALGORITHMS.

Algorithm	$P_{EL}(\%)$	$P_{NR}(\%)$	ΔSNR	Correct #
JADE	6.20	8.86	15.17	46/50
INFOMAX	6.07	8.62	15.23	46/50

Instantaneous mixtures consisting of four sources have been used.

TABLE V
EVALUATION WITH RANDOM INSTANTANEOUS MIXTURES CONSISTING OF N SIGNALS.

N	$P_{EL}(\%)$	$P_{NR}(\%)$	SNR_i	SNR_o	ΔSNR	Correct #
2	1.01	2.00	0	18.92	18.92	47/50
3	2.99	4.86	-3.95	12.50	16.45	46/50
4	6.07	8.62	-5.98	9.26	15.23	46/50
5	10.73	13.02	-7.40	5.56	14.27	44/50
6	14.31	13.63	-8.39	5.25	13.64	44/50
7	18.34	22.43	-9.24	4.24	13.48	41/50

The parameter $\tau = 1$.

As it can be seen, the proposed algorithm is capable of separating at least up to seven source signals. It can also be seen that the probability of recovering all N speech signals decreases as N increases. Also, the quality of the separated signals deteriorates when N increases. When N increases, the T-F domain becomes less sparse because of higher overlap between the source signals. When the performance for $\tau = 1$ in Table V is compared with that for $\tau = 2$ in Table VI, it can be seen that the performance is better for $\tau = 1$. However the algorithm with $\tau = 1$ uses more computation time compared to $\tau = 2$. As it can be seen in Table V, the algorithm fails to separate two sources from each other in three cases. This is probably because the masks at some point are merged due to a wrong decision by the merging criterion. In Fig. 9, the ideal binary masks for a source from an example mixture of three speech signals are shown, along with the estimated mask is shown. As it can be seen, the estimated mask is very similar to the ideal masks.

1) Stationarity assumption: The duration of the mixture is important for separation. It is required that the source signals remain at their positions while the data is recorded. Otherwise the mixing matrix will vary with time. Therefore, there is a tradeoff between the number of available samples and the time duration during which the mixing matrix can be assumed to be stationary. Mixtures containing four speech signals have been separated. The duration T is varied between 1 and 5 seconds. The average performance has been found from fifty different mixtures. Since the speech mixtures are randomly picked, one second is selected as the lower limit to ensure that all four speech signals are active in the selected time frame. The separation results are shown in Table VII. Fifty mixtures of four source signals have been separated and the average performance is shown. As it can be seen, the probability of recovering all the source signals decreases when less data is available. On the other hand, the performance does not increase further for data lengths above three seconds. By listening to the separated signals, we find that among the mixtures where all sources have been successfully recovered, there is no significant difference in the quality of the separated signals.

TABLE VI
EVALUATION WITH RANDOM INSTANTANEOUS MIXTURES CONSISTING OF N SIGNALS.

N	$P_{EL}(\%)$	$P_{NR}(\%)$	SNR_i	SNR_o	ΔSNR	Correct #
2	3.43	0.50	0	18.22	18.22	50/50
3	7.36	2.60	-3.96	11.10	15.06	46/50
4	12.26	4.17	-5.89	8.81	14.70	42/50
5	19.81	6.21	-7.32	6.59	13.91	40/50
6	25.91	8.81	-8.36	5.31	13.67	23/50
7	30.52	11.86	-9.12	3.00	13.46	4/50

The parameter $\tau = 2$.

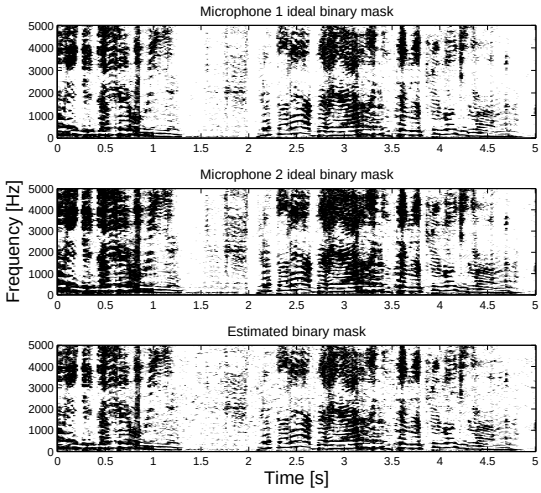

Fig. 9. Separation example. A segregated speech signal from a mixture of three speech signals. The two upper masks show the ideal binary mask for each of the two directional microphones. For this estimated signal, $P_{EL} = 1.38\%$, $P_{NR} = 0.46\%$, and $\Delta\text{SNR} = 20.98$ dB. Notice, unless the ideal masks from both microphones are exactly the same, P_{EL} and P_{NR} are always greater than zero. Perceptually, the segregated signal sounds clean without any artifacts. The separation quality is similar for the two other signals from the mixture.

TABLE VII
EVALUATION OF SEPARATION PERFORMANCE AS FUNCTION OF THE SIGNAL LENGTH T.

T	$P_{EL}(\%)$	$P_{NR}(\%)$	SNR_i	SNR_o	ΔSNR	Correct #
1	7.53	8.83	-6.38	9.44	15.83	34/50
2	7.85	8.23	-5.98	9.00	14.88	43/50
3	6.87	9.69	-6.04	8.80	14.85	46/50
4	7.57	9.05	-6.04	8.81	14.86	46/50
5	6.07	8.62	-5.98	9.26	15.23	46/50

Instantaneous mixtures consisting of four sources have been used.

2) Different loudness levels: In the previous simulations, all the speech signals are approximately equally strong. Now we test the separation performance in situations where the signals in the mixture have different levels of loudness. The mixtures consist of four speech signals, drawn from the pool of twelve signals. Before mixing, the first speech signal is

TABLE VIII

EVALUATION OF SEPARATION PERFORMANCE AS FUNCTION OF ADDITIVE MICROPHONE NOISE.

Noise	$P_{EL}(\%)$	$P_{NR}(\%)$	SNR_i	SNR_o	ΔSNR	Correct #
-10 dB	15.29	15.52	-6.51	5.95	12.43	19/50
-20 dB	7.42	10.26	-6.02	8.37	14.39	45/50
-30 dB	6.24	8.53	-5.99	9.27	15.26	46/50
-40 dB	6.23	8.72	-5.97	9.19	15.16	47/50
-50 dB	6.39	8.15	-5.98	9.29	15.27	45/50
-60 dB	6.04	8.62	-5.98	9.27	15.25	46/50

Instantaneous mixtures consisting of four sources have been used.

multiplied by 1, the second speech signal is multiplied by 0.5, and the remaining two speech sources are multiplied by 0.25. The average performance from fifty simulations is found. The two strongest sources are segregated in all the examples. In 25 of the 50 simulations, all of the four signals are segregated. On average ΔSNR is 16.57 dB, $P_{EL} = 6.65\%$ and $P_{NR} = 14.64\%$. When we compare to the more difficult case in Table V where all four speakers have equal loudness, we see that the average ΔSNR here is 1 dB better.

3) Microphone noise: In the previous simulations, noise is omitted. We now add white noise to the directional microphone signals with different noise levels. The simulation results are given in Table VIII. The noise level is calculated with respect to the level of the mixtures at the microphone. The mixtures without noise are normalized to 0 dB. As it can be seen from the table, noise levels of up to -20 dB can be well tolerated.

D. Separation results for anechoic mixtures

As mentioned in Section II, directional microphone gains can be obtained from two closely-spaced microphones. Signals impinging at a two-microphone array have been simulated and the directional microphone gains have been obtained as described in the Appendix. The distance between the microphones is chosen as $d = 1$ cm. Hereby an instantaneous mixture is approximated from delayed sources. With this setup, fifty mixtures each consisting of four speech signals drawn from the pool of twelve speakers have been evaluated. The results are given in Table IX. Because the microphone gain is slightly frequency-dependent, the performance deteriorates compared to the ideal case where the gain is frequency independent, especially for the frequencies above 4 kHz. This is illustrated in Fig. 10. This might be explained by the fact that the approximation $kd \ll 1$ (described in the Appendix) does not hold for higher frequencies. Fortunately, for the perception of speech, the higher frequencies are less important. It can also be seen that the number of times where the exactly four sources have been segregated is decreased. In many cases one source is segregated more than once, which is not merged in the merging stage because the correlation coefficient is too low.

E. Separation results for reverberant recordings

As described in Section III, the method can be applied to recordings of reverberant mixtures. We use recordings

TABLE IX

EVALUATION OF DIRECTIONAL MICROPHONE APPROXIMATION.

Mic. dist.	$P_{EL}(\%)$	$P_{NR}(\%)$	ΔSNR	Correct #
$d = 1$ cm	7.63	8.84	14.83	17/50
Ideal case	6.07	8.62	15.23	46/50

Anechoic mixtures consisting of four sources have been used.

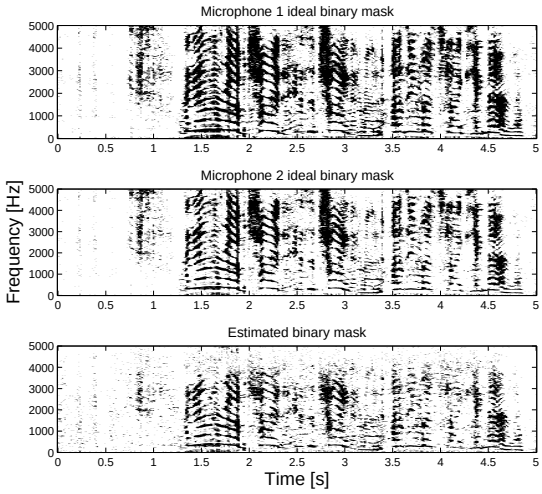

Fig. 10. Separation example. A segregated speech signal from a mixture of four speech signals. The speech signal impinges on an array consisting of two omnidirectional microphones spaced 1 cm apart. The two upper masks show the ideal binary masks for each of the two omnidirectional microphones. Because the directional gains are slightly frequency dependent, the performance for the high frequencies is deteriorated compared to the ideal case when the microphone gain is not frequency dependent, as shown in Fig. 9.

from a hearing aid with two closely-spaced, vertically placed omnidirectional microphones. The hearing aid is placed in the right ear of a Head and Torso Simulator (HATS) [61]. Room impulse responses are estimated from different loudspeaker positions. The source signals were then created by convolving the room impulses with the clean speech signals from the pool of twelve speakers.

The impulse responses are found in a reverberant room where the room reverberation time was $T_{60} = 400$ ms. Here reflections from the HATS and the room exist. The microphone distance is 12 mm. The room dimensions were $5.2\times7.9\times3.5$ m and the distance between the microphones and the loudspeakers were 2 m. Impulse responses from loudspeaker positions of 0°, 90°, 135°, and 180° are used. The configuration is shown in Figure 11. Fifty different mixtures consisting of four speakers from the pool of twelve speakers are created. The parameters of the algorithm have to be changed. When reverberation exists, the condition number never becomes as high as the chosen threshold of $\tau_C = 2000$. Therefore we need much lower thresholds. The separation performance is found for different values of τ_C. The remaining thresholds are set to $\tau_E = 25$, $\tau_{C1} = 0.1$ and $\tau_{C2} = 0.05$, with parameter $\tau = 1$. The separation results are provided in Table X. Four sources

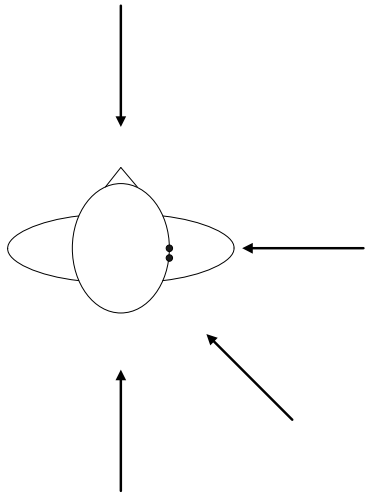

Fig. 11. Room configuration. The Head and Torso Simulator (seen from above) is placed in the middle of a room with a reverberation time of 400 ms. The two-microphone array is placed at the right ear. The distance between the microphones is 12 mm. The four sources arrive from positions of $0°$, $90°$, $135°$, and $180°$. The distance from the center of the head to each of the loudspeakers was 2 m. The room dimensions were $5.2 \times 7.9 \times 3.5$ m.

are not always segregated from a mixture. Therefore we count how many times the algorithm manages to segregate 0, 1, 2, 3 or all four sources from the mixture. This is denoted as 'freq.' in the table. We find the average P_{EL}, P_{NR} and ΔSNR for all these cases. It can be seen that often three of the four signals are segregated from the mixture. The average ΔSNR is around 6 dB. Even though the separation is not as good as in anechoic cases, it is worth noting that instantaneous ICA in the time domain may be used to segregate convolutive mixtures.

Another option is to apply a convolutive ICA algorithm [19] instead of an instantaneous ICA method. This was done in [42]. The advantage of using a convolutive algorithm compared to a instantaneous algorithm is that the convolutive algorithm is able to segregate sources, with larger microphone distances. Still, we have to assume that the convolutive algorithm at each step is able to segregate the sources into two groups, where some sources dominate in one group and other sources dominate in the other group. The stopping criterion from Section IV which is used to discriminate between one and more-than-one signal performs worse under the reverberant condition. Even though the criterion is applied to narrow frequency bands, the performance becomes worse as reported in [62]. In [42], we used a single-microphone criterion based on the properties of speech. There are some advantages of applying an instantaneous ICA as opposed to applying a convolutive ICA algorithm. The instantaneous algorithm is computationally less expensive. Further, frequency permutations which exist in many convolutive algorithms [19] are avoided.

The method used here cannot directly be compared to the method used in [42] which was applied with a much larger microphone distance. In [42], artificial room impulse responses were used with $T_{60} = 160$ ms, and here we have used recorded room impulses with $T_{60} = 400$ ms. The SNR gains obtained by the two methods are approximately the same.

TABLE X
SEPARATION OF CONVOLUTIVE MIXTURES CONSISTING OF FOUR SIGNALS.

# seg.	$P_{EL}(\%)$	$P_{NR}(\%)$	ΔSNR	Freq.
		$\tau_C = 200$		
0	–	–	–	0/50
1	–	–	–	0/50
2	–	–	–	0/50
3	56.30	45.74	6.22	29/50
4	65.21	49.85	5.57	21/50
		$\tau_C = 250$		
0	–	–	–	0/50
1	7.65	93.32	-5.20	1/50
2	45.61	49.19	6.73	1/50
3	56.42	48.90	6.01	30/50
4	62.90	50.32	5.62	18/50
		$\tau_C = 300$		
0	–	–	–	0/50
1	–	–	–	0/50
2	29.11	53.02	5.38	4/50
3	57.68	47.12	6.05	32/50
4	64.58	51.00	5.58	14/50
		$\tau_C = 350$		
0	–	–	–	0/50
1	–	–	–	0/50
2	36.86	53.85	5.56	9/50
3	54.83	47.63	5.97	30/50
4	65.02	49.55	5.71	11/50
		$\tau_C = 400$		
0	–	–	–	0/50
1	–	–	–	0/50
2	41.86	52.88	5.40	7/50
3	54.71	48.09	5.92	31/50
4	64.16	50.06	5.56	12/50

F. Comparison with other methods

Several other methods have been proposed for separation of an arbitrary number of speech mixtures with only two microphones by employing binary T-F masking [8], [24], [63]. In [24], speech signals were recorded binaurally and the interaural time difference (ITD) as well as the interaural intensity difference (IID) are extracted. The speech signals are separated by clustering in the joint ITD-IID domain. Separation results for three-source mixtures are given. An SNR gain of almost 14 dB is achieved. The gain also depends on the arrival directions of the source signals. Similarly, in the DUET algorithm described in [8], speech signals are separated by clustering speech signals in the amplitude/phase domain. In [8], the DUET algorithm was evaluated with synthetic anechoic mixtures, where amplitude and delay values are artificially chosen, as well as real reverberant recordings. These methods also have the advantage that the number of sources in the mixture need not be known in advance. In [24], the 128 frequency channels are (quasi) logarithmically distributed with center frequencies in the range of 80 Hz and 5000 Hz, while the frequency channels are linearly distributed in our proposed method and in [8] with a much higher frequency resolution.

In [38], the mask estimation is based on direction-of-arrival (DOA) techniques combined with ICA. The DOA technique is used to subtract $N - M$ sources, and the ICA algorithm is applied to the remaining M sources in the mixture. The method may be applied with binary masks, but in order to reduce musical noise, more continuous masks based on the directivity patterns have been applied. The method is shown for separation of mixtures containing up to four speech signals. In contrast to [38], our method separates speech mixtures by iteratively extracting individual source signals. Similar to other multi-microphone methods our method relies on spatially different source locations, but unlike the previous methods, our method uses ICA to estimate the binary masks by iteratively estimating independent subsets of the mixtures. While methods based on DOA may sweep all possible directions in order to estimate the null directions, our proposed ICA technique automatically steers the nulls. Our approach can be used to iteratively steer the nulls in settings with more sources than microphones. In [39], binary masks are also found based on the ICA outputs. Our method differs from the method in [39] for our method is able to segregate more sources than mixtures.

Another method for extraction of multiple speakers with only two microphones is presented in [64]. This method is based on localization of the source signals followed by a cancellation part where for each time frame different nulls are steered for each frequency. Simulations under anechoic conditions show subtraction of speech signals in mixtures containing up to six equally loud source signals. In [64] the SNR is found with the original signals as ground truth. An SNR gain of 7–10 dB was reported. Our method gives a significantly higher ΔSNR.

The microphone placement is different in our method compared to the microphone placement in the DUET algorithm [8]. Therefore, in order to provide a fair comparison between our proposed and the DUET algorithm, we have implemented the DUET algorithm for demixing approximately W-disjoint orthogonal sources by following the stepwise description in [8].

1) Comparison with DUET in the instantaneous case: The DUET algorithm has been applied to the same set of instantaneous mixtures that were used in Table V and VI. The results of the DUET algorithm for separation of 3–6 sources are reported in Table XI. When comparing the separation results in Table XI with the results from our proposed method in Table V and VI, it can be seen that our proposed method gives a better ΔSNR. Note that our ΔSNR is different from the signal-to-interference ratio used in [8] and tends to be more stringent. Furthermore, our method is better at estimating the exact number of sources, as the Correct # column indicates. The histogram smoothing parameter in the DUET algorithm provides a delicate trade-off. If the histogram is smoothed too much, it results in sources that merge together. If the histogram is smoothed too little, erroneous peaks appear resulting in too high an estimate of the number of sources. The best performing setting of the smoothing parameter is used in our implementation.

2) Comparison with DUET for convolutive mixtures: The DUET algorithm has been applied to the same synthetic reverberant data set that was used in Section V-E. The separation performance can be found in Table XII. When comparing the results of the first parte in Table X and Table XII we find that the performance of the DUET algorithm and our proposed method is generally similar. Both algorithms have difficulties in finding the exact number of sources under reverberant conditions. The DUET is able to extract all four sources in 43 of the 50 experiments, while our method is able to extract all sources in 21 of the 50 experiments. The lower number of extracted sources in our proposed method is caused by our merging criterion which often tends to merge different sources. On the other hand, the SNR gain is a little higher for our method. In the remaining 29 experiments we are able to segregate three of the four sources, again with a higher SNR gain than the DUET algorithm.

TABLE XI
EVALUATION OF THE DUET ALGORITHM WITH RANDOM INSTANTANEOUS MIXTURES CONSISTING OF N SIGNALS.

N	$P_{EL}(\%)$	$P_{NR}(\%)$	SNR_i	SNR_o	ΔSNR	Correct #
3	26.61	20.04	-3.94	3.17	7.11	11/50
4	36.44	23.21	-5.77	2.04	7.63	20/50
5	39.42	22.95	-7.25	1.73	8.98	10/50
6	52.80	40.97	-8.20	0.30	8.51	1/50

TABLE XII
SEPARATION OF CONVOLUTIVE MIXTURES CONSISTING OF FOUR SIGNALS WITH THE DUET ALGORITHM.

# seg.	$P_{EL}(\%)$	$P_{NR}(\%)$	ΔSNR	Freq.
0	–	–	–	0/50
1	–	–	–	0/50
2	–	–	–	0/50
3	65.28	29.92	5.80	7/50
4	82.56	37.79	5.55	43/50

In summary, our comparison with DUET suggests that the proposed method produces better results for instantaneous mixtures and comparable results for convolutive mixtures. By listening to our results and those published in [8], the quality of our results seems at least as good as the quality of the separated signals of [8]. In terms of computational complexity, our method depends on the number of sources in the mixtures, whereas the complexity of the DUET algorithm mainly depends on the histogram resolution. We have chosen a histogram resolution of 101×101 and a smoothing kernel of size 20×20. With this histogram resolution, the DUET algorithm and our proposed method take comparable amounts of computing time, for convolutive mixtures about 20 minutes per mixture on average on an HP 320 server. For the instantaneous case, our algorithm is faster; for example, with three sources, it takes about 4:30 min ($\tau = 1$) and 3:40 min ($\tau = 2$) to segregate all the sounds from a mixture, and about 10 min ($\tau = 1$) and 7 min ($\tau = 2$) to segregate all the sounds when the instantaneous mixture consists of seven sources.

VI. DISCUSSION

In this paper directional microphones placed at the same location are assumed. This configuration allows the mixing matrix to be delay-less, and any standard ICA algorithm can

therefore be applied to the problem. The configuration keeps the problem simple and still realistic. As shown in Section V-D, the algorithm may still be applied to delayed mixtures without significant changes. Alternatively, the ICA algorithm can be modified in order to separate delayed mixtures (see e.g. [4]). Since beamformer responses are used to determine the binary masks, the microphone distance cannot be too big. If the distance between the microphones is greater than half the wavelength, spatial aliasing occurs, and frequency-dependent null directions and sidelobes occur. An example of such multiple null directions and sidelobes is shown in Fig. 12. Therefore, for large microphone distances, the performance is expected to decrease, especially at high frequencies. A solution

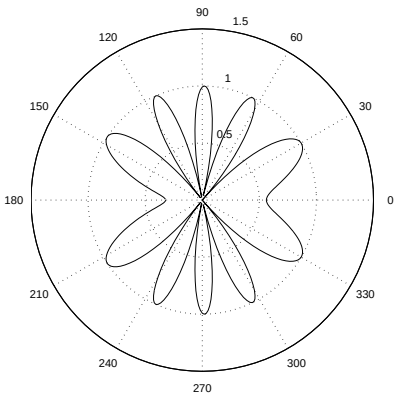

Fig. 12. A typical high-frequency microphone response. The response is given for the frequency of 4000 Hz, and a distance of 20 cm between the microphones. The half-wavelength at 4000 Hz is $\lambda/2 = 4.25$ cm. Since four whole half-wavelengths fit between the microphones, four nulls appear in the interval $0^\circ \leq \theta \leq 180^\circ$. Such a beampattern cannot efficiently be used to estimate the binary mask.

to this problem could be to use the envelope of the mixed high-frequency signal as ICA input directly.

By only using instantaneous ICA in the reverberant case, we assume that the sources can be divided into many independent components that can be merged afterwards. However, this assumption has some limitations. Sometimes, the independent components are very sparse, and hence it is difficult to apply reliable grouping. A way to better cope with this problem and the delays may be to apply a convolutive separation algorithm instead of an instantaneous separation step. Still, we believe it is an advantage to use instantaneous source separation compared to convolutive source separation because it is computationally much simpler - it only has four values to estimate, whereas convolutive ICA has thousands of filter coefficients to estimate.

When binary time-frequency masks are used, artifacts (musical noise) are sometimes audible in the segregated signals, especially when the masks are sparse. The musical noise degrades the perceptual quality of the segregated signal. Musical noise is caused by several factors. The binary mask can be regarded as a time-variant gain function multiplied to the mixture in the frequency domain. This corresponds to a circular convolution in the time domain. Therefore artifacts due to aliasing occur. From an auditory point of view, musical noise appears when separated T-F regions are isolated from each other. As a result, the sound of such an isolated region becomes an audible tone, which does not group with the other sounds in the auditory scene. In order to reduce musical noise, it has been suggested to use continuous masks [38]. By listening to the signals, we have observed that a mask created by combining masks produced with different thresholds and weighted by the thresholds results in less musical artifacts. In our case, a more graded mask could be obtained by finding masks using different parameters τ and weighting the T-F units of the masks with the corresponding thresholds or simply by smoothing the binary mask in time and in frequency.

Our method has also been applied to separate stereo music. Stereo signals are often constructed by applying different gains to the different instruments on the two channels. Sometimes stereo signals are created with directional microphones placed at the same location with an 90° angle between the directional patterns. Our method is able to segregate single instruments or vocal sounds from the stereo music mixture [41].

In the evaluation the source directions are limited to seven different directions uniformly distributed on a half-circle. In a real environment, speech signals may arrive from closer directions. Also, with only two microphones, it is not possible to distinguish the two half-planes divided by the microphone array. If two arrival angles become too close, the source signals can no longer be segregated and two spatially close sources may be considered as a single source by the stopping criterion. When two sources are treated as a single source depends on the number of sources in the mixture. In the evaluation, it becomes harder to segregate all N sources as N increases. Also the level of background/microphone noise influences the spatial resolution.

Several issues in our proposed method need further investigation. Different criteria have been proposed in order to decide when the iterations should stop and when different binary masks should be merged. These criteria need to set many parameters and many experiments are needed on order to optimize these parameters. Furthermore, the optimal parameters most likely depend on a given setting, e.g. the number of sources in the mixture or the amount of reverberation. The stopping criterion was proposed for the instantaneous mixing case but applied to reverberant mixtures too. A more robust stopping criterion in the convolutive case would be a subject for future work. Our grouping criterion in the convolutive case is based on correlation between different envelopes. One could interpret the grouping problem as a problem similar to a frequency permutation problem known in blind source separation (see e.g. [65]). The merging criterion may be more reliable if it is combined with other cues, such as DOA information.

VII. Conclusion

We have proposed a novel method for separating instantaneous and anechoic mixtures with an arbitrary number of speech signals of equal power with only two microphones.

We have dealt with underdetermined mixtures by applying ICA to produce independent subsets. The subsets are used to estimate binary T-F masks, which are then applied to separate original mixtures. This iterative procedure continues until the independent subsets consist of only a single source. The segregated signals are further improved by merging masks from correlated subsets. Extensive evaluation shows that mixtures of up to seven speech signals under anechoic conditions can be separated. The estimated binary masks are close to the ideal binary masks. The proposed framework has also been applied to speech mixtures recorded in a reverberant room. We find that instantaneous ICA applied iteratively in the time domain can be used to segregate convolutive mixtures. The performance of our method compares favorably with other methods for separation of underdetermined mixtures. Because the sources are iteratively extracted from the mixture the number of sources does not need to be assumed in advance; except for reverberant mixtures our method gives a good estimate of the number of sources. Further, stereo signals are maintained throughout the processing.

APPENDIX
DIRECTIONAL GAINS

The two directional gain patterns can be approximated from two closely-spaced omnidirectional microphones. The directional response from two microphones can be written as

$$r(\theta) = s_1 e^{j\frac{kd}{2}\cos(\theta)} + s_2 e^{-j\frac{kd}{2}\cos(\theta)}, \tag{25}$$

where s_1 and s_2 are the microphone sensitivities. $k = 2\pi/\lambda = 2\pi f/c$ is the wave number. f is the acoustic frequency and $c = 343$ m/s is the speed of sound traveling in the air at 20°C. θ is the angle between the microphone array line and the source direction of arrival and d is the distance between the two microphones. If $kd \ll 1$, the microphone response can be approximated by [66]

$$r(\theta) \approx A + B\cos(\theta), \tag{26}$$

where $A = s_1 + s_2$ and $B = \frac{j}{kd}(s_1 - s_2)$. Here,

$$s_1 = \frac{1}{2}A - \frac{j}{kd}B \tag{27}$$

$$s_2 = \frac{1}{2}A + \frac{j}{kd}B. \tag{28}$$

In the Laplacian domain, $s = j\omega$, we have

$$s_1 = \frac{1}{2}A + \frac{c}{sd}B \tag{29}$$

$$s_2 = \frac{1}{2}A - \frac{c}{sd}B. \tag{30}$$

For discrete signals, we use the bilinear transform [67]

$$s = 2f_s\frac{1-z^{-1}}{1+z^{-1}}, \tag{31}$$

where f_s is the sampling frequency. The two discrete microphone sensitivities are therefore

$$s_1 = \frac{(Af_sd + cB) + (cB - Af_sd)z^{-1}}{2f_sd(1-z^{-1})} \tag{32}$$

$$s_2 = \frac{(Af_sd - cB) - (cB + Af_sd)z^{-1}}{2f_sd(1-z^{-1})} \tag{33}$$

It can be seen that the denominators in (32) and (33) have a root on the unit circle. In order to ensure stability, we modify the denominator with a factor λ so that

$$s_1 = \frac{(Af_sd + cB) + (cB - Af_sd)z^{-1}}{2f_sd(1-\lambda z^{-1})} \tag{34}$$

$$s_2 = \frac{(Af_sd - cB) - (cB + Af_sd)z^{-1}}{2f_sd(1-\lambda z^{-1})} \tag{35}$$

We choose $\lambda = 0.75$. λ controls the gain that amplifies the low frequencies. The choice of λ is not very important, because the signals are used for comparison only.

In order to obtain the directional patterns in Fig. 1 we can find A and B by solving (26) for two different gains. For $r(0) = 1$ and $r(\pi) = 0.5$, we obtain $A = 0.75$ and $B = 0.25$. For $r(0) = 0.5$ and $r(\pi) = 1$, we obtain $A = 0.75$ and $B = -0.25$.

ACKNOWLEDGMENT

The work was performed while M.S.P. was a visiting scholar at The Ohio State University Department of Computer Science and Engineering. M.S.P was supported by the Oticon Foundation. M.S.P. and J.L. are partly also supported by the European Commission through the sixth framework IST Network of Excellence: Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL), contract no. 506778. D.L.W. was supported in part by an AFOSR grant (FA9550-04-1-0117) and an AFRL grant (FA8750-04-0093).

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, September 1953.

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, pp. 1875–1902, September 2005.

[3] L. K. Hansen, "Blind separation of noisy image mixtures," in *Advances in Independent Component Analysis, Perspectives in Neural Computing*, M. Girolami, Ed. Springer-Verlag, 2000, ch. 9, pp. 165–187.

[4] K. Torkkola, "Blind separation of delayed and convolved sources," in *Unsupervised Adaptive Filtering, Blind Source Separation*, S. Haykin, Ed. Wiley, John and Sons, Incorporated, January 2000, vol. 1, ch. 8, pp. 321–375.

[5] J. M. Peterson and S. Kadambe, "A probabilistic approach for blind source separation of underdetermined convolutive mixtures," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. (ICASSP '03)*, vol. 6, 2003, pp. VI–581–584.

[6] A. K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi, "Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 888–893, July 2002.

[7] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Blind separation of more speech than sensors with less distortion by combining sparseness and ICA," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, September 2003, pp. 271–274.

[8] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[9] R. K. Olsson and L. K. Hansen, "Blind separation of more sources than sensors in convolutive mixtures," in *ICASSP*, 2006. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?4321

[10] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001.

[11] C. Jutten and J. Herault, "Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture," *Signal Processing, Elsevier*, vol. 24, no. 1, pp. 1–10, 1991.

[12] P. Comon, E. Moreau, and L. Rota, "Blind separation of convolutive mixtures: A contrast-based joint diagonalization approach," in *3rd Int. Conf. Independent Component Analysis*, San Diego, CA, December 9–13 2001.

[13] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[14] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. (4–5), pp. 411–430, 2000.

[15] L. Molgedey and H. Schuster, "Separation of independent signals using time-delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994.

[16] T. Kolenda, L. Hansen, and J. Larsen, "Signal detection using ica: Application to chat room topic spotting," *In proc. ICA'2001*, pp. 540–545, 2001. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?826

[17] L. K. Hansen, J. Larsen, and T. Kolenda, "On independent component analysis for multimedia signals," in *Multimedia Image and Video Processing*, L. Guan, S. Kung, and J. Larsen, Eds. CRC Press, 2000, pp. 175–199. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?627

[18] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[19] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[20] D. Luengo, I. Santamaria, L. Vielva, and C. Pantaleon, "Underdetermined blind separation of sparse sources with instantaneous and convolutive mixtures," in *IEEE XIII Workshop on Neural Networks for Signal Processing*, 2003, pp. 279–288.

[21] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," in *ICA 2006*, 2006, pp. 536–543.

[22] D. B. Rowe, "A Bayesian approach to blind source separation," *Journal of Interdisciplinary Mathematics*, vol. 5, no. 1, pp. 49–76, 2002.

[23] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, pp. 2353–2362, 2001.

[24] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, October 2003.

[25] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete bss for convolutive mixtures based on hierarchical clustering," in *Proc. ICA'2004*, Granada, Spain, September 22–24 2004, pp. 652–660.

[26] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," Tech. Rep., April 4 2005, to appear in IJIST (International Journal of Imaging Systems and Technology), special issue on Blind Source Separation and Deconvolution in Imaging and Image Processing.

[27] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proceedings of Eurospeech03*, Geneva, September 2003, pp. 1009–1012.

[28] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. ICASSP'2000*, vol. V, Istanbul, Turkey, June 2000, pp. 2985–2988.

[29] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley & IEEE Press, 2006.

[30] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Networks*, vol. 10, no. 3, pp. 684–697, May 1999.

[31] A. S. Bregman, *Auditory Scene Analysis*, 2nd ed. MIT Press, 1990.

[32] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.

[33] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. Academic Press, 2003.

[34] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1135–1150, September 2004.

[35] S. Roweis, "One microphone source separation," in *NIPS'00*, 2000, pp. 793–799.

[36] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *IEEE Trans. Systems, Man, and Cybernetics – Part B: Cybernetics*, vol. 34, no. 4, pp. 1763–1773, August 2004.

[37] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proc. ICASSP2005*, vol. III, March 18–23 2005, pp. 81–84.

[38] ——, "Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA," in *Proc. ICA'2004*, September 22–24 2004, pp. 898–905.

[39] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," in *Proc. ICA'2004*, Granada, Spain, September 22–24 2004, pp. 832–839.

[40] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems, "Overcomplete blind source separation by combining ICA and binary time-frequency masking," in *Proceedings of the MLSP workshop*, Mystic, CT, USA, September 28–30 2005. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?3894

[41] M. S. Pedersen, T. Lehn-Schiøler, and J. Larsen, "BLUES from music: BLind Underdetermined Extraction of Sources from Music," in *ICA2006*, 2006, pp. 392–399. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?4060

[42] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems, "Separating underdetermined convolutive speech mixtures," in *ICA 2006*, 2006, pp. 674–681. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?4068

[43] M. Ito, Y. Takeuchi, T. Matsumoto, H. Kudo, M. Kawamoto, T. Mukai, and N. Ohnishi, "Moving-source separation using directional microphones," in *Proc. ISSPIT'2002*, December 2002, pp. 523–526.

[44] G. Cauwenberghs, M. Stanacevic, and G. Zweig, "Blind broadband source localization and separation in miniature sensor arrays," in *IEEE Int. Symp. Circuits and Systems (ISCAS'2001)*, vol. 3, May 6–9 2001, pp. 193–196.

[45] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays*, ser. Digital Signal Processing. Springer, 2001.

[46] S. Araki, S. Makino, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamformers," in *Consistent & Reliable Acoustic Cues for Sound Analysis (CRAC)*, September 2001.

[47] R. D. Patterson, "The sound of a sinusoid: Spectral models," *J. Acoust. Soc. Am.*, vol. 96, pp. 1409–1418, May 1994.

[48] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *J. Audio Eng. Soc*, vol. 48, no. 11, pp. 1011–1031, November 2000.

[49] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1–4, pp. 1–24, 2001.

[50] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind sourc separation," in *Proc. 2nd IEEE Int. Workshop Indep. Compon. Anal. Signal Separation*, Helsinki, Finland, June 2000, pp. 215–220.

[51] W. M. Hartmann, *Signals, Sound, and Sensation*. Springer, 1998.

[52] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. Baltimore: The Johns Hopkins University Press, 1996.

[53] D. P. W. Ellis, "Evaluating speech separation systems," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, ch. 20, pp. 295–304.

[54] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems, "Two-microphone separation of speech mixtures [demo]," January 2006. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?4400

[55] ——, "Two-microphone separation of speech mixtures [source code]," January 2006. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?4399

[56] H. Nielsen, "UCMINF - an algorithm for unconstrained, nonlinear optimization," IMM, Technical University of Denmark, Tech. Rep. IMM-TEC-0019, 2001. [Online]. Available: http://www.imm.dtu.dk/pubdb/p.php?642

[57] T. Kolenda, S. Sigurdsson, O. Winther, L. K. Hansen, and J. Larsen, "DTU:toolbox," Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark, 2002. [Online]. Available: http:/www.imm.dtu.dk/pubdb/p.php?4043

[58] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, January 1978.

[59] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, December 1993.

[60] "ICA central." [Online]. Available: http://www.tsi.enst.fr/icacentral/algos.html

[61] "Brüel & Kjær Head and Torso Simulator, Type 4128."
[62] F. Asano, Y. Motomura, H. Asoh, and T. Matsui, "Effect of PCA in blind source separation," in *Proceedings of the Second International Workshop on ICA and BSS*, P. Pajunen and J. Karhunen, Eds., Helsinki, Finland, June 19–22 2000, pp. 57–62.
[63] N. Roman, S. Srinivasan, and D. L. Wang, "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4040–4051, 2006.
[64] C. Liu, B. C. Wheeler, W. D. O'Brien Jr., C. R. Lansing, R. C. Bilger, D. L. Jones, and A. S. Feng, "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," *J. Acoust Soc. Amer.*, vol. 110, pp. 3218–3231, 2001.
[65] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, September 2004.
[66] S. C. Thompson, "Directional patterns obtained from two or three microphones," Knowles Electronics," Technical Report, September 29 2000.
[67] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*. Prentice-Hall, 1996.

Appendix G

A Survey of Convolutive Blind Source Separation Methods

A SURVEY OF CONVOLUTIVE BLIND SOURCE SEPARATION METHODS

Michael Syskind Pedersen[1], *Jan Larsen*[2], *Ulrik Kjems*[1], *and Lucas C. Parra*[3]

[1] Oticon A/S, 2765, Smørum, Denmark, {msp, uk}@oticon.dk
[2] Technical University of Denmark, Informatics and Mathematical Modelling, 2800 Kgs. Lyngby, Denmark, jl@imm.dtu.dk
[3] The City College of New York, Biomedical Engineering, New York, NY 10031, parra@ccny.cuny.edu

ABSTRACT

In this chapter, we provide an overview of existing algorithms for blind source separation of convolutive audio mixtures. We provide a taxonomy, wherein many of the existing algorithms can be organized, and we present published results from those algorithms that have been applied to real-world audio separation tasks.

1. INTRODUCTION

During the past decades, much attention has been given to the separation of mixed sources, in particular for the *blind* case where both the sources and the mixing process are unknown and only recordings of the mixtures are available. In several situations it is desirable to recover all sources from the recorded mixtures, or at least to segregate a particular source. Furthermore, it may be useful to identify the mixing process itself to reveal information about the physical mixing system.

In some simple mixing models each recording consists of a sum of differently weighted source signals. However, in many real-world applications, such as in acoustics, the mixing process is more complex. In such systems, the mixtures are weighted and delayed, and each source contributes to the sum with multiple delays corresponding to the multiple paths by which an acoustic signal propagates to a microphone. Such filtered sums of different sources are called convolutive mixtures. Depending on the situation, the filters may consist of a few delay elements, as in radio communications, or up to several thousand delay elements as in acoustics. In these situations the sources are the desired signals, yet only the recordings of the mixed sources are available and the mixing process is unknown.

There are multiple potential applications of convolutive blind source separation. In acoustics different sound sources are recorded simultaneously with possibly multiple microphones. These sources may be speech or music, or underwater signals recorded in passive sonar [1]. In radio communications, antenna arrays receive mixtures of different communication signals [2, 3]. Source separation has also been applied to astronomical data or satellite images [4]. Finally, convolutive models have been used to interpret functional brain imaging data and bio-potentials [5, 6, 7, 8].

This chapter considers the problem of separating linear convolutive mixtures focusing in particular on acoustic mixtures. The *cocktail-party problem* has come to characterize the task of recovering speech in a room of simultaneous and independent speakers [9, 10]. Convolutive blind source separation (BSS) has often been proposed as a possible solution to this problem as it carries the promise to recover the sources exactly. The theory on linear noise-free systems establishes that a system with multiple inputs (sources) and multiple output (sensors) can be inverted under some reasonable assumptions with appropriately chosen multi-dimensional filters [11]. The challenge lies in finding these convolution filters.

There are already a number of partial reviews available on this topic [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]. The purpose of this chapter is to pro-

vide a complete survey of convolutive BSS and identify a taxonomy that can organize the large number of available algorithms. This may help practitioners and researchers new to the area of convolutive source separation obtain a complete overview of the field. Hopefully those with more experience in the field can identify useful tools, or find inspiration for new algorithms. Figure 1 provides an overview of the different topics within convolutive BSS and in which section they are covered. An overview of published results is given in Section 8.

2. THE MIXING MODEL

First we introduce the basic model of convolutive mixtures. At the discrete time index t, a mixture of N source signals $\boldsymbol{s}(t) = (s_1(t), \ldots, s_N(t))$ are received at an array of M sensors. The received signals are denoted $\boldsymbol{x}(t) = (x_1(t), \ldots, x_M(t))$. In many real-world applications the sources are said to be *convolutively* (or dynamically) mixed. The convolutive model introduces the following relation between the m'th mixed signal, the original source signals, and some additive sensor noise $v_m(t)$:

$$x_m(t) = \sum_{n=1}^{N} \sum_{k=0}^{K-1} a_{mnk} s_n(t-k) + v_m(t) \quad (1)$$

The mixed signal is a linear mixture of filtered versions of each of the source signals, and a_{mnk} represents the corresponding mixing filter coefficients. In practice, these coefficients may also change in time, but for simplicity the mixing model is often assumed stationary. In theory the filters may be of infinite length (which may be implemented as IIR systems), however, again, in practice it is sufficient to assume $K < \infty$. In matrix form, the convolutive model can be written as:

$$\boldsymbol{x}(t) \;=\; \sum_{k=0}^{K-1} \boldsymbol{A}_k \boldsymbol{s}(t-k) + \boldsymbol{v}(t), \quad (2)$$

where $\boldsymbol{A}_k$ is an $M \times N$ matrix which contains the k'th filter coefficients. $\boldsymbol{v}(t)$ is the $M \times 1$ noise vector. In the z-domain the convolutive mixture (2) can be written as:

$$\boldsymbol{X}(z) = \boldsymbol{A}(z)\boldsymbol{S}(z) + \boldsymbol{V}(z), \quad (3)$$

where $\boldsymbol{A}(z)$ is a matrix with FIR polynomials in each entry [23].

2.1. Special cases

There are some special cases of the convolutive mixture which simplify Eq. (2):

Instantaneous Mixing Model: Assuming that all the signals arrive at the sensors at the same time without being filtered, the convolutive mixture model (2) simplifies to

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{v}(t). \quad (4)$$

This model is known as the *instantaneous* or delayless (linear) mixture model. Here, $\boldsymbol{A} = \boldsymbol{A}_0$, is an $M \times N$ matrix containing the mixing coefficients. Many algorithms have been developed to solve the instantaneous mixture problem, see e.g. [17, 24].

Delayed Sources: Assuming a reverberation-free environment with propagation delays the mixing model can be simplified to

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - k_{mn}) + v_m(t) \quad (5)$$

where k_{mn} is the propagation delay between source n and sensor m.

Noise Free: In the derivation of many algorithms, the convolutive model (2) is assumed to be noise-free, i.e.:

$$\boldsymbol{x}(t) \;=\; \sum_{k=0}^{K-1} \boldsymbol{A}_k \boldsymbol{s}(t-k). \quad (6)$$

Over and Under-determined Sources: Often it is assumed that the number of sensors equals (or exceeds) the number of sources in which case linear methods may suffice to invert the linear mixing. However, if the number of sources exceeds the number of sensors the problem is under-determined, and even under perfect knowledge of the mixing system linear methods will not be able to recover the sources.

2.2. Convolutive model in the frequency domain

The convolutive mixing process (2) can be simplified by transforming the mixtures into the frequency domain. The linear convolution in the time domain can

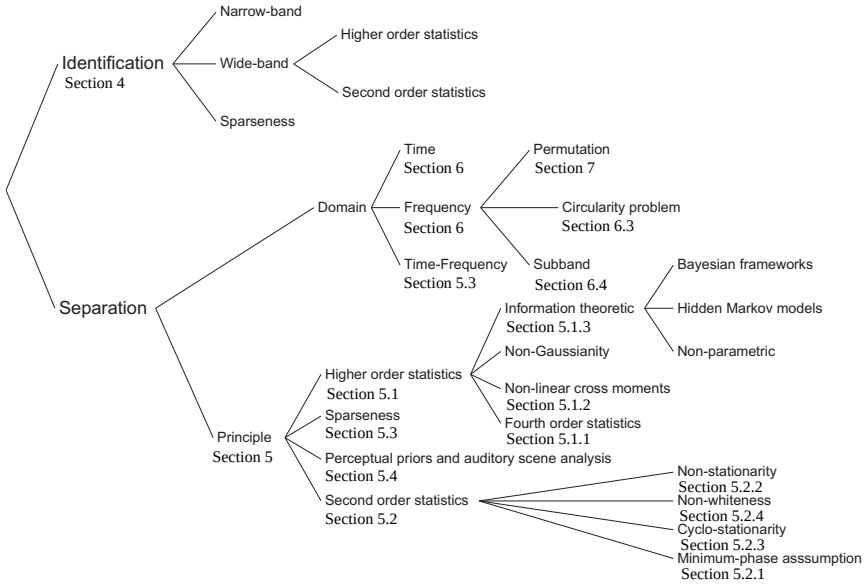

Figure 1: Overview of important areas within blind separation of convolutive sources.

be written in the frequency domain as separate multiplications for each frequency:

$$\boldsymbol{X}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{S}(\omega) + \boldsymbol{V}(\omega). \quad (7)$$

At each frequency, $\omega = 2\pi f$, $\boldsymbol{A}(\omega)$ is a complex $M \times N$ matrix, $\boldsymbol{X}(\omega)$ and $\boldsymbol{V}(\omega)$ are complex $M \times 1$ vectors, and similarly $\boldsymbol{S}(\omega)$ is a complex $N \times 1$ vector. The frequency transformation is typically computed using a discrete Fourier transform (DFT) within a time frame of size T starting at time t:

$$\boldsymbol{X}(\omega, t) = \text{DFT}([\boldsymbol{x}(t), \cdots, \boldsymbol{x}(t+T-1)]), \quad (8)$$

and correspondingly for $\boldsymbol{S}(\omega, t)$ and $\boldsymbol{V}(\omega, t)$. Often a windowed discrete Fourier transform is used:

$$\boldsymbol{X}(\omega, t) = \sum_{\tau=0}^{T-1} w(\tau)\boldsymbol{x}(t+\tau)e^{-j\omega\tau/T}, \quad (9)$$

where the window function $w(\tau)$ is chosen to minimize band-overlap due to the limited temporal aperture. By using the fast Fourier transform (FFT) convolutions can be implemented efficiently in the discrete Fourier domain, which is important in acoustics as it often requires long time-domain filters.

2.3. Block-based Model

Instead of modeling individual samples at time t one can also consider a block consisting of T samples. The equations for such a block can be written as follows:

$$\begin{aligned}
\boldsymbol{x}(t) &= \boldsymbol{A}_0\boldsymbol{s}(t) + \cdots + \boldsymbol{A}_{K-1}\boldsymbol{s}(t-K+1) \\
\boldsymbol{x}(t-1) &= \boldsymbol{A}_0\boldsymbol{s}(t-1) + \cdots + \boldsymbol{A}_{K-1}\boldsymbol{s}(t-K) \\
\boldsymbol{x}(t-2) &= \boldsymbol{A}_0\boldsymbol{s}(t-2) + \cdots + \boldsymbol{A}_{K-1}\boldsymbol{s}(t-K-1) \\
&\vdots
\end{aligned}$$

The M-dimensional output sequence can be written as an $MT \times 1$ vector:

$$\widehat{\boldsymbol{x}}(t) = \left[\boldsymbol{x}^T(t), \boldsymbol{x}^T(t-1), \cdots, \boldsymbol{x}^T(t-T+1)\right]^T, \tag{10}$$

where $\boldsymbol{x}^T(t) = [x_1(t), \cdots, x_N(t)]$. Similarly, the N-dimensional input sequence can be written as an $N(T+K-1) \times 1$ vector:

$$\widehat{\boldsymbol{s}}(t) = \left[\boldsymbol{s}^T(t), \boldsymbol{s}^T(t-1), \cdots, \boldsymbol{s}^T(t-T-K+2)\right]^T \tag{11}$$

From this the convolutive mixture can be expressed formally as:

$$\widehat{\boldsymbol{x}}(t) = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{s}}(t) + \widehat{\boldsymbol{v}}(t), \tag{12}$$

where $\widehat{\boldsymbol{A}}$ has the following form:

$$\widehat{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A}_0 & \cdots & \boldsymbol{A}_{K-1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \ddots & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_0 & \cdots & \boldsymbol{A}_{K-1} \end{bmatrix}. \tag{13}$$

The block-Toeplitz matrix $\widehat{\boldsymbol{A}}$ has dimensions $MT \times N(T+K-1)$. On the surface, Eq. (12) has the same structure as an instantaneous mixture given in Eq. (4), and the dimensionality has increased by a factor T. However, the models differ considerably as the elements within $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{s}}(t)$ are now coupled in a rather specific way.

The majority of the work in convolutive source separation assumes a mixing model with a finite impulse response (FIR) as in Eq. (2). A notable exception is the work by Cichocki which considers also an auto-regressive (AR) component as part of the mixing model [18]. The ARMA mixing system proposed there is equivalent to a first-order Kalman filter with an infinite impulse response (IIR).

3. THE SEPARATION MODEL

The objective of blind source separation is to find an estimate, $\boldsymbol{y}(t)$, which is a model of the original source signals $\boldsymbol{s}(t)$. For this, it may not be necessary to identify the mixing filters $\boldsymbol{A}_k$ explicitly. Instead, it is often sufficient to estimate separation filters $\boldsymbol{W}_l$ that remove the cross-talk introduced by the mixing process. These separation filters may have a feed-back structure with an infinite impulse response (IIR), or may have a finite impulse response (FIR) expressed as feed-forward structure.

3.1. Feed-forward Structure

An FIR separation system is given by

$$y_n(t) = \sum_{m=1}^{M} \sum_{l=0}^{L-1} w_{nml} x_m(t-l) \tag{14}$$

or in matrix form

$$\boldsymbol{y}(t) = \sum_{l=0}^{L-1} \boldsymbol{W}_l \boldsymbol{x}(t-l). \tag{15}$$

As with the mixing process, the separation system can be expressed in the z-domain as

$$\boldsymbol{Y}(z) = \boldsymbol{W}(z)\boldsymbol{X}(z), \tag{16}$$

and it can also be expressed in block Toeplitz form with the corresponding definitions for $\widehat{\boldsymbol{y}}(t)$ and $\widehat{\boldsymbol{W}}$ [25]:

$$\widehat{\boldsymbol{y}}(t) = \widehat{\boldsymbol{W}}\widehat{\boldsymbol{x}}(t). \tag{17}$$

Table 1 summarizes the mixing and separation equations in the different domains.

3.2. Relation between source and separated signals

The goal in source separation is not necessarily to recover identical copies of the original sources. Instead, the aim is to recover model sources without interferences from other sources, i.e., each separated signal $y_n(t)$ should contain signals originating from a single source only (see Figure 3). Therefore, each model source signal can be a filtered version of the original source signals, i.e.:

$$\boldsymbol{Y}(z) = \boldsymbol{W}(z)\boldsymbol{A}(z)\boldsymbol{S}(z) = \boldsymbol{G}(z)\boldsymbol{S}(z). \tag{18}$$

This is illustrated in Figure 2. The criterion for separation, i.e., interference-free signals, is satisfied if the recovered signals are permuted, and possibly scaled and filtered versions of the original signals, i.e.:

$$\boldsymbol{G}(z) = \boldsymbol{P}\boldsymbol{\Lambda}(z), \tag{19}$$

where $\boldsymbol{P}$ is a permutation matrix, and $\boldsymbol{\Lambda}(z)$ is a diagonal matrix with scaling filters on its diagonal. If one can identify $\boldsymbol{A}(z)$ exactly, and choose $\boldsymbol{W}(z)$ to be its (stable) inverse, then $\boldsymbol{\Lambda}(z)$ is an identity matrix, and one recovers the sources exactly. In source separation, instead, one is satisfied with convolved versions of the sources, i.e. arbitrary diagonal $\boldsymbol{\Lambda}(z)$.

Table 1: The convolutive mixing equation and its corresponding separation equation are shown for different domains in which blind source separation algorithms have been derived.

	Mixing Process	**Separation Model**
Time	$x_m(t) = \sum_{n=1}^{N} \sum_{k=0}^{K-1} a_{mnk} s_n(t-k) + v_m(t)$	$y_n(t) = \sum_{m=1}^{M} \sum_{l=0}^{L-1} w_{nml} x_m(t-l)$
	$\boldsymbol{x}(t) = \sum_{k=0}^{K-1} \boldsymbol{A}_k \boldsymbol{s}(t-k) + \boldsymbol{v}(t)$	$\boldsymbol{y}(t) = \sum_{l=0}^{L-1} \boldsymbol{W}_l \boldsymbol{x}(t-l)$
z-domain	$\boldsymbol{X}(z) = \boldsymbol{A}(z)\boldsymbol{S}(z) + \boldsymbol{V}(z),$	$\boldsymbol{Y}(z) = \boldsymbol{W}(z)\boldsymbol{X}(z)$
Frequency domain	$\boldsymbol{X}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{S}(\omega) + \boldsymbol{V}(\omega)$	$\boldsymbol{Y}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega)$
Block Toeplitz Form	$\widehat{\boldsymbol{x}}(t) = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{s}}(t)$	$\widehat{\boldsymbol{y}}(t) = \widehat{\boldsymbol{W}}\widehat{\boldsymbol{x}}(t)$

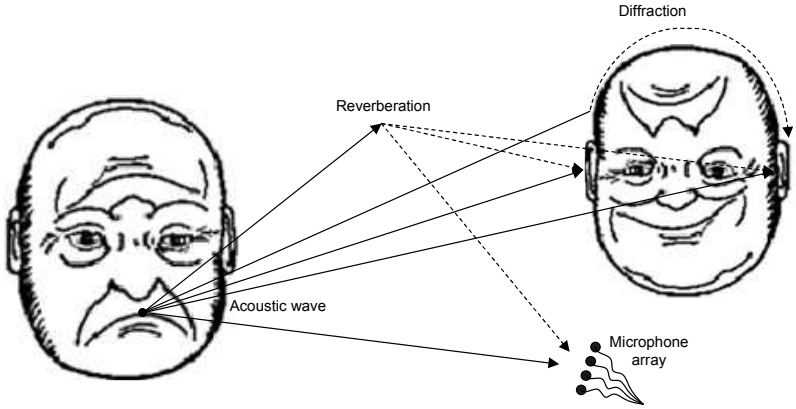

Figure 3: Illustration of a speech source. It is not always clear what the desired acoustic source should be. It could be the acoustic wave as emitted from the mouth. This corresponds to the signal as it would have been recorded in an anechoic chamber in the absence of reverberations. It could be the individual source as it is picked up by a microphone array. Or it could be the speech signal as it is recorded on microphones close to the two eardrums of a person. Due to reverberations and diffraction, the recorded speech signal is most likely a filtered version of the signal at the mouth. NOTE TO PUBLISHER: THIS FIGURE IS A PLACE HOLDER ONLY. IT WILL REQUIRE MODIFICATION BY YOUR PRODUCTION DEPARTMENT. THE FACES ARE TO BE REPLACED WITH ANY REASONABLE REPRESENTATION OF A "SOURCE" AND "RECEIVER" OF A SPEECH SIGNAL.

3.3. Feedback Structure

The mixing system given by (2) is called a feed-forward system. Often such FIR filters are inverted by a feedback structure using IIR filters. The esti-

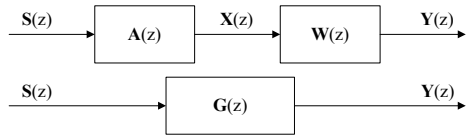

Figure 2: The source signals $\boldsymbol{Y}(z)$ are mixed with the mixing filter $\boldsymbol{A}(z)$. An estimate of the source signals is obtained through an unmixing process, where the received signals $\boldsymbol{X}(z)$ are unmixed with the filter $\boldsymbol{W}(z)$. Each estimated source signal is then a filtered version of the original source, i.e., $\boldsymbol{G}(z) = \boldsymbol{W}(z)\boldsymbol{A}(z)$. Note that the mixing and the unmixing filters do not necessarily have to be of the same order.

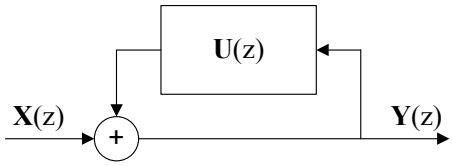

Figure 4: Recurrent unmixing (feedback) network given by equation (21). The received signals are separated by a IIR filter to achieve an estimate of the source signals.

mated sources are then given by the following equation, where the number of sources equals the number of receivers:

$$y_n(t) = x_n(t) + \sum_{l=0}^{L-1} \sum_{m=1}^{M} u_{nml} y_m(t-l), \quad (20)$$

and u_{nml} are the IIR filter coefficients. This can also be written in matrix form

$$\boldsymbol{y}(t) = \boldsymbol{x}(t) + \sum_{l=0}^{L-1} \boldsymbol{U}(l)\boldsymbol{y}(t-l). \quad (21)$$

The architecture of such a network is shown in Figure 4. In the z-domain, (21) can be written as [26]

$$\boldsymbol{Y}(z) = (\boldsymbol{I} + \boldsymbol{U}(z))^{-1}\boldsymbol{X}(z), \quad (22)$$

provided $(\boldsymbol{I}+\boldsymbol{U}(z))^{-1}$ exists and all poles are within the unit circle. Therefore,

$$\boldsymbol{W}(z) = (\boldsymbol{I} + \boldsymbol{U}(z))^{-1}. \quad (23)$$

The feed-forward and the feedback network can be combined to a so-called hybrid network, where a feed-forward structure is followed by a feedback network [27, 28].

3.4. Example: The TITO system

A special case, which is often used in source separation work is the two-input-two-output (TITO) system [29]. It can be used to illustrate the relationship between the mixing and unmixing system, feed-forward and feed-back structures, and the difference between recovering sources versus generating separated signals.

Figure 5 shows a diagram of a TITO mixing and unmixing system. The signals recorded at the two microphones are described by the following equations:

$$x_1(z) = s_1(z) + a_{12}(z)s_2(z) \quad (24)$$
$$x_2(z) = s_2(z) + a_{21}(z)s_1(z). \quad (25)$$

The mixing system is thus given by

$$\mathbf{A}(z) = \begin{bmatrix} 1 & a_{12}(z) \\ a_{21}(z) & 1 \end{bmatrix}, \quad (26)$$

which has the following inverse

$$[\mathbf{A}(z)]^{-1} = \frac{1}{1 - a_{12}(z)a_{21}(z)} \begin{bmatrix} 1 & -a_{12}(z) \\ -a_{21}(z) & 1 \end{bmatrix}. \quad (27)$$

If the two mixing filters $a_{12}(z)$ and $a_{21}(z)$ can be identified or estimated as $\bar{a}_{12}(z)$ and $\bar{a}_{21}(z)$, the separation system can be implemented as

$$y_1(z) = x_1(z) - \bar{a}_{12}(z)x_2(z) \quad (28)$$
$$y_2(z) = x_2(z) - \bar{a}_{21}(z)x_1(z). \quad (29)$$

A sufficient FIR separating filter is

$$\mathbf{W}(z) = \begin{bmatrix} 1 & -a_{12}(z) \\ -a_{21}(z) & 1 \end{bmatrix} \quad (30)$$

However, the exact sources are not recovered until this model sources $\boldsymbol{y}(t)$ are filtered with the IIR filter

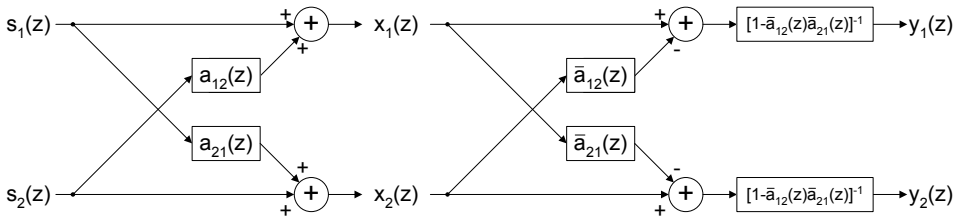

Figure 5: The two mixed sources s_1 and s_2 are mixed by a FIR mixing system. The system can be inverted by an alternative system, if the estimates $\bar{a}_{12}(z)$ and $\bar{a}_{21}(z)$ of the mixing filters $a_{12}(z)$ and $a_{12}(z)$ are known. Further, if the filter $[1 - \bar{a}_{12}(z)\bar{a}_{21}(z)]^{-1}$ is stable, the sources can be perfectly reconstructed as they were recorded at the microphones.

$[1-\bar{a}_{12}(z)\bar{a}_{21}(z)]^{-1}$. Thus, the mixing process is invertible, provided this inverse IIR filter is stable. If a filtered version of the separated signals is acceptable, we may disregard the potentially unstable recursive filter in (27) and limit separation to the FIR inversion of the mixing system with (30).

4. IDENTIFICATION

Blind identification deals with the problem of estimating the coefficients in the mixing process $\boldsymbol{A}_k$. In general, this is an ill-posed problem, and no unique solution exists. In order to determine the conditions under which the system is blindly identifiable, assumptions about the mixing process and the input data are necessary. Even though the mixing parameters are known, it does not imply that the sources can be recovered. Blind identification of the sources refers to the exact recovery of sources. Therefore one should distinguish between the conditions required to identify the mixing system and the conditions necessary to identify the sources. The limitations for the exact recovery of sources when the mixing filters are known are discussed in [30, 11, 31]. For a recent review on identification of acoustic systems see [32]. This review considers single and multiple input-output systems for the case of completely known sources as well as blind identification, where both the sources and the mixing channels are unknown.

5. SEPARATION PRINCIPLE

Blind source separation algorithms are based on different assumptions on the sources and the mixing system. In general, the sources are assumed to be *independent* or at least decorrelated. The separation criteria can be divided into methods based on higher order statistics (HOS), and methods based on second order statistics (SOS). In convolutive separation it is also assumed that sensors receive N linearly independent versions of the sources. This means that the sources should originate from different locations in space (or at least emit signals into different orientations) and that there are at least as many sources as sensors for separation, i.e., $M \geq N$.

Instead of spatial diversity a series of algorithms make strong assumptions on the statistics of the sources. For instance, they may require that sources do not overlap in the time-frequency domain, utilizing therefore a form of *sparseness* in the data. Similarly, some algorithms for acoustic mixtures exploit regularity in the sources such as common onset, harmonic structure, etc. These methods are motivated by the present understanding on the grouping principles of auditory perception commonly referred to as "Auditory Scene Analysis". In radio communications a reasonable assumption on the sources is cyclo-stationarity (see Section 5.2.3) or the fact that source signals take on only discrete values. By using such strong assumptions on the source statistics it is sometimes possible to relax the conditions on the number of sensors, e.g. $M < N$. The different

Table 2: Assumptions made for separation

$N < M$	$N = M$	$N > M$
• Subspace methods [25].	• Asymmetric sources by 2nd and 3rd order cumulants [33]	• Non-stationary, column-wise co-prime sources [34]
• Reduction of problem to instantaneous mixture [35, 36, 37, 25, 38, 39, 40]	• Separation criteria based on SOS and HOS for 2×2 system [41]	• Cross-cumulants [42, 43]
	• Uncorrelated sources with distinct power spectra [44].	• Sparseness in time and frequency [45, 46, 47]
	• 2×2, temporally colored sources [48] • Cumulants of order > 2, ML principle [49]. • Known cross filters [41] • 2×2, each with different correlation [50, 51], extended to $M \times M$ in [52] • Non-linear odd functions [53, 26, 54, 55, 56, 57, 58] • Non-linearity approximating the cdf see e.g. [59]	

criteria for separation are summarized in Table 5.

5.1. Higher Order Statistics

Source separation based on higher order statistics is based on the assumption that the sources are statistically independent. Many algorithms are based on minimizing second and fourth order dependence between the model signals. A way to express independence is that all the cross-moments between the model sources are zero, i.e.:

$$E[y_n(t)^\alpha, y_{n'}(t-\tau)^\beta] = 0 \quad , \\ n \neq n', \alpha, \beta = \{1, 2, \ldots\}, \forall \tau,$$

where $E[\cdot]$ denotes the statistical expectation. Successful separation using higher order moments requires that the underlying sources are non-Gaussian (with the exception of at most one), since Gaussian sources have zero higher cumulants [60] and therefore equations (31) are trivially satisfied without providing useful conditions.

5.1.1. 4th-order statistic

It is not necessary to minimize all cross-moments in order to achieve separation. Many algorithms are based on minimization of second and fourth order dependence between the model source signals. This minimization can either be based on second and fourth order cross-moments or second and fourth order cross-cumulants. Whereas off-diagonal elements of cross-cumulants vanish for independent signals the same is not true for all cross-moments [61]. Source separation based on cumulants has been used by several authors. Separation of convolutive mixtures by means of fourth order cumulants has been addressed by [62, 63, 41, 64, 65, 66, 67, 68, 61, 69, 70, 71]. In [72, 73, 74], the JADE algorithm for complex-valued signals [75] was applied in the frequency domain in order to separate convolved source signals. Other cumulant-based algorithms in the frequency domain are given in [76, 77]. Second and third order cumulants have been used by Ye et al. (2003) [33] for separation of asymmetric signals. Other algorithms based on higher order cumulants can be found in [78, 79]. For separation of more sources than sensors, cumulant-based approaches have been proposed in [80, 70]. Another popular 4th-order measure of non-Gaussianity is *kurtosis*. Separation of convolutive sources based on kurtosis has been addressed in [81, 82, 83].

5.1.2. Non-linear cross-moments

Some algorithms apply higher order statistics for separation of convolutive sources indirectly using non-linear functions by requiring:

$$E[f(y_n(t)), g(y_{n'}(t-\tau))] = 0. \qquad (31)$$

Here $f(\cdot)$ and $g(\cdot)$ are odd non-linear functions. The Taylor expansion of these functions captures higher order moments and this is found sufficient for separation of convolutive mixtures. This approach was among of the first for separation of convolutive mixtures [53] extending an instantaneous blind separation algorithm by Herault and Jutten (H-J) [84]. In Back and Tsoi (1994) [85], the H-J algorithm was applied in the frequency domain, and this approach was further developed in [86]. In the time domain, the approach of using non-linear odd functions has been used by Nguyen Thi and Jutten (1995) [26]. They present a group of TITO (2×2) algorithms based on 4th order cumulants, non-linear odd functions, and second and fourth order cross-moments. This algorithm has been further examined by Serviere (1996) [54], and it has also been used by Ypma et al. (2002) [55]. In Cruces and Castedo (1998) [87] a separation algorithm can be found, which can be regarded as a generalization of previous results from [26, 88]. In Li and Sejnowski (1995) [89], the H-J algorithm has been used to determine the delays in a beamformer. The H-J algorithm has been investigated further by Charkani and Deville (1997,1999) [90, 57, 58]. They extended the algorithm further to colored sources [56, 91]. Depending on the distribution of the source signals, also optimal choices of non-linear functions were found. For these algorithms, the mixing process is assumed to be minimum-phase, since the H-J algorithm is implemented as a feedback network. A natural gradient algorithm based on the H-J network has been applied in Choi et al. (2002) [92]. A discussion of the H-J algorithm for convolutive mixtures can be found in Berthommier and Choi (2003) [93]. For separation of two speech signals with two microphones, the H-J model fails if the two speakers are located on the same side, as the appropriate separating filters can not be implemented without delaying one of the sources and the FIR filters are constrained to be causal. HOS independence obtained by applying antisymmetric non-linear functions has also been used in [94, 95].

5.1.3. Information Theoretic

Statistical independence between the source signals can also be expressed in terms of the probability density functions (PDF). If the model sources $\boldsymbol{y}$ are independent, the joint probability density function can be written as

$$p(\boldsymbol{y}) = \prod_n p(y_n). \tag{32}$$

This is equivalent to stating that model sources y_n do not carry mutual information. Information theoretic methods for source separation are based on maximizing the entropy in each variable. Maximum entropy is obtained when the sum of the entropy of each variable y_n equals the total joint-entropy in $\boldsymbol{y}$. In this limit variables do not carry any mutual information and are hence mutually independent [96]. A well-known algorithm based on this idea is the Infomax algorithm by Bell and Sejnowski (1995) [97] which was significantly improved in convergence speed by the natural gradient method of Amari [98]. The Infomax algorithm can also be derived directly from model equation (32) using Maximum Likelihood [99], or equivalently, using the Kullback-Leibler divergence between the empirical distribution and the independence model [100].

In all instances it is necessary to assume or model the probability density function $p_s(s_n)$ of the underlying sources s_n. In doing so, one captures higher order statistics of the data. In fact, most information theoretic algorithms contain expressions rather similar to the non-linear cross-statistics in (31) with $f(y_n) = \partial \ln p_s(y_n)/\partial y_n$, and $g(y_n) = y_n$. The PDF is either assumed to have a specific form or it is estimated directly from the recorded data, leading to *parametric* and *non-parametric* methods respectively [16]. In non-parametric methods the PDF is captured implicitly through the available data. Such methods have been addressed in [101, 102, 103]. However, the vast majority of convolutive algorithms have been derived based on explicit parametric representations of the PDF.

Infomax, the most common parametric method, was extended to the case of convolutive mixtures by Torkkola (1996) [59] and later by Xi and Reilly (1997,1999) [104, 105]. Both feed-forward and feedback networks were shown. In the frequency domain it is necessary to define the PDF for complex variables. The resulting analytic non-linear functions can be derived with [106, 107]

$$\boldsymbol{f}(Y) = -\frac{\partial \ln p(|Y|)}{\partial |Y|} e^{j \arg(Y)}, \tag{33}$$

where $p(Y)$ is the probability density of the model source $Y \in \mathbb{C}$. Some algorithms assume circular sources in the complex domain, while other algorithms have been proposed that specifically assume non-circular sources [108, 109].

The performance of the algorithm depends to a certain degree on the selected PDF. It is important to determine if the data has super-Gaussian or sub-Gaussian distributions. For speech commonly a Laplace distribution is used. The non-linearity is also known as the Bussgang non-linearity [110]. A connection between the Bussgang blind equalization algorithms and the Infomax algorithm is given in Lambert and Bell (1997) [111]. Multichannel blind deconvolution algorithms derived from the Bussgang approach can be found in [112, 23, 111]. These learning rules are similar to those derived in Lee et al. (1997) [113].

Choi et al. (1999) [114] have proposed a *non-holonomic* constraint for multichannel blind deconvolution. Non-holonomic means that there are some restrictions related to the direction of the update. The non-holonomic constraint has been applied for both a feed-forward and a feedback network. The non-holonomic constraint was applied to allow the natural gradient algorithm by Amari et al. (1997) [98] to cope with over-determined mixtures. The non-holonomic constraint has also been used in [115, 116, 117, 118, 119, 120, 121, 122]. Some drawbacks in terms of stability and convergence in particular when there are large power fluctuations within each signal (e.g. for speech) have been addressed in [115].

Many algorithms have been derived from (32) directly using Maximum Likelihood (ML) [123]. The ML approach has been applied in [124, 125, 126, 127, 128, 129, 99, 130, 131, 132]. A method closely related to the ML is the Maximum a Posteriori (MAP) methods. In MAP methods, prior information about the parameters of the model are taken into account. MAP has been used in [23, 133, 134, 135, 136, 137, 138, 139, 140, 141].

The convolutive blind source separation problem has also been expressed in a Bayesian formulation [142]. The advantage of a Bayesian formulation is that one can derive an optimal, possibly non-linear estimator of the sources enabling the estimation of more sources than the number of available sensors. The Bayesian framework has also been applied in [143, 144, 145, 135, 137].

A strong prior on the signal can also be realized via Hidden Markov Models (HMMs). HMMs can incorporate state transition probabilities of different sounds [136]. A disadvantage of HMMs is that they require prior training and they carry a high computational cost [146]. HMMs have also been used in [147, 148].

5.2. Second Order Statistics

In some cases, separation can be based on second order statistics (SOS) by requiring only non-correlated sources rather then the stronger condition of independence. Instead of assumptions on higher order statistics these methods make alternate assumptions such as the non-stationarity of the sources [149], or a minimum phase mixing system [50]. By itself, however, second order conditions are not sufficient for separation. Sufficient conditions for separation are given in [150, 15]. The main advantage of SOS is that they are less sensitive to noise and outliers [13], and hence require less data for their estimation [50, 150, 151, 34, 152]. The resulting algorithms are often also easier to implement and computationally efficient.

5.2.1. Minimum-phase mixing

Early work by Gerven and Compernolle [88] had shown that two source signals can be separated by decorrelation if the mixing system is minimum phase. The FIR coupling filters have to be strictly causal and their inverses stable. The condition for stability is given as $|a_{12}(z)a_{21}(z)| < 1$, where $a_{12}(z)$ and $a_{21}(z)$ are the two coupling filters (see Figure 5). These conditions are not met if the mixing process is non-minimum phase [153]. Algorithms based on second order statistic assuming minimum-phase mixing can be found in [154, 38, 39, 51, 50, 155, 156, 52, 157, 158].

5.2.2. Non-stationarity

The fact that many signals are non-stationary has been successfully used for source separation. Speech signals in particular can be considered non-stationary on time scales beyond 10 ms [159, 160]).

The temporally varying statistics of non-stationarity sources provides additional information for separation. Changing locations of the sources, on the other hand, generally complicate source separation as the mixing channel changes in time. Separation based on decorrelation of non-stationary signals was proposed by Weinstein et al. (1993) [29] who suggested that minimizing cross-powers estimated during different stationarity times should give sufficient conditions for separation. Wu and Principe (1999) proposed a corresponding joint diagonalization algorithm [103, 161] extending an earlier method developed for instantaneous mixtures [162]. Kawamoto et al. (1998) extend an earlier method [163] for instantaneous mixtures to the case of convolutive mixtures in the time domain [164, 153] and frequency domain [165]. This approach has also been employed in [166, 167, 168, 169] and an adaptive algorithm was suggested by Aichner et al. (2003) [170]. By combining this approach with a constraint based on whiteness, the performance can be further improved [171].

Note that not all of these papers have used simultaneous decorrelation, yet, to provide sufficient second-order constraints it is necessary to minimize multiple cross-correlations simultaneously. An effective frequency domain algorithm for simultaneous diagonalization was proposed by Parra and Spence (2000) [149]. Second-order statistics in the frequency domain is captured by the cross-power spectrum,

$$\boldsymbol{R}_{yy}(\omega,t) = E\left[\boldsymbol{Y}(\omega,t)\boldsymbol{Y}^H(\omega,t)\right] \quad (34)$$

$$= \boldsymbol{W}(\omega)\boldsymbol{R}_{xx}(\omega,t)\boldsymbol{W}^H(\omega), \quad (35)$$

where the expectations are estimated around some time t. The goal is to minimize the cross-powers on the off-diagonal of this matrix, e.g. by minimizing:

$$J = \sum_{t,\omega} \|\boldsymbol{R}_{yy}(\omega,t) - \boldsymbol{\Lambda}_y(\omega,t)\|^2, \quad (36)$$

where $\boldsymbol{\Lambda}_y(\omega,t)$ is an estimate of the cross-power spectrum of the model sources and is assumed to be diagonal. This cost function simultaneously captures multiple times and multiple frequencies, and has to be minimized with respect to $\boldsymbol{W}(\omega)$ and $\boldsymbol{\Lambda}_y(\omega,t)$ subject to some normalization constraint. If the source signals are non-stationary the cross-powers estimated at different times t differ and provide independent conditions on the filters $\boldsymbol{W}(\omega)$. This algorithm has been successfully used on speech signals [172, 173] and investigated further by Ikram and Morgan (2000, 2001, 2002, 2005) [174, 175, 176] to determine the trade-offs between filter length, estimation accuracy, and stationarity times. Long filters are required to cope with long reverberation times of typical room acoustics, and increasing filter length also reduces the error of using the circular convolution in (35) (see Section 6.3). However, long filters increase the number of parameters to be estimated and extend the effective window of time required for estimating cross-powers thereby potentially loosing the benefit of non-stationarity of speech signals. A number of variations of this algorithm have been proposed subsequently, including time domain implementations [177, 178, 179], and other method that incorporate additional assumptions [180, 174, 181, 182, 183, 184, 185, 186, 187]. A recursive version of the algorithm was given in Ding et al. (2003) [188]. In Robeldo-Arnuncio and Juang (2005) [189], a version with non-causal separation filters was suggested. Based on a different way to express (35), Wang et al. (2003, 2004, 2005) [190, 191, 148, 192] propose a slightly different separation criterion, that leads to a faster convergence than the original algorithm by Parra and Spence (2000) [149].

Other methods that exploit non-stationarity have been derived by extending the algorithm of Molgedey and Schuster (1994) [193] to the convolutive case [194, 195] including a common two step approach of 'sphering' and rotation [159, 196, 197, 198, 199]. (Any matrix, for instance matrix $\boldsymbol{W}$, can be represented as a concatenation of a rotation with subsequent scaling (which can be used to remove second-order moments, i.e. sphering) and an additional rotation).

In Yin and Sommen (1999) [160] a source separation algorithm was presented based on non-stationarity and a model of the direct path. The reverberant signal paths are considered as noise. A time domain decorrelation algorithm based on different cross-correlations at different time lags is given in Ahmed et al. (1999) [200]. In Yin and Sommen (2000) [201] the cost function is based on minimization of the power spectral density between the

source estimates. The model is simplified by assuming that the acoustic transfer function between the source and closely spaced microphones is similar. The simplified model requires fewer computations. An algorithm based on joint diagonalization is suggested in Rahbar and Reilly (2003, 2005) [152, 152]. This approach exploits the spectral correlation between the adjacent frequency bins in addition to non-stationarity. Also in [202, 203] a diagonalization criterion based on non-stationarity has been used.

In Olsson and Hansen (2004) [139, 138] the non-stationary assumption has been included in a state-space Kalman filter model.

In Buchner et al. (2003) [204], an algorithm that uses a combination of non-stationarity, non-Gaussianity and non-whiteness has been suggested. This has also been applied in [205, 206, 207]. In the case of more source signals than sensors, an algorithm based on non-stationarity has also been suggested [70]. In this approach, it is possible to separate three signals: a mixture of two non-stationary source signals with short-time stationarity and one signal which is long-term stationary. Other algorithms based on the non-stationary assumptions can be found in [208, 209, 210, 211, 212, 213, 214].

5.2.3. Cyclo-stationarity

If a signal is assumed to be cyclo-stationary, the signals' cumulative distribution is invariant with respect to time shifts of some period T or any integer multiples of T. Further, a signal is said to be wide-sense cyclo-stationary if the signals mean and autocorrelation is invariant to shifts of some period T or any integer multiples of T [215], i.e.:

$$E[\boldsymbol{s}(t)] = E[\boldsymbol{s}(t+\alpha T)] \quad (37)$$

$$E[\boldsymbol{s}(t_1), \boldsymbol{s}(t_2)] = E[\boldsymbol{s}(t_1+\alpha T), \boldsymbol{s}(t_2+\alpha T)] \quad (38)$$

An example of a cyclo-stationary signal is a random amplitude sinusoidal signal. Many communication signals have the property of cyclo-stationarity, and voiced speech is sometimes considered approximately cyclo-stationary [216]. This property has been used explicitly to recover mixed source in e.g. [216, 217, 218, 55, 219, 220, 34, 118, 221, 222]. In [220] cyclo-stationarity is used to solve the frequency permutation problem (see Section 6.1) and in [118] it is used as additional criteria to improve separation performance.

5.2.4. Non-whiteness

Many natural signals, in particular acoustic signals, are temporally correlated. Capturing this property can be beneficial for separation. For instance, capturing temporal correlations of the signals can be used to reduce a convolutive problem to an instantaneous mixture problem, which is then solved using additional properties of the signal [35, 25, 36, 37, 38, 39, 40]. In contrast to instantaneous separation where decorrelation may suffice for non-white signals, for convolutive separation additional conditions on the system or the sources are required. For instance, Mei and Yin (2004) [223] suggest that decorrelation is sufficient provided the sources are an ARMA process.

5.3. Sparseness in the Time/Frequency domain

Numerous source separation applications are limited by the number of available microphones. It is in not always guaranteed that the number of sources is less than or equal to the number of sensors. With linear filters it is in general not possible to remove more than $M-1$ sources from the signal. By using non-linear techniques, in contrast, it may be possible to extract a larger number of source signals. One technique to separate more sources than sensors is based on sparseness. If the source signals do not overlap in the time-frequency (T-F) domain it is possible to separate them. A mask can be applied in the T-F domain to attenuate interfering signal energy while preserving T-F bins where the signal of interest is dominant. Often a binary mask is used giving perceptually satisfactory results even for partially overlapping sources [224, 225]. These methods work well for anechoic mixtures (delay-only) [226]. However, under reverberant conditions, the T-F representation of the signals is less sparse. In a mildly reverberant environment ($T_{60} \leq 200$ ms) under-determined sources have been separated with a combination of independent component analysis (ICA) and T-F masking [47]. The first $N-M$ signals are removed from the mixtures by applying a T-F mask estimated from the direction of arrival of the signal (cf. Section 7.1). The

remaining M sources are separated by conventional BSS techniques. When a binary mask is applied to a signal, artifacts (musical noise) are often introduced. In order to reduce the musical noise, smooth masks have been proposed [227, 47].

Sparseness has also been used as a post processing step. In [77], a binary mask has been applied as post-processing to a standard BSS algorithm. The mask is determined by comparison of the magnitude of the outputs of the BSS algorithm. Hereby a higher signal to interference ratio is obtained. This method was further developed by Pedersen et al. (2005, 2006) in order to segregate under-determined mixtures [228, 229]. Because the T-F mask can be applied to a single microphone signal, the segregated signals can be maintained as e.g. stereo signals.

Most of the T-F masking methods do not effectively utilize information from more than two microphones because the T-F masks are applied to a single microphone signal. However, some methods have been proposed that utilize information from more than two microphones [225, 230].

Clustering has also been used for sparse source separation [231, 232, 233, 234, 140, 141, 235, 236, 230]. If the sources are projected into a space where each source groups together, the source separation problem can be solved with clustering algorithms. In [46, 45] the mask is determined by clustering with respect to amplitude and delay differences.

In particular when extracting sources from single channels sparseness becomes an essential criterion. Pearlmutter and Zador (2004) [237] use strong prior information on the source statistic in addition to knowledge of the head-related transfer functions (HRTF). An *a priori* dictionary of the source signals as perceived through a HRTF makes it possible to separate source signals with only a single microphone. In [238], *a priori* knowledge is used to construct basis functions for each source signals to segregate different musical signals from their mixture. Similarly, in [239, 240] sparseness has been assumed in order to extract different music instruments.

Techniques based on sparseness are further discussed in the survey by O'Grady et al. (2005) [21].

5.4. Priors from Auditory Scene Analysis and Psycho-Acoustics

Some methods rely on insights gained from studies of the auditory system. The work by Bergman [241] on auditory scene analysis characterized the cues used by humans to segregate sound sources. This has motivated computational algorithms that are referred to as computational auditory scene analysis (CASA). For instance, the phenomenon of auditory masking, i.e., the dominant perception of the signal with largest signal power has motivated the use of T-F masking for many years [242]. In addition to the direct T-F masking methods outlined above, separated sources have been enhanced by filtering based on perceptual masking and auditory hearing thresholds [191, 243].

Another important perceptual cue that has been used in source separation is pitch frequency, which typically differs for simultaneous speakers [135, 244, 245, 137, 138, 147]. In Tordini and Piazza (2000) [135] pitch is extracted from the signals and used in a Bayesian framework. During unvoiced speech, which lacks a well-defined pitch they use an ordinary blind algorithm. In order to separate two signals with one microphone, Gandhi and Hasegawa-Johnson (2004) [137] have proposed a state-space separation approach with strong *a priori* information. Both pitch and Mel-frequency cepstral coefficients (MFCC) were used in their method. A pitch codebook as well as an MFCC codebook have to be known in advance. Olsson and Hansen [138] have used a Hidden-Markov Model, where the sequence of possible states is limited by the pitch frequency that is extracted in the process. As a pre-processing step to source separation, Furukawa et al. (2003) [245] use pitch in order to determine the number of source signals.

A method for separation of more sources than sensors is given in Barros et al. (2002) [244]. They combined ICA with CASA techniques such as pitch tracking and auditory filtering. Auditory filter banks are used in order to model the cochlea. In [244] wavelet filtering has been used for auditory filtering. Another commonly used auditory filter bank is the Gammatone filter-bank (see e.g. Patterson (1994) [246] or [247, 248]). In Roman et al. (2003) [248] binaural cues have been used to segregate sound sources, whereby inter-aural time and inter-aural intensity differences (ITD, IID) have been used to

group the source signals.

6. TIME VERSUS FREQUENCY DOMAIN

The blind source separation problem can either be expressed in the time domain

$$\boldsymbol{y}(t) = \sum_{l=0}^{L-1} \boldsymbol{W}_l \boldsymbol{x}(t-l) \tag{39}$$

or in the frequency domain

$$\boldsymbol{Y}(\omega, t) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega, t). \tag{40}$$

A survey of frequency-domain BSS is provided in [22]. In Nishikawa et al. (2003) [249] the advantages and disadvantages of the time and frequency domain approaches have been compared. This is summarized in Table 3.

An advantage of blind source separation in the frequency domain is that the separation problem can be decomposed into smaller problems for each frequency bin in addition to the significant gains in computational efficiency. The convolutive mixture problem is reduced to "instantaneous" mixtures for each frequency. Although this simplifies the task of convolutive separation a set of new problems arise: The frequency domain signals obtained from the DFT are complex-valued. Not all instantaneous separation algorithms are designed for complex-valued signals. Consequently, it is necessary to modify existing algorithms correspondingly [250, 251, 252, 5]. Another problem that may arise in the frequency domain is that there are no longer enough data points available to evaluate statistical independence [131]. For some algorithms [149] it is necessary that the frame size T of the DFT is much longer than the length of the room impulse response K (see Section 6.3). Long frames result in fewer data samples per frequency [131], which complicates the estimation of the independence criteria. A method that copes with this issue has been proposed by Servière (2004) [253].

6.1. Frequency Permutations

Another problem that arises in the frequency domain is the permutation and scaling ambiguity. If the convolutive problem is treated for each frequency as a separate problem, the source signals in each frequency bin may be estimated with an arbitrary permutation and scaling, i.e.:

$$\boldsymbol{Y}(\omega, t) = \boldsymbol{P}(\omega)\boldsymbol{\Lambda}(\omega)\boldsymbol{S}(\omega, t). \tag{41}$$

If the permutation $\boldsymbol{P}(\omega)$ is not consistent across frequency then converting the signal back to the time domain will combine contributions from different sources into a single channel, and thus annihilate the separation achieved in the frequency domain. An overview of the solutions to this permutation problem is given in Section 7. The scaling indeterminacy at each frequency – arbitrary solution for $\boldsymbol{\Lambda}(\omega)$ – will result in an overall filtering of the sources. Hence, even for perfect separation the separated sources may have a different frequency spectrum than the original sources.

6.2. Time-Frequency Algorithms

Algorithms that define a separation criteria in the time domain do typically not exhibit frequency permutation problems, even when computations are executed in the frequency domain. A number of authors have therefore used time-domain criteria combined with frequency domain implementations that speed up computations. [254, 113, 255, 256, 121, 101, 257, 179, 171]. However, note that second-order criteria may be susceptible to the permutation problem even if they are formulated in the time domain [184].

6.3. Circularity Problem

When the convolutive mixture in the time domain is expressed in the frequency domain by the DFT, the convolution becomes separate multiplications, i.e.:

$$\boldsymbol{x}(t) = \boldsymbol{A} * \boldsymbol{s}(t) \longleftrightarrow \boldsymbol{X}(\omega, t) \approx \boldsymbol{A}(\omega)\boldsymbol{S}(\omega, t). \tag{42}$$

However, this is only an approximation which is exact only for periodic $\boldsymbol{s}(t)$ with period T, or equivalently, if the time convolution is *circular*:

$$\boldsymbol{x}(t) = \boldsymbol{A} \circledast \boldsymbol{s}(t) \longleftrightarrow \boldsymbol{X}(\omega) = \boldsymbol{A}(\omega)\boldsymbol{S}(\omega). \tag{43}$$

For a *linear convolution* errors occur at the frame boundary, which are conventionally corrected with

Table 3: Advantages and disadvantages for separation in the time domain or separation in the frequency domain.

Time Domain		**Frequency Domain**	
Advantages	Disadvantages	Advantages	Disadvantages
• The independence assumption holds better for full-band signals	• Degradation of convergence in strong reverberant environment	• The convolutive mixture can be transformed into instantaneous mixture problems for each frequency bin	• For each frequency band, there is a permutation and a scaling ambiguity which needs to be solved
• Possible high convergence near the optimal point	• Many parameters need to be adjusted for each iteration step	• Due to the FFT, computations are saved compared to an implementation in the time domain	• Problem with too few samples in each frequency band may cause the independence assumption to fail
		• Convergence is faster	• Circular convolution deteriorates the separation performance.
			• Inversion of $\mathbf{W}$ is not guaranteed

the overlap-save method. However, a correct overlap-save algorithm is difficult to implement when computing cross-powers such as in (35) and typically the approximate expression (42) is assumed.

The problem of linear/circular convolution has been addressed by several authors [62, 149, 258, 171, 121]. Parra and Spence (2000) [149] note that the frequency domain approximation is satisfactory provided that the DFT length T is significantly larger than the length of the mixing channels. In order to reduce the errors due to the circular convolution, the filters should be at least two times the length of the mixing filters [131, 176].

To handle long impulse responses in the frequency domain, a frequency model which is equivalent to the time domain linear convolution has been proposed in [253]. When the time domain filter extends beyond the analysis window the frequency response is under-sampled [258, 22]. These errors can be mitigated by spectral smoothing or equivalently by windowing in the time domain. According to [259] the circularity problem becomes more severe when the number of sources increases.

Time domain algorithms are often derived using Toeplitz matrices. In order to decrease the complexity and improve computational speed, some calculations involving Toeplitz matrices are performed using the fast-Fourier transform. For that purpose, it is necessary to express the Toeplitz matrices in circulant Toeplitz form [23, 260, 261, 195, 121, 171]. A method that avoids the circularity effects but maintains the computational efficiency of the FFT has been presented in [262]. Further discussion on the circularity problem can be found in [189].

6.4. Subband filtering

Instead of the conventional linear Fourier domain some authors have used subband processing. In [142] a long time-domain filter is replaced by a set of short independent subband-filters, which results in faster convergence as compared to the full-band methods [214]. Different filter lengths for each subband filter have also been proposed motivated by the varying reverberation time of different frequencies (typically low-frequencies have a longer reverberation time) [263].

7. THE PERMUTATION AMBIGUITY

The majority of algorithms operate in the frequency domain due to the gains in computational efficiency, which are important in particular for acoustic mixtures that require long filters. However, in frequency domain algorithms the challenge is to solve the permutation ambiguity, i.e., to make the permutation

matrix $\boldsymbol{P}(\omega)$ independent of frequency. Especially when the number of sources and sensors is large, recovering consistent permutations is a severe problem. With N model sources there are $N!$ possible permutations in each frequency bin. Many frequency domain algorithms provide *ad hoc* solutions, which solve the permutation ambiguity only partially, thus requiring a combination of different methods. Table 4 summarizes different approaches. They can be grouped into two categories

1. Consistency of the filter coefficients
2. Consistency of the spectrum of the recovered signals

The first exploits prior knowledge about the mixing filters, and the second uses prior knowledge about the sources. Within each group the methods differ in the way consistency across frequency is established, varying sometimes in the metric they use to measure *distance* between solutions at different frequencies.

7.1. Consistency of the Filter Coefficients

Different methods have been used to establish consistency of filter coefficients across frequency, such as constraints on the length of the filters, geometric information, or consistent initialization of the filter weights.

Consistency across frequency can be achieved by requiring continuity of filter values in the frequency domain. One may do this directly by comparing the filter values of neighboring frequencies after adaptation, and pick the permutation that minimize the Euclidean distance between neighboring frequencies [269, 74]. Continuity (in a discrete frequency domain) is also expressed as smoothness, which is equivalent with a limited temporal support of the filters in the time domain. The simplest way to implement such a smoothness constraint is by zero-padding the time domain filters prior to performing the frequency transformation [264]. Equivalently, one can restrict the frequency domain updates to have a limited support in the time domain. This method is explained in Parra et al. [149] and has been used extensively [283, 161, 269, 174, 190, 188, 201, 119, 122, 192]. Ikram and Morgan [174, 176] evaluated this constraint and point out that there is a trade-off between the permutation alignment and the spectral resolution of the filters. Moreover, restricting the filter length may be problematic in reverberant environments where long separation filters are required. As a solution they have suggest to relax the constraint on filter length after the algorithm converges to satisfactory solutions [176].

Another suggestion is to assess continuity after accounting for the arbitrary scaling ambiguity. To do so, the separation matrix can be normalized as proposed in [265]:

$$\boldsymbol{W}(\omega) = \widetilde{\boldsymbol{W}}(\omega)\boldsymbol{\Lambda}(\omega), \tag{44}$$

where $\boldsymbol{\Lambda}(\omega)$ is a diagonal matrix and $\widetilde{\boldsymbol{W}}(\omega)$ is a matrix with unit diagonal. The elements of $\widetilde{\boldsymbol{W}}(\omega)$, $\widetilde{W}_{mn}(\omega)$ are the ratios between the filters and these are used to assess continuity across frequencies [48, 220].

Instead of restricting the *unmixing* filters, Pham et al. (2003) [202] have suggested to require continuity in the *mixing* filters, which is reasonable as the mixing process will typically have a shorter time constant. A specific distance measure has been proposed by Asano et al. (2003) [284, 267]. They suggest to use the cosine between the filter coefficients of different frequencies ω_1 and ω_2:

$$\cos\alpha_n = \frac{\boldsymbol{a}_n^H(\omega_1)\boldsymbol{a}_n(\omega_2)}{\|\boldsymbol{a}_n^H(\omega_1)\|\|\boldsymbol{a}_n(\omega_2)\|}, \tag{45}$$

where $\boldsymbol{a}_n(\omega)$ is the n'th column vector of $\boldsymbol{A}(\omega)$, which is estimated as the pseudo-inverse of $\boldsymbol{W}(\omega)$. Measuring distance in the space of separation filters rather than mixing filters was also suggested because these may better reflect the spacial configuration of the sources [285].

In fact, continuity across frequencies may also be assessed in terms of the estimated spatial locations of sources. Recall that the mixing filters are impulse responses between the source locations and the microphone locations. Therefore, the parameters of the separation filters should account for the position of the source in space. Hence, if information about the sensor location is available it can be used to address the permutation problem.

To understand this, consider the signal that arrives at an array of sensors. Assuming a distant

Table 4: Categorization of approaches to solve the permutation problem in the frequency domain.

Class	**Metric**	**Reference**
Consistency of the filter coefficients	Smooth spectrum	[264, 149]
	Source locations	[265]
	Directivity pattern	[266, 175, 73]
	Location vectors	[267]
	DOA	[184, 268, 72]
	Adjacent matrix distance	[269]
	Invariances	[48]
	Split spectrum	[270]
	Frequency link in update process	[127]
	Initialization	[250, 271]
	Moving sources	[167]
	Vision	[148]
Consistency of the spectrum of the recovered signals	Amplitude modulation	[159, 197, 272, 126, 203]
	Pitch	[135, 147]
	Psychoacoustics	[243, 243]
	Fundamental frequency	[244]
	Cyclo-stationarity	[273]
	Periodic signals	[221]
	Cross-correlation	[62, 274, 209]
	Cross-cumulants	[275]
	Kurtosis	[86]
	Source distribution	[276, 134]
	Multidimensional prior	[277, 278]
	Clustering	[230, 279]
Time-frequency information	FIR polynomial	[23, 254, 113, 255]
	TD cost function	[178]
	Apply ICA to whole spectrogram	[280]
Combined approaches		[106, 258, 281, 282]

source in an reverberation-free environment the signal approximates a plane-wave. If the plane-waves arrives at an angle to the microphone array it will impinge on each microphone with a certain delay (see Figure 6). This delay τ is given by the microphone distance d, the velocity of the wave c, and the direction-of-arrival (DOA) angle θ:

$$\tau = \frac{d}{c} \sin\theta, \tag{46}$$

Filters that compensate for this delay can add the microphone signals constructively (or destructively) to produce a maximum (or minimum) response in the DOA. Hence, the precise delay in filters (which in the frequency domain correspond to precise phase relationships) establishes a relationship between different frequencies that can be used to identify correct permutations. This was first considered by Soon et al. (1993) [286].

To be specific, each row in the separation matrix $\boldsymbol{W}(\omega)$ defines a *directivity pattern*, and therefore each row can be thought of as a separate beamformer. This directivity pattern is determined by the transfer function between the source and the filter output. The magnitude response of the n-th output is given by

$$\boldsymbol{r}_n(\omega, \boldsymbol{\theta}) = |\boldsymbol{w}_n^H(\omega)\boldsymbol{a}(\omega, \boldsymbol{\theta})|^2, \tag{47}$$

where $\boldsymbol{a}(\omega)$ is an $M \times 1$ vector representing the propagation of a distant source with DOA θ to the sensor array. When M sensors are available, it is possible to place $M-1$ nulls in each of the M *directivity patterns*, i.e., directions from which the arriving signal is canceled out. In an ideal, reverberation-free environment separation is achieved if these nulls point to the directions of the interfering sources. The lo-

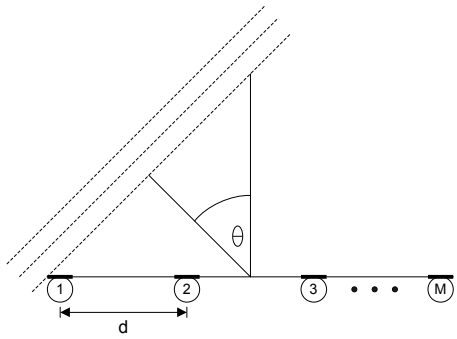

Figure 6: A sensor array consisting of M sensors linearly distributed with the distance d to the adjacent sensor. The sensors are placed in a free field. A source signal is considered coming from a point source of a distance r away from the sensor array. The source signal is placed in the far-field, i.e., $r \gg d$. Therefore the incident wave is planar and the arrival angle θ is the same for all the sensors.

cations of these nulls, as they may be identified by the separation algorithm, can be used to resolve the permutation ambiguity [266, 287, 288, 81, 77, 131, 289, 290]. These techniques draw strong parallels between source separation solutions and *beamforming*. The DOA's do not have to be known in advance and can instead be extracted from the resulting separation filters. Note, however, that the ability to identify source locations is limited by the physics of wave propagation and sampling: distant microphones will lead to grading lobes which will confuse the source locations, while small aperture limits its spatial resolution at low frequencies. Ikram and Morgan (2002) [175] extend the idea of Kurita et al. (2000) [266] to the case where the sensor space is wider than one half of the wavelength. Source locations are estimated at lower frequencies, which do not exhibit grating lobes. These estimates are then used to determine the correct nulls for the higher frequencies and hereby the correct permutations. In order to resolve permutations when sources arrive from the same direction, Mukai et al. (2004) [291] use a near-field model. Mitianoudis and Davies (2004) [268] suggested frequency alignment based on DOA estimated with the MuSIC algorithm [292]. A subspace method has been used in order to avoid constraints on the number of sensors. Knaak et al. (2003) [222] include DOA information as a part of the BSS algorithm in order to avoid the permutation. Although all these methods assume a reverberation-free environment they give reasonable results in reverberant environments as long as the source has a strong direct path to the sensors.

Two other methods also utilize geometry. In the case of moving sources, where only one source is moving, the permutation can be resolved by noting that only one of the parameters in the separation matrix changes [167]. If visual cues are available, they may also be used to solve the permutation ambiguity [148].

Instead of using geometric information as a separate step to solve the permutation problem Parra and Alvino (2002) include geometric information directly into the cost function [184, 185]. This approach has been applied to microphone arrays under reverberant conditions [187]. Baumann et al. (2003) [72] have also suggested a cost function, which includes the DOA estimation. The arrival angles of the signals are found iteratively and included in the separation criterion. Baumann et al. [73] also suggest a maximum likelihood approach to solve the permutation problem. Given the probability of a permuted or non-permuted solution as function of the estimated zero directions, the most likely permutation is found.

Gotanda et al. (2003) [270] have proposed a method to reduce the permutation problem based on the split spectral difference, and the assumption that each source is closer to one microphone. The split spectrum is obtained when each of the separated signals are filtered by the estimated mixing channels.

Finally, for iterative update algorithms a proper initialization of the separation filters can result in consistent permutations across frequencies. Smaragdis [250] proposed to estimate filter values sequentially starting with low frequencies and initializing filter values with the results of the previous lower frequency. This will tend to select solutions with filters that are smooth in the frequency domain, or equivalently, filters that are short in the time domain. Filter values may also be initialized to simple beamforming filters that point to estimated source locations. The separation algorithm will then tend

to converge to solutions with the same target source across all frequencies [184, 271].

7.2. Consistency of the Spectrum of the Recovered Signals

Some solutions to the permutation ambiguity are based on the properties of speech. Speech signals have strong correlations across frequency due to a common amplitude modulation.

At the coarsest level the power envelope of the speech signal changes depending on whether there is speech or silence, and within speech segments the power of the carrier signal induces correlations among the amplitude of different frequencies. A similar argument can be made for other natural sounds. Thus, it is fair to assumed that natural acoustic signals originating from the same source have a correlated amplitude envelope for neighboring frequencies. A method based on this co-modulation property was proposed by Murata et al. (1998) [159, 196]. The permutations are sorted to maximize the correlation between different envelopes. This is illustrated in Figure 7. This method has also been used in [293, 198, 199, 287, 263, 203]. Rahbar and Reilly (2001, 2005) [209, 152] suggest efficient methods for finding the correct permutations based on cross-frequency correlations.

Asano and Ikeda (2000) [294] report that the method sometimes fails if the envelopes of the different source signals are similar. They propose the following function to be maximized in order to estimate the permutation matrix:

$$\widehat{\boldsymbol{P}}(\omega) = \arg\max_{\boldsymbol{P}(\omega)} \sum_{t=1}^{T} \sum_{j=1}^{\omega-1} [\boldsymbol{P}(\omega)\bar{\boldsymbol{y}}(\omega,t)]^H \bar{\boldsymbol{y}}(j,t), \tag{48}$$

where $\bar{\boldsymbol{y}}$ is the power envelope of $\boldsymbol{y}$ and $\boldsymbol{P}(\omega)$ is the permutation matrix. This approach has also been adopted by Peterson and Kadambe (2003) [232]. Kamata et al. (2004) [282] report that the correlation between envelopes of different frequency channels may be small, if the frequencies are too far from each other. Anemüller and Gramms (1999) [127] avoid the permutations since the different frequencies are linked in the update process. This is done by serially switching from low to high frequency components while updating.

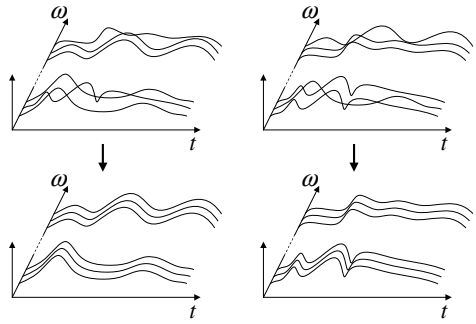

Figure 7: For speech signals, it is possible to estimate the permutation matrix by using information on the envelope of the speech signal (amplitude modulation). Each speech signal has a particular envelope. Therefore, by comparison with the envelopes of the nearby frequencies, it is possible to order the permuted signals.

Another solution based on amplitude correlation is the so-called Amplitude Modulation Decorrelation (AMDecor)–algorithm presented by Anemüller and Kollmeier (2000, 2001) [272, 126]. They propose to solve, the source separation problem and the permutation problems simultaneously. An amplitude modulation correlation is defined, where the correlation between the frequency channels ω_k and ω_l of the two spectrograms $\boldsymbol{Y}_a(\omega,t)$ and $\boldsymbol{Y}_b(\omega,t)$ is calculated as

$$\begin{aligned} &c(\boldsymbol{Y}_a(\omega,t), \boldsymbol{Y}_b(\omega,t)) = \\ &E[|\boldsymbol{Y}_a(\omega,t)||\boldsymbol{Y}_b(\omega,t)|] \\ &-E[|\boldsymbol{Y}_a(\omega,t)|]E[|\boldsymbol{Y}_b(\omega,t)|]. \end{aligned} \tag{49}$$

This correlation can be computed for all combinations of frequencies. This results in a square matrix $\boldsymbol{C}(\boldsymbol{Y}_a, \boldsymbol{Y}_b)$ with sizes equal to the number of frequencies in the spectrogram, whose k,lth element is given by (49). Since the unmixed signals $\boldsymbol{y}(t)$ have to be independent, the following decorrelation property must be fulfilled

$$C_{kl}(\boldsymbol{Y}_a, \boldsymbol{Y}_b) = 0 \quad \forall a \neq b, \forall k, l. \tag{50}$$

This principle also solves the permutation ambiguity. The source separation algorithm is then based on the

minimization of a cost function given by the Frobenius norm of the amplitude modulation correlation matrix.

A priori knowledge about the source distributions has also been used to determine the correct permutations. Based on assumptions of Laplacian distributed sources, Mitianopudis and Davies (2001, 2002) [251, 276, 134] propose a likelihood ratio test to test which permutation is most likely. A time-dependent function that imposes frequency coupling between frequency bins is also introduced. Based on the same principle, the method has been extended to more than two sources by Rahbar and Reilly (2003) [152]. A hierarchical sorting is used in order to avoid errors introduced at a single frequency. This approach has been adopted in Mertins and Russel (2003) [212].

Finally, one of the most effective convolutive BSS methods to-date (see Table 5) uses this statistical relationship of signal powers across frequencies. Rather than solving separate "instantaneous" source separation problems in each frequency band Kim et al. (2006) [295, 278, 277] propose a multi-dimensional version of the density estimation algorithms described in Section 5.1.3. The density function captures the power of the entire model source rather than the power at individual frequencies. As a result, the joint-statistics across frequencies are effectively captured and the algorithm converges to satisfactory permutations in each frequency.

Other properties of speech have also been suggested in order to solve the permutation indeterminacy. A *pitch*-based method has been suggested by Tordini and Piazza (2002) [135]. Also Sanei et al. (2004) [147] use the property of different pitch frequency for each speaker. The pitch and formants are modeled by a coupled hidden Markov model (HMM). The model is trained based on previous time frames.

Motivated by psycho-acoustics, Guddeti and Mulgrew (2005) [243] suggest to disregard frequency bands that are perceptually masked by other frequency bands. This simplifies the permutation problem as the number of frequency bins that have to be considered is reduced. In Barros et al. (2002) [244], the permutation ambiguity is avoided due to *a priori* information of the phase associated with the fundamental frequency of the desired speech signal.

Non-speech signals typically also have properties which can be exploited. Two proposals for solving the permutation in the case of cyclo-stationary signals can be found in Antoni et al. (2005) [273]. For machine acoustics, the permutations can be solved easily since machine signals are (quasi) periodic. This can be employed to find the right component in the output vector [221].

Continuity of the frequency spectra has been used by Capdevielle et al. (1995) [62] to solve the permutation ambiguity. The idea is to consider the sliding Fourier transform with a delay of one point. The cross correlation between different sources are zero due to the independence assumption. Hence, when the cross correlation is maximized, the output belongs to the same source. This method has also been used by Servière (2004) [253]. A disadvantage of this method is that it is computationally very expensive since the frequency spectrum has to be calculated with a window shift of one. A computationally less expensive method based on this principle has been suggested by Dapena and Servière (2001) [274]. The permutation is determined from the solution that maximizes the correlation between only two frequencies. If the sources have been whitened as part of separation, the approach by Capdevielle et al.(1995) [62] does not work. Instead, Kopriva et al. (2001) [86] suggest that the permutation can be solved by independence tests based on kurtosis. For the same reason, Mejuto et al. (2000) [275] consider fourth order cross-cumulants of the outputs at all frequencies. If the extracted sources belong to the same sources, the cross-cumulants will be non-zero. Otherwise, if the sources belong to different sources, the cross-cumulants will be zero.

Finally, Hoya et al. (2003) [296] use pattern recognition to identify speech pauses that are common across frequencies, and in the case of over-complete source separation, K-means clustering has been suggested. The clusters with the smallest variance are assumed to correspond to the desired sources [230]. Dubnov et al. (2004) [279] also address the case of more sources than sensors. Clustering is used at each frequency and Kalman tracking is performed in order to link the frequencies together.

7.3. Global permutations

In many applications only one of the source signals is desired and the other sources are only considered as interfering noise. Even though the local (frequency) permutations are solved, the global (external) permutation problem still exists. Only few algorithms address the problem of selecting the desired source signal from the available outputs. In some situations, it can be assumed that the desired signal arrives from a certain direction (e.g. the speaker of interest is in front of the array). Geometric information can determine which of the signals is the target [184, 171]. In other situations, the desired speaker is selected as the most dominant speaker. In Low et al. (2004) [289], the most dominant speaker is determined on a criterion based on kurtosis. The speaker with the highest kurtosis is assumed to be the dominant. In separation techniques based on clustering, the desired source is assumed to be the cluster with the smallest variance [230]. If the sources are moving it is necessary to maintain the global permutation by tracking each source. For block-based algorithm the global permutation might change at block-boundaries. This problem can often be solved by initializing the filter with the estimated filter from the previous block [186].

8. RESULTS

The overwhelming majority of convolutive source separation algorithms have been evaluated on simulated data. In the process, a variety of simulated room responses have been used. Unfortunately, it is not clear if any of these results transfer to real data. The main concerns are the sensitivity to microphone noise (often not better than -25 dB), non-linearity in the sensors, and strong reverberations with a possibly weak direct path. It is suggestive that only a small subset of research teams evaluate their algorithms on actual recordings. We have considered more than 400 references and found results on real room recordings in only 10% of the papers. Table 5 shows a complete list of those papers. The results are reported as signal-to-interference ratio (SIR), which is typically averaged over multiple output channels. The resulting SIR are not directly comparable as the results for a given algorithm are very likely to dependent on the recording equipment, the room that was used, and the SIR in the recorded mixtures. A state-of-the art algorithm can be expected to improve the SIR by 10-20 dB for two stationary sources. Typically a few seconds of data (2 s-10 s) will be sufficient to generate these results. However, from this survey nothing can be said about moving sources. Note that only 8 (of over 400) papers reported separation of more than 2 sources indicating that this remains a challenging problem.

9. CONCLUSION

We have presented a taxonomy for blind separation of convolutive mixtures with the purpose of providing a survey and discussion of existing methods. Further we hope that this might stimulate the development of new models and algorithms which more efficiently incorporate specific domain knowledge and useful prior information.

In the title of the BSS review by Torkkola (1999) [13], it was asked: *Are we there yet?* Since then numerous algorithms have been proposed for blind separation of convolutive mixtures. Many convolutive algorithms have shown good performance when the mixing process is stationary, but still only few methods work in real-world, time-varying environments. In real-time-varying environments, there are too many parameters to update in the separation filters, and too little data available in order to estimate the parameters reliably, while the less complicated methods such as null-beamformers may perform just as well. This may indicate that the long de-mixing filters are not the solution for real-world, time-varying environments such as the cocktail-party party situation.

Acknowledgments

M.S.P. was supported by the Oticon Foundation. M.S.P. and J.L. are partly also supported by the European Commission through the sixth framework IST Network of Excellence: Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL).

Table 5: An overview of algorithms applied in real rooms, where the SIR improvement has been reported.

Room size (approx.) $[m]$	T_{60} $[ms]$	N	M	SIR $[dB]$	Reference
$6 \times 3 \times 3$	300	2	2	13	[169, 170][1]
$6 \times 3 \times 3$	300	2	2	8–10	[271][1]
$6 \times 3 \times 3$	300	2	2	12	[249]
$6 \times 3 \times 3$	300	2	2	5.7	[290]
$6 \times 3 \times 3$	300	2	2	18–20	[297, 132][1]
	50	2	2	10	[207]
	250	2	2	16	[262]
$6 \times 6 \times 3$	200	2	2	< 16	[205][2]
$6 \times 6 \times 3$	150	2	2	< 15	[206]
$6 \times 6 \times 3$	150	2	2	< 20	[171]
	500	2	2	6	[262]
$4 \times 4 \times 3$	130	3	2	4–12	[298]
$4 \times 4 \times 3$	130	3	2	14.3	[227]
$4 \times 4 \times 3$	130	3	2	< 12	[47]
$4 \times 4 \times 3$	130	2	2	7–15	[130]
$4 \times 4 \times 3$	130	2	2	4–15	[22, 299][2]
$4 \times 4 \times 3$	130	2	2	12	[291]
$4 \times 4 \times 3$	130	6	8	18	[300]
$4 \times 4 \times 3$	130	4	4	12	[259]
	130	3	2	10	[140, 141]
Office		2	2	5.5–7.6	[142]
6×5	130	2	8	1.6–7.0	[269]
8×7	300	2	2	4.2–6.0	[73]
15×10	300	2	2	5–8.0	[72]
		2	2	< 10	[57, 91]
Office		2	2	6	[122]
Many rooms		2	2	3.1–27.4	[115]
Small room		2	2	4.7–9.5	[252]
$4 \times 3 \times 2$		2	2	< 10	[181]
$4 \times 3 \times 2$		2	2	14.4	[183]
$4 \times 3 \times 2$		2	2	4.6	[182]
		2	2	< 15	[245]
6×7	580	2	3	< 73	[31][3]
	810	2	2	< 10	[167][2]
Conf. room		4	4	14	[278]
	150	3	3	10	[222]
$15 \times 10 \times 4$	300	2	2	10	[77]
	360	2	2	5	[266]
5×5	200	2	2	6–21	[301]
	300	2	2–12	8–12	[302]
3×6		3	8	10	[184]
$4 \times 3 \times 2$		2	2	15	[149]
$5 \times 5 \times 3$		2	2	5	[187]
8×4	700	2	4	16	[152]
$7 \times 4 \times 3$	250	2	2	9.3	[253][1]
4×4	200	2	2	15	[303]
Office	500	3	2	4.3–10.6	[45]
	300	2	6	< 15	[213]

[1] Sources convolved with real impulse responses.
[2] Moving sources.
[3] This method is not really blind. It requires that sources are on one at a time.

10. REFERENCES

[1] A. Mansour, N. Benchekroun, and C. Gervaise, "Blind separation of underwater acoustic signals," in *ICA'06*, 2006, pp. 181–188.

[2] S. Cruces-Alvarez, A. Cichocki, and L. Castedo-Ribas, "An iterative inversion approach to blind source separation," *IEEE Trans. Neural Networks*, vol. 11, no. 6, pp. 1423–1437, Nov 2000.

[3] K. I. Diamantaras and T. Papadimitriou, "MIMO blind deconvoluition using subspace-based filter deflation," in *ICASSP'04*, vol. IV, 2004, pp. 433–436.

[4] D. Nuzillard and A. Bijaoui, "Blind source separation and analysis of multispectral astronomical images," *Astron. Astrophys. Suppl. Ser.*, vol. 147, pp. 129–138, Nov. 2000.

[5] J. Anemüller, T. J. Sejnowski, and S. Makeig, "Complex independent component analysis of frequency-domain electroencephalographic data," *Neural Networks*, vol. 16, no. 9, pp. 1311–1323, Nov 2003.

[6] M. Dyrholm, S. Makeig, and L. K. Hansen, "Model structure selection in convolutive mixtures," in *ICA'06*, 2006, pp. 74–81.

[7] C. Vayá, J. J. Rieta, C. Sánchez, and D. Moratal, "Performance study of convolutive BSS algorithms applied to the electrocardiogram of atrial fibrillation," in *ICA'06*, 2006, pp. 495–502.

[8] L. K. Hansen, "ICA of fMRI based on a convolutive mixture model," in *Ninth Annual Meeting of the Organization for Human Brain Mapping (HBM 2003)*, 2003.

[9] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, Sep 1953.

[10] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, pp. 1875–1902, Sep 2005.

[11] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acout. Speech. Sig. Proc.*, vol. 36, no. 2, pp. 145–152, Feb 1988.

[12] K. J. Pope and R. E. Bogner, "Blind Signal Separation II. Linear Convolutive Combinations," *Digital Signal Processing*, vol. 6, pp. 17–28, 1996.

[13] K. Torkkola, "Blind separation for audio signals – are we there yet?" in *ICA'99*, 1999, pp. 239–244.

[14] ——, "Blind separation of delayed and convolved sources," in *Unsupervised Adaptive Filtering, Blind Source Separation*, S. Haykin, Ed. Wiley, John and Sons, Incorporated, Jan 2000, vol. 1, ch. 8, pp. 321–375.

[15] R. Liu, Y. Inouye, and H. Luo, "A system-theoretic foundation for blind signal separation of MIMO-FIR convolutive mixtures - a review," in *ICA'00*, 2000, pp. 205–210.

[16] K. E. Hild, "Blind separation of convolutive mixtures using renyi's divergence," Ph.D. dissertation, University of Florida, 2003.

[17] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001.

[18] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. Wiley, 2002.

[19] S. C. Douglas, "Blind separation of acoustic signals," in *Microphone Arrays*, M. S. Brandstein and D. B. Ward, Eds. Springer, 2001, ch. 16, pp. 355–380.

[20] ——, "Blind signal separation and blind deconvolution," in *Handbook of neural network signal processing*, ser. Electrical engineering and applied signal processing, Y. H. Hu and J.-N. Hwang, Eds. CRC Press LLC, 2002, ch. 7.

[21] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *IJIST*, vol. 15, pp. 18–33, 2005.

[22] S. Makino, H. Sawada, R. Mukai, and S. Araki, "Blind source separation of convolutive mixtures of speech in frequency domain," *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1640–1655, Jul 2005.

[23] R. Lambert, "Multichannel blind deconvolution: FIR matrix algebra and separation of multipath mixtures," Ph.D. dissertation, University of Southern California, Department of Electrical Engineering, May 1996.

[24] S. Roberts and R. Everson, *Independent Components Analysis: Principles and Practice*. Cambridge University Press, 2001.

[25] A. Gorokhov and P. Loubaton, "Subspace based techniques for second order blind separation of convolutive mixtures with temporally correlated sources," *IEEE Trans. Circ. Syst.*, vol. 44, no. 9, pp. 813–820, Sep 1997.

[26] H.-L. N. Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Processing, Elsevier*, vol. 45, no. 2, pp. 209–229, 1995.

[27] S. Choi and A. Cichocki, "Adaptive blind separation of speech signals: Cocktail party problem," in *ICSP'97*, 1997, pp. 617–622.

[28] ——, "A hybrid learning approach to blind deconvolution of linear MIMO systems," *Electronics Letters*, vol. 35, no. 17, pp. 1429–1430, Aug 1999.

[29] E. Weinstein, M. Feder, and A. Oppenheim., "Multichannel signal separation by decorrelation." *IEEE Trans. Speech Audio Proc.*, vol. 1, no. 4, pp. 405–413, Oct 1993.

[30] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 165–169, Jul 1979.

[31] Y. A. Huang, J. Benesty, and J. Chen, "Blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. Speech Audio Proc.*, vol. 13, no. 5, pp. 882–895, Sep 2005.

[32] Y. Huang, J. Benesty, and J. Chen, "Identification of acoustic MIMO systems: Challenges and opportunities," *Signal Processing, Elsevier*, vol. 86, pp. 1278–1295, 2006.

[33] Z. Ye, C. Chang, C. Wang, J. Zhao, and F. H. Y. Chan, "Blind separation of convolutive mixtures based on second order and third order statistics," in *ICASSP'03*, vol. 5, 2003, pp. V–305–308.

[34] K. Rahbar, J. P. Reilly, and J. H. Manton, "Blind identification of MIMO FIR systems driven by quasistationary sources using second-order statistics: A frequency domain approach," *IEEE Trans. Sig. Proc.*, vol. 52, no. 2, pp. 406–417, Feb 2004.

[35] A. Mansour, C. Jutten, and P. Loubaton, "Subspace method for blind separation of sources and for a convolutive mixture model," in *Signal Processing VIII, Theories and Applications*. Elsevier, Sep 1996, pp. 2081–2084.

[36] W. Hachem, F. Desbouvries, and P. Loubaton, "On the identification of certain noisy FIR convolutive mixtures," in *ICA'99*, 1999.

[37] A. Mansour, C. Jutten, and P. Loubaton, "Adaptive subspace algorithm for blind separation of independent sources in convolutive mixture," *IEEE Trans. Sig. Proc.*, vol. 48, no. 2, pp. 583–586, Feb 2000.

[38] N. Delfosse and P. Loubaton, "Adaptive blind separation of convolutive mixtures," in *ICASSP'96*, 1996, pp. 2940–2943.

[39] ——, "Adaptive blind separation of independent sources: A second-order stable algorithm for the general case," *IEEE Trans. Circ. Syst.–I: Fundamental Theory and Applications*, vol. 47, no. 7, pp. 1056–1071, Jul 2000.

[40] L. K. Hansen and M. Dyrholm, "A prediction matrix approach to convolutive ICA," in *NNSP'03*, 2003, pp. 249–258.

[41] D. Yellin and E. Weinstein, "Multichannel signal separation: Methods and analysis," *IEEE Trans. Sig. Proc.*, vol. 44, no. 1, pp. 106–118, Jan 1996.

[42] B. Chen and A. P. Petropulu, "Frequency domain blind MIMO system identification based on second- and higher order statistics," *IEEE Trans. Sig. Proc.*, vol. 49, no. 8, pp. 1677–1688, Aug 2001.

[43] B. Chen, A. P. Petropulu, and L. D. Lathauwer, "Blind identification of complex convolutive MIMO systems with 3 sources and 2 sensors," in *ICASSP'02*, vol. II, 2002, pp. 1669–1672.

[44] Y. Hua and J. K. Tugnait, "Blind identifiability of FIR-MIMO systems with colored input using second order statistics," *IEEE Sig. Proc. Lett.*, vol. 7, no. 12, pp. 348–350, Dec 2000.

[45] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, Jul 2004.

[46] N. Roman, "Auditory-based algorithms for sound segregatoion in multisource and reverberant environtments," Ph.D. dissertation, The Ohio State University, Columbus, OH, 2005.

[47] A. Blin, S. Araki, and S. Makino, "Underdetermined blind separation og convolutive mixtures of speech using time-frequency mask and mixing matrix estimation," *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1693–1700, Jul 2005.

[48] K. I. Diamantaras, A. P. Petropulu, and B. Chen, "Blind Two-Input-TwoOutput FIR Channel Identification Based on Frequency Domain Second- Order Statistics," *IEEE Trans. Sig. Proc.*, vol. 48, no. 2, pp. 534–542, Feburary 2000.

[49] E. Moulines, J.-F. Cardoso, and E. Cassiat, "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," in *ICASSP'97*, vol. 5, 1997, pp. 3617–3620.

[50] U. A. Lindgren and H. Broman, "Source separation using a criterion based on second-order statistics," *IEEE Trans. Sig. Proc.*, vol. 46, no. 7, pp. 1837–1850, Jul 1998.

[51] H. Broman, U. Lindgren, H. Sahlin, and P. Stoica, "Source separation: A TITO system identification approach," *Signal Processing, Elsevier*, vol. 73, no. 1, pp. 169–183, 1999.

[52] H. Sahlin and H. Broman, "MIMO signal separation for FIR channels: A criterion and performance analysis," *IEEE Trans. Sig. Proc.*, vol. 48, no. 3, pp. 642–649, Mar 2000.

[53] C. Jutten, L. Nguyen Thi, E. Dijkstra, E. Vittoz, and J. Caelen, "Blind separation of sources: An algorithm for separation of convolutive mixtures," in *Higher Order Statistics. Proceedings of the International Signal Processing Workshop*, J. Lacoume, Ed. Elsevier, 1992, pp. 275–278.

[54] C. Servière, "Blind source separation of convolutive mixtures," in *SSAP'96*, 1996, pp. 316–319.

[55] A. Ypma, A. Leshem, and R. P. Duina, "Blind separation of rotating machine sources: Bilinear forms and convolutive mixtures," *Neurocomp.*, vol. 49, no. 1–4, pp. 349–368, 2002.

[56] N. Charkani, Y. Deville, and J. Herault, "Stability analysis and optimization of time-domain convolutive source separation algorithms," in *SPAWC'97*, 1997, pp. 73–76.

[57] N. Charkani and Y. Deville, "A convolutive source separation method with self-optimizing nonlinearities," in *ICASSP'99*, vol. 5, 1999, pp. 2909–2912.

[58] ——, "Self-adaptive separation of convolutively mixed signals with a recursive structure. part I: Stability analysis and optimization of asymptotic behaviour," *Signal Processing, Elsevier*, vol. 73, no. 3, pp. 225–254, 1999.

[59] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *NNSP'96*, 1996, pp. 423–432.

[60] P. Comon, "Independent Component Analysis, a new concept ?" *Signal Processing, Elsevier*, vol. 36, no. 3, pp. 287–314, Apr 1994, special issue on Higher-Order Statistics.

[61] P. Comon and L. Rota, "Blind separation of independent sources from convolutive mixtures," *IEICE Trans. on Fundamentals*, vol. E86-A, no. 3, pp. 542–549, Mar 2003.

[62] V. Capdevielle, C. Servire, and J. L. Lacoume, "Blind separation of wide-band sources in the frequency domain," in *ICASSP95*, vol. III, Detroit, MI, USA, May 9–12 1995, pp. 2080–2083.

[63] S. Icart and R. Gautier, "Blind separation of convolutive mixtures using second and fourth order moments," in *ICASSP'96*, vol. 5, 1996, pp. 3018–3021.

[64] M. Girolami and C. Fyfe, "A temporal model of linear anti-hebbian learning," *Neural Processing Letters*, vol. 4, no. 3, pp. 139–148, 1996.

[65] J. K. Tugnait, "On blind separation of convolutive mixtures of independent linear signals in unknown additive noise," *IEEE Trans. on Sig. Proc.*, vol. 46, no. 11, pp. 3117–3123, Nov 1998.

[66] C. Simon, P. Loubaton, C. Vignat, C. Jutten, and G. d'Urso, "Separation of a class of convolutive mixtures: A contrast function approach," in *ICASSP'99*, 1999, pp. 1429–1432.

[67] Y. Su, L. He, and R. Yang, "An improved

cumulant-based blind speech separation method," in *ICASSP'00*, 2000, pp. 1867–1870.

[68] P. Baxter and J. McWhirter, "Blind signal separation of convolutive mixtures," in *AsilomarSSC*, vol. 1, 2003, pp. 124–128.

[69] S. Hornillo-Mellado, C. G. Puntonet, R. Martin-Clemente, and M. Rodríguez-Álvarez, "Characterization of the sources in convolutive mixtures: A cumulant-based approach," in *ICA'04*, 2004, pp. 586–593.

[70] Y. Deville, M. Benali, and F. Abrard, "Differential source separation for underdetermined instantaneous or convolutive mixtures: Concept and algorithms," *Signal Processing*, vol. 84, no. 10, pp. 1759–1776, Oct 2004.

[71] M. Ito, M. Kawamoto, N. Ohnishi, and Y. Inouye, "Eigenvector algorithms with reference signals for frequency domain BSS," in *ICA'06*, 2006, pp. 123–131.

[72] W. Baumann, D. Kolossa, and R. Orglmeister, "Beamforming-based convolutive source separation," in *ICASSP'03*, vol. V, 2003, pp. 357–360.

[73] ——, "Maximum likelihood permutation correction for convolutive source separation," in *ICA'03*, 2003, pp. 373–378.

[74] M. S. Pedersen and C. M. Nielsen, "Gradient flow convolutive blind source separation," in *MLSP'04*, 2004, pp. 335–344.

[75] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, Dec 1993.

[76] D. Yellin and E. Weinstein, "Criteria for multichannel signal separation," *IEEE Trans. Sig. Proc.*, vol. 42, no. 8, pp. 2158–2168, Aug 1994.

[77] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation," in *ICA'04*, 2004, pp. 832–839.

[78] P. Comon, E. Moreau, and L. Rota, "Blind separation of convolutive mixtures: A contrast-based joint diagonalization approach," in *ICA'01*, 2001, pp. 686–691.

[79] E. Moreau and J. Pesquet, "Generalized contrasts for multichannel blind deconvolution of linear systems," *IEEE Sig. Proc. Lett.*, vol. 4, no. 6, pp. 182–183, Jun 1997.

[80] Y. Li, J. Wang, and A. Cichocki, "Blind source extraction from convolutive mixtures in ill-conditioned multi-input multi-output channels," *IEEE Trans. Circ. Syst. I: Regular Papers*, vol. 51, no. 9, pp. 1814–1822, Sep 2004.

[81] R. K. Prasad, H. Saruwatari, and K. Shikano, "Problems in blind separation of convolutive speech mixtures by negentropy maximization," in *IWAENC'03*, 2003, pp. 287–290.

[82] X. Sun and S. Douglas, "Adaptive paraunitary filter banks for contrast-based multichannel blind deconvolution," in *ICASSP'01*, vol. 5, 2001, pp. 2753–2756.

[83] J. Thomas, Y. Deville, and S. Hosseini, "Time-domain fast fixed-point algorithms for convolutive ICA," *IEEE Sig. Proc. Lett.*, vol. 13, no. 4, pp. 228–231, Apr 2006.

[84] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing, Elsevier*, vol. 24, no. 1, pp. 1–10, 1991.

[85] A. D. Back and A. C. Tsoi, "Blind deconvolution of signals using a complex recurrent network," in *NNSP'94*, 1994, pp. 565–574.

[86] I. Kopriva, Željko Devčić, and H. Szu, "An adaptive short-time frequency domain algorithm for blind separation of nonstationary convolved mixtures," in *IJCNN'01*, 2001, pp. 424–429.

[87] S. Cruces and L. Castedo, "A Gauss-Newton method for blind source separation of convolutive mixtures," in *ICASSP'98*, vol. IV, 1998, pp. 2093–2096.

[88] S. V. Gerven and D. V. Compernolle, "Signal separation by symmetric adaptive decorrelation: Stability, convergence, and uniqueness," *IEEE Trans. Sig. Proc.*, vol. 43, no. 7, pp. 1602–1612, Jul 1995.

[89] S. Li and T. J. Sejnowski, "Adaptive separation of mixed broadband sound sources with delays by a beamforming Herault-Jutten network," *IEEE J. Ocean. Eng.*, vol. 20, no. 1, pp. 73–79, Jan 1995.

[90] N. Charkani and Y. Deville, "Optimization of the asymptpotic preformance of time-domain convolutive source separation algorithms," in *ESANN'97*, 1997, pp. 273–278.

[91] ——, "Self-adaptive separation of convolutively mixed signals with a recursive structure. part II: Theoretical extensions and application to synthetic and real signals," *Signal Processing, Elsevier*, vol. 75, no. 2, pp. 117–140, 1999.

[92] S. Choi, H. Hong, H. Glotin, and F. Berthommier, "Multichannel signal separation for cocktail party speech recognition: A dynamic recurrent network," *Neurocomp.*, vol. 49, no. 1–4, pp. 299–314, Dec 2002.

[93] F. Berthommier and S. Choi, "Several improvements of the Hérault-Jutten model for speech segregation," in *ICA'03*, 2003, pp. 1089–1094.

[94] M. Cohen and G. Cauwenbergh, "Blind separation of linear convolutive mixtures through parallel stochastic optimization," in *ISCAS'98*, vol. 3, 1998, pp. 17–20.

[95] M. Stanacevic, M. Cohen, and G. Cauwenberghs, "Blind separation of linear convolutive mixtures using orthogonal filter banks," in *ICA'01*, 2001, pp. 260–265.

[96] G. Deco and D. Obradovic, *An Information-Theoretic Approach to Neural Computing*. New York: Springer Verlag, 1996.

[97] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[98] S. Amari, S. Douglas, A. Cichocki, and H.H.Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in *IEEE International Workshop on Wireless Communication*, 1997, pp. 101–104.

[99] B. A. Pearlmutter and L. C. Parra, "Maximum likelihood blind source separation: A context-sensitive generalization of ICA," *NIPS'97*, pp. 613–619, 1997.

[100] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 9, no. 10, pp. 2009–2025, Oct 1998.

[101] K. Kokkanikis and A. K. Nandi, "Optimal blind separation of convolutive audio mixtures without temporal constraints," in *ICASSP'04*, vol. I, 2004, pp. 217–220.

[102] K. Kokkinakis and A. K. Nandi, "Multichannel speech separation using adaptive parameterization of source pdfs," in *ICA'04*, 2004, pp. 486–493.

[103] H.-C. Wu and J. C. Principe, "Generalized anti-hebbian learning for source separation," in *ICASSP'99*, vol. 2, 1999, pp. 1073–1076.

[104] J. Xi and J. P. Reilly, "Blind separation and restoration of signals mixed in convolutive environment," in *ICASSP'97*, 1997, pp. 1327 – 1330.

[105] J. P. Reilly and L. C. Mendoza, "Blind signal separation for convolutive mixing environments using spatial-temporal processing," in *ICASSP'99*, 1999, pp. 1437–1440.

[106] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Proc.*, vol. 12, no. 5, pp. 530–538, Sep 2004.

[107] ——, "Polar coordinate based nonlinear function for frequency domain blind source separation," in *ICASSP'02*, 2002, pp. 1001–1004.

[108] J.-F. Cardoso and T. Adali, "The maximum likelihood approach to complex ICA," in *ICASSP*, vol. V, 2006, pp. 673–676.

[109] M. Novey and T. Adali, "Adaptable nonlinearity for complex maximization of nongaussianity and a fixed-point algorithm," in *MLSP*, 2006.

[110] M. Joho, H. Mathis, and G. S. Moschytz, "An FFT-based algorithm for multichannel blind deconvolution," in *ISCAS'99*, vol. 3, 1999, pp. 203–206.

[111] R. H. Lambert and A. J. Bell, "Blind separation of multiple speakers in a multipath environment," in *ICASSP'97*, vol. 1, 1997, pp. 423–426.

[112] R. H. Lambert, "A new method for source separation," in *ICASSP'95*, vol. 3, 1995, pp. 2116–2119.

[113] T.-W. Lee, A. J. Bell, and R. Orglmeister, "Blind source separation of real world signals," in *ICNN'97*, 1997, pp. 2129–2135.

[114] S. Choi, S. Amari, A. Cichocki, and R. wen Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," in *ICA'99*, 1999, pp. 371–376.

[115] S. C. Douglas and X. Sun, "Convolutive blind separation of speech mixtures using the natural gradient," *Speech Communication, Elsevier*, vol. 39, pp. 65–78, 2003.

[116] K. Matsuoka, Y. Ohba, Y. Toyota, and S. Nakashima, "Blind separation for convolutive mixture of many voices," in *IWAENC'03*, 2003, pp. 279–282.

[117] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *ICA'01*, 2001, pp. 722–727.

[118] W. Wang, M. G. Jafari, S. Sanei, and J. A. Chambers, "Blind separartion of convolutive mixtures of cyclostationary signals," *International Journal of Adaptive Control and Signal Processing*, vol. 18, pp. 279–298, Mar 2004.

[119] G.-J. Jang, C. Choi, Y. Lee, and Y.-H. Oh, "Adaptive cross-channel interference cancellation on blind signal separation outputs using source absence/presence detection and spectral subtraction," in *ICLSP'04*, vol. IV, 2004, pp. 2865–2868.

[120] S. H. Nam and S. Beack, "A frequency-domain normalized multichannel blind deconvolution algorithm for acoustical signals," in *ICA'04*, 2004, pp. 524–531.

[121] M. Joho and P. Schniter, "Frequency domain realization of a multichannel blind deconvolution algorithm based on the natural gradient," in *ICA'03*,

2003, pp. 543–548.

[122] C. Choi, G. Jang, Y. Lee, and S. R. Kim, "Adaptive cross-channel interference cancellation on blind source separation outputs," in *ICA'04*, 2004, pp. 857–864.

[123] L. Parra, C. Spence, and B. de Vries, "Convolutive source separation and signal modeling with ML," in *ISIS'97*, 1997.

[124] L. C. Parra, "Temporal models in blind source separation," *Lecture Notes in Computer Science*, vol. 1387, pp. 229–247, 1998.

[125] K. Yamamoto, F. Asano, W. van Rooijen, E. Ling, T. Yamada, and N. Kitawaki, "Estimation of the number of sound sources using support vector machines and its application to sound source separation," in *ICASSP'03*, vol. 5, 2003, pp. 485–488.

[126] J. Anemüller, "Across-frequency processing in convolutive blind source separation," Ph.D. dissertation, Oldenburg, Univ., Jul 30st 2001.

[127] J. Anemüller and T. Gramms, "On-line blind separation of moving sound sources," in *ICA'99*, 1999, pp. 331–334.

[128] S. Deligne and R. Gopinath, "An EM algorithm for convolutive independent component analysis," *Neurocomp.*, vol. 49, pp. 187–211, 2002.

[129] J. Rosca, C. Borss, and R. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *ICASSP'04*, vol. III, 2004, pp. 877–880.

[130] S. C. Douglas, H. Sawada, and S. Makino, "Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters," *IEEE Trans. Speech Audio Proc.*, vol. 13, no. 1, pp. 92–104, Jan 2005.

[131] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 2, pp. 109–116, Mar 2003.

[132] S. Ukai, T. Takatani, T. Nishikawa, and H. Saruwatari, "Blind source separation combining SIMO-model-based ICA and adaptive beamforming," in *ICASSP'05*, vol. III, 2005, pp. 85–88.

[133] R. H. Lambert and C. L. Nikias, "Polynomial matrix whitening and application to the multichannel blind deconvolution problem," in *MILCOM'95*, vol. 3, 1995, pp. 988–992.

[134] N. Mitianoudis and M. Davies, "Audio source separation of convolutive mixtures," 2002, iEEE Trans. Speech Audio Proc.

[135] F. Tordini and F. Piazza, "A semi-blind approach to the separation of real world speech mixtures," in *IJCNN'02*, vol. 2, 2002, pp. 1293–1298.

[136] H. Attias, "Source separation with a sensor array using graphical models and subband filtering," in *NIPS'02*, vol. 15, 2002, pp. 1229–1236.

[137] M. A. Gandhi and M. A. Hasegawa-Johnson, "Source separation using particle filters," in *ICLSP'04*, vol. III, 2004, pp. 2449–2452.

[138] R. K. Olsson and L. K. Hansen, "A harmonic excitation state-space approach to blind separation of speech," in *NIPS*, Dec 2004.

[139] ——, "Probabilistic blind deconvolution of nonstationary sources," in *EUSIPCO'04*, 2004, pp. 1697–1700.

[140] S. Winter, H. Sawada, S. Araki, and S. Makino, "Hierarchical clustering applied to overcomplete BSS for convolutive mixtures," in *SAPA'04*, 2004.

[141] ——, "Overcomplete BSS for convolutive mixtures based on hierarchical clustering," in *ICA'04*, 2004, pp. 652–660.

[142] H. Attias, "New EM algorithms for source separation and deconvolution," in *ICASSP'03*, vol. V, 2003, pp. 297–300.

[143] C. Andrieu and S. Godsill, "A particle filter for model based audio source separation," in *ICA'00*, 2000, pp. 381–386.

[144] J. R. Hopgood, "Bayesian blind MIMO deconvolution of nonstationary subband autoregressive sources mixed through subband all-pole channels," in *SSP'03*, 2003, pp. 422–425.

[145] S. J. Godsill and C. Andrieu, "Bayesian separation and recovery of convolutively mixed autoregressive sources," in *ICASSP'99*, vol. III, 1999, pp. 1733–1736.

[146] K. Abed-Meraim, W. Qiu, and Y. Hua, "Blind system identification," *Proc. IEEE*, vol. 85, no. 8, pp. 1310–1322, Aug 1997.

[147] S. Sanei, W. Wang, and J. A. Chambers, "A coupled HMM for solving the permutation problem in frequency domain BSS," in *ICASSP'04*, vol. V, 2004, pp. 565–568.

[148] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. Chambers, "Video assisted speech source seperation," in *ICASSP'05*, vol. V, 2005, pp. 425–428.

[149] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Proc.*, vol. 8, no. 3, pp. 320–327, May 2000.

[150] D. W. E. Schobben and P. C. W. Sommen, "On the indeterminacies of convolutive blind signal separa-

tion based on second-order statistics," in *ISSPA'99*, 1999, pp. 215–218.

[151] J. Liang and Z. Ding, "Blind MIMO system identification based on cumulant subspace decomposition," *IEEE Trans. Sig. Proc.*, vol. 51, no. 6, pp. 1457–1468, Jun 2003.

[152] K. Rahbar and J. P. Reilly, "A frequency domain method for blind source separation of convolutive audio mixtures," *IEEE Trans. Speech Audio Proc.*, vol. 13, no. 5, pp. 832–844, Sep 2005.

[153] M. Kawamoto, K. Matsuoka, and N. Ohnishi, "A method of blind separation for convolved non-stationary signals," *Neurocomp.*, vol. 22, no. 1–3, pp. 157–171, Nov 1998.

[154] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Sig. Proc.*, vol. 45, no. 2, pp. 434–444, Feb 1997.

[155] A. G. Lindgren, T. P. von Hoff, and A. N. Kaelin, "Stability and performance of adaptive algorithms for multichannel blind source separation and deconvolution," in *EUSIPCO'00*, vol. 2, 2000, pp. 861–864.

[156] H. Sahlin and H. Broman, "Separation of real-world signals," *Signal Processing, Elsevier*, vol. 64, pp. 103–113, 1998.

[157] D. C. B. Chan, P. J. W. Rayner, and S. J. Godsill, "Multi-channel blind signal separation," in *ICASSP'96*, 1996, pp. 649–652.

[158] C. Simon, C. Vignat, P. Loubaton, C. Jutten, and G. d'Urso, "On the convolutive mixture - source separation by the decorrelation approach," in *ICASSP'98*, vol. 4, 1998, pp. 2109–2112.

[159] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," RIKEN Brain Science Institute, Japan, BSIS Technical Reports 98-2, Apr 1998.

[160] B. Yin and P. Sommen, "Adaptive blind signal separation using a new simplified mixing model," in *ProRISC'99*, J. Veen, Ed., 1999, pp. 601–606.

[161] H.-C. Wu and J. C. Principe, "Simultaneous diagonalization in the frequency domain (SDIF) for source separation," in *ICA'99*, 1999, pp. 245–250.

[162] A. Souloumiac, "Blind source detection and separation using second order non-stationarity," in *ICASSP'95*, vol. III, 1995, pp. 1912–1915.

[163] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[164] M. Kawamoto, A. K. Barros, A. Mansour, K. Matsuoka, and N. Ohnishi, "Blind separation for convolutive mixtures of non-stationary signals," in *Int. Conf. Neural Inf. Proc.*, 1998, pp. 743–746.

[165] M. Kawamoto, A. K. Barros, K. Matsuoka, and N. Ohnishi, "A method of real-world blind separation implemented in frequency domain," in *ICA'00*, 2000, pp. 267–272.

[166] M. Ito, M. Maruyoshi, M. Kawamoto, T. Mukai, and N. Ohneshi, "Effectiveness of directional microphones and utilization of source arriving directions in source separation," in *ICONIP'02*, 2002, pp. 523–526.

[167] M. Ito, Y. Takeuchi, T. Matsumoto, H. Kudo, M. Kawamoto, T. Mukai, and N. Ohnishi, "Moving-source separation using directional microphones," in *ISSPIT'02*, 2002, pp. 523–526.

[168] Y. Katayama, M. Ito, Y. Takeuchi, T. Matsumoto, H. Kudo, N. Ohnishi, and T. Mukai, "Reduction of source separation time by placing microphones close together," in *ISSPIT'02*, 2002, pp. 540–544.

[169] R. Aichner, S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Time domain blind source separation of non-stationary convolved signals by utilizing geometric beamforming," in *NNSP'02*, 2002, pp. 445–454.

[170] R. Aichner, H. Buchner, S. Araki, and S. Makino, "On-line time-domain blind source separation of nonstationary convolved signals," in *ICA'03*, 2003, pp. 987–992.

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

[172] E. Visser and T.-W. Lee, "Speech enhancement using blind source separation and two-channel energy based speaker detection," in *ICASSP'03*, vol. 1, 2003, pp. 884–887.

[173] E. Visser, K. Chan, S. Kim, and T.-W. Lee, "A comparison of simultaneous 3-channel blind source separation to selective separation on channel pairs using 2-channel BSS," in *ICLSP'04*, vol. IV, 2004, pp. 2869–2872.

[174] M. Z. Ikram and D. R. Morgan, "A multiresolution approach to blind separation of speech signals in a reverberant environment," in *ICASSP'01*, vol. V, 2001.

[175] M. Ikram and D. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *ICASSP'02*, 2002, pp. 881–884.

[176] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: Investigation and solutions," *IEEE Trans. Speech Audio Proc.*, vol. 13, no. 1, pp. 1–13, Jan 2005.

[177] L. Parra and C. Spence, "On-line convolutive source separation of non-stationary signals," *IEEE J. VLSI Sig. Proc.*, vol. 26, no. 1/2, pp. 39–46, Aug 2000.

[178] M. Joho, "Blind signal separation of convolutive mixtures: A time-domain joint-diagonalization approach," in *ICA'04*, 2004, pp. 578–585.

[179] ——, "Convolutive blind signal separation in acoustics by joint approximate diagonalization of spatiotemporal correlation matrices," in *Asilomar SSC*, 2004.

[180] S. Araki, S. Makino, R. Mukai, and H. Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive beamformers," in *CRAC'01*, 2001.

[181] C. L. Fancourt and L. Parra, "The coherence function in blind source separation of convolutive mixtures of non-stationary signals," in *NNSP*, 2001, pp. 303–312.

[182] ——, "The generalized sidelobe decorrelator," in *WASPAA'01*, 2001.

[183] C. Fancourt and L. Parra, "A comparison of decorrelation criteria for the blind source separation of nonstationary signals," in *SAM'02*, 2002.

[184] L. Parra and C. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Proc.*, vol. 10, no. 6, pp. 352–362, Sep 2002.

[185] L. Parra and C. Fancourt, *An Adaptive Beamforming Perspective on Convolutive Blind Source Separation*. CRC Press LLC, 2002, book chapter in: Noise Reduction in Speech Applications, Ed. Gillian Davis.

[186] E. Visser and T.-W. Lee, "Blind source sepoaration in mobile environments usinbg a priori knowledge," in *ICASSP'04*, vol. III, 2004, pp. 893–896.

[187] M. S. Pedersen, L. K. Hansen, U. Kjems, and K. B. Rasmussen, "Semi-blind source separation using head-related transfere functions," in *ICASSP'04*, vol. V, 2004, pp. 713–716.

[188] S. Ding, T. Hikichi, T. Niitsuma, M. Hamatsu, and K. Sugai, "Recursive method for blind source separation and its applications to real-time separations of acoustic signals," in *ICA'03*, 2003, pp. 517–522.

[189] E. Robledo-Arnuncio and B. F. Juang, "Issues in frequency domain blind source separation - a critical revisit," in *ICASSP'05*, vol. V, 2005, pp. 281–284.

[190] W. Wang, J. A. Chambers, and S. Sanei, "A joint diagonalization method for convolutive blind separation of nonstationary sources in the frequency domain," in *ICA'03*, 2003, pp. 939–944.

[191] ——, "A novel hybrid approach to the permutation problem of frequency domain blind source separation," in *ICA'04*, 2004, pp. 532–539.

[192] W. Wang, S. Sanei, and J. A. Chambers, "Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Trans. Sig. Proc.*, vol. 53, no. 5, pp. 1654–1669, May 2005.

[193] L. Molgedey and H. Schuster, "Separation of independent signals using time-delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994.

[194] T. J. Ngo and N. Bhadkamkar, "Adaptive blind separation of audio sources by a physically compact device using second order statistics," in *ICA'99*, 1999, pp. 257–260.

[195] D. W. Schobben and P. C. W. Sommen, "A frequency domain blind signal separation method based on decorrelation," *IEEE Trans. Sig. Proc.*, vol. 8, no. 50, pp. 1855–1865, Aug 2002.

[196] S. Ikeda and N. Murata, "An approach to blind source separation of speech signals," in *ICANN'98*, vol. 2, 1998, pp. 761–766.

[197] ——, "A method of blind separation on temporal structre of signals," in *ICONIP'98*, vol. 2, 1998, pp. 737–742.

[198] ——, "A method of ICA in time-frequency domain," in *ICA'99*, 1999, pp. 365–371.

[199] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomp.*, vol. 41, no. 1–4, pp. 1–24, 2001.

[200] A. Ahmed, P. J. W. Rayner, and S. J. Godsill, "Considering non-stationarity for blind signal separation," in *WASPAA'99*, 1999, pp. 111–114.

[201] B. Yin and P. Sommen, "A new convolutive blind signal separation algorithm based on second order statistics using a simplified mixing model," in *EUSIPCO'00*, vol. 4, 2000, pp. 2049–2053.

[202] D.-T. Pham, C. Servière, and H. Boumaraf, "Blind separation of convolutive audio mixtures using nonstationarity," in *ICA'03*, 2003, pp. 981–986.

[203] C. Servière and D. Pham, "A novel method for permutation correction in frequency-domain in blind separation of speech mixtures," in *ICA'04*, 2004, pp. 807–815.

[204] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures exploiting nongaussianity, nonwhiteness, and nonstationarity," in *IWAENC'03*, 2003, pp. 275–278.

[205] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "Real-time convolutive blind source separation based on a broadband approach," in *ICA'04*, 2004, pp. 840–848.

[206] H. Buchner, R. Aichner, and W. Kellermann, "Trinicon: A versatile framework for multichannel blind signal processing," in *ICASSP'04*, vol. III, 2004, pp. 889–892.

[207] R. Aichner, H. Buchner, and W. Kellermann, "On the causality problem in time-domain blind source separation and deconvolution algorithms," in *ICASSP'05*, vol. V, 2005, pp. 181–184.

[208] B. S. Krongold and D. L. Jones, "Blind source separation of nonstationary convolutively mixed signals," in *SSAP'00*, 2000, pp. 53–57.

[209] K. Rahbar and J. P. Reilly, "Blind source separation algorithm for MIMO convolutive mixtures," in *ICA'01*, 2001.

[210] A. Holobar and D. Zazula, "A novel approach to convolutive blind separation of close-to-orthogonal pulse sources using second-order statistics," in *EUSIPCO'04*, 2004, pp. 381–384.

[211] K.-C. Yen and Y. Zhao, "Adaptive co-channel speech separation and recognition," *IEEE Trans Speech Audio Proc.*, vol. 7, no. 2, pp. 138–151, Mar 1999.

[212] A. Mertins and I. Russel, "An extended ACDC algorithm for the blind estimation of convolutive mixing systems," in *ISSPA'03*, vol. 2, 2003, pp. 527–530.

[213] Y. Zhao and R. Hu, "Fast convergence speech source separation in reverberant acoustic environment," in *ICASSP'04*, vol. III, 2004, pp. 897–900.

[214] I. Russell, J. Xi, A. Mertins, and J. Chicharo, "Blind source separation of nonstationary convolutively mixed signals in the subband domain," in *ICASSP'04*, vol. V, 2004, pp. 484–484.

[215] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, 2nd ed. Addison Wesley, May 1994.

[216] S. Shamsunder and G. B. Giannakis, "Multichannel blind signal separation and reconstruction," *IEEE Trans. Speech, Audio Proc.*, vol. 5, no. 6, pp. 515–528, Nov 1997.

[217] L. Deneire and D. T. Slock, "A Schur method for multiuser multichannel blind identification," in *ICASSP'99*, 1999, pp. 2905–2908.

[218] C. T. Ma, Z. Ding, and S. F. Yau, "A two-stage algorithm for MIMO blind deconvolution of nonstationary colored signals," *IEEE Trans. Sig. Proc.*, vol. 48, no. 4, pp. 1187–1192, Apr 2000.

[219] I. Bradaric, A. P. Petropulu, and K. I. Diamantaras, "On blind identifiability of FIR-MIMO systems with cyclostationary inputs using second order statistics," in *ICASSP'02*, vol. II, 2002, pp. 1745–1748.

[220] ——, "On blind identifiability of FIR-MIMO systems with cyclostationary inputs using second order statistics," *IEEE Trans. Sig. Proc.*, vol. 51, no. 2, pp. 434–441, Feburary 2003.

[221] M. Knaak, M. Kunter, and D. Filbert, "Blind source separation for acoustical machine diagnosis," in *DSP'02*, 2002.

[222] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained ICA for robust separation of sound mixtures," in *ICA'03*, 2003, pp. 951–956.

[223] T. Mei and F. Yin, "Blind separation of convolutive mixtures by decorrelation," *Signal Processing, Elsevier*, vol. 84, no. 12, pp. 2297–2213, Nov 2004.

[224] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *ICASSP'02*, vol. I, 2002, pp. 529–532.

[225] S. Rickard, T. Melia, and C. Fearon, "DESPRIT - histogram based blind source separation of more sources than sensors using subspace methods," in *WASPAA'05*, 2005, pp. 5–8.

[226] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *ICASSP'00*, vol. V, 2000, pp. 2985–2988.

[227] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *ICASSP'05*, vol. III, 2005, pp. 81–84.

[228] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Overcomplete blind source separation by combining ICA and binary time-frequency masking," in *MLSP'05*, 2005.

[229] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems, "Separating underdetermined convolutive speech mixtures," in *ICA'06*, 2006, pp. 674–681.

[230] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of a dominant source signal from mixtures of many sources," in *ICASSP'05*, vol. III, 2005, pp. 61–64.

[231] H.-C. Wu, J. C. Principe, and D. Xu, "Exploring the time-frequency microstructure of speech for blind

source separation," in *ICASSP'98*, vol. 2, 1998, pp. 1145–1148.

[232] J. M. Peterson and S. Kadambe, "A probabilistic approach for blind source separation of underdetermined convolutive mixtures," in *ICASSP'03*, vol. 6, 2003, pp. VI–581–584.

[233] D. Luengo, I. Santamaria, L. Vielva, and C. Pantaleon, "Underdetermined blind separation of sparse sources with instantaneous and convolutive mixtures," in *NNSP'03*, 2003, pp. 279–288.

[234] S. A. Abdallah and M. D. Plumbley, "Application of geometric dependency analysis to the seperation of convolved mixtures," in *ICA'04*, 2004, pp. 540–547.

[235] M. Babaie-Zadeh, A. Mansour, C. Jutten, and F. Marvasti, "A geometric approach for separating several speech signals," in *ICA'04*, 2004, pp. 798–806.

[236] Y. Li, A. Cichocki, and L. Zhang, "Blind source estimation of FIR channels for binary sources: A grouping decision approach," *Signal Processing, Elsevier*, vol. 84, no. 12, pp. 2245–2263, Nov 2004.

[237] B. A. Pearlmutter and A. M. Zador, "Monaural source separation using spectral cues," in *ICA'04*, 2004, pp. 478–485.

[238] P. Smaragdis, "Non negative matrix factor deconvolution, extraction of multiple sound sources from monophonic inputs," in *ICA'04*, 2004, pp. 494–499.

[239] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *SAPA'04*, 2004.

[240] M. S. Pedersen, T. Lehn-Schiøler, and J. Larsen, "BLUES from music: BLind Underdetermined Extraction of Sources from Music," in *ICA'06*, 2006, pp. 392–399.

[241] A. S. Bregman, *Auditory Scene Analysis*, 2nd ed. MIT Press, 1990.

[242] M. Weintraub, "The GRASP sound separation system," in *ICASSP'84*, 1984, pp. 69–72.

[243] R. R. Guddeti and B. Mulgrew, "Perceptually motivated blind source separation of convolutive mixtures," in *ICASSP'05*, vol. V, 2005, pp. 273–276.

[244] A. K. Barros, T. Rutkowski, F. Itakura, and N. Ohnishi, "Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 888–893, Jul 2002.

[245] M. Furukawa, Y. Hioka, T. Ema, and N. Hamada, "Introducing new mechanism in the learning process of FDICA-based speech separation," in *IWAENC'03*, 2003, pp. 291–294.

[246] R. D. Patterson, "The sound of a sinusoid: Spectral models," *J. Acoust. Soc. Am.*, vol. 96, pp. 1409–1418, May 1994.

[247] T. Rutkowski, A. Cichocki, and A. K. Barros, "Speech enhancement from interfering sounds using CASA techniques and blind source separation," in *ICA'01*, 2001, pp. 728–733.

[248] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236–2252, Oct 2003.

[249] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA," *IEICE Trans. Fundamentals*, vol. E86-A, no. 4, pp. 846–858, Sep 2003.

[250] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," in *WASPAA'97*, Oct 19–22 1997.

[251] M. Davies, "Audio source separation," in *Mathematics in Signal Processing V*. Oxford University Press, 2001.

[252] F. Duplessis-Beaulieu and B. Champagne, "Fast convolutive blind speech separation via subband adaptation," in *ICASSP'03*, vol. 5, 2003, pp. V–513–516.

[253] C. Servière, "Separation of speech signals under reverberant conditions," in *EUSIPCO'04*, 2004, pp. 1693–1696.

[254] T.-W. Lee, A. J. Bell, and R. H. Lambert, "Blind separation of delayed and convolved sources," in *NIPS*, vol. 9, 1997, pp. 758–764.

[255] T.-W. Lee, A. Ziehe, R. Orglmeister, and T. J. Sejnowski, "Combining time-delayed decorrelation and ICA: towards solving the cocktail party problem," in *ICASSP'98*, vol. 2, 1998, pp. 1249–1252.

[256] A. Westner and V. M. Bove Jr., "Blind separation of real world audio signals using overdetermined mixtures," in *ICA'99*, 1999.

[257] K. Kokkinakis and A. K. Nandi, "Multichannel blind deconvolution for source separation in convolutive mixtures of speech," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 14, no. 1, pp. 200–212, Jan. 2006.

[258] H. Sawada, R. Mukai, S. F. G. M. de la Kethulle de Ryhove, S. Araki, and S. Makino, "Spectral smoothing for frequency-domain blind source separation," in *IWAENC'03*, 2003, pp. 311–314.

[259] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Convolutive blind source separation for more than two sources in the frequency domain," in *ICASSP'04*, vol. III, 2004, pp. 885–888.

[260] D. W. E. Schobben and P. C. W. Sommen, "A new blind signal separation algorithm based on second-order statistics," in *IASTED SIP'06*, 1998, pp. 564–569.

[261] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," *NIPS'01*, vol. 13, 2001.

[262] R. Aichner, H. Buchner, and W. Kellermann, "A novel normalization and regularization scheme for broadband convolutive blind source separation," in *ICA'06*, 2006, pp. 527–535.

[263] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind source separation with different sensor spacing and filter length for each frequency range," in *NNSP'02*, 2002, pp. 465–474.

[264] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," in *Neurocomp.*, ser. 1–3, vol. 22, Nov 1998, pp. 21–34.

[265] V. C. Soon, L. Tong, Y. F. Huang, and R. Liu, "A wideband blind identification approach to speech acquisition using a microphone array," in *ICASSP'92*, vol. 1, 1992, pp. 293–296.

[266] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of frequency-domain blind signal separation using directivity pattern under reverberant conditions," in *ICASSP'00*, 2000, pp. 3140–3143.

[267] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 3, pp. 204–215, May 2003.

[268] N. Mitianoudis and M. E. Davies, "Permutation alignment for frequency domain ICA using subspace beamforming methods," in *ICA'04*, 2004, pp. 669–676.

[269] W. Baumann, B.-U. Köhler, D. Kolossa, and R. Orglmeister, "Real time separation of convolutive mixtures," in *ICA'01*, 2001, pp. 65–69.

[270] H. Gotanda, K. Nobu, T. Koya, K. ichi Kaneda, T. aki Ishibashi, and N. Haratani, "Permutation correction and speech extraction based on split spectrum through fastICA," in *ICA'03*, 2003, pp. 379–384.

[271] S. Araki, S. Makino, R. Aichner, T. Nishikawa, and H. Saruwatari, "Subband based blind source separation with appropriate processing for each frequency band," in *ICA'03*, 2003, pp. 499–504.

[272] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind sourc separation," in *ICA'00*, 2000, pp. 215–220.

[273] J. Antoni, F. Guillet, M. El Badaoui, and F. Bonnardot, "Blind separation of convolved cyclostationary processes," *Signal Processing, Elsevier*, vol. 85, no. 1, pp. 51–66, Jan 2005.

[274] A. Dapena and C. Serviere, "A simplified frequency-domain approach for blind separation of convolutive mixtures," in *ICA'01*, 2001, pp. 569–574.

[275] C. Mejuto, A. Dapena, and L. Castedo, "Frequency-domain infomax for blind separation of convolutive mixtures," in *ICA'00*, 2000, pp. 315–320.

[276] N. Mitianoudis and M. Davies, "New fixed-point ICA algorithms for convolved mixtures," in *ICA'01*, 2001, pp. 633–638.

[277] I. Lee, T. Kim, and T.-W. Lee, "Complex fastiva: A robust maximum likelihood approach of mica for convolutive bss," in *ICA'06*, 2006, pp. 625–632.

[278] T. Kim, H. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 15, no. 1, Jan. 2007.

[279] S. Dubnov, J. Tabrikain, and M. Arnon-Targan, "A method for directionally-disjoint source separation in convolutive environment," in *ICASSP'04*, vol. V, 2004, pp. 489–492.

[280] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *ICA'06*, 2006, pp. 601–608.

[281] D. Kolossa, B. uwe Khler, M. Conrath, and R. Orgelmeister, "Optimal permutation correlation by multiobjective genetic algorithms," in *ICA'01*, 2001, pp. 373–378.

[282] K. Kamata, X. Hu, and H. Kobataka, "A new approach to the permutation problem in frequency domain blind source separation," in *ICA'04*, 2004, pp. 849–856.

[283] H. Attias and C. E. Schreiner, "Blind source separation and deconvolution: The dynamic component analysis algorithm," *Neural Computation*, vol. 11, pp. 803–852, 1998.

[284] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Blind source separation in reflective sound fields," in *HSC'01*, Apr 9–11 2001.

[285] H. Sawada, S. Araki, R. Mukai, and S. Makino, "On calculating the inverse of separation matrix in frequency-domain blind source separation," in *ICA'06*, 2006, pp. 691–699.

[286] V. C. Soon, L. Tong, Y. F. Huang, and R. Liu, "A robust method for wideband signal separation," in *ISCAS'93*, 1993, pp. 703–706.

[287] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Re-

moval of residual cross-talk components in blind source separation using LMS filters," in *NNSP'02*, 2002, pp. 435–444.

[288] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, and K. Shikano, "Blind source separation based on subband ICA and beamforming," in *ICSLP'00*, vol. III, 2000, pp. 94–97.

[289] S. Y. Low, S. Nordholm, and R. Togneri, "Convolutive blind signal separation with post-processing," *IEEE Trans. Speech, Audio Proc.*, vol. 12, no. 5, pp. 539–548, Sep 2004.

[290] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 14, no. 2, pp. 666–678, Mar 2006.

[291] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Near-field frequency domain blind source separation for convolutive mixtures," in *ICASSP'04*, vol. IV, 2004, pp. 49–52.

[292] R. O. Schmidt and R. E. Franks, "Multiple Source DF Signal Processing: An Experimental System," *IEEE Trans. Ant. Prop.*, vol. AP-34, no. 3, pp. 281–290, Mar 1986.

[293] N. Murata and S. Ikeda, "An on-line algorightm for blind source separation on speech signals," in *International Symposium on Theory and its Applications*, vol. 3, 1998, pp. 923–926.

[294] F. Asano and S. Ikeda, "Evaluation and real-time implementation of blind source separation system using time-delayed decorrelation," in *ICA'00*, 2000, pp. 411–416.

[295] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *ICA'06*, 2006, pp. 165–172.

[296] T. Hoya, A. K. Barros, T. Rutkowski, and A. Cichocki, "Speech extraction based upon a combined subband independent component analysis and neural memory," in *ICA'03*, 2003, pp. 355–360.

[297] S. Ukai, H. Saruwatari, T. Takatani, and R. Mukai, "Multistage SIMO-model-based blind source separation combining frequenct-domain ICA and time-domain ICA," in *ICASSP'04*, vol. IV, 2004, pp. 109–112.

[298] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Blind separation of more speech than sensors with less distortion by combining sparseness and ICA," in *IWAENC'03*, 2003, pp. 271–274.

[299] R. Mukai, H. Sawada, S. Arakiand, and S. Makino, "Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction," *IEICE Trans. Fundamentals*, vol. E87-A, no. 8, pp. 1941–1948, Aug 2004.

[300] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Frequency domain blind source separation for many speech signals," in *ICA'04*, 2004, pp. 461–469.

[301] Y. Mori, H. Saruwatari, T. Takatani, K. Shikano, T. Hiekata, and T. Morita, "ICA and binary-mask-based blind source separation with small directional microphones," in *ICA'06*, 2006, pp. 649–657.

[302] T. Nishikawa, H. Saruwatari, and K. Shikano, "Stable learning algorithm for blind separation of temporally correlated signals combining multistage ICA and linear prediction," in *ICA'03*, 2003, pp. 337–342.

[303] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind separation of binaural sound mixtures using SIMO-model-based independent component analysis," in *ICASSP'04*, vol. IV, 2004, pp. 113–116.

Index

Bibliography

[1] P. Aarabi and G. Shi. Phase-based dual-microphone robust speech enhancement. *IEEE Trans. Systems, Man, and Cybernetics – Part B: Cybernetics*, 34(4):1763–1773, August 2004.

[2] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada. Blind separation of more speech than sensors with less distortion by combining sparseness and ICA. In *International Workshop on Acoustic Echo and Noise Control (IWAENC2003)*, pages 271–274, Kyoto, Japan, September 2003.

[3] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada. Underdetermined blind separation for speech in real environments with sparseness and ICA. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 881–884, Montreal, Quebec, Canada, May 17–21 2004. IEEE.

[4] S. Araki, S. Makino, H. Sawada, and R. Mukai. Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA. In C. G. Puntonet and A. Prieto, editors, *Proceedings of ICA'2004*, pages 898–905, Granada, Spain, September 22–24 2004. Springer.

[5] S. Araki, S. Makino, H. Sawada, and R. Mukai. Underdetermined blind speech separation with directivity pattern based continuous mask and ICA. In *EUSIPCO2004*, pages 1991–1995, Vienna, Austria, September 6–10 2004.

[6] S. Araki, S. Makino, H. Sawada, and R. Mukai. Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask. In *IEEE International Conference on Acoustics, Speech,*

and Signal Processing (ICASSP'05), volume III, pages 81–84, Philadelphia, PA, March 18–23 2005.

[7] L. Atlas and C. Janssen. Coherent modulation spectral filtering for single channel music source separation. In *Proc. ICASSP'05*, volume IV, pages 461–464, Philadelphia, PA, USA, March 2005.

[8] F. R. Bach and M. I. Jordan. Blind one-microphone speech separation: A spectral learning approach. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 65–72. MIT Press, Cambridge, MA, 2005.

[9] R. Balan and J. Rosca. Statistical properties of STFT ratios for two channel systems and applications to blind source separation. In *Proceedings of 2nd ICA and BSS Conference*, Helsinki, Finland, June 2000.

[10] R. Balan and J. Rosca. Convolutive demixing with sparse discrete prior models for markov sources. In *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *LNCS*, pages 544–551. Springer, 2006.

[11] D.-C. Balcan and J. Rosca. Independent component analysis for speech enhancement with missing tf content. In *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *LNCS*, pages 552–560. Springer, 2006.

[12] J. Barker, M. P. Cooke, and D. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45:5–25, 2005.

[13] J. Benesty, S. Makino, and J. Chen, editors. *Speech enhancement*. Signals and Communication Technology. Springer, 2005.

[14] J. Bitzer and K. U. Simmer. Superdirective microphone arrays. In Brandstein and Ward [20], chapter 2, pages 19–38.

[15] J. Blauert. *Spatial hearing. The Psychophysics of human sound localization*. MIT Press, Cambridge, USA, 1999.

[16] A. Blin, S. Araki, and S. Makino. Blind source separation when speech signals outnumber sensors using a sparseness-mixing matrix estimation (SMME). In *International Workshop on Acoustic Echo and Noise Control (IWAENC2003)*, pages 211–214, Kyoto, Japan, September 2003.

[17] A. Blin, S. Araki, and S. Makino. A sparseness - mixing matrix estimation (SMME) solving the underdetermined BSS for convolutive mixtures. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, volume IV, pages 85–88, Montreal, Quebec, Canada, May 17–21 2004. IEEE.

[18] A. Blin, S. Araki, and S. Makino. Underdetermined blind separation og convolutive mixtures of speech using time-frequency mask and mixing matrix estimation. *IEICE Trans. Fundamentals*, E88–A(7):1693–1700, July 2005.

[19] P. Bofill and M. Zibulevsky. Blind separation of more sources than mixtures using sparsity of their short-time fourier transform. In *Proc. ICA2000*, pages 87–92, Helsinki, June 2000.

[20] M. S. Brandstein and D. B. Ward, editors. *Microphone Arrays*. Digital Signal Processing. Springer, 2001.

[21] A. S. Bregman. *Auditory Scene Analysis*. MIT Press, 2 edition, 1990.

[22] A. W. Bronkhorst and R. Plomb. Effect on multiple speechlike maskers on binaural speech recognition and normal impaired hearing. *J. Acoust. Soc. Am*, 92(6):3132–3139, December 1992.

[23] G. J. Brown and M. P. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8:297–336, 1994.

[24] O. Cappé. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, 2(2):345–349, April 1994.

[25] J.-F. Cardoso. Blind signal separation: Statistical principles. *Proceedings of the IEEE*, 9(10):2009–2025, October 1998.

[26] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, December 1993.

[27] G. Cauwenberghs, M. Stanacevic, and G. Zweig. Blind broadband source localization and separation in miniature sensor arrays. In *IEEE Int. Symp. Circuits and Systems (ISCAS'2001)*, volume 3, pages 193–196, May 6–9 2001.

[28] A. B. Cavalcante, D. P. Mandic, T. M. Rutkowski, and A. K. Barros. Speech enhancement based on the response features of facilitated ei neurons. In *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *LNCS*, pages 585–592. Springer, 2006.

[29] E. C. Cherry. Some experiments on the recognition of speech, with one and two ears. *The Journal of the Acoustical Society of America*, 25(5):975–979, September 1953.

[30] P. Comon. Independent Component Analysis, a new concept ? *Signal Processing, Elsevier*, 36(3):287–314, April 1994. Special issue on Higher-Order Statistics.

[31] M. Cooke and D. P. W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35:141–177, 2001.

[32] B. A. Cray and A. H. Nuttall. Directivity factors for linear arrays of velocity sensors. *J. Acoust. Soc. Am.*, 110(1):324–331, July 2001.

[33] T. Dau, B. Kollmeier, and A. Kohlrausch. Modeling auditory processing of amplitude modulation. I. Modulation detection and masking with narrowband carriers. *J. Acoust Soc. Am*, 102:2892–2905, 1997.

[34] G. W. Elko. Superdirectional Microphone Arrays. In S. L. Gay and J. Benesty, editors, *Acoustic Signal Processing for Telecommunication*, The Kluwer International Series in Engineering and Computer Science, chapter 10, pages 181–237. Kluwer Academic Publishers, 2000.

[35] G. W. Elko and A.-T. N. Pong. A simple adaptive first-order differential microphone. In *Proceedings of 1995 Workshop on Applications of Single Processing to Audio and Accoustics*, pages 169–172, October 15–18 1995.

[36] D. P. W. Ellis. Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures. *Speech Communication*, 27:281–298, 1999.

[37] D. P. W. Ellis. Evaluating speech separation systems. In P. Divenyi, editor, *Speech Separation by Humans and Machines*, chapter 20, pages 295–304. Kluwer, Norwell, MA, 2004.

[38] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, 32(6):1109–1121, December 1984.

[39] S. L. Gay and J. Benesty. An introduction to acoustic echo and noise control. In S. L. Gay and J. Benesty, editors, *Acoustic Signal Processing for Telecommunication*, The Kluwer International Series in Engineering and Computer Science, chapter 1, pages 1–19. Kluwer Academic Publishers, 2000.

[40] S. Greenberg and B. E. D. Kingsbury. The modulation spectrogram: in pursuit of an invariant representation of speech. In *Proc. ICASSP'97*, pages 1647–1650, 1997.

[41] S. Gustafsson, R. Martin, P. Jax, and P. Vary. A psychoacoustic approach to combined acoustic echo cancellation and noise reduction. *IEEE Transactions on Speech and Audio Processing*, 10(5):245–256, July 2002.

[42] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi. Frequency-warped signal processing for audio applications. *J. Audio Eng. Soc*, 48(11):1011–1031, November 2000.

[43] W. M. Hartmann. *Signals, Sound, and Sensation.* Springer, 1998.

[44] S. Haykin and Z. Chen. The cocktail party problem. *Neural Computation*, 17:1875–1902, September 2005.

[45] M. J. Hewitt and R. Meddis. Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse. *Journal of the Acoustical Society of America*, 87(4):1813–1816, April 1990.

[46] G. Hu and D. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on Neural Networks*, 15(5):1135–1150, September 2004.

[47] G. Hu and D. L. Wang. Auditory segmentation based on onset and offset analysis. OSU-CISRC 04, The Ohio State University, Department of Computer Science and Engineering, Columbus, OH, 2005.

[48] G. Hu and D. L. Wang. Separation of fricatives and affricatives. In *ICASSP'05*, volume I, pages 1101–1104, Philadelphia, PA, USA, March 2005.

[49] N. Hurley, N. Harte, C. Fearon, and S. Rickard. Blind source separation of speech in hardware. In *Proceedings of SIPS*, pages 442–445, 2005.

[50] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley, 2001.

[51] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In *IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP2000)*, volume V, pages 2985–2988, Istanbul, Turkey, June 2000.

[52] D. Kolossa, A. Klimas, and R. Orglmeister. Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 82–85, New Paltz, NY, October 2005.

[53] D. Kolossa and R. Orglmeister. Nonlinear postprocessing for blind speech separation. In C. G. Puntonet and A. Prieto, editors, *Proceedings of ICA'2004*, pages 832–839, Granada, Spain, September 22–24 2004. Springer.

[54] G. Kubin and W. B. Kleijn. On speech coding in a perceptual domain. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99*, volume 1, pages 205–208. IEEE, March 15–19 1999.

[55] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 6(4):87–90, April 1999.

[56] A. Leon-Garcia. *Probability and Random Processes for Electrical Engineering*. Addison Wesley, 2nd edition, May 1994.

[57] L. Lin, E. Ambikairajah, and W. H. Holmes. Perceptual domain based speech and audio coder. In *Proc. of the third international symposium DSPCS 2002*, Sydney, January 28–31 2002.

[58] L. Lin, W. Holmes, and E. Ambikairajah. Auditory filter bank inversion. In *ISCAS 2001. The 2001 IEEE International Symposium on Circuits and Systems*, volume 2, pages 537–540, 2001.

[59] Z. Lin and R. Goubran. Musical noise reduction in speech using two-dimensional spectrogram enhancement. In *2nd IEEE Internatioal Workshop on Haptic, Audio and Visual Environments and Their Applications, HAVE'03*, pages 61–64, September 20–21 2003.

[60] R. F. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *ICASSP'82*, pages 1282–1285, 1982.

[61] R. Meddis. Simulation of mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America*, 79(3):702–711, March 1986.

[62] B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, 5 edition, 2003.

[63] Y. Mori, H. Saruwatari, T. Takatani, K. Shikano, T. Hiekata, and T. Morita. Ica and binary-mask-based blind source separation with small directional microphones. In *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *LNCS*, pages 649–657. Springer, 2006.

[64] A. H. Nuttall and B. A. Cray. Approximations to directivity for linear, planar, and volumetric apertures and arrays. *IEEE Journal of Oceanic Engineering*, 26(3):383–398, July 2001.

[65] H. G. Okuno, T. Nakatani, and T. Kawabata. A new speech enhancement: Speech stream segregation. In *Proc. ICSLP '96*, volume 4, pages 2356–2359, Philadelphia, PA, 1996.

[66] R. K. Olsson and L. K. Hansen. Blind separation of more sources than sensors in convolutive mixtures. In *ICASSP*, volume V, pages 657–660, 2006.

[67] L. Parra and C. Spence. Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, 8(3):320–327, May 2000.

[68] B. A. Pearlmutter and A. M. Zador. Monaural source separation using spectral cues. In C. G. Puntonet and A. Prieto, editors, *Proceedings of ICA'2004*, pages 478–485, Granada, Spain, September 22–24 2004. Springer.

[69] M. S. Pedersen. Matricks. Technical report, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, January 2004.

[70] M. S. Pedersen, L. K. Hansen, U. Kjems, and K. B. Rasmussen. Semi-blind source separation using head-related transfere functions. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, volume V, pages 713–716, Montreal, Quebec, Canada, May 17–21 2004. IEEE.

[71] M. S. Pedersen, T. Lehn-Schiøler, and J. Larsen. BLUES from music: BLind Underdetermined Extraction of Sources from Music. In *ICA2006*, pages 392–399, 2006.

[72] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems. Overcomplete blind source separation by combining ICA and binary time-frequency masking. In *Proceedings of the MLSP workshop*, pages 15–20, Mystic, CT, USA, September 28–30 2005.

[73] M. S. Pedersen, D. L. Wang, J. Larsen, and U. Kjems. Separating underdetermined convolutive speech mixtures. In *ICA 2006*, pages 674–681, 2006.

[74] K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2006.

[75] S. Rickard, R. Balan, and J. Rosca. Real-time time-frequency based blind source separation. In *3rd International Conference on Independent Component Analysis and Blind Source Separation*, San Diego, CA, December 9–12 2001.

[76] S. Rickard, T. Melia, and C. Fearon. DESPRIT - histogram based blind source separation of more sources than sensors using subspace methods. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'05)*, pages 5–8, New Paltz, NY, October 16–19 2005.

[77] S. Rickard and O. Yilmaz. On the approximate W-disjoint orthogonality of speech. In *Proceedings of ICASSP'02*, volume I, pages 529–532, Orlando, FL, USA, 2002.

[78] N. Roman. *Auditory-based algorithms for sound segregatoion in multisource and reverberant environtments.* PhD thesis, The Ohio State University, Columbus, OH, 2005.

[79] N. Roman, D. Wang, and G. J. Brown. Speech segregation based on sound localization. *Journal of the Acoustical Society of America*, 114(4):2236–2252, October 2003.

[80] J. Rosca, T. Gerkmann, and D.-C. Balcan. Statistical interference of missing data in the ICA domain. In *Proc. ICASSP'06*, Toulouse, May 2006.

[81] S. Roweis. One microphone source separation. In *NIPS'00*, pages 793–799, 2000.

[82] S. T. Roweis. Factorial models and refiltering for speech separation and denoising. In *Proceedings of Eurospeech03*, pages 1009–1012, Geneva, September 2003.

[83] S. Schimmel and L. Atlas. Coherent envelope detection for modulation filtering of speech. In *Proc. ICASSP'05*, volume I, pages 221–224, Philadelphia, PA, USA, March 2005.

[84] S. Shamma. Encoding sound timbre in the auditory system. *IETE Journal of Research*, 49(2):193–205, March-April 2003.

[85] M. Slaney. An efficient implementation of the patterson-holdsworth auditory filter bank. Technical Report 35, Apple Computer, Perception Group – Advanced Technology Group, 1993.

[86] M. Slaney. Auditory toolbox. Technical Report 1998-010, Interval Research Corporation, 1998. Version 2.

[87] M. Slaney, D. Naar, and R. F. Lyon. uditory model inversion for sound separation. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume ii, pages 77–80, Adelaide, Australia, April 19–22 1994. IEEE.

[88] T. Takatani, T. Nishikawa, H. Saruwata, and K. Shikano. High-fidelity blind separation for convolutive mixture of acoustic signals using simemodel-based independent component analysis. In *Seventh International Symposium on Signal Processing and Its Applications. Proceedings*, volume 2, pages 77–80. IEEE, 2003.

[89] Y. Takenouchi and N. Hamada. Time-frequency masking for BSS problem using equilateral trianguilar microphone array. In *Proceedings of 2005 International Symposium on Intelligent Signal Processing and Communication Systems*, pages 185–188, Hong Kong, December 13–16 2005.

[90] S. C. Thompson. Directional patterns obtained from two or three microphones. Technical report, Knowles Electronics, September 29 2000.

[91] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.

[92] B. D. V. Veen and K. M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, pages 4–24, April 1988.

[93] D. Wang. On ideal binary mask as the computational goal of auditory scene analysis. In P. Divenyi, editor, *Speech Separation by Humans and Machines*, pages 181–197. Kluwer, Norwell, MA, 2005.

[94] D. Wang and G. J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, 10(3):684–697, May 1999.

[95] D. Wang and G. J. Brown, editors. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Publisher: Wiley-IEEE Press, September 2006.

[96] M. Weintraub. The GRASP sound separation system. In *Proc. ICASSP'84*, pages 69–72, March 1984.

[97] M. Weintraub. A computational model for separating two simultaneous talkers. In *Proc. ICASSP'86*, pages 81–84, April 1986.

[98] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, July 2004.