PhD Dissertation

International Doctorate School in Information and Communication Technologies

DIT - University of Trento

Techniques for robust source separation and localization in adverse environments

Issues and performance of a new framework of emerging techniques for frequency-domain convolutive blind/semi-blind separation and localization of acoustic sources

Francesco Nesta

Advisor:

Prof. Maurizio Omologo

FBK-irst Fondazione Bruno Kessler

March 2010

”Doubt is the only certainty”

...someone I don’t remember

Acknowledgment

I would like to express my gratitude to all the people who walked with me during this adventure. In particular, to my advisor Maurizio Omologo for his moral support and totally confidence in my work even when I felt to be ”in a dead-end street”. To the FBK and the University of Trento which gave me an outstanding support to present my work and funny demos to a wide international audience. To all my colleagues for their helpfulness in solving my technical problems and willingness to listen my boring and strange theories. To prof. Fred Juang and all the colleagues of Georgia Tech which made me grow professionally with them. To all my room-mates for acting interest in my enthusiasm every time I thought to have discovered a new wheel. To all the new international friends, which made these years of my life an exciting time and to all my old friends for never letting me forget my origin. And last, but not for importance, I want to thank my parents and my brothers for always encouraging me when something was wrong and Geiza, for keeping me sane during the last frenetic months. Finally, a special thanks to the spirits that guided me on this challenging but always worthful experience.

Abstract

Acoustic source separation is a relatively recent topic of signal processing which aims to simultaneously separate many acoustic sources recorded through one or more microphones. Such a problem was formulated to emulate the natural capability of the human auditory system which is able to recognize and enhance the sound coming from a particular source. Addressing this problem is of high interest in the automatic speech recognition (ASR) community since it would improve the effectiveness of a natural human-machine interaction. Among numerous methods of multichannel blind source separation techniques, those based on the Independent Component Analysis (ICA) applied in the frequency-domain [81] are the most investigated, due to their straightforward physical interpretation and computational efficiency. In spite of recent developments many issues still need to be address to make such techniques robust in adverse conditions, such as high reverberation, ill-conditioning and occurrence of permutations. Furthermore, most of the proposed BSS methods are computationally expensive and not feasible for a real-time implementation.

This PhD thesis describes a research activity in the robust separation of acoustic sources in adverse environment. A new framework of blind and semi-blind techniques is proposed which allows source localization and separation even in highly reverberant environment and with real-time constraint. For each proposed technique, theoretical and practical issues are discussed and a comparison with alternative state-of-art methods is provided. Furthermore, the robustness of the proposed framework is validated implementing two real-time blind and semi-blind systems which are tested in challenging real-world scenarios.

Keywords [blind source separation, source localization, acoustic echo cancellation, independent component analysis, permutation problem]

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The Problem of Source Enhancement

Sound source enhancement is a relatively recent topic of signal processing which aims to simulate the natural capability of the human auditory system to extract a sound source of interest from a complex noisy auditory scene. Human brain is successfully able to perform this task every day and this capability allows people to hold a conversation even in a noisy environment like that of a ”Cocktail party”. The solution to the so called ”cocktail party problem” is one of the challenging goals of acoustic research community and gives rise to a lot of interest. The emulation of the human capability is an essential step towards a new generation of machines which are able to naturally interact with humans. In our society, such an interaction have been increasing due to the wide spread of information technologies. In fact, the way human beings live has been drastically changing due to the increasing importance of the information exchange in the everyday life.

Nowadays, the interaction is still oriented to a computer perspective and does not use the same methods naturally adopted by humans to interact with the environment. An important evolution to reduce the gap between machines and humans is making the interaction as natural as possible. One important step to reach this goal is to build systems, which are able to interact with people using verbal communication. Much progress has been achieved in the speech processing community, which made it possible to build real-life applications, based on human-machine speech interactions. For instance, dictating into favorite word processing application, asking for information in a museum, controlling device without usage of hands, writing notes on a mobile phone, finding paths in big cities while driving, etc. These applications are possible with the progress of speech recognition and synthesis technologies.

In spite of the progress, speech technologies are still far from the goal of allowing a robust natural speech-based communication in real life applications. The main obstacle to the performance of such systems is the inability to discriminate and isolate a source of interest from other

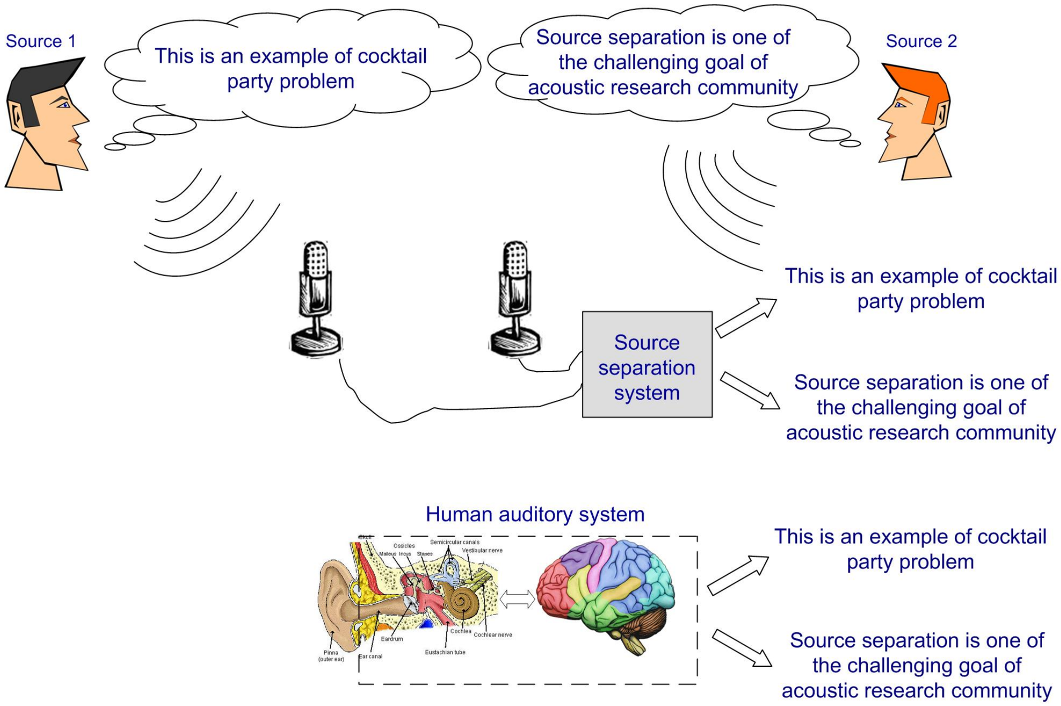

Source 1
This is an example of cocktail party problem
Source separation is one of the challenging goal of acoustic research community
Source 2
Source separation system
This is an example of cocktail party problem
Source separation is one of the challenging goal of acoustic research community
Human auditory system
Skull
Ossicles
Malleus
Incus
Stapes
Semicircular canals
Vestibular nerve
Eardrum
Cochlea
Cochlear nerve
Eustachian tube
Pinna (outer ear)
Ear canal
This is an example of cocktail party problem
Source separation is one of the challenging goal of acoustic research community

Figure 1.1: Example of cocktail party problem

interfering sources, such as music, vehicular noise or speech of other humans. It is clear that techniques of source enhancement become essential to make the overall goal achievable.

In general source enhancement can be achieved by segregating a real-world sound mixture in their basic individual components. For the human auditory system addressing this task is relatively easy in spite of the few sensors (i.e. two ears) used to sense the environment. On the other hand, the robustness of artificial systems is more sensitive to the quantity of recorded data. In the last two decades promising results have been obtained in the solution of the ”cocktail party problem” by means of statistical methods under the name of blind source separation (BSS). Such techniques aim to separate and extract sources of interest with only a small knowledge on their statistical properties [72] and differ from previous methods for their high efficiency and applicability to real-life problem. In general BSS methods can be classified in two main categories:

- single channel methods;
- multi channel methods.

The fist category methods are most challenging since they use only a single mixture of sources to retrieve that of interest. Separation is generally performed by exploiting information on the spectral redundancy and diversity and/or different source statistical properties. Independent Subspace Analysis (ISA)[15] and Non-Negative Matrix Factorization (NMF)[93] are some of

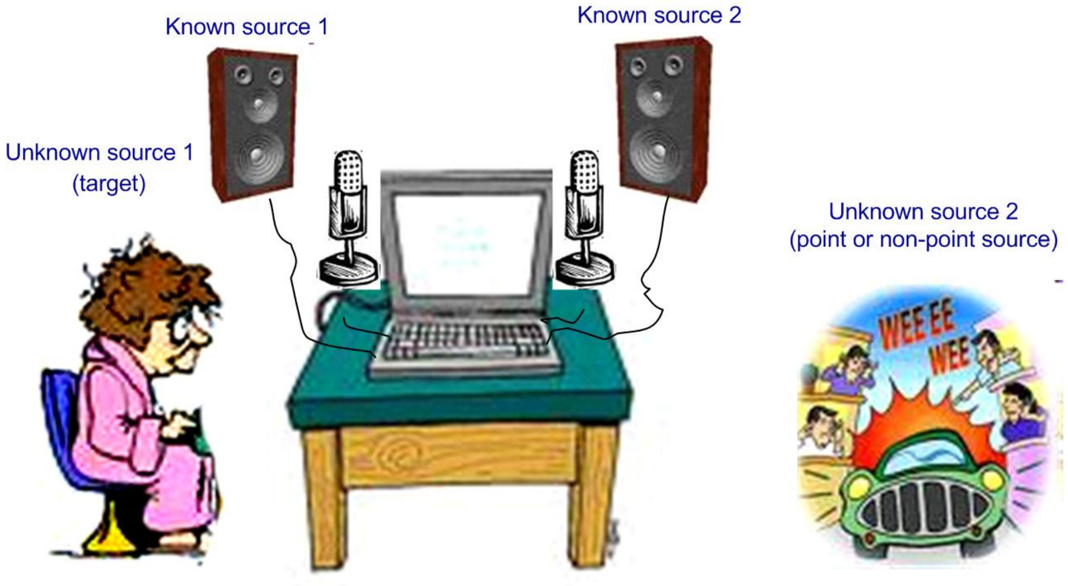


Figure 1.2: Typical scenarion for a speech controlled system in presence of background interfering noise and loudspeaker echos

the most popular techniques on which many methods have been formulated. However, their applicability to real-life problems is still limited due to the poor robustness and performance, if a good prior knowledge on the target properties is not available. Furthermore, the high computational complexity hampers the adoption of these algorithms for real-time application. On the other hand, multi-channel BSS is the most robust one since it uses geometrical knowledge on the acoustic scene. Recent progress in the field indicates that those techniques can be applied to real-world problems with real-time constraints [57][2]. For such a reason, these techniques form the basis of this thesis work which will be presented in the next sections.

1.2 Multichannel Source Enhancement

Techniques of multichannel source enhancement exploit spatial diversity in the location of sources and sensors. Their fundamentals are based on the theory of array signal processing which was mainly on the basis of telecommunication devices such as antennas and sonars and only afterwards, at beginning of 80's, was adopted in audio applications. The main goal of array processing is to exploit spatial redundancy in order to enhance the energy of the acoustic wave impinging on the array, propagating from a particular direction or spatial location. According to the geometry of the array, it is possible to opportunely combine the signals recorded by different sensors so that a *beam* is steered (*beamforming*) in the direction or location of the target source. Alternatively, the overall gain pattern of the array can be designed in order to suppress the energy of the acoustic wave coming from the interfering sources (*null beamforming*).

Beamformers generally require a high number of sensors to obtain gain pattern with an acceptable spatial rejection of the interfering sources from the target source. More important, they need to be supervised by other techniques for source localization and activity detection. On the other hand, blind source separation methods do not require additional information more than the knowledge of the statistical characterization of the sources. However, BSS methods based on sensor array exploit the same spatial redundancy to perform the source separation and can be considered a class of unsupervised beamformers. For such reasons in the next section a brief introduction on the fundamentals of beamforming techniques is given since it is essential for the BSS methods which will be presented in the next chapters.

1.2.1 Model

Let us consider a time-domain signal recorded in open field condition by the sensors of the array $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_M(t)]^T$ and for sake of simplicity let us assume to have in the auditory scene one narrow-band source which can be modeled as:

$$s(t) = ksin(2\pi ft) \tag{1.1}$$

where k is the amplitude and f is the carrier frequency. In complex-domain the source can be represented as:

$$s(t) = ke^{j2\pi ft} \tag{1.2}$$

The signals recorded by the array can be considered as delayed and scaled versions of the orginal source signals.

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ \cdots \\ x_M(t) \end{bmatrix} = \begin{bmatrix} \alpha_0 ks(t-T_0) \\ \cdots \\ \alpha_{M-1} ks(t-T_{M-1}) \end{bmatrix} = ke^{j2\pi ft} \begin{bmatrix} \alpha_0 e^{-j2\pi fT_0} \\ \cdots \\ \alpha_{M-1} e^{-j2\pi fT_{M-1}} \end{bmatrix} \tag{1.3}$$

Note that the attenuation α_m only depends on the distance between the microphone and the source if we ideally assume each microphone having unitary gain. Assuming to be in far-field condition and the array to have linear geometry with uniform microphone spacing (Uniform Linear Array) the time of arrivals T_{m-1} between the source and the *m-th* microphone can be expressed as:

$$T_{m-1} = T_0 + \Delta t_{m-1} = T_0 + m\frac{dcos\theta}{c}, \quad \forall m > 1 \tag{1.4}$$

where d is the microphone spacing, θ is the Direction Of the Arrival of the source (DOA) and c is the speed of the acoustic wave (i.e. about 340 m/s in open air). Equation (1.3) can be rewritten as:

$$\mathbf{x}(t) = x_1 \begin{bmatrix} 1 \\ \cdots \\ \hat{\alpha}_{M-1}e^{-j2\pi f\Delta T_{M-1}} \end{bmatrix} \simeq x_1 \begin{bmatrix} 1 \\ \cdots \\ e^{-j2\pi f(M-1)d\frac{cos\theta}{c}} \end{bmatrix} = x_1\mathbf{a}(f,\theta) \qquad (1.5)$$

where $\hat{\alpha}_m$ is the attenuation ratio $\overline{\alpha}_m = \frac{\alpha_m}{\alpha_0}$ (which in far-field is approximated to 1 for each m). The vector $\mathbf{a}(f,\theta)$ is defined as steering vector of the ULA array for a source located at the angle θ. The goal of a beamformer is to estimate a row vector of coefficients $\mathbf{w}(f)$ so that the output signal $\mathbf{w}(f)\mathbf{x}(t)$ is an estimate of the original source $s(t)$:

$$y(t) = x_1(t)\mathbf{w}(f)\mathbf{a}(f,\theta) = x_1(t)\mathbf{b}(f,\theta) \simeq s(t) \qquad (1.6)$$

where $\mathbf{b}(f,\theta)$ indicates the transfer function between a source at location θ and the beamformer output. The squared magnitude of $\mathbf{b}(f,\theta)$ defines the beampattern which indicates the gain of the global transfer function of the beamformer in a given direction θ.

If $s(t)$ is a broadband source the signal recorded at each microphone can be considered the convolution of the source and the impulse response between source and microphones:

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ \cdots \\ x_M(t) \end{bmatrix} = \begin{bmatrix} s(t) * h_1(t) \\ \cdots \\ s(t) * h_M(t) \end{bmatrix}. \qquad (1.7)$$

Here $*$ indicates the convolution operator. There are generally two methods of beamforming. One is to deconvolve $\mathbf{x}(t)$ by an FIR filter of length L which approximate the inverse response h_m for each sensor and to form the output by summing over each contribution. Alternatively the beamforming can be applied in frequency domain by transforming the signal $\mathbf{x}(t)$ in narrow-band signals $\mathbf{x}(f,t)$ (i.e. by means of band pass filtering or short-time Fourier analysis) and applying beamforming techniques for narrow-band sources. In this work we will focus on frequency-domain approaches since it is the same used for the BSS methods proposed in the next chapters.

1.2.2 Fixed Beamformers

Fixed beamformers optimize the filter weights $\mathbf{w}(f)$ according to the source and array geometry, which is assumed to be known in advance. For ULA the easiest way to enhance a source at given direction is to compute the weight $\mathbf{w}(f)$ so that each signal in $\mathbf{x}(t)$ is opportunely scaled and delayed before to be summed. The delays are computed in order to sum constructively the acoustic waves coming from a particular direction θ. Since θ is known the coefficients are obtained deterministically as:

$$w_m(f) = \frac{1}{M} e^{j2\pi f c^{-1}(m-1)dcos(\theta)} \tag{1.8}$$

Assuming to have an ideal knowledge of the DOA θ the performance of a fixed delay-and-sum beamformer depends on the microphone distance and on the wave length of each narrow-band source. It is interesting to plot for each frequency the *beampattern* of a beamformer steered in a given direction θ_1. It is clearly observed that the resolution of the beamformer and the gain attenuation in each direction is not constant. Therefore, applying narrow-band techniques to broadband sources would result in a non uniform attenuation of the interfering sources across the frequencies. In particular the residual of the interfering source would sound as its low-passed filtered since the spatial attenuation increases with the frequencies.

A simple approach to impose a constant directivity is to perform a least-square optimization of the coefficients in order to impose the average pattern across frequencies to have a given shape.

$$\overline{\mathbf{w}}(f) = \underset{\mathbf{w}(f)}{argmin} \int |b(\theta) - \mathbf{w}(f)\mathbf{a}(f,\theta)|^2 d\theta \tag{1.9}$$

where $b(\theta)$ is the desidered frequency-invariant response. One major drawback of delay-and-sum beamformer is the low directivity which is approximatively proportional to the number of microphones. It has been found that the directivity of linear endfire arrays theoretically approaches M^2 as the spacing approach to 0 in isotropic noise [92]. This capability is exploited by *superdirective beamformers* where the filters $\mathbf{w}(f)$ are optimized in order to maximize the factor of directivity (or array gain).

Fixed beamformers are easy to implement and do not introduce any computational load since the coefficients can be predetermined for some directions θ and switched according to the location of the sources. As major drawback their effectiveness is strictly dependent on the number of sensors and on their calibration. In fact a pre-calibration procedure is essential to compensate the variation on the gain of the microphones (theoretically assumed to be equal to 1) and to have an exact geometrical knowledge of the array. Therefore, fixed-beamformers are very sensitive to calibration errors which make the resulting *beampattern* different from the true optimal one.

1.2.3 Adaptive Beamformers

To overcome the limitations of fixed-beamformers adaptive weight estimation is generally preferred. While common fixed beamformers enhance a target sound coming from a given direction by forming directional beampatterns, Adaptive BeamFormers (ABF) estimate the weights $\mathbf{w}(f)$ according to particular optimization rule. Among many techniques the Linear Constrained Minimum Variance (LCMV)[30] is one of the most popular. It is based on an optimization which aims to reduce the energy of the acoustic wave, impinging to the array, related to the interfering sources. We briefly review this technique since it was shown that under certain conditions it is strictly connected with the frequency-domain BSS [4].

From now on we represent the signals in frequency-domain. Let us consider the vector the sources $\mathbf{s}(f) = [s_1(f), s_1(f), \cdots, s_N(f)]^T$. The signals recorded at each microphone are modeled as:

$$\mathbf{x}(f) = \mathbf{H}(f)\mathbf{s}(f) \tag{1.10}$$

where the generic element $h_{mn}(f)$ of the matrix $\mathbf{H}(f)$ is the frequency response between the *m-th* microphone and *n-th* source. An estimation of the signals $\mathbf{s}(f)$ can be obtained by means of a demixing matrix $\mathbf{W}(f)$:

$$\mathbf{y}(f) = \mathbf{W}(f)\mathbf{x}(f). \tag{1.11}$$

Assuming $s_1(f)$ to be the source of interest and the $N-1$ remaining sources to be the interferences, the enhancement can be achieved by the determining the weights which ensure the following equivalence:

$$\mathbf{w}_1(f)\mathbf{h}_n(f) = 0, \quad \forall n = 2, \cdots, N \tag{1.12}$$

where $\mathbf{w}_1(f)$ and $\mathbf{h}_n(f)$ are the sub vectors defined as:

$$\mathbf{w}_1 = [w_{11}(f), w_{12}(f), \cdots, w_{1M}(f)], \quad \mathbf{h}_n(f) = [h_{1n}(f), h_{2n}(f), \cdots, h_{Mn}(f)]^T \tag{1.13}$$

Ideally, the resulting output signals would be equivalent to:

$$\mathbf{y}_1(f) = \mathbf{w}_1(f)\mathbf{x}(f) \simeq \mathbf{w}_1(f)\mathbf{h}_1(f)s_1(f) \tag{1.14}$$

The mixing coefficient $\mathbf{h}_n(f)$ are not known in advanced and the weights $\mathbf{w}_1(f)$ need to be adaptively estimated. If the target source is not active, the weights can be adapted in order to minimize the output energy:

$$\overline{\mathbf{w}}_1(f) = \underset{\mathbf{w}_1(f)}{\operatorname{argmin}} E[|\mathbf{w}_1(f)\mathbf{x}(f)|^2] \tag{1.15}$$

where $E[\cdot]$ indicates the expectation operator. In the above notation we omitted the time dependencies for sake of simplicity. However in a practical scenario the optimization is performed by first transforming the signals in a time-frequency representation (e.g. by means of a Short-time Fourier analysis). If a Double-talk-detector (DTD) is available the expectation can be substituted with time-averaging over the instants where the target source is not active. Therefore we can rewrite the optimization rule as:

$$\overline{\mathbf{w}}_1(f) = \underset{\mathbf{w}_1(f)}{\operatorname{argmin}} \langle |\mathbf{w}_1(f)\mathbf{x}(f,t)|^2 \rangle_{t \in \mathtt{T}} \tag{1.16}$$

where $\langle \cdot \rangle_t$ is time-averaging operator and $\mathtt{T}$ is the time subset for which s_1 is inactive. The optimization in (1.16) has a degenerate solution for $\overline{\mathbf{w}}_1(f)$ which is that to set all the elements equal to 0. Therefore to prevent such degeneration we need to constrain the optimization in order to keep invariant the energy coming from the target source. A common way is to impose the following constraint:

$$\mathbf{w}_1(f)\mathbf{a}(f,\theta) = 1 \tag{1.17}$$

where $\mathbf{a}(f,\theta_1)$ is the steering vector of the array for the source $s_1(f)$ (located in θ_1). If the source is in far-field for an ULA the vector $\mathbf{a}(f,\theta_1)$ can be defined as in (1.5). Note in (1.5), in far-field the relative attenuation $\hat{\alpha}_m$ is approximated to 1 but the direction of arrival θ needs to be known or estimated in advance. Finally, the solution which optimizes (1.16) under the constraint (1.17) is obtained as:

$$\mathbf{w}_1(f) = \frac{\mathbf{a}(f,\theta_1)^H \mathbf{R}^{-1}(f)}{\mathbf{a}(f,\theta_1)^H \mathbf{R}^{-1}(f)\mathbf{a}(f,\theta_1)} \tag{1.18}$$

where $\mathbf{R}(f)$ is the covariance matrix of the recorded signals $\mathbf{x}(f,t)$:

$$\mathbf{R}(f) = \langle \mathbf{x}(f,t)\mathbf{x}(f,t)^H \rangle_{t \in \mathtt{T}} \tag{1.19}$$

An adaptive beamformer ideally defines an upper bound on the performance of a multichannel source separation system [5] since it directly attempts to minimize the output energy of an interfering source. However, its optimality is subject to the condition that the target source is not continuously active in order that the adaptation based on the optimization rule in (1.15) is meaningful. When the locations and the activity of the sources is not known, ABF are inappropriate and BSS techniques must be considered.

1.3 Scope and Thesis Outline

The scope of this thesis is to suggest a set of new techniques aimed to address the problem of the source separation in adverse conditions such as high reverberation and limited amount of observed data. The physical interpretation of the mixing and separation system is the main concept on which all the theory is derived. Therefore the separation problem is addressed together with that of the spatial localization of the sources. Many issues have been considered such as:

- high reverberation
- real-time feasibility
- multiple source localization in multiple dimensions
- separation of short mixtures
- robust solution to the permutation problem of frequency-domain BSS
- extension to semi-blind case
- applicability of BSS to distributed array networks

The thesis is organized as follows:

- chapter 2 recalls the fundamentals of the Blind Source Separation problem and of the Independent Component Analysis (ICA). Moreover, it focuses on the separation of convolutive mixtures of acoustic sources and on its state-of-art.
- chapter 3 gives a deep description of the frequency-domain BSS and of the related issues.
- chapter 4 proposes an extension of the traditional ICA based frequency-domain BSS to overcome its limitation when short signals are processed.
- chapter 5 proposes the Generalized State Coherence Transform (GSCT). The GSCT is a multidimensional spatial likelihood function which exploits the phase coherence of demixing matrices estimated by an ICA algorithm, in order to evaluate the parameters of the wave propagation of acoustic sources. The GSCT is a useful tool to allow source separation and localization of acoustic sources.
- chapter 6 discusses and proposes a new method based on a coherent source spectral estimation for solving the permutation problem of frequency-domain blind source separation (BSS).
- chapter 7 discusses the problem of Semi-blind source separation (SBSS) which is a special case of the blind source separation when some partial knowledge of the source signals is available to the system. Theoretical and practical issues are deeply analyzed and an SBSS algorithm is proposed and evaluated.
- chapter 8 discusses and evaluates in a preliminary analysis a new scheme, under the name of Collaborative Wiener ICA, to perform the BSS using distributed arrays.
- chapter 9 describes and evaluate two real-time demo prototypes for BSS and SBSS, which combine together the proposed techniques.

1.4 Publications

1.4.1 International Journal Publications

- Francesco Nesta, Piergiorgio Svaizer, Maurizio Omologo, "**Convolutive BSS by recursively regularized ICA across frequencies**" (accepted for publication in the IEEE Transactions in Audio Speech and Language processing)
- Francesco Nesta, Ted Wada, Biing-Hwang (Fred) Juang, "**Batch-Online Semi-Blind Source Separation Applied to Multi-Channel Acoustic Echo Cancellation**" (accepted for publication in the IEEE Transactions in Audio Speech and Language processing)
- Francesco Nesta, Maurizio Omologo, "**Estimation of the multidimensional acoustic propagation parameters of multiple sources with the Generalized State Coherence Transform**" (in preparation for submission to IEEE Transactions in Audio Speech and Language processing).

1.4.2 International Conference Publications

- Francesco Nesta, Maurizio Omologo, "**Cooperative Wiener-ICA for source localization and separation by distributed microphone arrays**", (in proceedings) ICASSP2010
- Benedikt Loesch, Francesco Nesta, and Bin Yang, "**On the robustness of the multidimensional State Coherence Transform for solving the permutation problem of frequency-domain ICA**, (in proceedings) ICASSP2010
- Francesco Nesta, Ted Wada, Shigeki Miyabe, Biing-Hwang (Fred) Juang, "**On the non-uniqueness problem and the semi-blind source separation**", (in proceedings), WASPAA 2009
- Francesco Nesta, Maurizio Omologo "**Generalized State Coherence Transform for multidimensional localization of multiple sources**", (in proceedings), WASPAA 2009
- Francesco Nesta, Ted Wada, Biing-Hwang (Fred) Juang, "**Coherent spectra estimation for a robust solution to the permutation problem**", (in proceedings) to WASPAA 2009
- Francesco Nesta, Piergiorgio Svaizer, and Maurizio Omologo, "**Cumulative State Coherence Transform for a Robust Two-Channel Multiple Source Localization**", (in proceedings), ICA 2009
- Francesco Nesta, Piergiorgio Svaizer, and Maurizio Omologo, "**Robust two-channel TDOA estimation for multiple speaker localization by using recursive ICA and a state coherence transform**", (in proceedings), ICASSP 2009

- Francesco Nesta, Maurizio Omologo, Piergiorgio Svaizer, ”**Multiple TDOA estimation by using a state coherence transform for solving the permutation problem in frequency-domain BSS**”, (in proceedings), MLSP 2008
- Francesco Nesta, Maurizio Omologo, Piergiorgio Svaizer, ”**A novel robust solution to the permutation problem based on a joint multiple TDOA estimation**”, (in proceedings), IWAENC 2008
- Francesco Nesta, Piergiorgio Svaizer, Maurizio Omologo, ”**Separating short signals in highly reverberant envirnonment by a recursive frequency-domain BSS**”, (in proceedings), HSCMA 2008
- Francesco Nesta, Maurizio Omologo, ”**A BSS method for short utterances by a recursive solution to the permutation problem**”, (in proceedings), SAM2008

Chapter 2

An Introduction to Blind Source Separation

Blind Source Separation (BSS) is a relatively recent digital signal processing technique that has given rise to a lot of interest in the last two decades [72]. Its main objective is to separate multiple sources mixed through unknown channels using only the observations of their mixtures. Such techniques have been applied in many scientific fields such as biology, biomedical signal processing, digital communication and speech processing. The term “blind” indicates that separation is performed without using any information about the mixing channels or the sources.

Let us consider a set of N sources $\mathbf{s}(t) = [s_1(t), s_2(t), \cdots, s_N(t)]^T$ and a set of M mixture observations $\mathbf{x}(t) = [x_1(t), x_2(t), \cdots, x_M(t)]^T$ obtained as:

$$\mathbf{x}(t) = \boldsymbol{H}[\mathbf{s}(t)] + \mathbf{e}(t) \tag{2.1}$$

where $\boldsymbol{H}[\cdot]$ is a generic transfer function and $\mathbf{e}(t)$ is a column vector of additive noise terms. Assuming $\boldsymbol{H}[\cdot]$ invertible the goal of a Source Separation system is to estimate a function $\boldsymbol{W}[\cdot]$ in order to reconstruct the original source signals $\mathbf{s}(t)$:

$$\mathbf{y}(t) = \boldsymbol{W}[\mathbf{x}(t) + \mathbf{e}(t)] \tag{2.2}$$

It is common to refer to many source separation methods as Blind Source Separation. This term is used to refer to the fact that the function $\boldsymbol{W}[\cdot]$ is estimated without any further knowledge other than the observed signals $\mathbf{x}(t)$. In practice the separation can never be completely blind since at least knowledge on the statistical properties of the sources is essential. However, BSS techniques differ from other supervised methods of speech enhancement since the separation process is more general and less sensitive to modeling errors of mixing/demixing systems as

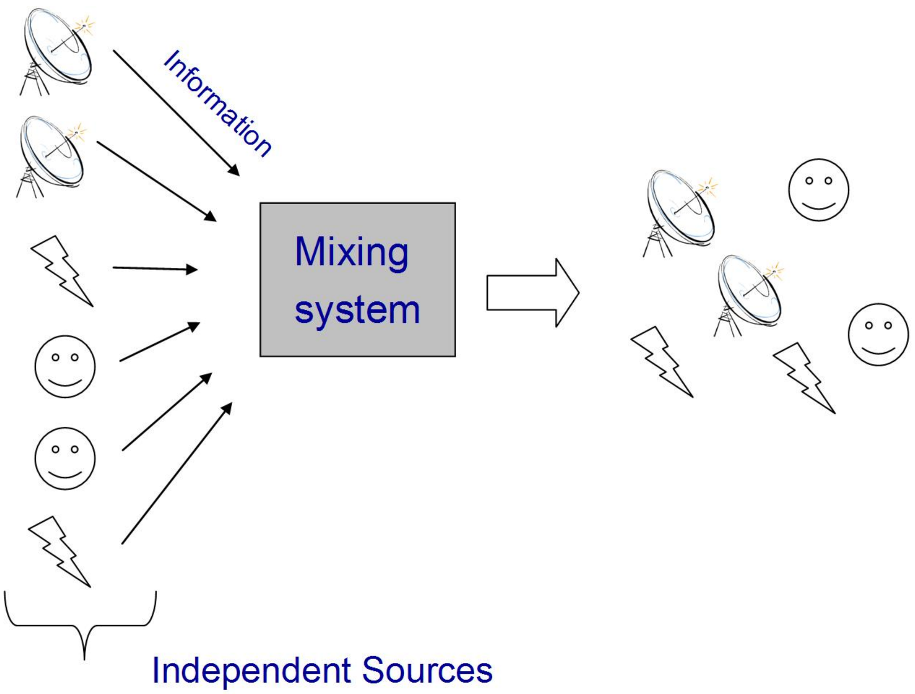


Figure 2.1: Generic mixing model for a source separation system

well as to the characteristic of the sources. The complexity of a BSS system strictly depends on the model used to represent the generic transfer function. For the case of acoustic sources two main models give rise of interest:

- instantaneous mixing model
- anechoic mixing model
- convolutive mixing model

The simplest case of BSS regards the separation of instantaneous mixtures when the observed signals are linear combinations of the original source signals. To a large extent, this case is solved by means of the Independent Component Analysis (ICA)[38]. The goal of the ICA methods is to find a representation of nongaussian data so that the estimated components are as statistically independent as possible.

As a second case, BSS can be applied to sources mixed in anechoic conditions where observed mixtures are a linear combination of scaled and delayed versions of the original source signals. Most of the ICA algorithms can be adapted to this situation, where components are mixed using a complex-valued basis. Natural gradient [3] or Infomax [8] and complex Fast-ICA[40] are some of the most popular ICA algorithms for the complex-domain but many other methods are described in [18].

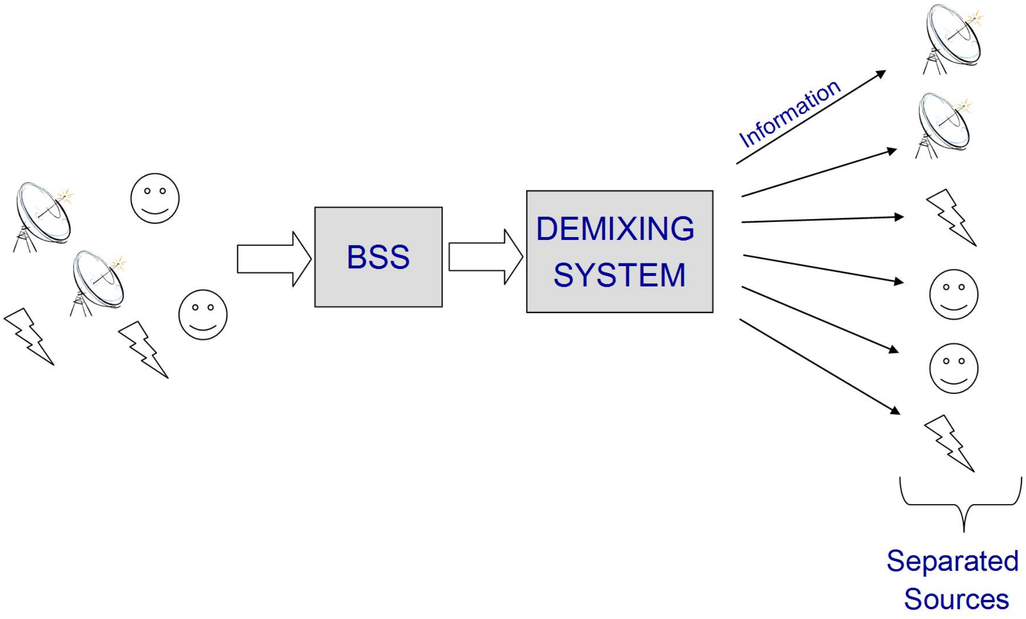


Figure 2.2: Generic demixing model for a source separation system

A more challenging case regards sources which are mixed through convolutive channels. This situation is quite common for audio applications where signals are generally recorded in reverberant environments.

In the next sections we discuss the instantaneous model and the basis of the ICA theory. After that we consider the convolutive model, treating the anechoic one as a special case.

2.1 Instantaneous Linear Model

The simplest way to model the observed signals is to assume $\mathbf{x}(t)$ to be a vector of instantaneous linear mixtures of the source signals:

$$\mathbf{x}(t) = \mathbf{H}\mathbf{s}(t) \tag{2.3}$$

where $\mathbf{H}$ is the mixing matrix which is assumed for simplicity to be time-invariant. Assuming the number of the mixture observations to be greater than the number of the sources ($M \geq N$) the problem is defined as *overdeterminated* (or *overcomplete*) and the original sources can be retrieved by estimating the pseudo inverse of $\mathbf{H}$:

$$\mathbf{W} = \mathbf{H}^{+} = (\mathbf{H}^{T}\mathbf{H})^{-1}\mathbf{H}^{T} \tag{2.4}$$

which for the case of $N = M$ is equivalent to the inverse of $\mathbf{H}$. Therefore the original signals are demixed as:

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) \tag{2.5}$$

If $N > M$, known as *underdeterminated* case, the matrix $\mathbf{H}^T\mathbf{H}$ is not fully ranked and thus a pseudo inverse of $\mathbf{W}$ does not exit. Therefore, the estimation of the output signals cannot be addressed by linear demixing of the observations vector even though theoretically it is still possible to estimate a mixing matrix $\mathbf{H}$.

2.1.1 Principal Component Analysis

The Principal Component Analysis (PCA), often named Karhunen-Loéve transform, is a well-known statistical tool used to reduce the representation of a signal by means of a set of variables with less redundancy [37]. The method is based on Second-Order Statistic (SOS) and has the main goal to represent the observed data series with uncorrelated series. Assuming $\mathbf{x}(t)$ to be a set of zero-mean random variables we can define the autocorrelation matrix as:

$$\mathbf{R}_{\mathbf{xx}} = E[\mathbf{x}\mathbf{x}^H] \tag{2.6}$$

where we removed the time dependency to simplify the notation. The PCA decomposition associated with $\mathbf{x}$ is defined as:

$$\mathbf{R}_{\mathbf{xx}} = \mathbf{Q}\mathbf{D}\mathbf{Q}^H \tag{2.7}$$

where $\mathbf{D}$ is a diagonal matrix whose elements d_{ii} are the principal eigenvalues of $\mathbf{R}_{\mathbf{xx}}$ corresponding to the eigenvectors $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ...]$. Thus, the principal components of $\mathbf{x}(t)$ can be obtained by the whitening transformation:

$$\mathbf{V} = \sqrt{\mathbf{D}}\mathbf{Q}^H \tag{2.8}$$

$$\mathbf{z}(t) = \mathbf{V}\mathbf{x}(t) \tag{2.9}$$

Multiplying the elements of $\mathbf{x}(t)$ by $\mathbf{Q}^H$ the outputs would represent an orthogonal base, while the term $\sqrt{\mathbf{D}}$ normalizes the variance of the signals which become also orthonormal.

Note that if the signals $\mathbf{x}(t)$ are obtained as in (2.3), $\mathbf{z}(t)$ do not necessarily correspond to the original components $\mathbf{s}(t)$ since uncorrelatedness corresponds to independence only if the sources $\mathbf{s}(t)$ are normally distributed. Furthermore, it is possible to demonstrate that only by means of SOS correlation the identification of $\mathbf{W}$ is not possible. For example, if the sources $\mathbf{s}(t)$ are mutually independent their covariance matrix is diagonal:

$$\mathbf{R}_{\mathbf{ss}} = E[\mathbf{s}\mathbf{s}^H] = \mathbf{D}_{\mathbf{s}} \tag{2.10}$$

where $\mathbf{D}_s$ is a generic diagonal matrix. For the goal of the source separation we want to estimate a demixing matrix $\mathbf{W}$ in order to recover the original components $\mathbf{s}$. According to SOS the matrix $\mathbf{W}$ must decorrelate the estimated output signals $\mathbf{y}(t)$:

$$\underset{\mathbf{W}}{arg}(E[\mathbf{W}\mathbf{x}\mathbf{x}^H\mathbf{W}^H] = \mathbf{D}_\mathbf{s}) \tag{2.11}$$

which can be rewritten as:

$$\underset{\mathbf{W}}{arg}(E[\mathbf{W}\mathbf{H}\mathbf{s}\mathbf{s}^H\mathbf{H}^H\mathbf{W}^H] = \mathbf{D}_\mathbf{s}) \tag{2.12}$$

$$\underset{\mathbf{W}}{arg}(\mathbf{W}\mathbf{H}\mathbf{D}_\mathbf{s}(\mathbf{W}\mathbf{H})^H = \mathbf{D}_\mathbf{s}) \tag{2.13}$$

Assuming for simplicity that the sources have unitary autocorrelation, $\mathbf{D}_\mathbf{s}$ is equivalent to the identity matrix $\mathbf{I}$ and the matrix $\mathbf{U} = \mathbf{W}\mathbf{H}$ is unitary:

$$\underset{\mathbf{W}}{arg}(\mathbf{U}\mathbf{U}^H = \mathbf{I}) \tag{2.14}$$

If $\mathbf{W}$ and $\mathbf{H}$ do not have a particular structure the matrix $\mathbf{W}$ can only be determined up to a unitary transformation $\mathbf{U}$:

$$\mathbf{W} = \mathbf{U}\mathbf{H}^+ \tag{2.15}$$

where $+$ is the Moonre-Penrose pseudo inverse. Therefore the uncorrelatedness is only a necessarily but not sufficient condition for independence and therefore cannot be used to separate stationary sources mixed through a linear mixing system. Nevertheless, the whitening procedure is widely used in many High-Order-Statistics (HOS) methods as preprocessing tool to limit the search of the solution in a bounded space, consequently increasing the convergence rate and the adaptation stability.

2.1.2 Independent Component Analysis

Independent Component Analysis (ICA) is a recently developed method, whose goal is to find a linear representation of nongaussian data so that the components are statistically independent, or as independent as possible [38]. ICA can be viewed as an extension of PCA, which is based on SOS, to the case of HOS. As for the PCA the goal of ICA is to decompose a set of data series in a reduced number of components which in this case are not only required to be uncorrelated but also mutual independent. As opposed to the PCA a measure of HOS independence is not unequivocally defined and for this reason a huge number ICA algorithms have been proposed during the last two decades. Besides the cocktail-party problem, ICA theory has been applied

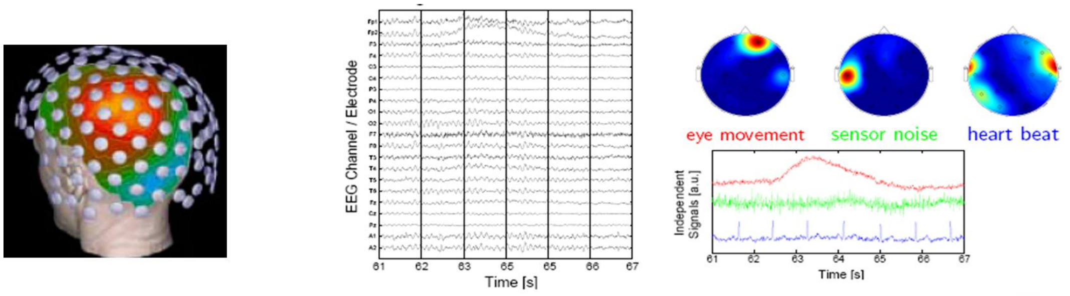


Figure 2.3: Example of ICA applied to EEG

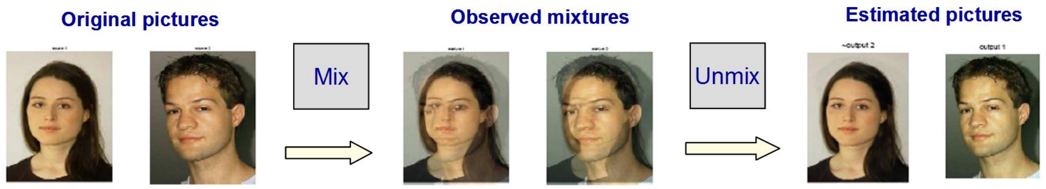


Figure 2.4: Example of ICA applied to image separation

to a number of scientific fields such as astronomic image analysis [31][51], biomedical signal separation [56][79] and video coding [77],[69]. The scope of this thesis is to provide a set of robust methods for acoustic source separation in convolutive adverse environments based on ICA. The proposed methods do not require a deep knowledge of the huge mathematical formulations of ICA theory which goes behind the scope of this thesis. Therefore, we limit the discussion only to the fundamentals of the ICA and we give a description of the Natural Gradient algorithm which is on the bases of the methods proposes in the next chapters.

Two main hypotheses need to be verified to successfully apply the ICA:

- The source signals need to be mutual independent which means that, for continuous variable, the joint probability density function of the sources $\mathbf{s}(t)$ can be marginalized as:

$$p(s_1, s_2, \cdots, s_N) = p(s_1)p(s_2)\cdots p(s_N) \tag{2.16}$$

  An equivalent way to express such an independence is by means of high order cross moments. Considering a set of time-domain signals $\mathbf{s}(t) = [s_1(t)...s_N(t)]$ we can define with $S_i(t)$ the random variable associated to the source signal observed at time instant t. Sources are assumed independent if the following relationship holds:

$$E[S_{n_1}(t)^\alpha S_{n_2}(t-\tau)^\beta] = 0 \quad \forall n_1, n_2, \alpha, \beta, \tau \tag{2.17}$$

  Note, considering explicitly the random variables associated to signals observed at particular time instants we also assume that sources may be non-stationary. This is a typical situation for acoustic sources such as speech and music.

- According to the PCA analysis, stationary sources cannot be separated only by means of second-order-statistics. Therefore, if the sources have a Gaussian distribution all the high-order moments are null and thus the separation is neither possible by means of ICA. In general it is required that at most one source can have Gaussian statistics if we want to uniquely identify an inverse of the mixing matrix $\mathbf{H}$ [28].

The first assumption is probably the biggest limitation of ICA. In fact, though independent sources are expected to generate uncorrelated signals, the HOS independence has sense only if evaluated over a sufficient data observations (theoretically infinite). However, in practical situation the estimation of the mixing system is performed with a limited amount of data which may correspond to highly correlated signal sources. Therefore, the high statistical bias in the evaluation of the HOS could compromise the accuracy of the solution if the observed data is not sufficiently large.

As for the PCA also the ICA solution has some ambiguities:

- scaling ambiguity (or variance indeterminacy). From equation (2.3) we note that since both $\mathbf{s}(t)$ and $\mathbf{H}$ are unknown, any scalar multiplier in one of the sources $s_n(t)$ could be always cancelled by dividing the corresponding *n-th* column of $\mathbf{H}$ by the same scalar. Therefore, if one of the variables, $\mathbf{H}$ or $\mathbf{s}(t)$ is unknown, it is not possible to retrieve the correct variance of the original sources. Note that the scalar can also be negative and therefore there is also an ambiguity of sign. One of the methods used to solve this problem is to impose the variance of sources to be unitary. However, we will see later that for the frequency-domain convolutive BSS a better method is used to reduce the spectral distortion in the output signals.

- order ambiguity (or permutation indeterminacy). Still, since the structure of the mixing matrix $\mathbf{H}$ is unknown one could freely invert the order of the rows of $\mathbf{W}$ (which would correspond to change the order of the column of $\mathbf{H}$) still obtaining at output independent signals. This ambiguity is well-known as permutation indeterminacy. In other terms, given a demixing matrix $\mathbf{W}$ for any permutation matrix $\mathbf{\Pi}$, the resulting permuted demixing matrix $\mathbf{\Pi W}$ is still a solution.

To summarize the above indeterminacies, we can say that the separation matrix $\mathbf{W}$ estimated by the ICA is an estimate of the inverse $\mathbf{H}$ up to a scaling and permutation ambiguity:

$$\mathbf{W} = \mathbf{\Lambda \Pi} \overline{\mathbf{H}}^{-1} \tag{2.18}$$

where $\mathbf{\Lambda}$ is an arbitrary diagonal matrix of scalar multiplier and $\mathbf{\Pi}$ is a permutation matrix and $\overline{\mathbf{H}}^{-1}$ is the inverse estimate of $\mathbf{H}$.

2.1.3 ICA Based on Nongaussianity

As already mentioned one of the cues used in ICA is that signal have nongaussian statistics. According to the Central Limit Theorem (CLT) the distribution of a sum of independent random variables tends toward a normal distribution. If an observed signal is a mixtures of two or more independent signal sources its distribution is expected to be more Gaussian than that of each single source. Thus, one criteria to estimate $\mathbf{W}$ is to maximize the non-gaussianity of the output signals distributions.

$$\mathbf{W} = \underset{\tilde{\mathbf{W}}}{argmax} Ng[\tilde{\mathbf{W}}\mathbf{x}(t)] \tag{2.19}$$

where $Ng[\cdot]$ indicates a function which measures the nongaussianity of the output signals. Given s being a zero-mean random variable a classical measure of nongaussianity is the kurtosis or the fourth-order cumulant defined as:

$$kurt(s) = E[s^4] - 3E[s^2]^2 \tag{2.20}$$

which is generally simplified with $E[s^4] - 3$ if the variable is normalized to unit variance. The kurtosis is 0 if s is Gaussian, while assumes negative and positive values for subgaussian and supergaussian distribution, respectively. Subgaussian distributions have typically a flat pdf (i.e. uniform distribution). On the other hand, supergaussian distribution have a *spiky* or heavy tailed pdf; an important example is the Laplace distribution defined as:

$$p(s) = \frac{1}{\sqrt{2}} e^{\sqrt{2}|s|} \tag{2.21}$$

Laplace pdf is widely use to model the distribution of acoustic sources in frequency-domain. Acoustic sources such as speech or music have a sparse representation in time-frequency domain which means that the signals assume values close to zero in most of the time-frequency points. Therefore an effective method to extract a speech from a mixture of signals is by maximizing the supergaussianity of the output signals $\mathbf{y}$, which means to increase their sparseness in time-frequency domain.

The kurtosis is an interesting function since it is easy to compute and have linear properties which simplify the formulation of adaptation methods:

$$kurt(s_1 + s_2) = kurt(s_1) + kurt(s_2) \tag{2.22}$$

$$kurt(\alpha s_1) = \alpha^4 kurt(s_1) \tag{2.23}$$

Typically, sources are separated by means of the kurtosis applying a deflation procedure which extracts sequentially the sources with the highest kurtosis [21]. However, the kurtosis is not a robust measure since it is high sensitive to outliers [87]. For this reason other measures of non-gaussianity have been proposed. An information theoretic quantity to measure the nongaussianity is given by the negentropy which is based on the entropy $H(\cdot)$ defined as:

$$H(\mathbf{s}) = -\int pdf(\mathbf{s})log[pdf(\mathbf{s})]d\mathbf{s} \tag{2.24}$$

where $\mathbf{s}$ is a vector of random variables. The entropy of a random variable represents the degree of information given by their observations. It is well-known that among the distributions with the same variance, a Gaussian random variable is that with the highest entropy or in other terms that with the lowest information. Therefore a measure of non gaussianity may be performed by defining the negentropy J as:

$$J(\mathbf{s}) = H(\mathbf{s}_g) - H(\mathbf{s}) \tag{2.25}$$

where $\mathbf{s}_g$ is a vector of random variables of the same covariance matrix of $\mathbf{s}$. The practical evaluation of the negentropy is not an easy task since there is not any closed form to evaluate the entropy H. Therefore approximation are generally made as:

$$J(s) \simeq (E[G(v)] - E[G(s)])^2 \tag{2.26}$$

where v is a standardized Gaussian variable and G is some nonquadratic *contrast* function. Many contrast functions have been proposed which increase the robustness of the negentropy when compared to kurtosis based measures [39]. A popular method which uses the approximation in (2.26) is the popular Fast-ICA algorithm [40].

2.2 Natural Gradient and Kullback-Liebler Divergence

A popular method proposed in [3] is to estimate $\mathbf{W}$ by minimizing the mutual dependence between the sources according to a Natural Gradient optimization. It was shown in [38] that the minimization of the mutual dependence is connected to the Infomax principle which is a popular method for BSS based on the maximization of the output entropy of a neural network with non-linear outputs [8]. Furthermore it was shown in [13] that under certain conditions the Infomax principle is connected to the maximum-likelihood estimation of $\mathbf{W}$ proposed in [14] and [20].

According to the model in (2.1) we want to design an adaptive learning algorithm in order to estimate a demixing matrix $\mathbf{W}$ which minimizes the mutual dependence of the output $\mathbf{y}(t)$. For the sake of simplicity, it is assumed that the number of the sources *N* is equal to that of the

observations *M* and the sources are mutually independent with zero mean. For the basic derivation of the Natural gradient it is also assumed that the noise $\mathbf{e}(t)$ is negligible but extensions to the noisy case are provided by other methods such as Robust ICA [18].
The statistical independence between the output sources can be achieved by minimizing different statistical measures of mutual dependence. A popular method is to derive an adaptive learning rule which minimizes the Kullback-Leibler divergence between the joint probability distribution of the output sources $p(\mathbf{y}(t))$ and the product of the marginal probability distributions $\prod_{n=1}^{N} p(y_n(t))$, which is defined for the output sources $\mathbf{y}(t)$ as:

$$I_{KL} = \int p(\mathbf{y}(t))\text{log}\frac{p(\mathbf{y}(t))}{\prod_{n=1}^{N} p(y_n(t))}\mathbf{d}\mathbf{y}(t) \tag{2.27}$$

After some mathematical manipulation, (2.27) can be rewritten as:

$$I_{KL} \simeq E[\log p(\mathbf{y}(t))] - E\left[\sum_{n=1}^{N} \log p(y_n(t))\right] \tag{2.28}$$

where $E[\cdot]$ is the expectation operator.

By means of the steepest descent gradient the matrix $\mathbf{W}$ is updated in the direction which minimizes $\frac{\partial I_{KL}}{\partial \mathbf{W}}$. The derivation of the gradient can be split considering separately the two terms in the equation (2.28). Applying the partial derivative to the first term in the RHS of the equation leads to:

$$\frac{\partial E[\log p(\mathbf{y}(t))]}{\partial \mathbf{W}} = \frac{\partial}{\partial \mathbf{W}} E\left[\log \frac{p(\mathbf{x}(t))}{|det(\mathbf{W})|}\right] = -\frac{\partial E\left[\log|det(\mathbf{W})|\right]}{\partial \mathbf{W}} = -E[\mathbf{W}^{-1}]^T \tag{2.29}$$

where T is the matrix transpose. Deriving the partial derivative of each term in the summation of the second term of equation (2.28) leads to:

$$\frac{\partial E[\log p(y_n(t))]}{\partial W_{nm}} = E\left[\frac{\frac{\partial p(y_n(t))}{\partial y_n(t)}}{p(y_n(t))}\frac{\partial y_n(t)}{\partial W_{nm}}\right] = E\left[\frac{\frac{\partial p(y_n(t))}{\partial y_n(t)}}{p(y_n(t))}x_m(t)\right] \tag{2.30}$$

The pdf derivative cannot be analytically derived but it can be approximated as:

$$\frac{\frac{\partial p(y_n(t))}{\partial y_n(t)}}{p(y_n(t))} \simeq -\phi(y_n) \tag{2.31}$$

where $\phi(y_n(t))$ is a proper non-linear function of $y_n(t)$. Then, assuming that the output source distributions can be modeled with the same non-linearity (i.e. the distribution of the sources can be represented by the same model) we define the non-linearity vector function $\mathbf{\Phi}$ as:

$$\mathbf{\Phi}(\mathbf{y}(t)) = [\phi(y_1(t)), \cdots, \phi(y_N(t))]^T \tag{2.32}$$

Therefore the gradient can be approximated to:

$$\frac{\partial I_{KL}}{\partial \mathbf{W}} = -E[\mathbf{W}^{-1}]^T + E[\Phi(\mathbf{y}(t))\mathbf{y}(t)^T] \tag{2.33}$$

The gradient $-\frac{\partial I_{KL}}{\partial \mathbf{W}}$ represents the steepest decreasing direction of function I_{KL} when the parameter space is Euclidean. In our case the separation is possible only if the matrix $\mathbf{W}$ is not singular and therefore the space of the solutions is represented by all the $N \times N$ non singular matrices $\mathbf{W}$ [18]. By means of such an evidence in [3] it was introduced a Riemannian metric to the space of $\mathbf{W}$. It was shown that the true steepest-descent direction in the Riemannian space of parameters is $-\frac{\partial I_{KL}}{\partial \mathbf{W}}\mathbf{W}^T\mathbf{W}$. Finally the adaptive learning rule takes the following form:

$$\mathbf{W} \longleftarrow \mathbf{W} + \eta(\mathbf{I} - E[\Phi(\mathbf{y}(t))\mathbf{y}(t)^T])\mathbf{W}, \qquad \mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) \tag{2.34}$$

The matrix $\Phi(\mathbf{y}(t))\mathbf{y}(t)^T$ is known as *generalized covariance matrix* and is expected to be diagonal if the output signals are mutual independent. If the demixing matrix $\mathbf{W}$ is estimated on-line with the incoming data $\mathbf{x}(t)$, the expectation of the gradient is substituted with its instantaneous value. On the other hand, if the learning is applied to a batch of observed data $\mathbf{y}(t)$, the expectation may be substituted with the average over time.

2.3 BSS for convolutive mixtures

The linear model is rarely useful for the separation of acoustic sources in real environment. In fact, the mixing system is not well represented by a simple linear mixing model. In indoor environments, microphones record many replicas of the same acoustic source which propagate according to the different reflection paths and thus reach the array with different time-delays. Therefore, a better model for the recorded signals is to consider the mixtures $\mathbf{x}(t)$ as sum of convolved version of the original sources $\mathbf{s}(t)$ according to the impulse response between each source and microphone.

$$\mathbf{x}(t) = \mathbf{h}(t) * \mathbf{s}(t) \tag{2.35}$$

where $*$ indicates the convolution operator and $\mathbf{h}(t)$ is the matrix of the impulse responses:

$$\mathbf{h}(t) = \begin{pmatrix} h_{11}(t) & \cdots & h_{1N}(t) \\ \cdots & \cdots & \cdots \\ h_{M1}(t) & \cdots & h_{MN}(t) \end{pmatrix} \tag{2.36}$$

If $M \geq N$ the output signals would be estimated as:

$$\mathbf{y}(t) = \mathbf{w}(t) * \mathbf{x}(t) \simeq \mathbf{s}(t) \tag{2.37}$$

where $\mathbf{w}(t)$ is a matrix of deconvolution filters:

$$\mathbf{w}(t) = \begin{pmatrix} w_{11}(t) & \cdots & w_{1N}(t) \\ \cdots & \cdots & \cdots \\ w_{M1}(t) & \cdots & w_{MN}(t) \end{pmatrix} \tag{2.38}$$

In a practical implementation of a convolutive BSS, the demixing filters $\mathbf{w}(t)$ are discrete and have a finite length L which needs to be chosen according to the reverberation time of the room and to the frequency f_s used to sample the signals $\mathbf{x}(t)$. The number of parameters that need to be estimated is $M \times N \times L$ and thus the complexity increases with the length of the demixing filters needed to compensate the impulse responses. Therefore, the formulation of BSS algorithm for convolutive mixtures is much more complex than that of instantaneous mixtures since many parameters need to be jointly optimized by the same adaptation. Such a complexity represents also an obstacle to robustness of convolutive time-domain BSS since its convergence behavior becomes more unstable as L increases due to the sensitivity of gradient-based adaptation to the presence of local minima.

2.3.1 Main framework for convolutive BSS

Defining the state-of-art for the convolutive blind source separation field is not a trivial operation. In fact in the last decade a lot of ideas have been proposed but there is not a common evaluation framework that would allow one to have a clear comparison between the methods. Some trials to define an objective comparison have been done in the last two SIgnal source Separation Evaluation Campaigns (SISEC, [95]). However, a the lack of an objective evaluation criteria, coherent with the quality perceived by human listeners, hampered a clear interpretation of the results. Moreover, the optimality of the separation have been often overstressed against the robustness among environmental conditions, which would have a stronger impact in real-world applications. Furthermore, there is still a number of theoretical limits that still must be addressed and no method is able to address the source separation for all the cases. In this thesis, rather than defining the state-of-art for the convolutive BSS, we preferred to select three main BSS frameworks which, according to the author, are the most interesting according to the following evaluation criteria:

- straightforward physical interpretation of the problem
- computational cost and applicability in a real-time context

- capability to solve the separation for challenging situations
- innovation of the proposed work.

The chosen methods are: 1) Independent Vector Analysis 2) TRINICON ("Triple-N ICA for convolutive mixtures") 3) Frequency-domain BSS based on ICA. In the next subsections we briefly recall the bases of the first two methods while we will deeply analyze the third one in a separated chapter, since it defines the starting point for the innovative methods proposed in this thesis work.

IVA: Independent Vector Analysis

The Independent Vector Analysis is a recent method proposed by Intae Lee et al. [45] which aims to extend the ICA to the multivariate case. According to the property of the Fourier transform convolutive mixtures of time-domain signals can be represented by instantaneous mixture in frequency domain. This property is exploited by methods of BSS which perform the source separation in frequency-domain. However, applying the separation to each frequency independently the permutation problem [81] must be solved, which means find for the optimal permutation of the output at each frequency in order to group together all the components related to the same source. Especially in highly reverberant environment and when we work with a lot of sources the solution to the permutation problem is not trivial. The Independent Vector Analysis introduces a frequency coupling directly in the adaptive learning rule, in order to reduce the occurrence of wrong permutations.

Compared with ICA which works with uni-variate components signals, IVA extends the separation to multivariate components. The signal statistics is modeled with a multivariate distribution which consider all the frequencies at the same time. The adaptive optimization aims to find the set of the demixing matrices which makes the output multivariate signals as much independent as possible. Since all the frequency components are considered at the same time, the output source dependencies are minimized with correct permutations of the demixing matrices across the frequencies. Therefore, inner dependency between frequency components of each output are intrinsically considered. IVA is equivalent to a multiple layers of standard ICA where the demixing matrix of all the frequencies are jointly optimized. Figure 2.5 shows the mixing model for a generic case of N sources and M microphones. Each original source is modeled as a multivariate component as $\mathbf{s}_n = \left[\begin{array}{cccc} s_n^1, & s_n^2, & \dots & s_n^K \end{array}\right]^T$ where K is the number of the frequency components according to the short-time Fourier transform analysis. Similarly, the observed mixtures are modeled as: $\mathbf{x}_1 = \left[\begin{array}{cccc} x_m^1, & x_m^2, & \dots & x_m^K \end{array}\right]^T$ We define the multivariate

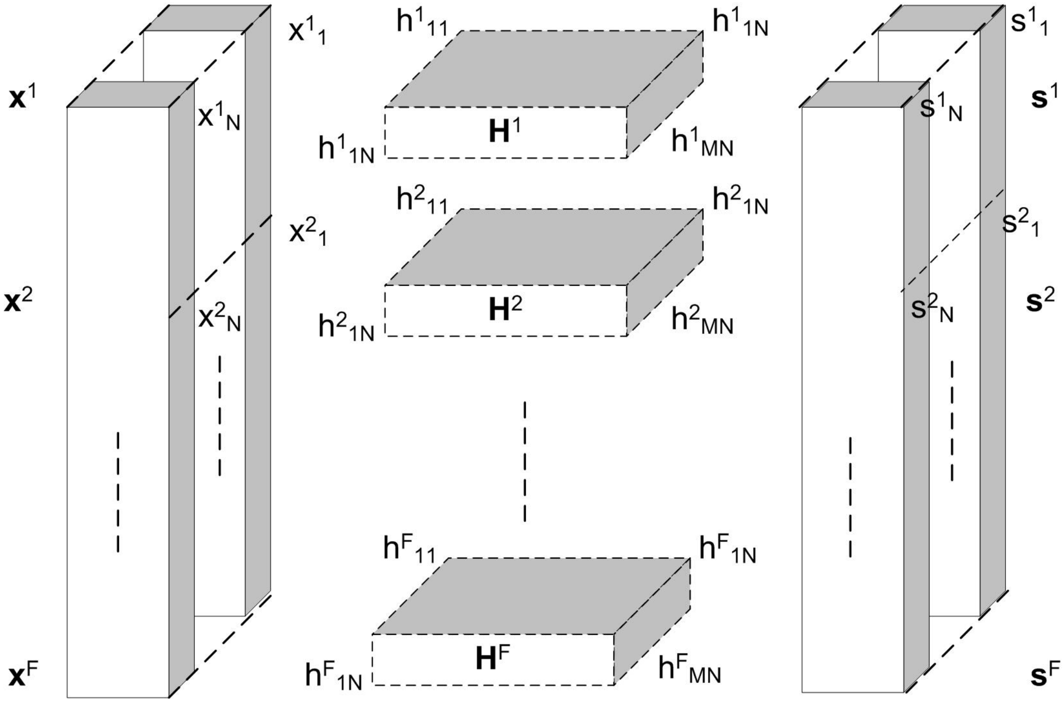


Figure 2.5: IVA mixing model for the case of 2 sources and 2 microphones[45].

output source as $\mathbf{y}_n = \left[\begin{array}{cccc} y_n^1, & y_n^2, & ... & y_n^K \end{array} \right]^T$ where each component is estimated as:

$$y_n^k = \sum_{m=1}^{M} w_{nm}^k x_m^k, \tag{2.39}$$

where w_{nm}^k are the coefficients of the demixing matrix $\mathbf{W}_k$ which separates the components of the *k-th* frequency. IVA aims to estimate the demixing matrices $\mathbf{W}_k$ which minimizes the mutual dependence between the output sources. Similarly to ICA an adaptive learning rule can be derived in order to minimize the Kullback-Leibler divergence between the joint probability distribution of the output sources $p(\mathbf{y}_1, \cdots, \mathbf{y}_N)$ and the product of approximated marginal probability distribution $q(\mathbf{y}_n)$:

$$\mathcal{C} = KL\left(p(\mathbf{y}_1, \cdots, \mathbf{y}_N) || \prod_{n=1}^{N} q(\mathbf{y}_n)\right) \tag{2.40}$$

After some mathematical manipulation (2.40) can be rewritten as:

$$\mathcal{C} = const + \sum_{f=1}^{F} log|det(\mathbf{W}_d^{-1})| - \sum_n E[log\ q(\mathbf{y}_n)] \tag{2.41}$$

where $\mathbf{W}_d^{-1}$ is the inverse estimate of $\mathbf{A}^d$ and $E[\cdot]$ is the expectation operator. Note that in this case the marginal probability distribution of $\mathbf{y}_n$ is a multivariate distribution of his K components. By differentiating the cost function $\mathcal{C}$ with respect to the coefficients w_{nm}^k and according to the Natural Gradient rule we obtain the following adaptation:

$$\mathbf{W}_k^{(i+1)} = \mathbf{W}_k^{(i)} + \eta(\mathbf{I} - E[\mathbf{\Phi}^k(\mathbf{y}^1, \cdots, \mathbf{y}^K)\mathbf{y}_k^H])\mathbf{W}_k^{(i)} \tag{2.42}$$

where i is the index of the iteration, $\mathbf{y}_k^H$ is the Hermitian transpose of $\mathbf{y}^k$ and $\mathbf{\Phi}^k(\mathbf{y}^1, \cdots, \mathbf{y}^K)$ is a vector of contrast function of all the output components $\mathbf{y}_n^k$. Defining the density model of $q(\mathbf{y}_n)$ as a scale mixture of Gaussian distributions, the elements of the vector approximates to:

$$\phi_n^k(y_n^1, \cdots, y_n^K) = \frac{y_n}{\sqrt{\sum_{k=1}^K |y_n^k|}} \tag{2.43}$$

The IVA algorithm has been applied to some challenging situations. The reported performance are very interesting and show that using constraints that introduce frequency dependence is a good strategy to improve the standard frequency-domain approach. However, the intrinsic complexity due to the high dimensionality of the contrast function slow-down the convergence of the algorithm and the computational cost is still too expensive to be implemented in real-time. Furthermore, similarly to time-domain methods the IVA updates all the variables within the same adaptation and is highly sensitive to the convergence into local minima. This drawback results in a high instability which lowers the average separation performance even when compared with tradition frequency-domain methods. A better discussion of such a problem is discussed in the next sections where the IVA has been used as a comparison algorithm for many experiments on simulated and real-world data.

TRINICON Blind Source Separation

The TRINICON based BSS is an alternative approach to perform the separation in time-domain. The framework presented by Buchner et al. [12] introduces a new model to deal with both blind source separation and multichannel blind deconvolution (MCBD). Multivariate models are used in the cost function to consider the entire temporal structure of the original source signals. The structure is analyzed by a block processing which is a simplification of the optimal formulation for the BSS in time-domain. The observed signals are segmented in blocks of length L_B. Each sample in time-domain is modeled as $x(b, shift)$ where b is the the block index and $shift$ is the time-shift index within the block, which can assume value from 0 to $L_B - 1$. The separated signals are modeled as:

$$\mathbf{y}(b, shift) = \mathbf{x}(b, shift)\mathbf{W}(b) \tag{2.44}$$

The vectors are composed by N elements according to the number of the sources (assumed to be equal to the number of microphones M).

$$\mathbf{x}(b, shift) = \begin{bmatrix} x_1(b, shift), & \cdots & , x_M(b, shift) \end{bmatrix} \tag{2.45}$$

$$\mathbf{y}(b, shift) = \begin{bmatrix} y_1(b, shift), & \cdots & , y_N(b, shift) \end{bmatrix} \tag{2.46}$$

The separation matrix for a given block can be represented as:

$$\mathbf{W}(b) = \begin{bmatrix} \mathbf{W}_{11}(b), & \cdots & , \mathbf{W}_{1M}(b) \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{N1}(b), & \cdots & , \mathbf{W}_{NM}(b) \end{bmatrix} \tag{2.47}$$

The *m-th* mixture is modeled as:

$$\mathbf{x}_m(b, shift) = \begin{bmatrix} x_m(bL + shift), & \cdots & , x_M(bL - 2L + 1 + shift) \end{bmatrix} \tag{2.48}$$

where L is the length of the FIR filters used to model the mixing system. At same way the output signals are modeled as:

$$\mathbf{y}_r(b, shift) = \begin{bmatrix} y_n(bL + shift), & \cdots & , y_N(bL - D + 1 + shift) \end{bmatrix} = \sum_{m=1}^{M} x_m(b, shift)\mathbf{W}_{mn}(b) \tag{2.49}$$

where D is the number of the time-lags taken into account to exploit the non-whiteness of the source signals. Each $\mathbf{W}_{nm}(b)$ is a Sylvester $2L \times D$ matrix which contains the filter coefficients of the deconvolution filters between the *n-th* source and the *m-th* microphone. The TRINICON method aims to estimate the demixing matrices $\mathbf{W}_{nm}(b)$ which are an exact estimate of the inverse of each $\mathbf{H}_{mn}(b)$ in order to have a perfect separation and deconvolution of the sources:

$$\mathbf{C} = \mathbf{H}_{nm}(b)\mathbf{W}_{mn}(b) = \mathbf{I}, \quad \forall n, m, b \tag{2.50}$$

The demixing matrix in TRINICON is estimated according to three different optimization criteria:

1. **Nongaussianity**: already used by ICA methods that aim to minimize the mutual dependence of the outputs.
2. **Nonwhiteness**: exploited minimizing output cross-correlations over multiple time-lags.
3. **Nonstationarity**: exploited by simultaneous minimization of output cross-correlation at

different time-instants.

According to these criteria a general cost function and an optimization rule are defined. The TRINICON framework is versatile since it realizes an estimation of the demixing system with a straightforward physical interpretation of the problem, according to the model of the mixing system in time-domain. Theoretically, the framework can be applied to any number of sources and for any reverberation time. However, increasing the length of the deconvolution filters the number of the parameters that have to be estimated is too high. To make possible the separation in a real-time context some approximations have been implemented, mainly based on the Second Order Statistics (SOS). These approximations introduce further reduction to the convergence speed and robustness of the algorithm and limit the possibility to apply the TRINICON-BSS in a real adverse conditions.

Frequency-Domain BSS based on ICA

Frequency-domain approaches to the BSS of convolutive mixtures have been widely investigated in the literature [72]. The main assumption on which such approach is based is that a convolutive mixture can be approximated by instantaneous mixtures of narrow frequency components. Among many algorithms proposed in the last years, in this work we refer to the approach based on the application of the ICA in frequency-domain which is well summarized in [81]. All the aspects of this approach are analyzed in the next chapter.

Chapter 3

Frequency-Domain BSS

A popular method to simplify the BSS for real-world scenario is to decompose the convolutive BSS into independent simpler tasks by transforming the signals in frequency-domain [81]. For the well known property of Fourier transform $F(\cdot)$ the convolutive model can be represented in frequency-domain as:

$$F(\mathbf{x}(t)) = \mathbf{x}(f) = F(\mathbf{h}(t) * \mathbf{s}(t)) = \mathbf{H}(f)\mathbf{s}(f) \tag{3.1}$$

where we indicate with $\mathbf{x}(f)$ and $\mathbf{s}(f)$ the Fourier transform of the signals and with $\mathbf{H}(f)$ the frequency response of $\mathbf{h}(t)$:

$$\mathbf{H}(f) = \begin{pmatrix} h_{11}(f) & \cdots & h_{1N}(f) \\ \cdots & \cdots & \cdots \\ h_{M1}(f) & \cdots & h_{MN}(f) \end{pmatrix} \tag{3.2}$$

Therefore, each frequency can be independently separated by methods for linear mixtures minimizing the risk to converge into local minima during the stochastic adaptation. To retrieve a sufficient statistics for each frequency in practice the observed signals sampled at the discrete time instants n_s, according to the sampling frequency f_s, are transformed in time-frequency series $\mathbf{x}(k, l)$ by a short Fourier transform (STFT) applied to L samples overlapped according to the shifting S

$$x_m(k, l) = \sum_{\mathsf{n}_s} x_n(\mathsf{n}_s) win_a \left(\frac{\mathsf{n}_s - lS}{f_s} \right) e^{-j2\pi f_k \frac{\mathsf{n}_s}{f_s}} \tag{3.3}$$

computed for all the discrete frequencies f_k:

$$f_k \in \left[0, \frac{1}{L} f_s, \cdots, \frac{k}{L} f_s, \cdots, \frac{L-1}{L} f_s \right] \tag{3.4}$$

and for each frame l. Here we indicate with $win_a(\mathsf{n}_s)$ an analysis windowing function with nonzero values only in the range $\left[-\frac{L}{2}\frac{1}{f_s}, \left(\frac{L}{2}-1\right)\frac{1}{f_s}\right]$ and tapers smoothly to zero at end of the interval. Typically for acoustic signals an Hanning window is chosen which is defined as:

$$win_a^H(\mathsf{n}_s) = \frac{1}{2}\left(1 + cos\frac{2\pi\mathsf{n}_s}{L}\right) \tag{3.5}$$

In real-life the impulse responses $\mathbf{h}(t)$ are theoretically infinite. However, we can assume that the length of the impulse responses equal the time needed to reduce the sound energy of $60dB$ when the source is not longer active. Such a time is defined as reverberation time T_{60} and corresponds to the time perceived for the sound to die away after the sound source ceases. Then, it is generally assumed that if the time-length of the analysis frame L is sufficiently larger than T_{60}, the convolutive model can be approximated with an instantaneous model as:

$$\mathbf{x}(k,l) = \mathbf{H}(k)\mathbf{s}(k,l) \tag{3.6}$$

where k is the frequency bin index, l is the index related to the analysis frame (e.g. in a short-time Fourier analysis), $\mathbf{x}(k,l)$ is a column vector of the observed mixtures (in frequency-domain), $\mathbf{s}(k,l)$ is the column vector of the original signals (in frequency-domain) and $\mathbf{H}(k)$ is a $M \times N$ mixing matrix. By means of the above approximation the convolutive model is simplified to L instantaneous models. Assuming $N = M$, by applying a complex-valued ICA to the time-series $\mathbf{x}(k,l)$, one can retrieve the original components $\mathbf{y}(k,l)$ by means of a set of demixing matrices $\mathbf{W}(k)$:

$$\mathbf{y}(k,l) = \mathbf{W}(k)\mathbf{x}(k,l) \tag{3.7}$$

For the well-known symmetry it follows that:

$$x_m(k,l) = x_m'(L-k,l) \quad \forall k = 1, ..., \frac{L}{2} - 1 \tag{3.8}$$

where $'$ indicates the complex-conjugate operator. Thus, it is sufficient to compute the demixing matrices for the first $\frac{L}{2}$ frequency bins and the symmetric part is reconstructed a *posteriori*.

3.0.2 Output Source Estimation

After the estimation of the demixing matrices $\mathbf{W}(k)$ the separation can be performed in two different ways:

- by estimating time-domain demixing filters and apply (2.37);
- by reconstructing the time-domain sources from the separated time-frequency series.

For the first case, after having applied the separation to the first $L/2$ bins, the symmetric part of the demixing filters are reconstructed as:

$$w_{nm}(k) = w'_{nm}(L-k) \quad \forall k = 1, \cdots, \frac{L}{2} - 1 \tag{3.9}$$

where $'$ denotes the complex-conjugate. After that, the demixing filters $\mathbf{w}(l)$ are obtained as:

$$\tilde{w}_{nn}(\mathsf{n}_s) = IDFT[w_{nm}(k)] \tag{3.10}$$

$$w_{nm}(\mathsf{n}_s) = cshift[L/2, \tilde{w}_{nm}(\mathsf{n}_s)] \tag{3.11}$$

where $IDFT$ indicates the inverse discrete Fourier transform and $cshift$ is an operator which introduces a circular shift of $L/2$ in the impulse response. The circular shifting is essential since the estimated filters in frequency-domain are generally acausal while the demixing model in (2.37) assumes causal demixing filters. In the second case, after having applied the separation to the first $L/2$ frequency components the symmetric part is obtained as in (3.8) and the time-domain signals are obtained by the inverse operation of the STFT:

$$y_n(\mathsf{n}_s) = \sum_l \left[win_{synth}\left(\frac{\mathsf{n}_s - lS}{f_s}\right) \sum_k y(k,l) e^{j2\pi f_k \frac{\mathsf{n}_s}{f_s}} \right] \tag{3.12}$$

where $win_{synth}(\cdot)$ is a synthesis window with nonzero values only in the range $\left[-\frac{L}{2}\frac{1}{f_s}, \left(\frac{L}{2}-1\right)\frac{1}{f_s}\right]$. According to the perfect reconstruction theorem, if we want to apply the STFT and inverse STFT without introducing any distortion, the analysis and synthesis window must verify the condition:

$$\sum_l \left[win_{synth}\left(\frac{\mathsf{n}_s - lS}{f_s}\right) win_a\left(\frac{\mathsf{n}_s - lS}{f_s}\right) \right] = 1 \quad \forall \mathsf{n}_s \tag{3.13}$$

For example, this condition is verified if we use an Hanning and rectangular window for the analysis and synthesis, respectively. Note that we used the term *demixing* matrix instead of deconvolution matrix. In fact we will see in the next section that a perfect deconvolution of the sources is not possible since even filtered versions of the original sources are possible solutions.

3.0.3 Scaling Ambiguity and Minimal Distortion Principle (MDP)

As mentioned in section 2.1 the solution of ICA has an intrinsic scaling ambiguity. In a frequency-domain BSS for such ambiguity we can write a generic solution for $\mathbf{y}(t)$ in the form:

$$\mathbf{y}(t) = \mathbf{c}(t)\ (*)\ [\mathbf{w}(l) * \mathbf{x}(t)] \simeq \mathbf{c}(t)\ (*)\ \mathbf{s}(t) \tag{3.14}$$

where $*$ indicates the convolution operator, $(*)$ indicates the element-wise convolution between the matrices (e.g. equivalent to the Hadamard product but with the convolution operator) and $\mathbf{c}(t)$ is a vector of arbitrary scaling filters. Equation (3.14) means that the output signals are identified up to an arbitrary filtering. The exact estimation of $\mathbf{c}(t)$ is not an easy task since it would turn into a perfect de-reverberation of the original sources which is well-know to be an ambitious problem even for the case of single source and multiple microphones.

An approximated solution can be found in frequency-domain rescaling the demixing matrix in order to make the energy of the output signals closer to that of the image of the source at microphones. For the sake of simplicity we assume that there is no permutation indeterminacy. At each frequency bin the output signals are obtained by the demixing matrix $\mathbf{W}(k)$ as:

$$\mathbf{y}(k,l) = \mathbf{W}(k)\mathbf{x}(k,l) \simeq \mathbf{\Lambda}(k)\mathbf{s}(k,l) \tag{3.15}$$

The output signals are scaled versions of the original sources according to the scaling matrix $\mathbf{\Lambda}(k)$ since the demixing matrix $\mathbf{W}(k)$ is an estimate of the inverse of $\mathbf{H}(k)$ up to a scaling ambiguity:

$$\mathbf{W}(k) = \mathbf{\Lambda}(k)\mathbf{H}(k)^{-1} \tag{3.16}$$

Therefore multiplying both the sides of (3.15) by $\mathbf{W}(k)^{-1}$ we have:

$$\mathbf{W}(k)^{-1}\mathbf{y}(k,l) = \mathbf{W}(k)^{-1}\mathbf{\Lambda}(k)\mathbf{s}(k,l) \simeq \mathbf{H}(k)\mathbf{s}(k) \tag{3.17}$$

If only the *n-th* signal in $\mathbf{y}(k,l)$ is not null, the multiplication by $\mathbf{W}(k)^{-1}$ would give the components of the corresponding source at each microphone. Therefore a common method to rescale the signals in $\mathbf{y}(k,l)$ is to return to the separated components of $\mathbf{s}(k,l)$ in $\mathbf{x}(k,l)$ by projecting-back the separated components as:

$$\mathbf{x}^{S_n}(k,l) = \mathbf{W}(k)^{-1}\mathbf{y}^n(k,l) \tag{3.18}$$

where $\mathbf{y}^n(k,l)$ indicates the vector $\mathbf{y}(k,l)$ where all the elements, except the *n-th*, are null and $\mathbf{x}^{S_n}(k,l)$ represents the projection of the *n-th* source to the microphones. Rather than estimating all the source images at the microphones in a more compact way the output signals are rescaled as:

$$\overline{\mathbf{y}}(k,l) = diag(\mathbf{W}(k)^{-1})\mathbf{y}(k,l) \tag{3.19}$$

which means that the *n-th* source is scaled with respect to the image at the *n-th* microphone. This rescaling strategy is known as Minimal Distortion Principle [53].

3.0.4 Circularity of FFT

Another important limitation of FD-BSS methods relies on the circularity effect of the Discrete Fourier Transform (DFT). Theoretically the Fourier analysis requires that time-domain signals are observed for an infinite time. In practical situation the time-domain signals are transformed in frequency-domain by means of DFTs computed over short-time blocks. Such approximation assumes that each signal, observed in a time block of length L, is periodical with period equal to f_s/L. Analogously, the discrete frequency response of a filter which length is L corresponds to a periodical time-domain filter. Such a filter is not realizable and therefore a one period realization is generally used. This simplification generates filters that have spikes in the global transfer function between the sources $\mathbf{s}(t)$ and the outputs $\mathbf{y}(t)$. This problem was arisen in [85] and [58]. The reasons of such spikes can be found in two main approximations made in the formulation of the FD-BSS:

- the simplification of the convolutive mixtures to a set of instantaneous mixtures in frequency-domain is valid only theoretically, when a continuous representation of the frequency responses is used. However, in a practical scenario the approximation performed by the DFT becomes worst as the frequency resolution is reduced. Consequently since the frequency responses are under-sampled the corresponding time-domain filters have an overlap with another period. The limit of the FD-BSS based on the STFT analysis was also pointed out in [76] where it is shown that the approximation has sense if the DFT window size is much longer than the reverberation time;

- another reason is that limiting the time-domain filters to a one-period realization we change the meaning of the filtering that is performed in frequency domain.

There are two possible solutions to such a drawback. One is to increase the length L of the DFT analysis but this method should be avoided for the intrinsic limitation of the FD-BSS [5]. Another method proposed in [85] is to modify the frequency-response so that the corresponding time-domain filters has a support equal to L and are smooth at both the sides of the window. This characteristic can be forced by imposing the time-domain filters to be windowed by a smooth taping window:

$$\overline{w}_{nm}(t) = w_{nm}(t) \cdot sw(t) \tag{3.20}$$

where $\overline{w}_{nm}(t)$ indicates the smoothed time-domain filter and $sw(t)$ the smoothing window. In frequency domain the windowing corresponds to:

$$\overline{w}_{nm}(f) = \sum_{\phi=0}^{f_s-\Delta f} w_{nm}(f) \cdot sw(f) \tag{3.21}$$

where $\Delta f = f_s/L$, and $sw(f)$ is the Fourier transform of $sw(l)$. If $sw(t)$ is an Hanning window the frequency responses must be smoothed as:

$$\overline{w}_{nm}(f) = [w_{nm}(f - \Delta f) + 2w_{nm}(f) + w_{nm}(f + \Delta f)]/4 \tag{3.22}$$

3.1 Permutation Problem

Permutation ambiguity is a problem which mostly affects the global performance of a FD-BSS system. In BSS based on ICA in the frequency-domain, if no connections are established between the outputs estimated at each frequency, the time-domain signals cannot be correctly reconstructed. Many methods have been proposed to correctly group a *posteriori* the frequency components related to the same source but permutation is still an open issue. We can identify two main categories of permutation solvers:

1. Methods based on spatial information such as TDOA or DOA (or Direction of Arrivals). They exploit the inherent phase information of the demixing matrices by assuming a coherence across frequencies [81]. More advanced approaches exploit the coherence of the demixing matrices across frequencies [25], or estimate the entire propagation model [83]..

2. method based on the inter-frequency magnitude correlation: they exploit the non-stationarity of sources as well as the spectral continuity between adjacent frequencies, which is typical of acoustic signals such as speech [75],[67].

Methods of the first category are more robust and reliable than that of the second category, especially when a short amount of data is available. On the other hand they are not optimal since they assume a simplified model for the acoustic propagation of the sources, which often does not well reflect the true conditions of many real-world scenarios. On the contrary, methods of the second category are less sensitive to the environmental conditions but lack of reliability and robustness if very short signals are analyzed [80] and for a higher number of sources. With this regard, hybrid approaches have been investigated to overcome the limitations of each single method [84], [88]. A novel method is also proposed in the chapter 6.

An alternative method to reduce the permutations is to avoid them by means of constraints which interconnect ICA adaptations related to different frequencies. As an example the Independent Vector Analysis (IVA) discussed in the previous chapter introduces these constraints by means of inter-frequency high order dependencies. Alternatively, another mechanism is to impose an inter-frequency coupling binding each demixing matrix with a frequency dependent model. An indirect method is to initialize the optimization by means of geometrical information such as source and microphone locations [70]. However, in these procedures the adaptation is strictly constrained by the model and often the geometrical information needs to be known in advance. A strategy which partially relaxes the constraint on the geometrical model imposes continuity only between the separation matrices of adjacent frequencies. The latter constraint is justified by the physical interpretation of section 3.2, and it also corresponds to continuity in the minima of ICA cost function [62]. This evidence was exploited in [23] and [74] by methods that use a recursive initialization of the separation matrix $\mathbf{W}_0(k)$ for the *k-th* frequency bin, with the solution $\mathbf{W}(k-1)$ obtained at the previous adjacent frequency bin. In fact, if the variation of two subsequent iterations in (3.31) is small enough, the convergence point and the permutations depend only on the initialization of the matrix $\mathbf{W}_0(k)$, which derives from the steepest-descent behavior of the natural gradient. However, such recursive approaches can lack of reliability when ICA converges towards poor solutions which introduce discontinuities across adjacent frequencies. The robustness of these systems can be improved by considering two main cues. First of all, the late reverberation introduces discontinuities between adjacent estimated demixing matrices, which affect the reliability of the recursion. Therefore, the recursive approach can be improved if the initializing matrix is a filtered version of the estimated demixing matrix. Second of all, the occurrence of poor solutions, which introduce discontinuities, can be minimized also by reducing the high statistical bias of ICA by means of *a priori* information regarding the time activity of the sources.

Since the work of this thesis was focussed on the robustness of source separation algorithms, particular attention have been given to the permutation problem and some reliable methods are proposed in the next chapters.

3.2 Physical Interpretation of FD-BSS

The mixing matrix associated to each frequency bin can be modeled as:

$$\mathbf{H}(k) = \begin{pmatrix} |h_{11}(k)|e^{-j\varphi_{11}(k)} & \cdots & |h_{1N}(k)|e^{-j\varphi_{1N}(k)} \\ \cdots & \cdots & \cdots \\ |h_{M1}(k)|e^{-j\varphi_{M1}(k)} & \cdots & |h_{MN}(k)|e^{-j\varphi_{MN}(k)} \end{pmatrix} \tag{3.23}$$

$$\varphi_{mn}(k) = 2\pi f_k T_{mn}(k) \tag{3.24}$$

where $T_{mn}(k)$ represents the propagation delay from the *n-th* source to the *m-th* microphone for the *k-th* frequency bin, $|h_{mn}(k)|$ is the magnitude of the frequency response between the *n-th* source and the *m-th* microphone for the *k-th* frequency bin, and f_k is the real frequency (in Hz) corresponding to the *k-th* frequency bin. For the intrinsic ambiguity of the ICA the demixing matrix $\mathbf{W}(k)$ estimated in (3.7), can be modeled as:

$$\mathbf{W}(k) = \mathbf{\Lambda}(k)\mathbf{\Pi}(k)\overline{\mathbf{H}}^{-1}(k). \tag{3.25}$$

Here $\mathbf{\Lambda}(k)$ is an arbitrary diagonal scaling matrix, $\mathbf{\Pi}(k)$ is a permutation matrix and $\overline{\mathbf{H}}^{-1}(k)$ is an estimate of the inverse of the true mixing matrix $\mathbf{H}(k)$. By applying the inversion to the matrix $\mathbf{W}(k)$ we obtain:

$$\mathbf{W}^{-1}(k) = \overline{\mathbf{H}}(k)\mathbf{\Pi}^T(k)\mathbf{\Lambda}^{-1}(k) \tag{3.26}$$

The matrix $\mathbf{\Lambda}^{-1}(k)$ is still diagonal and the *p-th* ratio between the elements of the *n-th* column of two generic rows a_p and b_p of the matrix $\mathbf{W}^{-1}(k)$ is scaling invariant:

$$r^p_{nk} = \frac{\{w(k)\}^{-1}_{a_pn}}{\{w(k)\}^{-1}_{b_pn}} \simeq \frac{h_{a_pn}(k)}{h_{b_pn}(k)} \tag{3.27}$$

Assuming the permutation problem to be solved (i.e., $\mathbf{\Pi}(k) = \mathbf{I}$), each ratio represents the acoustic propagation of the *n-th* source with respect to the microphone pair (a_p, b_p) at the *k-th* frequency bin. According to the model in (3.23) we can rewrite (3.27) as:

$$r^p_{nk} = |r^p_{nk}|e^{-j2\pi f_k \Delta t^p_{nk}} \tag{3.28}$$

where Δt^p_{nk} is the *n-th* time-delay at *k-th* frequency bin of the acoustic wave related to the *n-th* source recorded by the microphone pair p. In anechoic conditions the magnitude $|r^p_{nk}|$ and the TDOA Δt^p_{nk} are expected to be invariant with respect to the frequency, and consequently the phase of (3.28) must vary linearly. Hence, up to the scaling indeterminacy, the separation matrices can be approximated using only the parameters of the anechoic propagation model. Each state r^p_{nk} could be theoretically represented by the ideal model:

$$c_n(f_k)^p = \beta^p_n e^{-j2\pi f_k \overline{\Delta t^p_n}} \tag{3.29}$$

where β^p_n is the inter-microphone attenuation ratio and $\overline{\Delta t^p_n}$ is the true TDOA related to the *n-th* source, with respect to the microphone pair p. A generic demixing matrix for the *k-th* frequency bin can be parameterized as:

$$\hat{\mathbf{H}}(k) = \begin{pmatrix} 1 & \cdots & 1 \\ \cdots & \cdots & \cdots \\ c^p_1(f_k) & \cdots & c^p_N(f_k) \\ \cdots & \cdots & \cdots \end{pmatrix} \tag{3.30}$$

Once defined $\hat{\mathbf{W}}(k) = \hat{\mathbf{H}}^{-1}(k)$, the *n-th* row of $\hat{\mathbf{W}}(k)$ is equivalent to the steering-vector of a null-beamformer which removes all the $N-1$ interfering sources with respect to the *n-th* source of interest.

3.3 ICA for FD-BSS

Following the model in (3.23)-(3.26), $\mathbf{W}(k)$ can be estimated in a batch adaptation by a gradient descent as follows:

$$\mathbf{W}_{(i+1)}(k) \longleftarrow \mathbf{W}_{(i)}(k) + \eta \Delta \mathbf{W}_{(i)}(k) \tag{3.31}$$

where $\mathbf{W}_{(i)}(k)$ is the demixing matrix estimated at *i-th* iteration and $\Delta\mathbf{W}_{(i)}(k)$ is the gradient which updates the solution according to the step-size η. The gradient could assume many forms according to the cost function that ICA attempts to minimize. In the following we will consider the application of the Natural Gradient and the minimization of the Kullback-Liebler divergence. In frequency-domain it can be applied generalizing the adaptive learning derived in (2.34) to the complex-domain. In such a case, $\Delta\mathbf{W}_{(i)}(k)$ is updated at each iteration i as follows:

$$\mathbf{y}_{(i)}(k) = \mathbf{W}_{(i)}(k)\mathbf{x}(k) \tag{3.32}$$

$$\Delta\mathbf{W}_{(i)}(k) \longleftarrow (\mathbf{I} - E[\Phi(\mathbf{y}_{(i)}(k))\mathbf{y}_{(i)}(k)^H])\mathbf{W}_{(i)}(k) \tag{3.33}$$

where $\mathbf{y}_{(i)}(k)$ and $\mathbf{x}(k)$ indicates the output and input signal vectors, $\Phi(\cdot)$ is a non-linear function and $E[\cdot]$ is the expectation operator. In a batch implementation, the expectation operator is approximated by averaging the instantaneous generalized covariance matrix over l:

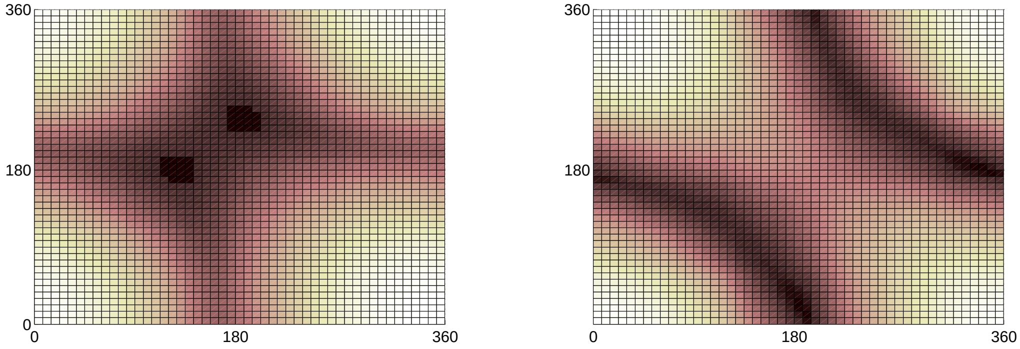
$$E[\Phi(\mathbf{y}_{(i)}(k))\mathbf{y}_{(i)}(k)^H] \simeq \langle \Phi(\mathbf{y}_{(i)}(k,l))\mathbf{y}_{(i)}(k,l)^H \rangle_l \tag{3.34}$$

where $\mathbf{y}_{(i)}(k,l)$ indicates the output signal vector at instant l and $\langle \cdot \rangle_l$ is the average operator. However, the statistical bias due to local dependencies between narrow-band signals increases as the number of the time-observations reduces; this occurs especially when the signals $\mathbf{x}(k,l)$ are obtained by means of a high resolution FFT. Such a bias is the main reason for the poor performance of standard frequency-domain BSS techniques, when long demixing filters are used to separate short signals.

To show the effect of this bias we can consider again the Natural Gradient applied to the Kullback-Liebler divergence and we measure the dependence of the output signals in the whole solution space. The Kullback-Liebler divergence between the output sources cannot be directly evaluated. However, it is possible to equivalently measure the convergence of the gradient search by a direct search of the minima in the function:

$$\Gamma(k) = ||\mathit{off}(A[\Phi(\mathbf{y}(k,l))\mathbf{y}(k,l)^H])||_F \tag{3.35}$$

where *off*$(\cdot)$ returns the off-diagonal element of the matrix, $|| \cdot ||$ indicates the Frobenius norm



(a) Surface computed using 9 second of data.

(b) Surface computed using 1 second of data.

Figure 3.1: Cost function computed for the case of two sources

and the estimated separated signals $\overline{\mathbf{y}}(k,l)$ are evaluated as:

$$\mathbf{y}(k,l) = \hat{\mathbf{H}}(k)\mathbf{x}(k,l) \tag{3.36}$$

where $\hat{\mathbf{H}}(k)$ is a test matrix defined as in (3.30).
To simplify the analysis we consider the case of two microphones and two sources. Therefore $\hat{\mathbf{H}}(k)$ can be simplified as:

$$\hat{\mathbf{H}}(k) = \begin{pmatrix} 1 & 1 \\ c_1(f_k) & c_2(f_k) \end{pmatrix}^{-1} \tag{3.37}$$

where the superscript p have been removed since for the two channel case there is only a single microphone pair combination. Thus for a given frequency, it is possible to plot the surface by evaluating the cost function in (3.35) for all the admissible values in the solution space of T_1 and T_2. Figure 3.1 shows a comparison of the surfaces obtained for a typical scenario: two sources are recorded in a real environment and the signals are sampled at fs=16kHz. Signals in time domain are transformed in time-frequency series by means of short-time Fourier analysis with 75 % overlapped frames of 4096 points. For a given frequency bin k (i.e. $k = 1200$), the surfaces have been evaluated by computing the function (3.35) using 9s and 1s of data which correspond to 11 and 136 time observations, respectively. The axes x and y are associated to the phases computed as $\varphi_1 = 2\pi f_k T_1$ and $\varphi_2 = 2\pi f_k T_2$. First of all, for each figure it is possible to observe two symmetric regions because, for the permutation ambiguity of the ICA, we can permute the row of the separation matrix, which means exchanging the axes x and y, still maintaining a valid solution. We observe that the surface in 3.1(a) is more convex than the one in 3.1(b) and it is possible to identify two minima that are the admissible solutions according to all the possible permutations. According to the propagation model in (3.23) such phases are expected to be associated to the true TDOAs of the sources. When the cost function is computed by only using 1s of data the surface does not present two well localized global minima because in a short-time the time-frequency series associated to the source signals are highly correlated. The separation based on the ICA is an ill-conditioned problem and, regardless of the scaling and permutation ambiguities, there is not a unique solution. Consequently, it means that the ICA could converge to several points in the solution space that are not always coherent with the physical interpretation of the problem. To overcome the intrinsic limitation of the ICA a prior knowledge can be exploited to better constrain its convergence behavior. In particular, in the next chapter, we focus on the convergence effect of the ICA when a proper initialization is adopted and when a better estimation of the covariance matrix is available.

Chapter 4

Recursive Convolutive ICA

In this chapter we present a new frequency-domain BSS method able to perform the separation of short-mixtures using long demixing filters, to cope with a high reverberation. A *priori* knowledge on the mixing system and on the source spectra is included directly in the ICA adaptation, in order to regularize the convergence and increase the overall robustness of the source separation. The work on this chapter is on the bases of an article accepted for publication in the IEEE Transactions in Audio Speech and Language processing, with the name of ”Convolutive BSS of short mixtures by ICA recursively regularized across frequencies” [65].

4.1 Introduction

Separation of acoustic sources in a real environment with the FD-BSS approach has some drawbacks. If the reverberation time can not be neglected, it is required to estimate a high number of demixing matrices. This task becomes hard if the mixtures of the sources are short. The capability to separate short mixtures is essential in real-life application since the mixing conditions could be non-stationary in a long-time. Under non-stationary mixing conditions, time-domain on-line BSS algorithm can be adopted [2]. However, time-domain methods suffer of convergence problem and of high computational complexity as the number of the parameters is increased. Alternatively a batch on-line FD-BSS algorithm was defined in [57] where the separation was obtained applying the ICA to disjoint block of data and iterating the estimated demixing matrices over the time.

For real-world application, one may also assume that the mixing conditions are *quasi-stationary*. Here we indicate with this term the conditions for which the sources do not likely change their locations in a short-time window. Thus, FD-BSS can be independently applied to the mixtures over a sliding window of short-time under the assumption that the used ICA method is accurate enough with the only data observed within the given window. Alternatively, a method to solve the separation problem with a short amount of data was proposed in [44].

It applies the ICA in time-domain to a signal subspace composed by delayed version of the observed mixtures and the output signals are obtained by clustering and reconstructing the separated components belonging to the same source. The method can be in principle applied to a subspace of any number of delayed replicas in order to describe a more accurate demixing model. However, as commonly met in many other time-domain methods, the computational complexity and the increasing dimensionality of the overall model limit its application to the estimation of short filters. For this reason by now on, we will focus on frequency-domain methods. In section 3.2 we showed as the FD-BSS is strongly affected by the problem of the statistical bias which is related to the number of frames of the STFT analysis. As pointed out in [78] a way to improve the accuracy is to decrease the number of the mixing matrices estimated with the available data, which means to limit the number of parameters that describe our mixing system. For instance, one may reduce the frame size of the short-time Fourier analysis, but in this way the ICA would be less accurate since an instantaneous mixing model cannot describe exactly the mixtures observed at each frequency. An alternative method to mitigate these drawbacks was proposed in [78] where a normalization procedure was proposed in order to apply the same ICA adaptation to a group of frequencies and without reducing the STFT frame size. However, since this method approximates the acoustic propagation by an anechoic model, its performance degrades as the distance between microphones and sources increases (i.e. as the DRR decreases).

We identify three major key points which must be considered to make an FD-BSS algorithm feasible for real applications:

- when only a short amount of data is available the accuracy of the ICA estimation must be controlled;
- the permutation correction method must be robust against variation of parameters, such as signals length, reverberation time and source dynamic;
- the computational power required for ICA adaptation must be reduced.

Following the above points, in this chapter a novel procedure is proposed which overcomes the limitations of the previous methods and improves the FD-BSS accuracy, allowing to estimate a large number of demixing matrices (i.e. long demixing filters) even with a short amount of data. The proposed algorithm has the following advantages:

1. the demixing matrices are not constrained by any anechoic model with clear benefit in presence of higher reverberation;
2. under certain conditions, the permutation problem is intrinsically reduced;
3. it is computationally efficient in terms of the number of ICA operations necessary at each frequency.

The chapter is organized as follows: Section 4.2 describes how to include *a priori* knowledge in the ICA and proposes the resulting structure of the Regularized Recursive-ICA (RR-ICA). In Section 4.3.1 a new performance evaluation criterion is proposed in order to compare the accuracy of RR-ICA with other well-known algorithms. Finally, experimental results on real-data are shown in Section 4.3.2, supported by a comparison with other popular BSS approaches.

4.2 Recursively regularized ICA

The novelty of RR-ICA is based on the use of a prior knowledge regarding the acoustic propagation and the spectral characteristic of the sources. In this section we analyze how to include this knowledge in the ICA, in order to regularize its convergence behavior.

4.2.1 Wiener-like weighting

Standard ICA-based BSS methods implicitly assume that the sources are always overlapping in time. On the other hand, acoustic sources have a sparse representation in time-frequency domain which implies that only one source in each time-frequency point has a not negligible energy. Furthermore, since acoustic sources such as speech and music have a common amplitude modulation (AM), continuity of the temporal activity is often observable in the frequency-domain. Here, we show how to exploit this continuity and the sparseness of the sources in order to improve the ICA accuracy, without sacrificing the resolution of the estimated frequency-responses.

First of all, we need to reformulate the optimization in (3.31)-(3.33) in terms of mixing matrix. Still according to the Natural Gradient optimization, the mixing matrix is updated [18] as follows:

$$\mathbf{y}_{(i)}(k) = [\overline{\mathbf{H}}_{(i)}(k)]^{-1}\mathbf{x}(k) \tag{4.1}$$

$$\Delta\overline{\mathbf{H}}_{(i)}(k) \longleftarrow \overline{\mathbf{H}}_{(i)}(k)(\mathbf{I} - E[\Phi(\mathbf{y}_{(i)}(k))\mathbf{y}_{(i)}(k)^H]) \tag{4.2}$$

$$\overline{\mathbf{H}}_{(i+1)}(k) \longleftarrow \overline{\mathbf{H}}_{(i)}(k) + \eta\Delta\overline{\mathbf{H}}_{(i)}(k) \tag{4.3}$$

During the adaptation all the coefficients of the matrix $\overline{\mathbf{H}}_{(i+1)}(k)$ are updated with the estimated gradient, i.e. computed by means of the expectation of the generalized covariance matrix $E[\Phi(\mathbf{y}_{(i)}(k))\mathbf{y}_{(i)}(k)^H]$. From a theoretical point of view, according to (4.2) the convergence is reached when the gradient becomes null, i.e., when the generalized covariance matrix $E[\Phi(\mathbf{y}_{(i)}(k))\mathbf{y}_{(i)}(k)^H]$ becomes an identity matrix. However, the diagonality of this matrix depends not only on the estimated $\overline{\mathbf{H}}(k)$ but even on the intrinsic correlation between the source signals which is not necessarily null in a short-time.

According to the mixing model in (3.6), each column of the matrix $\overline{\mathbf{H}}_{(i+1)}(k)$ is an estimation (up to a scaling ambiguity) of the mixing coefficients related to one of the sources. Hence, under the assumption of sparseness, the estimation of the gradient expectation can be improved by weighting the instantaneous gradient as follows:

$$\begin{array}{c}\Delta\overline{\mathbf{H}}_{(i)}(k) = \overline{\mathbf{H}}_{(i)}(k)(\mathbf{I} - E[\Phi(\mathbf{y}_{(i)}(k))\mathbf{y}_{(i)}(k)^H]) = \\ E[\overline{\mathbf{H}}_{(i)}(k)(\mathbf{I} - \Phi(\mathbf{y}_{(i)}(k)\mathbf{y}_{(i)}(k)^H)] \simeq \\ \sum_l \mathbf{\Psi}(k,l) \odot [\overline{\mathbf{H}}_{(i)}(k)(\mathbf{I} - \Phi(\mathbf{y}_{(i)}(k,l))\mathbf{y}_{(i)}(k,l)^H)]\end{array} \tag{4.4}$$

where $\odot$ is the Hadamard product (i.e. element-wise), and $\mathbf{\Psi}(k,l)$ is a weighting matrix constructed as

$$\mathbf{\Psi}(k,l) = [\boldsymbol{\psi}_1(k,l), \boldsymbol{\psi}_2(k,l), \cdots, \boldsymbol{\psi}_N(k,l)] \tag{4.5}$$

and the generic weighting column vector $\boldsymbol{\psi}_n(k,l)$ is defined as:

$$\boldsymbol{\psi}_n(k,l) = \frac{\psi_n(k,l)}{N_l}[1, 1, \cdots, 1]^T_{1\times N} \tag{4.6}$$

Here T indicates the vector transpose, N_l is the number of time frames on which the gradient is averaged and each $\psi_n(k,l)$ is a weight indicating the plausibility that the *n-th* source is active in the given time-frequency point. If the weights $\psi_n(k,l)$ are constrained in the range between 0 and 1, according to the definition in (4.6) the resulting gradient in (4.4) is equivalent to a weighted average over time. Under the assumption of ideal sparseness, for each point (k,l) all the weights $\psi_n(k,l)$ but one must be equal to 0, which means that only one source is assumed to be active. If the ideal assumption of sparseness is relaxed, a more accurate model is obtained imposing $\psi_n(k,l)$ to be a continuous weight which indicates the dominance of the *n-th* source over the other sources. An estimate of $\psi_n(k,l)$ can be obtained through a Bayesian estimation framework (see the *Appendix* in section 4.5), by means of the gain used to compute the minimum mean-square-error estimation of $y_n(k,l)$ in the sum $\sum_q y_q(k,l)$ [22]. Assuming the sources to be normally distributed, for the sake of simplicity, the weights are obtained as:

$$\psi_n(k,l) = \frac{E[|y_n(k,l)|^2]}{E[|y_n(k,l)|^2] + \sum_{q\neq n} E[|y_q(k,l)|^2]} \tag{4.7}$$

Thus, by the proposed weighting the generalized covariance matrix is not evaluated on the basis of the time-instants for which the source is likely to be silent, so reducing the bias introduced by the presence of interfering sources. Assuming no permutation occurs, and by exploiting the spectral smoothness of acoustic sources, the source energy can be recursively estimated across frequencies. For instance, $E[|\mathbf{y}(k,l)|^2]$ can be estimated from the highest to the lowest frequency bins, based on an autoregressive moving average, as:

$$E[|\mathbf{y}(k,l)|^2] = \alpha E[|\mathbf{y}(k+1,l)|^2] + (1-\alpha)|\mathbf{y}(k,l)|^2 \tag{4.8}$$

where α is a smoothing factor (e.g. $\alpha = 0.8 - 0.9$).

Note, this procedure cannot be applied if the adaptation is formulated as in (3.31). In the general case, the coefficients of the demixing matrix $\mathbf{W}(k)$ are a function of the coefficients of $\overline{\mathbf{H}}(k)$ and the matrix can not be partitioned to isolate specific sub-matrices (i.e. columns or rows) selectively related to a single source. In fact it makes sense to associate a different weight to each column of the gradient since different columns update the mixing coefficients of different sources. This is the reason for which we need the formulation of the Natural Gradient adaptation in terms of mixing matrix $\mathbf{H}(k)$.

4.2.2 Relationship between Wiener-like weighted ICA and Adaptive BeamFormers

Here we show that the Wiener weighting leads to an improved approximation of FD-BSS based on ICA to a set of independent Adaptive BeamFormers (ABF). According to the optimization in (3.33), the ICA adaptation converges when $E[\Phi(\mathbf{y}(k))\mathbf{y}(k)^H]$ is diagonalized. The generalized covariance matrix can be approximated with a sum of high-order cross-cumulants, in conformity with the Taylor expansion of the non-linearity $\Phi(\cdot)$.

Therefore, for zero-mean random variables the adaptation converges if all the high-order covariance matrices are diagonalized. In particular, considering only second-order-statistics (SOS), for $\alpha = 1$ one should obtain:

$$\begin{aligned} E[\mathbf{y}(k)\mathbf{y}(k)^H] &= E[\mathbf{W}(k)\mathbf{x}(k)\mathbf{x}(k)^H\mathbf{W}(k)^H] \\ &= E[\mathbf{W}(k)\mathbf{R}_{xx}(k)\mathbf{W}(k)^H] \\ &= \mathbf{W}(k)\mathbf{H}(k)E[\mathbf{R}_{ss}(k)]\mathbf{H}(k)^H\mathbf{W}(k)^H = \mathbf{D} \end{aligned} \tag{4.9}$$

where $\mathbf{D}$ is a diagonal matrix. Since the sources are uncorrelated $E[\mathbf{R}_{\mathbf{ss}}(k)]$ is ideally diagonal and the optimization in (4.9) is equivalent to compute the weights of a set of N null-beamformers steered to the direction of each source. However, this hypothesis is unlikely due to the intrinsic correlation between the source in a short-time.

Under the assumption of sparseness, the source vector $\mathbf{s}(k,l) = [s_1(k,l), \cdots, s_N(k,l)]$ can be approximated with:

$$\mathbf{s}(k,l) \simeq [0, \cdots, s_{n(l)}(k,l), \cdots, 0] \tag{4.10}$$

where $n(l)$ indicates the index of the unique active source, at frame l, which means that the instantaneous covariance matrix $\mathbf{R}_{\mathbf{ss}}(k,l)$ has all the off-diagonal elements equal to zero.

In this ideal case, according to (4.7), the weights $\psi_n(k,l)$ are equal to 1 in the frames l when only the *n-th* source is active. By means of the above-described weighting procedure, the *n-th* column of the gradient $\Delta\mathbf{W}(k)$, which updates the null-steering weights for the *n-th* source, is computed by averaging the instantaneous gradient over the time instants l when the source is active. Therefore, the expectation of $\mathbf{R}_{\mathbf{ss}}(k)$, obtained averaging $\mathbf{R}_{\mathbf{ss}}(k,l)$ over all the time instants, is a diagonal matrix. This means that the adaptation for the *n-th* source is equivalent to that of a null-beamformer when only the *n-th* source is present.

4.2.3 Least Mean Square tracking of the demixing matrix

Blind sources separation based on the frequency-domain is based on the assumption that a sufficiently large STFT window is used. In this conditions the convolutive time-domain mixtures can be well approximated by instantaneous frequency-domain mixtures. However, in real-world scenarios due to a high reverberation time, this cannot be satisfied. Furthermore, the statistical bias introduced by intrinsic dependencies between the sources observed in a short-time makes the estimated demixing matrix $\mathbf{W}(k)$ a noisy version of the true separation matrix $\widetilde{\mathbf{W}}(k)$. We can describe this approximation by the model:

$$\mathbf{W}(k) = \widetilde{\mathbf{W}}(k) \odot \mathbf{N}(k) \tag{4.11}$$

where $\odot$ is the Hadamard product (i.e. element-wise) and $\mathbf{N}(k)$ is a noise term representing the uncertainty of the ICA estimator. Due both to the difficulty in modeling statistics of $\mathbf{N}(k)$ and to the highly non linear phase distortion introduced by the reverberation, $\widetilde{\mathbf{W}}(k)$ cannot be accurately estimated. However, we found that the risk of convergence to wrong local minima, and then to inaccurate solutions, can be drastically reduced by initializing each ICA with a tracked demixing matrices which vary smoothly with the frequency. In fact, according to the propagation model of the acoustic wave, we expect a local continuity of the mixing system across frequencies, as long as the acoustic propagation is dominant along the direct path. With this approach, a direct estimation of $\widetilde{\mathbf{W}}(k)$ is avoided, while we assume that the stochastic optimization performed by ICA converges towards a solution very close to the true $\widetilde{\mathbf{W}}(k)$. The separation is performed recursively from the highest to the lowest frequency, which implicitly includes the knowledge about the continuity of the demixing matrices in the ICA adaptation. For each frequency a matrix $\mathbf{W}_{tracked}$ is obtained as a filtered version of the $\mathbf{W}(k)$ estimated at previous frequencies. However, $\mathbf{W}_{tracked}$ is not used as the separation matrix for the *k-th* frequency but only to initialize the matrix $\mathbf{W}_0(k)$ for the ICA of the new frequency.

Note, many methods can be adopted to estimated $\mathbf{W}_{tracked}$. In this work we propose to determine $\mathbf{W}_{tracked}$ by constraining its determinant to vary smoothly across frequencies. Smoothing the discontinuities in the determinant value has also the effect to remove discontinuities in $\mathbf{W}(k)$ (and then to have a smoother phase). In fact, the determinant of $\mathbf{W}(k)$ is strictly related to the spatial diversity of the sources and varies slowly with frequency under anechoic conditions.

In this work as a filtering procedure we employ an efficient ε- Normalized Least Mean Square (*ε-NLMS*) predictor, implemented with the variable ε approach as proposed in [16]. Figure 4.1 shows a typical plot of the determinant values, computed for the observed noisy

$\mathbf{W}(k)$ and the ε-*NLMS* smoothed versions. It is important to remind that the demixing matrices are still affected by the scaling ambiguity which needs to be mitigate before the filtering procedure. For this purpose we use the following normalization procedure:

$$\overline{\mathbf{W}}(k) = \mathbf{U}(k)\mathbf{W}(k) \tag{4.12}$$

where $\mathbf{U}(k)$ is defined as:

$$\mathbf{U}(k) = \begin{pmatrix} u_1(k) & 0 & \cdots & 0 \\ 0 & u_2(k) & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & u_N(k) \end{pmatrix}, \tag{4.13}$$

$$u_n(k) = \frac{e^{-j\cdot arg(w_{nn}(k))}}{\sum_{r=1}^{N} |w_{nr}(k)|} \tag{4.14}$$

Then, after the normalization the tracked matrices are computed according to:

$$\mathbf{W}_{tracked}(k) = \sum_{tap=1}^{L} \overline{\mathbf{W}}(k+tap)h_{tap}(k) \tag{4.15}$$

$$\mathbf{h}(k) = \mathbf{h}(k+1) + \mu \frac{e(k)^{*}\mathbf{D}(k)}{\mathbf{D}(k)^{H}\mathbf{D}(k) + \varepsilon} \tag{4.16}$$

$$e(k) = |\overline{\mathbf{W}}(k)| - \mathbf{h}(k+1)^{H}\mathbf{D}(k) \tag{4.17}$$

where L is the order of the filter, $\mathbf{h}(k)$ is the vector $[h_1(k), h_2(k), ..., h_L(k)]^T$ of the complex-valued coefficients of the smoothing filter evaluated at frequency k, μ is the step-size, ε is the normalization factor, and $\mathbf{D}(k)$ is the vector of the observed determinant values:

$$\mathbf{D}(k) = [d_1(k), d_2(k), ..., d_L(k)]^T, \quad d_{tap} = |\overline{\mathbf{W}}(k+tap)| \tag{4.18}$$

4.2.4 Proposed algorithm

The following pseudo code summarizes the main steps of the proposed RR-ICA procedure:

```
***INITIALIZATION***
$\mathbf{W}_{tracked}(highest\_frequency) = \mathbf{I}$
$\mathbf{\Psi}(highest\_frequency, l) = \mathbf{I}, \ \forall l$
***ITERATION***
for k=highest_frequency to 1
```

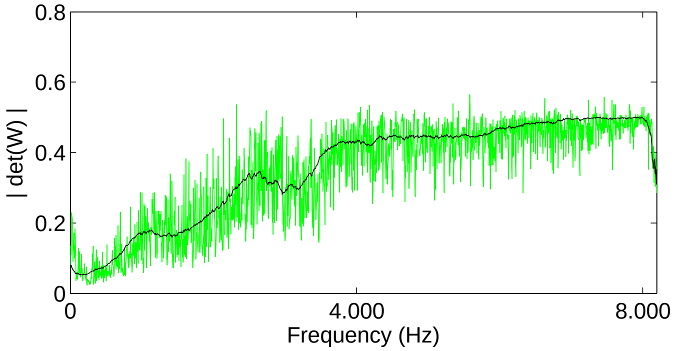


Figure 4.1: Magnitude of the determinant of $\mathbf{W}(k)$ and $\mathbf{W}_{tracked}(k)$ across frequencies.

```
  ***WIENER-LIKE WEIGHTS ESTIMATION***
  if ($k < highest\_frequency$)
    estimate $E[|\mathbf{y}(k,l)|^2]$ as in (4.8) and compute $\mathbf{\Psi}(k,l)$ as in (4.5)
  end
  ***ICA ADAPTATION***
  $\overline{\mathbf{H}}_0(k)$=$\mathbf{W}^{-1}_{tracked}(k)$,  $\mathbf{y}(k,l) = \mathbf{W}_{tracked}(k)\mathbf{x}(k,l)$
  compute $\overline{\mathbf{H}}(k)$ as in (4.1)-(4.4), from $\overline{\mathbf{H}}_0(k)$
  ***LMS DEMIXING MATRIX TRACKING***
  $\mathbf{W}(k)$=$\overline{\mathbf{H}}^{-1}(k)$
  if ($k == highest\_frequency$)
    $\mathbf{W}_{tracked}(k) = \mathbf{W}(k)$
  else
    normalize $\mathbf{W}(k)$ as in (4.12)-(4.14)
    $L = \min(L_{max}, highest\_frequency - k + 1)$
    estimate $\mathbf{W}_{tracked}(k)$ as in (4.12)-(4.18)
  end
end
```

4.2.5 Relationship Between RR-ICA and Others BSS Algorithms

It is important to underline that the recursive ICA is not equivalent to a frequency-domain implementation of a time-domain algorithm. Time-domain methods intrinsically introduces constraint in the global solution which theoretically avoid permutations and inter-frequency scaling ambiguities. On the other hand, the recursive regularization of RR-ICA partially constrains the solution of ICA at each frequency but the optimization is still performed by maximizing the mutual independence between narrow-band signals. Similarly, comparing RR-ICA to IVA, the phase and magnitude inter-frequency dependencies are only locally confined. Specifically, neglecting the permutations and assuming to have a sufficient amount of data, the final solution would be equivalent to that of a traditional frequency-domain approach in which the activity of the sources is ideally estimated. A graphical interpretation of the relationship between RR-ICA and other approaches is provided in Figure 4.2. As shown in previous works, such as [24],

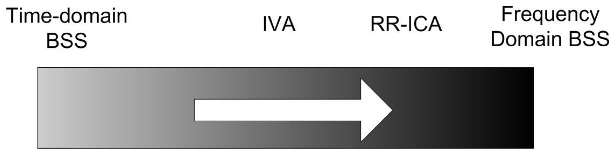


Figure 4.2: Graphical representation of the relationship between the degree of freedom of the solution and the BSS approaches.

the initialization of ICA plays an important role in conditioning the convergence speed and the accuracy of its solution. In frequency-domain, a typical way to improve the ICA convergence behavior is to initialize the iteration with a matrix computed as in (3.30), which can be parameterized only according to a prior knowledge of the source positions as well as of the array geometry. Alternatively, an efficient way is to use such spatial information by combining null-beamforming with ICA to get benefit of both deterministic and probabilistic criteria [80]. In all the cases it is implicitly assumed that the spatial information can be roughly estimated *a priori*, for example by using a preliminary ICA step. However, when very short signals are observed, even the use of a preliminary ICA stage would not be useful, as the reliability of the estimated demixing matrices is not sufficient and would compromise the refining of the solution with a further ICA stage.

The proposed approach implicitly solves that problem by exploiting frequencies redundancies, where the recursive initialization can be interpreted as a recursive propagation of the source statistics across frequencies. Furthermore, the tracked demixing matrix is only constrained to be continuous across frequency and does not have to approximate any ideal anechoic model. This guarantees a high degree of freedom of the final solution which makes the estimation of long filters possible even in a highly reverberant environment.

4.3 Experimental results

In this section we asses and analyze the performance of the proposed method through two groups of experiments::

- experiments with simulated data: impulse responses have been generated by the Lehmann & Johansson's image source method [46] according to the setup shown in Figure 4.3, where sources and microphones are assumed to be omnidirectional. Microphones are spaced apart of 0.02 m to avoid spatial aliasing for any source position and frequency, given a sampling frequency of $f_s = 16$kHz. Two clean speech utterances are convolved with the resulting impulse responses in order to generate the signals received at the two microphones. Then the mixture signals are obtained as sum, over each channel, of the signals generated for all the sources. The performance is evaluated averaging the results over

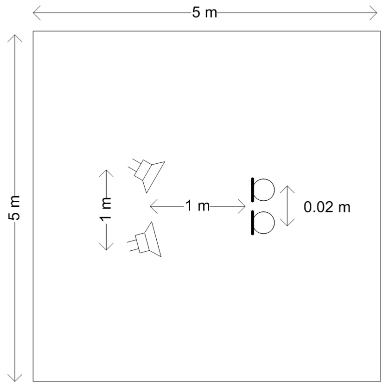


Figure 4.3: Simulated test setup, for a room characterized by $T_{60} = 300$ms.

15 different mixtures, which are generated by considering all the possible combinations of 6 sentences uttered by italian speakers (3 males + 3 females).

- experiments with real-world data: in this evaluation, two sources were recorded at $f_s = 16$kHz with two microphones spaced of $d = 0.02\ m$, according to the configurations showed in Figure 4.8. The mixtures are generated by summing the images of the source individually recorded at each microphone. Two rooms have been considered with reverberation time $T_{60} = 160ms$ and $T_{60} = 700ms$, respectively. The performance are determined averaging the results over 15 speaker combinations (3 males + 3 females).

4.3.1 Experiments with simulated data

In this experiment we are not interested in the evaluation of the overall performance of the BSS algorithm but on the accuracy of the estimated demixing matrices. For this purpose we define in this section a new performance measure which is invariant to the permutation and scaling ambiguities and allows us to measure only the global accuracy of RR-ICA in estimating the demixing matrices. This new measure is named as Propagation Model Error (PME), which measures the mean square error between estimated and true propagation models. First, since we are considering closely-spaced microphones, the attenuation ratios of the acoustic wave can be neglected and the observed propagation models are normalized as follows:

$$\overline{r}_{nk} = \frac{r_{nk}}{|r_{nk}|} \tag{4.19}$$

In the same way, one can compute the true normalized propagation models:

$$\overline{o}_{nk} = \frac{o_{nk}}{|o_{nk}|} \tag{4.20}$$

where o_{nk} are the models obtained by the simulated impulse responses $h_{mn}(k)$ as:

$$o_{nk} = \frac{h_{1n}(k)}{h_{2n}(k)} \tag{4.21}$$

Then the error for the *n-th* source PME is defined as:

$$PME_n = 10 \cdot log_{10} \frac{\sum_{k=1}^{N_{bin}} |\overline{r}_{nk} - \overline{o}_{nk}|^2}{N_{bin}} \tag{4.22}$$

where N_{bin} is the number of frequency bins of the FFT analysis. Note, applying a direct distance measure between the estimated and true impulse responses would not be meaningful in the BSS case because ICA is still affected by a scaling ambiguity. Then, using the ratios r_{nk}, which are scaling invariant, allow a better evaluation of the accuracy of the estimated demixing filters. For each frequency bin, the separation matrices are obtained by applying a Natural Gradient ICA to the time-series produced by a short-time Fourier analysis with windows of 4096 points and overlapping factor of 75%. To better show the robustness of the proposed method, the time observations used by ICA were obtained analyzing only a signal portion of 1 second.

To better asses the performance we evaluate the PME under two conditions:

1. permutation are ideally solved: the matrices $\mathbf{W}(k)$ are permuted with the optimal permutation which minimizes the total error:

$$\overline{\mathbf{\Pi}}(k) = \underset{\mathbf{\Pi}}{\operatorname{argmin}} \sum_{n} |\overline{r}_{nk}^{\mathbf{\Pi}} - \overline{o}_{nk}|^2 \tag{4.23}$$

   where $\overline{r}_{nk}^{\mathbf{\Pi}}$ are the states obtained from the matrix $\mathbf{W}(k)$ after the permutation $\mathbf{\Pi}$;

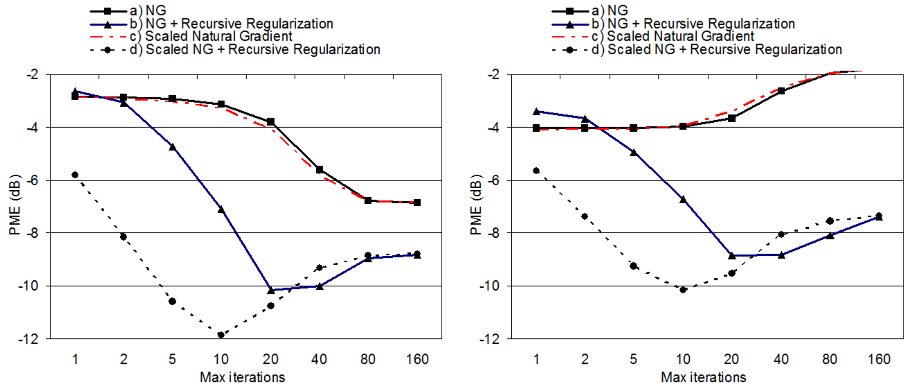
2. no additional permutation correction method is applied;

Furthermore, PME is computed imposing a different maximum number of iterations in the ICA adaptation, in order to better underline the benefit of the recursive regularization. This result is shown in Figure 4.4(a), for PME (averaged over both sources) obtained with each of the following ICA strategies: a) Natural Gradient (NG) b) Natural Gradient with the recursive regularization c) Scaled Natural Gradient [24] d) Scaled Natural Gradient with the recursive regularization. The Scaled Natural gradient is implemented in the algorithm in section 4.2.1 by modifying the equations (4.2) and (4.3) as:

$$\Delta\overline{\mathbf{H}}_{(i)}(k) \longleftarrow \overline{\mathbf{H}}_{(i)}(k)(\mathbf{I} - d_{(i)}(k)^{-1}E[\Phi(\mathbf{y}_{(i)}(k))\mathbf{y}_{(i)}(k)^H]) \tag{4.24}$$

$$\overline{\mathbf{H}}_{(i+1)}(k) \longleftarrow c_{(i)}(k)^{-1}[\overline{\mathbf{H}}_{(i)}(k) + \eta\Delta\overline{\mathbf{H}}_{(i)}(k)] \tag{4.25}$$

where the scaling factors $d_{(i)}(k)$ and $c_{(i)}(k)$ are computed as in [24] with respect to the chosen



(a) With ideal permutation correction (b) Without ideal permutation correction

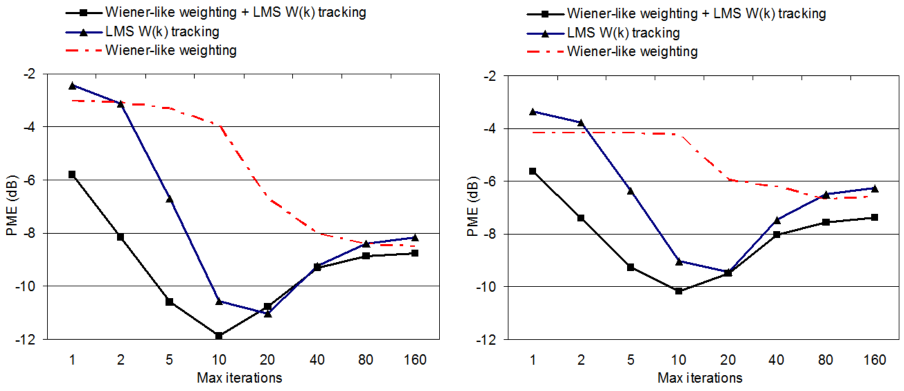
Figure 4.4: PME between the estimated and true propagation models. The ICA is applied to signal segments of just 1 second and with different number of ICA iterations.

nonlinear function $\Phi(\cdot)$.

First, we consider the PME results with the ideal permutation correction. Figure 4.4(a) shows that the recursive approach always reduces PME for any value of Max iterations. This means that the imposed deterministic constraint improves the accuracy of the demixing matrices estimation. Another important aspect is that when RR-ICA is applied a low PME value is obtained after only very few iterations. In particular when RR-ICA and Scaled Natural Gradient are combined together the computational cost is drastically reduced since a very good PME is obtained with only 10-20 iterations per frequencies. This suggest that algorithm d) is a good solution for real-time applications.

Second, we observed in figure 4.4 the PME results obtained without any ideal permutation correction. It is clear that PME is low only when the recursive strategy is applied, which indicates that the permutation problem is less critical when the deterministic knowledge is propagated through the recursion.

An increased number of iterations does not necessarily correspond to a lower PME; in fact, the minimum is obtained with 10-20 iterations. Note that the ICA optimization is still unconstrained at each frequency, and thus the final solution depends only on the observed data. When the number of iterations is limited, the final solution of ICA is more dependent on the initialization provided by $\mathbf{W}_{tracked}$ and the filters are induced to comply with continuity across frequency. Therefore, the effect of discontinuities due to poor solutions in the ICA optimization is reduced. In other words, as the number of ICA iterations decreases the solution is closer to the expected demixing matrix according to the introduced deterministic knowledge (see figure 4.6). On the other hand, as the number of ICA iterations increases the final solution depends more on the statistics of the observed data and the adaptation may converge to a poor solution if the data do not sufficiently describe the statistics of the sources. However, when both the



(a) With ideal permutation correction (b) Without ideal permutation correction

Figure 4.5: PME between the estimated and true propagation models. The Scaled NG is applied to signal segments of just 1 second. The performance is compared applying jointy or individually the two proposed regularization strategies.

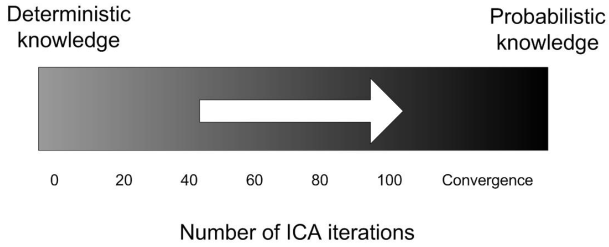


Figure 4.6: Graphical representation of the relationship between number of ICA iterations and the type of knowledge which defines the final solution.

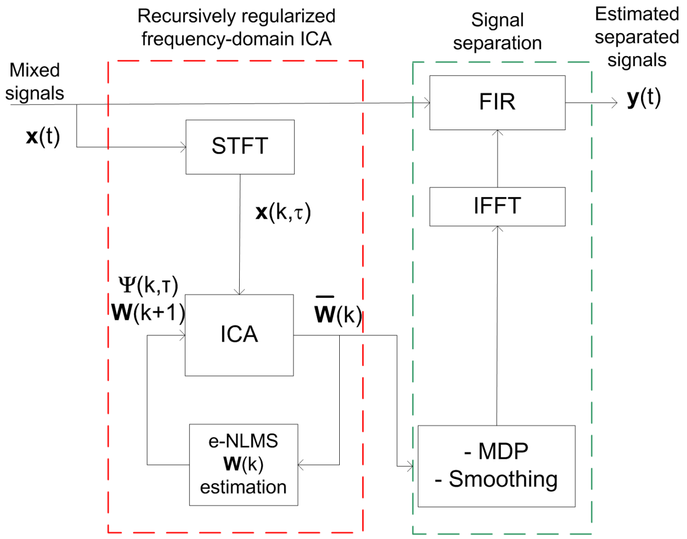


Figure 4.7: Block diagram of the implemented BSS algorithm.

regularization methods are applies, the whole deterministic knowledge is better exploited and PME further diminishes to values less dependent on the number of the ICA iterations.

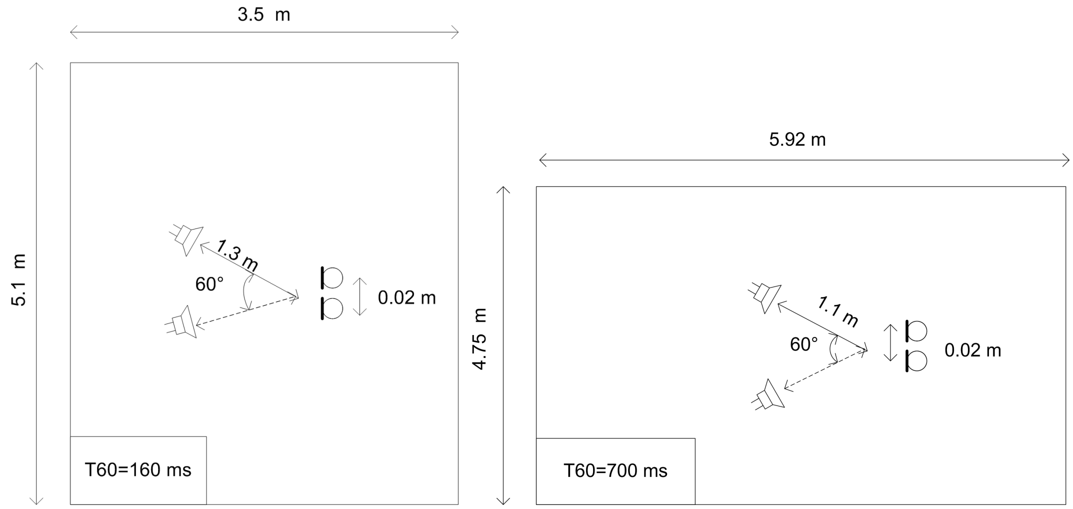
4.3.2 Experiments with real-world data

The block diagram in Figure 4.7 summarizes the main steps of the whole algorithm used in this evaluation. As a first step, a short-time Fourier analysis is performed in order to obtain a time-frequency representation of the observed mixtures $\mathbf{x}(t)$. For each frequency bin, the demixing matrices $\mathbf{W}(k)$ are obtained by using the Scaled Natural Gradient [24]. ICA is applied recursively from the highest to the lowest frequency according to the estimation/initialization method explained in Section 4.2.4. The Minimal Distortion Principle (MDP) [53] is applied to solve the scaling ambiguity and the smoothing method proposed in [85] is adopted in order to reduce the spikes due to the circularity effect of FFT. By means of the inverse Fourier transform, the resulting demixing matrices are finally used to obtain the demixing filters in the time-domain.

According to the FFT window size, the length of the filter used by the ε-NLMS procedure was chosen between 40 and 80 taps, while the step-size was set to 0.01. As for the Scaled Natural Gradient, the fixed step-size was set to 0.1.

The proposed algorithm was compared with other popular BSS methods:

- ALG1: the Independent Vector Analysis (IVA)[45], parameterized with step size 0.1 and a maximum number of 1000 iterations;
- ALG2: the time-domain Parra's method [71], applied by using time-domain filters whose size is half of the chosen FFT frame size, a number of matrices to diagonalize equal to 5,



(a) Test1 (b) Test2

Figure 4.8: Configuration of the two experimental setups. In the first case (Test1), the room is characterized by a moderate reverberation time of 160 ms, while in the second case (Test2) the room is larger and with a higher ceiling, which corresponds to quite challenging reverberant conditions.

and a maximum number of 1000 iterations;

- ALG3: the frequency-domain Pham's algorithm based on [75] and [74], parameterized with a FFT overlapping factor equal to 75% and a window size equal to 5.

Performance is evaluated with the BSS_EVAL toolbox using time-invariant filters of 1024 taps to represent the family of allowed distortions. The Source-to-Interferences Ratio (SIR) and Source-to-Distortion Ratio (SDR) are evaluated using the whole separated signals (whose length is about 9s); however, the demixing filters are computed using time observations within portions of different length ranging from 0.5 to 9 s. The experiments were conducted by using FFT analysis of different window size. All the related results are reported in Tables (I-VIII) in terms of average performance and standard deviation over all the separated sources. The symbol div indicates the divergence of the given algorithm.

To asses the performance, we do not take into account output signals for which the algorithms diverged and we perform the following evaluations:

- best performance over the FFT size
- average performance over FFT size and signal length

The best performance reported in Figures 4.9 and 4.10, for Test1 and Test2 configurations, respectively, shows that the proposed method performs well for any signal length. Moreover, the standard deviation (indicated by the black lines around the bars) indicates that the performance does not vary a lot with different speaker combinations and is more stable than the other methods

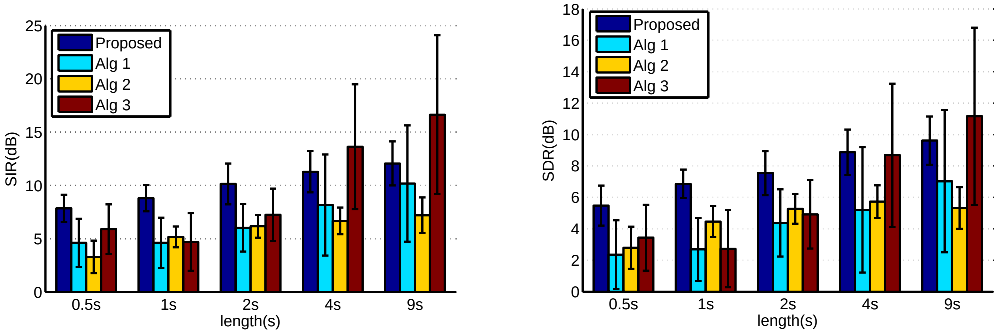


Figure 4.9: Results obtained in the Test1 experiments. Best performance is reported in terms of SIR and SDR, by applying the given algorithms with different signal lengths.

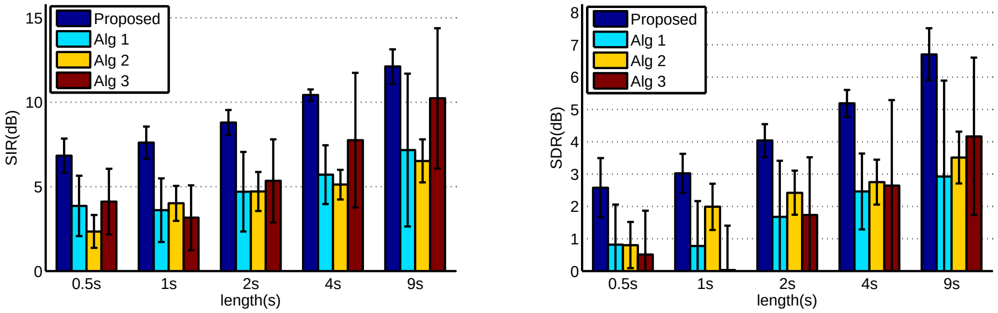
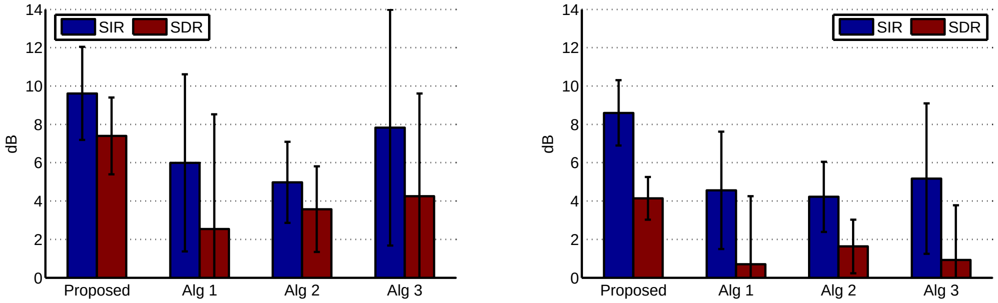


Figure 4.10: Results obtained in the Test2 experiments. Best performance is reported in terms of SIR and SDR, by applying the given algorithms with different signal lengths.



(a) Test1 experiments

(b) Test2 experiments

Figure 4.11: Average performance computed from all the possible combinations of FFT size and signal length.

used in the comparison. The IVA method (ALG1) provides good performance for some test files when the observed signals are long enough but does not have acceptable performance for short signals where the performance varies a lot with different speaker combinations. The time-domain method ALG2 is more stable but with very limited performance. The frequency-domain method ALG3 produced better results for Test2 and for longer signals but is even more unstable than IVA and does not provide a good average performance when applied to separate short signals (less than 2 seconds in Test2). On the other hand, the average performance in figure 4.11 shows that the proposed method is stable against variation of parameters and speaker combinations.

The improvement in the separation of RR-ICA is better underlined by the results for Test2 where the performance gap between the proposed and standard methods becomes even more evident. This is allowed by the improved stability of the source separation algorithm when long demixing filters are adopted to tackle a higher reverberation time.

4.4 Concluding Remarks

In this chapter we propose RR-ICA as a new frequency-domain approach to separate acoustic sources under highly reverberant conditions, even when acoustic signals of limited length are processed.

RR-ICA introduces a link between different frequencies by a recursive procedure which exploits the knowledge about filter coherence and time-activity of the sources. This information is estimated and propagated across frequencies ensuring a robust filter estimation even when a little amount of data is available.

The performance assessment confirms the superiority of the proposed method when compared to other popular BSS algorithms described in the literature.

RR-ICA is not affected by a severe problem of permutations even when short signals are separated. However, in case of intersections of the state trajectories permutation errors may occur. In the experimental evaluation the intersection of the state trajectories was avoided by tracking the states for frequencies not affected by spatial aliasing. For this reason an array with a small microphone spacing was used to avoid spatial aliasing for all the frequencies of the sampled mixtures. Among future directions of investigation, a relevant issue is to extend the given method in order to account for permutations in presence of state trajectory intersections. For example, the expected ideal phase-linearity of the demixing matrices, according to their physical interpretation, can be exploited a posteriori or directly included in the recursion to avoid permutations also in presence of trajectory intersections. The estimation of such linearity, which is related to the TDOA of the sources will be discussed in deep in the next chapter.

Another interesting direction of investigation is the analysis of the probability of intersection when a higher number of microphones is used. In this case multiple microphones would define multiple state observations of the acoustic propagation. Therefore, the acoustic propagation can be tracked in multiple dimensions and the trajectories are defined in a multidimensional space. The effect of different array geometries may be studied, in order to determine which is the best geometries which would make the intersection of the resulting multidimensional trajectories a rare occurrence.

4.5 Appendix

Given an observed signal $b = a_1 + a_2$, and the corresponding random variables A_1,A_2 and B, the conditional probability density functions of A_1 given B when A_1 and A_2 are independently distributed, is given by

$$p_{A_1|B}(a_1|b) = \frac{p_{B|A_1}(b|a_1)p_{A_1}(a_1)}{\int p_{B|A_1}(b|a_1)p_{A_1}(a_1)da_1} = \frac{p_{A_2}(b-a_1)p_{A_1}(a_1)}{\int p_{A_2}(b-a_1)p_{A_1}(a_1)da_1}, \tag{4.26}$$

where $p_{A_2}(\cdot)$ and $p_{A_1}(\cdot)$ are the probability density functions of A_2 and A_1, respectively. Then if A_2 and A_1 are zero-mean Gaussian-distributed with corresponding variances of $E[a_2^2]$ and $E[a_1^2]$, the conditional expected value for A_1 is given by [22]

$$E[A_1|B=b] = \frac{E[A_1^2]}{E[A_2^2]+E[A_1^2]}b, \tag{4.27}$$

That is, the ratio $E[A_1^2]/(E[A_2^2]+E[A_1^2])$ can be interpreted as the average amount of a_1 present in the sum $b = a_1 + a_2$ after accounting for the randomness in a_1 and a_2. Then the weights $\psi_n(k,l)$ in (4.7) may be defined imposing $a_1 = |y_n(k,l)|$ and $a_2 = \sum_{q\neq n} |y_q(k,l)|$.

SIR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	7.49(1.64)	8.11(1.47)	8.54(1.32)	8.77(1.38)	8.88(1.34)
2048	7.84(1.27)	8.79(1.23)	9.59(1.23)	10.36(1.36)	10.43(1.42)
4096	x	8.77(1.37)	10.15(1.91)	11.29(1.95)	12.06(2.07)
8192	x	x	x	10.15(3.51)	11.87(3.02)
16384	x	x	x	9.19(3.20)	10.77(4.01)

SDR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	5.88(1.43)	6.61(1.19)	7.06(1.05)	7.32(1.10)	7.43(1.08)
2048	5.48(1.28)	6.86(0.91)	7.86(0.91)	8.61(1.01)	8.74(1.05)
4096	x	5.67(0.88)	7.54(1.40)	8.87(1.44)	9.61(1.53)
8192	x	x	x	7.26(2.73)	9.10(2.28)
16384	x	x	x	5.40(2.31)	7.90(3.15)

TABLE I

Performance Proposed (test1)

SIR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	4.60(2.26)	4.61(2.37)	6.02(2.23)	7.19(2.98)	7.90(2.77)
2048	3.65(2.99)	3.25(2.72)	5.49(2.64)	8.02(3.65)	8.96(3.83)
4096	x	div	div	8.16(4.75)	10.18(5.46)
8192	x	x	x	5.25(5.42)	9.70(7.04)
16384	x	x	x	2.23(3.11)	7.60(6.98)

SDR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	2.35(2.20)	2.68(2.01)	4.37(2.14)	5.57(2.67)	6.28(2.63)
2048	-2.42(6.44)	-1.06(5.26)	3.36(2.37)	5.88(3.14)	6.79(3.43)
4096	x	div	div	5.21(3.99)	7.03(4.52)
8192	x	x	x	0.91(6.56)	6.02(5.58)
16384	x	x	x	-6.40(8.43)	2.98(7.21)

TABLE II

Performance ALG1 algorithm (test 1)

SIR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	2.92(1.07)	5.17(0.97)	6.16(1.06)	6.66(1.25)	7.17(1.53)
2048	3.29(1.52)	3.29(1.52)	5.37(1.13)	6.23(1.32)	7.01(1.86)
4096	x	div	2.78(0.96)	5.72(1.27)	7.21(1.67)
8192	x	x	x	2.88(1.24)	5.84(1.64)
16384	x	x	x	3.45(1.59)	3.45(1.59)

SDR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	2.45(1.01)	4.46(0.98)	5.27(0.95)	5.74(1.04)	6.14(1.22)
2048	2.79(1.34)	2.79(1.34)	4.57(0.95)	5.43(1.11)	5.97(1.51)
4096	x	div	0.90(0.73)	3.74(1.02)	5.32(1.32)
8192	x	x	x	0.42(0.84)	3.51(1.05)
16384	x	x	x	0.63(1.05)	0.63(1.05)

TABLE III

PERFORMANCE ALG2 ALGORITHM (TEST 1)

SIR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	5.89(2.33)	4.69(2.71)	7.25(2.46)	10.30(2.19)	10.45(1.71)
2048	2.48(1.80)	2.50(1.96)	3.84(2.99)	11.81(4.42)	13.79(3.82)
4096	x	1.85(1.00)	3.52(1.95)	13.63(5.86)	16.64(7.44)
8192	x	x	x	9.38(5.68)	13.89(9.00)
16384	x	x	x	3.56(2.76)	5.38(5.19)

SDR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	3.43(2.10)	2.73(2.45)	4.92(2.18)	8.07(2.00)	8.30(1.46)
2048	-1.28(2.08)	-0.05(1.41)	1.52(2.66)	8.50(3.86)	10.40(3.16)
4096	x	-2.27(1.11)	0.27(1.57)	8.68(4.56)	11.16(5.65)
8192	x	x	x	4.11(4.06)	8.61(6.56)
16384	x	x	x	-1.96(1.76)	1.41(3.35)

TABLE IV

PERFORMANCE ALG3 ALGORITHM (TEST 1)

SIR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	6.69(0.92)	7.17(0.68)	7.51(0.59)	7.57(0.57)	7.60(0.57)
2048	6.83(1.01)	7.58(0.72)	8.18(0.67)	8.27(0.68)	8.36(0.66)
4096	x	7.61(0.95)	8.80(0.74)	9.45(0.55)	9.70(0.67)
8192	x	x	x	10.43(0.33)	11.21(0.97)
16384	x	x	x	9.70(1.84)	12.11(1.02)

SDR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	3.14(0.54)	3.60(0.36)	3.77(0.38)	3.80(0.39)	3.83(0.38)
2048	2.57(0.92)	3.52(0.41)	3.96(0.41)	4.01(0.41)	4.06(0.41)
4096	x	3.02(0.60)	4.03(0.51)	4.45(0.37)	4.61(0.45)
8192	x	x	x	5.19(0.42)	5.72(0.76)
16384	x	x	x	4.59(1.26)	6.70(0.81)

TABLE V

Performance Proposed (test2)

SIR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	3.86(1.79)	3.60(1.89)	4.69(2.35)	5.71(1.74)	6.76(1.54)
2048	2.23(2.33)	2.88(1.36)	4.54(1.70)	5.40(2.05)	6.41(2.64)
4096	x	div	2.92(2.09)	5.24(2.55)	6.54(3.51)
8192	x	x	x	4.04(2.89)	7.16(4.53)
16384	x	x	x	2.31(2.19)	6.24(5.00)

SDR(dB)

FFT size \ Length(s)	**0.5**	**1**	**2**	**4**	**9**
1024	0.81(1.24)	0.77(1.39)	1.67(1.74)	2.46(1.17)	3.14(1.08)
2048	-3.74(4.75)	-0.08(1.02)	1.35(1.20)	1.95(1.45)	2.61(1.80)
4096	x	div	-0.21(1.58)	1.62(1.78)	2.47(2.31)
8192	x	x	x	0.52(2.00)	2.93(2.96)
16384	x	x	x	-4.23(6.90)	2.25(3.45)

TABLE VI

Performance ALG1 algorithm (test2)

SIR(dB)

FFT size \ Length(s)	0.5	1	2	4	9
1024	2.34(0.98)	4.01(1.04)	4.71(1.16)	5.12(0.88)	6.25(1.27)
2048	1.98(1.05)	1.98(1.03)	3.85(0.86)	4.90(0.92)	6.52(1.27)
4096	x	div	2.56(0.76)	4.62(0.92)	6.47(1.07)
8192	x	x	x	3.23(1.31)	5.22(1.26)
16384	x	x	x	3.99(1.83)	3.99(1.84)

SDR(dB)

FFT size \ Length(s)	0.5	1	2	4	9
1024	0.80(0.72)	1.98(0.71)	2.42(0.68)	2.75(0.69)	3.41(0.86)
2048	0.89(0.82)	0.90(0.80)	2.00(0.59)	2.66(0.65)	3.51(0.80)
4096	x	div	0.04(0.48)	1.52(0.58)	2.86(0.46)
8192	x	x	x	-0.01(0.90)	1.74(0.78)
16384	x	x	x	0.15(1.31)	0.15(1.31)

TABLE VII

PERFORMANCE ALG2 ALGORITHM (TEST2)

SIR(dB)

FFT size \ Length(s)	0.5	1	2	4	9
1024	4.11(1.95)	3.15(1.92)	5.34(2.47)	6.62(2.04)	7.30(1.76)
2048	2.04(1.34)	1.48(0.88)	2.54(2.34)	6.75(3.03)	8.25(2.91)
4096	x	1.56(0.87)	2.12(0.99)	7.74(3.99)	10.23(4.15)
8192	x	x	x	6.55(3.58)	9.13(6.13)
16384	x	x	x	3.34(2.21)	4.78(4.53)

SDR(dB)

FFT size \ Length(s)	0.5	1	2	4	9
1024	0.51(1.36)	0.03(1.37)	1.73(1.79)	2.80(1.54)	3.31(1.21)
2048	-1.83(1.26)	-1.56(0.68)	-0.72(1.76)	2.45(2.25)	3.33(1.81)
4096	x	-2.29(0.70)	-1.42(0.90)	2.64(2.65)	4.17(2.44)
8192	x	x	x	1.38(2.48)	3.55(3.84)
16384	x	x	x	-1.99(1.43)	0.47(3.10)

TABLE VIII

PERFORMANCE ALG3 ALGORITHM (TEST2)

Chapter 5

Generalized State Coherence Transform

In this chapter we present a new method to estimate the parameters of the acoustic propagation of multiple sources. This estimation is useful to reduce the permutation problem of FD-BSS and to allow the localization of multiple sources in multiple dimensions. Furthermore, the estimated ideal acoustic propagation is helpful to address the problem of the underdetermined source separation. The chapter includes some results of the State Coherence Transform (SCT) and of the Generalized SCT presented in [63][64],[61],[60],[59],[47].

5.1 Introduction

The problem of multiple TDOA estimation gives rise to high interest in the field of acoustic signal processing. Multiple speaker localization and blind source separation (BSS) are the fields where this problem has the most relevant impact. Especially when few microphones are used, a high reverberation time, strong environmental noise and spatial ambiguity make this task hard to address. In the last years multiple TDOA estimation was also investigated by the BSS community. First of all, it was shown in [83] that demixing matrices estimated in the frequency-domain BSS are strictly related to the parameters of the acoustic propagation of the sources and, under certain conditions, are expected to be coherent with the frequency. This knowledge is useful to reduce the permutation ambiguity of the frequency-domain approaches [72],[94],[82]. Second, the estimation of the mixing parameters or of the corresponding source locations is useful to make easier the separation in the underdetermined case (i.e. number of the sources greater than number of the microphones) [101],[54],[73].

Our earlier work [59] showed that a joint TDOA estimation for multiple sources and multiple directions of propagation can be performed by using the demixing matrices, estimated for different frequencies and time instants by an Independent Component Analysis algorithm. Such an estimation is accomplished by a proper transform of the state space associated with the demixing matrices which defines a multivariate likelihood function for the parameters of the acoustic propagation. The proposed transform, named Generalized State Coherence Transform (GSCT), is effective to estimate the propagation parameters of multiple sources even in the presence of spatial aliasing and therefore multiple sources can be localized if the sensor geometry is known. The GSCT extended the previous works presented in [64] and [61] to the multidimensional case allowing the source localization in multiple dimensions. In fact, due to the use of a multivariate likelihood of the acoustic propagation, the location estimation is not affected by the intrinsic inter-dimensional source ambiguity which arises if each dimension is treated independently [49] (i.e. presence of *ghost* locations).

The transform was originally formulated from a heuristic perspective. An alternative interpretation was provided in [48] where the SCT was compared to a kernel-density estimator of the complex-valued propagation model. However, in such an explanation no model for the acoustic propagation was explicitly considered, which limited the interpretation of the rule of the kernel-function used as non-linearity.

In this chapter we revise the formulation of the SCT/GSCT for both the univariate and multivariate cases. It will be shown that, with an appropriate choice for the non-linearities, the SCT is an approximated kernel-density estimator of the parameters of the acoustic propagation. According to this analysis new optimal definitions of the non-linearity are proposed, whose effectiveness is confirmed by numerical examples.

The chapter is organized as follows: the FD-BSS and the ideal physical interpretation of the states is recalled in section 5.2. A deep theoretical discussion of the SCT and its numerical evaluation is presented in section 5.3; analogously, the Generalized SCT is formulated and evaluated in section 5.4. Section 5.5 shows some possible application of the GSCT to real-world problems of multiple acoustic source localization and separation. Finally, concluding remarks ends this chapter.

5.2 Frequency-Domain BSS and Ideal Acoustic Propagation

First, let us recall the notation and the formulation of the SCT which will be used in the following sections. We assume to observe a number of N sources by an array of M elements. We start from the model defined in chapter 3 in the equations (3.23)-(3.27). Neglecting the magnitude, the acoustic propagation of the sources can be approximated as:

$$\overline{r}_{nk}^{p} = \frac{r_{nk}^{p}}{|r_{nk}^{p}|} = e^{-j2\pi f_k \Delta t_{nk}^{p}} \tag{5.1}$$

We remind that n, k and p are the state, frequency bin and microphone-pair indexes, respectively. In anechoic conditions, where the reverberation is absent, the propagation time-delays

Δt^p_{nk} are constant and the phase of (5.1) is expected to vary linearly across the frequency. In such a case the resulting time-delays equal the TDOA which can be deterministically derived by the geometrical configuration. In reverberant condition the time-delays Δt^p_{nk} would deterministically depend on the impulse responses between sources and microphones which sensibly vary with their locations. However, as long as the acoustic waves related to the propagation along the direct paths dominate the secondary reflections, the reverberation can be approximated by a statistical model. Let us start with the model used in [33]. The key assumption is that a generic impulse response between the *n-th* source and the *m-th* microphone can be decomposed as:

$$h_{mn}(t) = h^d_{mn}(t) + h^r_{mn}(t) \tag{5.2}$$

where $h^d_{mn}(t)$ and $h^r_{mn}(t)$ indicates the part of the impulse response related to the direct-path propagation and to the reverberation, respectively. While the propagation along the direct-path can be easily modeled according to the geometrical setup, modeling the reverberation deterministically is not trivial since many high order reflections travel from different directions and encounter the wall at different angles of incidence. However, if some conditions are fulfilled the reverberation can be modeled statistically as a random diffuse sound field [33]. If the source and microphone locations do not change, the randomness of $H^r_{mn}(f)$ (which is the Fourier transform of $h^r_{mn}(t)$) is not related to the time of observation. Thus, we can model $H^r_{mn}(f)$ as a random impulse response $\mathrm{H}^r_{mn}(f, \boldsymbol{\theta})$, where we indicate with $\boldsymbol{\theta}$ the vector of source and microphone location. Thus, the whole frequency response of $h_{mn}(t)$ can be modeled statistically with a random variable $\mathrm{H}_{mn}(f, \boldsymbol{\theta})$ as:

$$\mathrm{H}_{mn}(f, \boldsymbol{\theta}) = H^d_{mn}(f, \boldsymbol{\theta}) + \mathrm{H}^r_{mn}(f, \boldsymbol{\theta}) \tag{5.3}$$

where $H^d_{mn}(f, \boldsymbol{\theta})$ is the geometrically derived frequency response related to the direct-path propagation. The ratio between the generic random frequency responses related to the *p-th* microphone pair can modeled with the random variable:

$$\mathrm{r}(f, \boldsymbol{\theta}) = \frac{H^d_{a_p n}(f, \boldsymbol{\theta}) + \mathrm{H}^r_{a_p n}(f, \boldsymbol{\theta})}{H^d_{b_p n}(f, \boldsymbol{\theta}) + \mathrm{H}^r_{b_p n}(f, \boldsymbol{\theta})} \tag{5.4}$$

$$\simeq \frac{|H^d_{a_p n}(f, \boldsymbol{\theta})| e^{-j2\pi f T_{a_p}(\boldsymbol{\theta})} + \mathrm{H}^r_{a_p n}(f, \boldsymbol{\theta})}{|H^d_{b_p n}(f, \boldsymbol{\theta})| e^{-j2\pi f T_{b_p}(\boldsymbol{\theta})} + \mathrm{H}^r_{b_p n}(f, \boldsymbol{\theta})}$$

Here we indicated with $T_{a_p}(\boldsymbol{\theta})$ and $T_{b_p}(\boldsymbol{\theta})$ the time of arrival (TOA) between the source and the a_p-*th* and b_p-*th* microphone, respectively. Equation (5.4) can be rewritten as:

$$\simeq \frac{|H^d_{a_pn}(f,\boldsymbol{\theta})\cdot \mathcal{D}^r_{a_pn}(f,\boldsymbol{\theta})|e^{-j2\pi f T_{a_p}(\boldsymbol{\theta})+\phi_{a_p}(f,\boldsymbol{\theta})}}{|H^d_{b_pn}(f,\boldsymbol{\theta})\cdot \mathcal{D}^r_{b_pn}(f,\boldsymbol{\theta})|e^{-j2\pi f T_{b_p}(\boldsymbol{\theta})+\phi_{b_p}(f,\boldsymbol{\theta})}} \tag{5.5}$$

where $\mathcal{D}_{a_pn}(f,\boldsymbol{\theta})$, $\mathcal{D}_{b_pn}(f,\boldsymbol{\theta})$ and $\phi_{a_p}(f,\boldsymbol{\theta})$, $\phi_{b_p}(f,\boldsymbol{\theta})$ are the random variables representing the distortion introduced by the reverberation in the magnitude and phase of the frequency responses, respectively.

To simplify the model we normalize (5.5) by its magnitude assuming that the acoustic propagation is well described only by the time-delays, which is a reasonable assumption if the microphone spacing is not too high. Furthermore, because for the physical interpretation the reverberation introduces a distortion in the resulting direction of the acoustic propagation, the randomness of the phase distortion can be modeled as randomness in the time-delay:

$$\bar{\mathrm{r}}(f,\boldsymbol{\theta}) \simeq \frac{e^{-j2\pi f[T_{a_p}(\boldsymbol{\theta})+\mathrm{dT}_{a_p}(f,\boldsymbol{\theta})]}}{e^{-j2\pi f[T_{b_p}(\boldsymbol{\theta})+\mathrm{dT}_{b_p}(f,\boldsymbol{\theta})]}} \tag{5.6}$$

where we indicate with $\mathrm{dT}_{a_p}(f,\boldsymbol{\theta})$ and $\mathrm{dT}_{b_p}(f,\boldsymbol{\theta})$ the noise in the time-delay. Still, since we modeled the reverberation as a diffuse sound field, it can be demonstrated that given two generic frequencies f_1 and f_2 such that $|f_2 - f_1| > \frac{1}{T_{60}}$, $\mathrm{dT}_{a_p}(f,\boldsymbol{\theta})$ and $\mathrm{dT}_{b_p}(f,\boldsymbol{\theta})$ may be considered uncorrelated [33]. Assuming $\mathrm{dT}_{a_p}(f,\boldsymbol{\theta})$ and $\mathrm{dT}_{b_p}(f,\boldsymbol{\theta})$ to be identically distributed across the frequencies the ratio in (5.6) is rewritten as:

$$\bar{\mathrm{r}}^p(f,\boldsymbol{\theta}) = e^{-j2\pi f[T_{a_p}(\boldsymbol{\theta})-T_{b_p}(\boldsymbol{\theta})+\mathrm{dT}_{a_p}(f,\boldsymbol{\theta})-\mathrm{dT}_{a_b}(f,\boldsymbol{\theta})]} \simeq e^{-j2\pi f[\tau^p(\boldsymbol{\theta})]} \tag{5.7}$$

where $\tau^p(\boldsymbol{\theta})$ is the random variable of the TDOA of the *n-th* source with respect to the *p-th* microphone pair. From now on we consider the TDOA to be the only parameter of the acoustic propagation with respect to a chosen microphone pair and to simplify the notation we remove the dependence from the source and microphone locations $\boldsymbol{\theta}$. Considering the discrete-frequencies f_k obtained by the STFT analysis, for the *p-th* microphone pair, the normalized propagation can be modeled statistically as:

$$\mathrm{r}^p(k) \sim e^{-j2\pi f_k \tau^p} \tag{5.8}$$

where τ^p is a random variable of given pdf. Hence, the states in (5.1) can be considered a sample of observations of $\mathrm{r}^p(k)$. If a sample of τ^p is available, the parameter of the underlying distribution can be in principle estimated by a Maximum Likelihood Estimator (MLE). However, a sample of τ^p cannot be always derived from $\bar{r}^p_{nk}$ since the function (5.8) is not invertible for all the values k.

Assuming the sources to be in anechoic environment and defining c the sound speed and d the distance between the microphones, independently on the location of the source, for frequencies smaller than $f_{max} = \frac{c}{2d}$ the function in (5.1) is invertible and for each state it is possible to have an observation of τ^p as:

$$\tau^p_{nk} = \frac{arg(\overline{r}^p_{nk})}{2\pi f_k}. \tag{5.9}$$

Since the time-delays τ^p_{nk} are assumed to be an independent and identically distributed (i.i.d) sample of a random variable τ^p, a MLE of the parameters of its pdf can be easily derived. From the central limit theorem (CLT) one may argue that the impulse response parts due to the diffuse reverberation is Gaussian and consequently, after the above approximation, also τ^p can be approximated with a Gaussian variable of parameters (μ^p, σ^p). Thus, if the normalized states $\overline{r}^p_{nk}$ represents the acoustic propagation of a single source, the MLE of τ^p would lastly correspond to averaging the value of τ^p_{nk} for each n and k.

In a general case, the states $\overline{r}^p_{nk}$ represent the propagation of N different sources and τ^p can be modeled with a random variable with a multimodal distribution (i.e. a mixture of Gaussians). Therefore, there is not a closed-form solution for the parameters and clustering techniques should be considered as an alternative (i.e. the K-means used in [86]).

It is worth noting that, by considering a model of the propagation of multiple sources with a single multimodal variable τ^p, we avoid to explicitly solve the permutation problem since the randomness of the permutation is implicitly modeled in the randomness of τ^p.

Note, the equivalence in (5.9) is still an approximated method to estimate the time-delays since the phase $2\pi f_k \tau^p$ varies in the range $[0, \pi]$ for frequency smaller than f_{max} only if we assume anechoic propagation. However, in reverberant environments, especially for frequencies close to f_{max} such a condition is not guaranteed. Furthermore, using only the states estimated at frequencies smaller than f_{max} limits the estimation accuracy. On this regard, in the next section we propose an alternative function of the observed $\overline{r}^p_{nk}$ which is able to estimate an approximated pdf of τ^p, still using states corresponding to frequencies higher than f_{max}.

5.3 State Coherence Transform

We first recall the definition of the State Coherence Transform in the form proposed in [61] and after a discussion on possible statistical interpretations we generalize the transform to the multidimensional case. The monodimensional State Coherence Trasform is formulated as follows:

$$SCT^p(\tau) = \sum_k \sum_{n=1}^{N} g\left(\epsilon^p_{nk}(\tau)\right) \tag{5.10}$$

where $\epsilon^p_{nk}(\tau) = |\overline{r}^p_{nk} - c(f_k, \tau)|$, N is the number of states (i.e. number of the sources), $g(\cdot)$ is a generic decreasing monotonic function of $\epsilon^p_{nk}(\tau)$, $c(f_k, \tau)$ is the ideal, geometry-derived propagation model of a source for which the TDOA is equal to τ:

$$c(f_k, \tau) = e^{-j2\pi f_k \tau} \tag{5.11}$$

The idea behind the formula in (5.10) is that we can measure how likely an ideal anechoic model defined by a parameter τ fits the observed states $\overline{r}^p_{nk}$. In other words the SCT evaluates the coherence of the observed states along the direction spanned by the model $c(f_k, \tau)$. Each state may represent the propagation of a different source and therefore such a coherence is expected to be maximized for different models, which means that the likelihood of the SCT should present two clear peaks. Therefore, we have a multi-fit problem which can be addressed by properly controlling the function g. With a linear function $g(\epsilon^p_{nk}(\tau)) = -\epsilon^p_{nk}(\tau)$ the coherence would be measured by simply integrating the error between a candidate model $c(f_k, \tau)$ and the observed states. Then, the states that do not belong to the source for which we want estimate the coherence are outliers which would bias the error measure. It may result an SCT envelope which resolution is not sufficient to discriminate the two expected peaks. Nevertheless, if we assume the phase of the states belonging to different sources to be sparsely distributed in the frequency-phase space, the outliers can be better removed from the summation in (5.10) by defining g as a non-linear monotonic function which rapidly decreases as the distance $\epsilon^p_{nk}(\tau)$ increases.

5.3.1 Equivalence Between SCT and MLE

The SCT was derived heuristically but with a proper definition of $g(\cdot)$ it can be interpreted as an approximated maximum likelihood estimator (MLE) of the TDOA parameters of the acoustic propagation. Note, with TDOA parameters we refer to the TDOAs of the propagation of all the sources, with respect to different microphone pairs.

To simplify the analysis, let us consider the case of a single microphone pair and define with $\mathbf{r}$ the following vector:

$$\mathbf{r} = [\overline{r}_{nk}]_{n,k} \tag{5.12}$$

where $\overline{r}_{nk}$ is the *n-th* normalized state observed at frequency k, without the superscript p to simplify the notation. According to the model in (5.8), $\mathbf{r}$ can be considered a transformed sample of the random variable τ. Let us define with $\boldsymbol{\Psi}$ the TDOA parameters of the pdf of τ. The distribution of $\overline{r}_{nk}$ has a given pdf which can be still parameterized by $\boldsymbol{\Psi}$. Therefore the maximum likelihood estimator of $\mathbf{\Psi}$ is derived as:

$$MLE(\mathbf{\Psi}) = \operatorname*{argmax}_{\mathbf{\Psi}} \mathbf{p}(\mathbf{r}|\mathbf{\Psi}) \tag{5.13}$$

where $\mathbf{p}(\cdot)$ is the joint conditional probability density function of $\mathbf{r}$. If $\mathbf{r}$ is assumed to be a sample of independent random variables the MLE simplifies to:

$$MLE(\mathbf{\Psi}) \simeq \prod_{k,n} \mathrm{p}(\overline{r}_{nk}|\mathbf{\Psi}) \tag{5.14}$$

where $\mathrm{p}(\overline{r}_{nk}|\mathbf{\Psi})$ is the pdf of $\overline{r}_{nk}$ given the parameters $\mathbf{\Psi}$ and the transformation in (5.8). We consider again the approximation of the reverberation as a diffuse noise field and assume τ being a Gaussian random variable whose pdf parameters are (μ, σ). Here μ denotes the expected TDOA according to the anechoic propagation and the geometrical setup and σ is a variance introduced by the diffuse reverberation. To derive the MLE in (5.13) we need to consider the conditional pdf of $\overline{r}_{nk}$. Due to the non-monotonicity of the transformation in (5.8) the resulting pdf is modeled as a sum of single pdfs evaluated in all the possible solutions for τ_{nk} for the equation in (5.8)

$$\tau^i_{nk} = \frac{arg(\overline{r}_{nk})}{2\pi f_k} + \frac{i}{f_k}, \quad \forall i \in \mathbb{N} \tag{5.15}$$

Then (5.14) is derived as:

$$MLE(\tau) = \operatorname*{argmax}_{\tau} \prod_{k,n} \sum_{i} |(2\pi f_k)^{-1}| \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(\tau^i_{nk}-\mu)^2}{2\sigma^2}} \tag{5.16}$$

In other words the resulting pdf of $\overline{r}_{nk}$ is a wrapped Gaussian distribution similarly to the model used in [6, 89]. To compare the full MLE derivation to the SCT in (5.10), equation in (5.16) can be simplified to:

$$MLE(\tau) \simeq \operatorname*{argmax}_{\tau} \prod_{k,n} |(2\pi f_k)^{-1}| \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(\tau^{Min}_{nk}-\mu)^2}{2\sigma^2}} \tag{5.17}$$

where τ^{Min}_{nk} is one of the solution in (5.15) for which the distance $(\tau^{Min}_{nk} - \mu)^2$ is minimized. In other terms, we assume that only one Gaussian gives a not negligible contribute to the likelihood. Thus the MLE can be equivalently derived maximizing the log-likelihood:

$$\begin{aligned} MLE(\mu) &\simeq \operatorname*{argmax}_{\tau}\ \log \prod_{n,k} \mathrm{p}(\overline{r}_{nk}|\mu) \\ &\simeq \operatorname*{argmax}_{\mu} \left[\mathrm{const} + \sum_{k} \sum_{n} -\frac{(\tau^{Min}_{nk} - \mu)^2}{2\sigma^2} \right] \end{aligned} \tag{5.18}$$

where const is a generic constant. To directly compare the SCT with (5.18) we define $g(\cdot)$ as:

$$g(\epsilon_{nk}(\tau)) = -\frac{\epsilon_{nk}(\tau)^2}{(2\pi f_k)^2} \tag{5.19}$$

Equation (5.10) can be rewritten as:

$$SCT(\tau) \simeq \sum_k \sum_n \left[-\left(\frac{(|e^{-j2\pi f_k \tau_{nk}^{Min}} - e^{-j2\pi f_k \tau}|)^2}{(2\pi f_k)^2} \right) \right] \tag{5.20}$$

According to the geometrical interpretation, even when the microphones are spaced of some meters the expected TDOA is a small value that approaches 0 as the microphone distance decreases. Therefore, by a Taylor series expansion and after some manipulation (5.20) approximates to:

$$SCT(\tau) \simeq \sum_k \sum_n \left[-(\tau_{nk}^{Min} - \tau)^2 \right] \tag{5.21}$$

Therefore, regardless of σ, the value of τ which maximizes (5.21) is equivalent to the approximated MLE of μ in (5.18). Note, if there is no spatial aliasing, there is a single solution for the function in (5.15) corresponding to $i = 0$. In this case the only approximation is that from (5.20) to (5.21) and the SCT is almost equivalent to the true log-likelihood in (5.16), up to a scaling factor.

5.3.2 Equivalence Between SCT and Kernel pdf Estimation

In the above analysis we discusses the meaning of the SCT if the estimated states represent the acoustic propagation of one source in diffuse reverberation. Further on, in a more general case, the SCT can be interpreted as an approximated non-parametric kernel pdf estimation of τ. Assuming to have an i.i.d sample $[\tau_{nk}]_{n,k}$ of a random variable τ, the kernel approximation of its probability density function is:

$$f(\tau) = \frac{1}{nkh} \sum_k \sum_n \Upsilon\left(\tau - \tau_{nk}, h\right) \tag{5.22}$$

where Υ is a kernel function and h is a parameter that controls the bandwidth of the kernel. However, similarly to the MLE analysis, we do not observe τ_{nk} but $\overline{r}_{nk}$ which is a non-monotonic function of τ_{nk}. Therefore, we define the kernel of transformed variable as:

$$f(\tau) = \frac{1}{nkhi} \sum_k \sum_n \sum_i \Upsilon\left(\tau - \tau_{nk}^i, h\right) \tag{5.23}$$

where τ_{nk}^{i} is the *i-th* possible solution for τ_{nk} given $\overline{r}_{nk}$ according to the inverse function (5.15). As for the MLE analysis we assume that with a proper selection of the bandwidth h only the kernel of one solution τ_{nk}^{Min} gives a not negligible contribute in the summation. Hence (5.23) approximates to:

$$f(\tau) \simeq \frac{1}{nkh} \sum_{k} \sum_{n} \Upsilon \left(\tau - \tau_{nk}^{Min}, h \right) \tag{5.24}$$

where τ_{nk}^{Min} is the closest solution to τ. Note, this approximation is equivalent to that in the MLE derivation if we assume a Gaussian kernel which is defined as:

$$\Upsilon(\tau - \tau_{nk}^{i}, h) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\tau - \tau_{nk}^{i})^2}{2h^2}} \tag{5.25}$$

Being the observed and expected time-delay values very close to 0, we can perform the same approximation between (5.20) and (5.21). Then (5.10) can be approximated as:

$$SCT(\tau) \simeq \sum_{k} \sum_{n} \left[g \left(\frac{|\tau - \tau_{nk}^{Min}|}{\frac{1}{2\pi f_k}} \right) \right] \tag{5.26}$$

Thus, we define the function $g(\cdot)$ as dependent on k:

$$g(\epsilon_{nk}(\tau)) = \Upsilon \left(\frac{\epsilon_{nk}(\tau)}{2\pi f_k}, h \right) \tag{5.27}$$

where Υ is a standard kernel function of bandwidth h. Hence, (5.26) can be rewritten as

$$SCT(\tau) \simeq \sum_{k} \sum_{n} \Upsilon \left(\tau - \tau_{nk}^{Min}, h \right) \tag{5.28}$$

Therefore, if the function g is a locally confined kernel computed on the frequency normalized error $\frac{\epsilon_{nk}(\tau)}{2\pi f_k}$ the SCT is equivalent to an approximated kernel density estimator of the time-delay τ.

We want stress out that, due to the uncertainty introduced by the phase aliasing, the SCT envelope is only comparable to the approximated kernel pdf in (5.24) and then cannot give a precise description of the underlying true pdf for any value of τ. However, we will show in the next section that it gives a good approximation in the corresponding modal values of the distributions which, under certain conditions, indicates the most likely location of the sources.

5.3.3 Definition of $g(\cdot)$ and Adaptive Estimation of α

An effective function for $g(\cdot)$ proposed in [61] is:

$$g(\epsilon_{nk}(\tau)) = 1 - tanh\left(\alpha \cdot \frac{\epsilon_{nk}(\tau)}{2}\right) \tag{5.29}$$

with α denoting a factor that controls the degree of spatial state rejection. In our previous works [61, 63] the parameter α, which directly modifies the resolution of the SCT envelope, was manually optimized. Alternatively, a practical empirical method was proposed in [60]:

$$\alpha = \frac{v \times 10}{f_s \cdot d_{max}} \tag{5.30}$$

where v is the sound speed, f_s is the sampling frequency and d_{max} is the distance between microphones. Practically, we assumed that reducing the microphone spacing the diversity of the acoustic propagation described by the states r_{nk} would reduce and therefore we need to increase the resolution of the SCT. An alternative is to use a "best-fit" function (adopted in [94] for a similar purpose):

$$g(x_n) = \begin{cases} x_n, & if \quad \underset{i}{argmin}(x_i) = n \\ \\ 0, & \qquad otherwise \end{cases} \tag{5.31}$$

where $x_i = |c(f,\tau) - \overline{r}_i^p(f)|$.

In this discussion, as opposed to the previously heuristic methods, we exploit the analogy between SCT and kernel density estimation to derive an optimal formulation for $g(\cdot)$ and α. According to the equivalence between SCT and kernel density estimation the function $g(\cdot)$ should be evaluated in the normalized error $\frac{\epsilon_{nk}(\tau)}{2\pi f_k}$ rather than in $\epsilon_{nk}(\tau)$ which equivalently correspond to adapt the parameter α of the function (5.29) according to the frequency k. Therefore, from the analogy in (5.27) $g(\cdot)$ must be a kernel function of bandwidth h whose parameter can be optimized defining a proper model for the underlining time-delay distribution.

If we assume the normalized states $\overline{r}_{nk}$ to represent the acoustic propagation of multiple sources propagating over the direct path, the shape of the time-delay distribution can be approximated by a Gaussian mixture model (GMM):

$$p(\tau_{nk}|\mathbf{\Psi}) = \sum_j N(\mu_j, \sigma_j) \tag{5.32}$$

where $N(\mu_j, \sigma_j)$ is the *j-th* individual component with mean and variance μ_j and σ_j, respectively. Note, a GMM is considered for the sake of simplicity, but the analysis can be extended to other heavy tailed distributions such as a Laplacian. Following (5.32) the bandwidth of the kernel should be adapted according to the variance of each mixture. For example, if $g(\cdot)$ is a Gaussian kernel defined as in (5.25) the bandwidth should correspond to the variance of each mixture and therefore should be locally adapted for different values of τ,[32]. A simplification of the model that well reflects the acoustic propagation in a real-world scenario is to assume that (5.32) is a sparse mixture of Gaussians, which means that each Gaussian slightly overlap to each other. The sparseness is based on two main assumptions:

- even when the sources are relatively close to each other, with a sufficient microphone spacing their TDOAs (μ_j) are sufficiently diverse;
- as long as the direct path propagation is dominant, the standard deviation in the acoustic propagation is sufficiently smaller than the difference between the TDOAs of two generic sources.

A further approximation is to assume that all the Gaussian components have the same variance, equal to σ. Thus, the mixtures can be assumed to be sparse if:

$$\sigma \leq \frac{\min_{i,j} |\mu_j - \mu_i|}{2\beta} \tag{5.33}$$

where β defines the amount of admissible overlapping between each individual distribution. For example, if $\beta = 2$ the mixtures are sparse if the individual distributions overlap for less than 2.1%. Therefore, the bandwidth can be safely chosen to be equal to:

$$h = \frac{\overline{\min_{i,j} |\mu_j - \mu_i|}}{2\beta} \tag{5.34}$$

where $\overline{\min_{i,j} |\mu_j - \mu_i|}$ is an estimation of $\min_{i,j} |\mu_j - \mu_i|$. By means of the same approximation between (5.20) and (5.21) the minimum distance can be estimated as:

$$\overline{\min_{i,j} |\mu_j - \mu_i|} \simeq \frac{1}{k} \sum_k \min_{i,j} \left(\frac{|\overline{r}_{jk} - \overline{r}_{ik}|}{2\pi f_k} \right) \tag{5.35}$$

Therefore, we propose two equivalent frequency-adapted non-linear functions $g(\cdot)$ which approximates the SCT to an effective kernel density estimator of τ:

$$g_1(\epsilon_{nk}(\tau)) = 1 - tanh(\alpha_k \epsilon_{nk}(\tau)), \quad \alpha_k = \frac{1}{h4\pi f_k} \tag{5.36}$$

$$g_2(\epsilon_{nk}(\tau)) = e^{-\epsilon_{nk}(\tau)^2/(2\alpha_k^2)}, \quad \alpha_k = h2\pi f_k \tag{5.37}$$

where h is defined as in (5.34) with $\beta \geq 2$. With the above definition of $g(\cdot)$, the SCT is an approximation of the time-delay distribution up to a scaling factor. Note, the function $g_2(\cdot)$ defines a Gaussian kernel and thus is optimal if the time-delay distribution is well represented by the model in (5.32). On the other hand, the function $g_1(\cdot)$ is a better kernel if the time-delay distribution is modeled by a Laplacian Mixture Model (LMM). However, in a similar analysis, it was shown that modifying the kernels do not introduce sensibly differences in the envelopes [7] estimated in a real-world scenario.

It is worth to underline that in general the number of the Gaussian components does not necessarily correspond to the number of the true sources if the reverberation is not diffuse. In other words, without any knowledge on the room geometry, a virtual location corresponding to a strong reflection cannot be distinguished by a true location by means of the solely directional information. Nevertheless, the robustness of the TDOA estimation can be improved if the acoustic propagation is analyzed by means of multiple microphone pairs, in order to exploit a multidimensional spatial coherence [11]. Such generalization is provided in the section (5.4)

5.3.4 Connections Between SCT and GCC-PHAT

Traditionally, the instantaneous TDOA at sensor pair is estimated by evaluating a coherence measure which expresses the similarities of the signals when are shifted of a give delay. One of the most common coherence measure is the Generalized Cross Correlation (GCC) introduced by Knapp and Carter in [43]. Among several different variants of the GCC, the authors proposed the GCC with PHAse Transform (GCC-PHAT) which is also known as Crosspower Spectrum Phase analysis (CSP) and was shown to be an effective estimator of the TDOA for real-world reverberant environments [68].

Let us consider the SCT formula defined in (5.10) and assume that there is no permutation problem. Then, we consider only the states representing the propagation of a single source (e.g. $N = 1$). With the kernel function g defined as

$$g(\epsilon_{nk}(\tau)) = 1 - \frac{\epsilon_{nk}^2(\tau)}{2}, \tag{5.38}$$

it can be shown that the SCT is equivalent to the GCC-PHAT. According to such definitions the SCT in (5.10) simplifies to:

$$SCT(\tau) = \sum_{k=0}^{(N_k-1)/2} \left(1 - \frac{|c(f_k,\tau) - r_{1k}|^2}{2}\right) \tag{5.39}$$

where N_k is the number of the frequency bins according to the STFT analysis. Note, the summation is limited to half of the frequency bins because of the well-known symmetry of the DFT.

By a simple mathematical manipulation (5.39) can be written as:

$$\begin{aligned} SCT(\tau) &= \sum_{k=0}^{(N_k-1)/2} \left(1 - \frac{[c(f_k,\tau) - r_{1k}][c(f_k,\tau) - r_{1k}]^*}{2}\right) \\ &= \sum_{k=0}^{(N_k-1)/2} \left(Re[c(f_k,\tau)^* r_{1k}]\right) \end{aligned} \tag{5.40}$$

Using the signals observed with a pair of microphones, the GCC-PHAT is computed as:

$$GCC-PHAT(n) = IDFT\left(\frac{x_1(k)x_2(k)^*}{|x_1(k)x_2(k)^*|}\right) = \frac{1}{N_k}\sum_{k=0}^{N_k-1}\left(\frac{x_1(k)x_2(k)^*}{|x_1(k)x_2(k)^*|}\right) e^{j2\pi\frac{k}{N_k}n} \tag{5.41}$$

where n is the index of the discrete time sample, $x_1(k)$ and $x_2(k)$ are the discrete Fourier transforms of the signals recorded by the first and the second microphone, respectively. Under ideal conditions each microphone observes a delayed version of the original acoustic wave according to the position of the source. Hence, for each frequency the product $x_1(k)x_2(k)^*$ can be rewritten as:

$$\begin{aligned} (x_1(k)x_2(k)^*) &= |x_1|e^{-j2\pi f_k T_1}|x_2|e^{j2\pi f_k T_2} \\ &= |x_1|e^{-j2\pi f_k T_1}|x_2|e^{[j2\pi f_k(T_1+\delta\tau)]} \\ &= |x_1||x_2|e^{j2\pi f_k \delta\tau} \end{aligned} \tag{5.42}$$

where $\delta\tau$ is the relative TDOA and T_1 and T_2 are the time of arrivals (TOA) of the direct wavefront recorded by the microphones 1 and 2, respectively. Thus (5.41) can be simplified to:

$$GCC - PHAT(n) = \frac{1}{N_k}\sum_{k=0}^{N_k-1} e^{j2\pi f_k\delta\tau} e^{j2\pi\frac{k}{N_k}n} \tag{5.43}$$

To compare the above equation with (5.40) we substitute the index n with τ multiplying and dividing the argument of the second exponential by the sampling frequency f_s.

$$GCC - PHAT(\tau) = \frac{1}{N_k}\sum_{k=0}^{N_k-1} e^{j2\pi f_k\delta\tau} e^{j2\pi\frac{k\times f_s}{N_k}\tau} = \frac{1}{N_k}\sum_{k=0}^{N_k-1} e^{j2\pi f_k\delta\tau} e^{j2\pi f_k\tau} \tag{5.44}$$

Because of the symmetry of the DFT, the summation in (5.44) is equivalent to

$$GCC - PHAT(\tau) = \frac{2}{N_k}\sum_{k=0}^{(N_k-1)/2} Re[e^{j2\pi f_k\delta\tau} e^{j2\pi f_k\tau}] \tag{5.45}$$

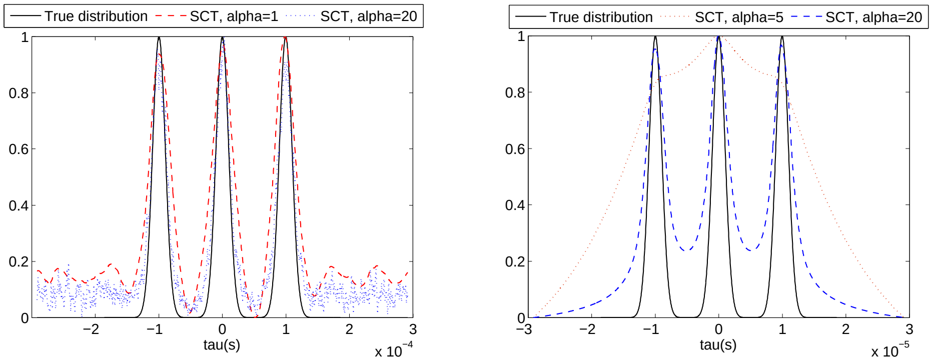
The exponential $e^{j2\pi f_k \delta\tau}$ has the same meaning of the observed state $\overline{r}_{1k}$ since it represents the sound delay from the source to the two microphones at frequency f_k. It is worth noting that PHAT normalization has a similar effect as the state normalization in (5.1). The exponential $e^{j2\pi f_k \tau}$ is exactly the conjugate of $c(f_k, \tau)$. Then, equations (5.45) and (5.40) are equivalent up to the scaling factor $\frac{2}{N_k}$ and the GCC-PHAT can be viewed as a particular case of SCT. Note that the GCC-PHAT uses the normalized cross-correlation between the input signals as state observations of the propagation models. On the other hand, in the SCT the states come from the resulting demixing matrix obtained by ICA. Therefore, the SCT does not only extend the GCC-PHAT to the case of multiple sources but also extends the statistics exploited to estimate the TDOAs, from second order to high order.

5.3.5 Numerical Evaluation of the SCT

It is possible to evaluate the effect of the proposed SCT function by simulating the states in (5.1) according to the model in (5.32). Considering a source located at the angular direction θ_j from the center of the array with microphones spaced of d_{max}, the time-delay observations are randomly generated according to a normal distribution with the following parameters:

$$\mu_j = \frac{v \times sin(\theta_j)}{d_{max}}, \quad \sigma = \frac{\min_{i,j} |\mu_j - \mu_i|}{10}, \tag{5.46}$$

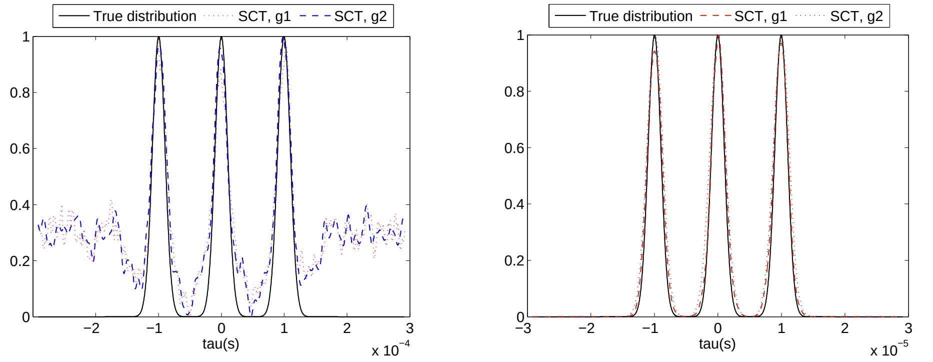
The above definition of σ comes from the assumption of a sparse mixtures. The complex-valued states for three sources located at $-20°$,$0°$ and $20°$ have been generated according to the model in (5.11) for 1000 discrete frequencies uniformly sampled between 0 and 16KHz and assuming microphone spacing d_{max}=0.1m or d_{max}=0.01m. The SCT is first evaluated with $g(\cdot)$ defined as in (5.29) manually modifying the parameter α and then applying the new frequency-adapted function defined as in (5.36). Figure 5.1 shows the resulting normalized SCT envelopes with two values of alpha. Note that a correct choice of the parameter α is essential to obtain an SCT envelope with clear peaks located at the corresponding modal values of the true distribution. If the parameter α is to high the SCT envelope is noisy and the peak selection becomes difficult (see figure 5.1a). On the other hand if α is too small the resolution of the SCT may not be sufficient to clearly distinguish the expected peaks (see figure 5.1b). As opposed to a fixed value of alpha, figure 5.2 shows that the resulting SCT estimates smooth and clear envelopes for both the cases, with either a Gaussian or Laplacian frequency-adapted kernel. Furthermore, as expected from the theoretical discussion, for the case of $d = 0.01m$ the SCT is almost equivalent to the true time-delay distribution while for the case of $d = 0.1m$ the distribution is well approximated only around the modal values because of the intrinsic uncertainty introduced by the phase wrapping.



(a) Microphone spacing of d_{max} = 0.1 m and $g(\epsilon_{nk}(\tau)) = -\epsilon_{nk}(\tau)$

(b) Microphone spacing of d_{max} = 0.01 m and $g(\epsilon_{nk}(\tau)) = -\epsilon_{nk}(\tau)$

Figure 5.1: Simulated time-delay distribution and estimated SCT with fixed α for three source located at $-20°$,$0°$ and $20°$.



(a) Microphone spacing of d_{max} = 0.1 m and $g(\epsilon_{nk}(\tau)) = -\epsilon_{nk}(\tau)$

(b) Microphone spacing of d_{max} = 0.01 m and $g(\epsilon_{nk}(\tau)) = -\epsilon_{nk}(\tau)$

Figure 5.2: Simulated TDOA distribution and estimated SCT with adapted α for three source located at $-20°$,$0°$ and $20°$.

5.4 Generalized State Coherence Transform

The SCT can be generalized introducing two main extensions:

(a) the number of sources N may be greater than the microphones;

(b) the acoustic propagation is modeled by a multivariate distribution exploiting the signals recorded by more sensor pairs.

First of all, in the original formulation of the SCT there is not an explicit dependence on the number of the sources since the distribution of τ is estimated non-parametrically. However, it is implicitly assumed that at each frequency only N sources are present otherwise the states estimated by ICA would be meaningless. In a more general case, if the number of the sources is greater than the number of the microphones, the ICA cannot in principle estimate their acoustic propagation since the separation problem is underdetermined. However, if the sources are non-white and non-stationary, their dominance in a particular time-frequency point can be considered sparse. In other words, the estimated states obtained ideally applying the ICA to different time-frequency points, are expected to belong to the most N dominant sources. Thus the SCT can be extended evaluating the coherence of the states even in multiple time-instants as:

$$SCT^{p}(\tau) = \sum_{d}\sum_{k}\sum_{n=1}^{N} g\left(\epsilon_{nkd}(\tau)^{p}\right) \tag{5.47}$$

where $\epsilon^{p}_{nkd}(\tau) = |\overline{r}^{p}_{nkd} - c(f_k, \tau)|$ is the error evaluated with the *n-th* state at frequency k and time-instant d. The states $\overline{r}^{p}_{nkd}$ can be evaluated by framing the mixtures in (3.6) in short blocks (i.e. corresponding to signal length of 0.5-1s) and applying the ICA to each block independently. To this purpose, it was shown in [62] that a frequency-recursive normalized implementation of the ICA, such as RR-ICA (chapter 4) is useful to improve the estimation of the demixing filters $\mathbf{W}(k)$ for such short signals. To complete the generalization we extend the SCT to the multidimensional case where more state observations are exploited in order to describe the acoustic propagation with respect to multiple microphone pairs. Hence, the monodimensional propagation model $c(k, \tau)$ is substituted by a multidimensional model $\mathbf{c}(k, \mathbf{T})$ defined as:

$$\mathbf{c}(k, \mathbf{T}) = \begin{bmatrix} c(f_k, \tau_1) \\ c(f_k, \tau_2) \\ . \\ . \\ c(f_k, \tau_P) \end{bmatrix} \tag{5.48}$$

where $\mathbf{T}$ is the vector of the time-delays:

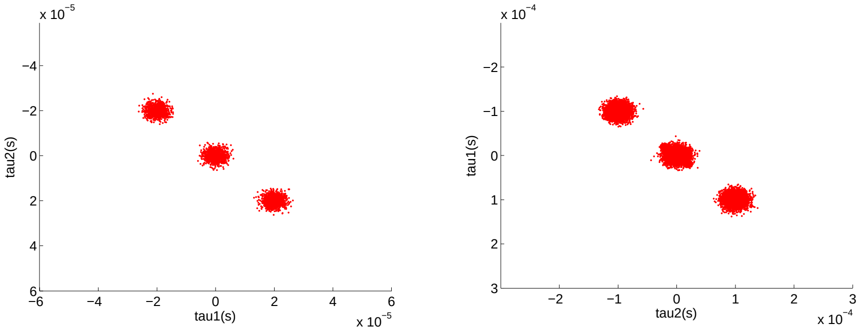
$$\mathbf{T} = [\tau_1, \tau_2, ..., \tau_P]^{T} \tag{5.49}$$

with P denoting the number of the used microphone pairs that is limited by:

$$P \leq \frac{M!}{2!(M-2)!} \tag{5.50}$$

Similarly, the state in (3.27) is substituted with a state vector

$$\overline{\mathbf{r}}_{nkd} = \begin{bmatrix} \overline{r}^{1}_{nkd} \\ . \\ . \\ \overline{r}^{P}_{nkd} \end{bmatrix} \tag{5.51}$$



(a) d=0.02cm (b) d=0.1cm

Figure 5.3: Original time-delay distributions

where each element $\overline{r}^{p}_{nkd}$ is the normalized state related to the microphone pair (a_p, b_p). Hence, the SCT can be generalized as:

$$GSCT(\mathbf{T}) = \sum_{d}\sum_{k}\sum_{n=1}^{N} \mathbf{g}\left(E_{nkd}(\mathbf{T})\right) \tag{5.52}$$

where $\mathbf{g}(\cdot)$ is a multivariate kernel function and $E_{nkd}(\mathbf{T})$ is an error measure between the vectors $\mathbf{c}(k, \mathbf{T})$ and $\overline{\mathbf{r}}_{nkd}$. The error can be computed with any metric in the vectors space. We consider here a generic L^{γ}-norm:

$$E_{nkd}(\mathbf{T}) = ||\mathbf{c}(k, \mathbf{T}) - \overline{\mathbf{r}}_{nkd}||_{\gamma} \tag{5.53}$$

In particular, if $\gamma = 2$ (e.g. Euclidean norm) $\mathbf{g}(\cdot)$ can be defined as a spherical Kernel which would make the GSCT equivalent to a kernel density estimation of a mixture of multivariate distributions. Similarly to the monodimensional case the kernel bandwidth needs to be defined according to the underlying distribution of $\mathbf{T}$ which can be modeled for sake of simplicity as a mixture of multivariate Gaussians:

$$p(\mathbf{T}|\mathbf{\Psi}) = \sum_{j} N(\boldsymbol{\mu}_j, \mathbf{\Sigma}_j) \tag{5.54}$$

where $\boldsymbol{\mu}_j$ is a vector of TDOAs representing the center of the *j-th* distribution and $\mathbf{\Sigma}_j$ is its covariance matrix. In the multivariate case the kernel is defined with a bandwidth matrix which should be locally adapted according to the covariance matrix of each individual Gaussian. A further simplification to the model can be done assuming $\mathbf{T}$ to be a sparse mixture of spherical Gaussians whose covariance matrices are defined as:

$$\mathbf{\Sigma}_j = \sigma\mathbf{I}, \quad \forall j \tag{5.55}$$

where $\mathbf{I}$ being an identity matrix. Under such hypothesis the bandwidth matrix can be modeled as a diagonal matrix $h\mathbf{I}$ where h is determined as:

$$h = \frac{\overline{\min_{i,j} ||\boldsymbol{\mu}_j - \boldsymbol{\mu}_i||}}{2\beta} \tag{5.56}$$

where β is a number defining the amount of admissible distribution overlapping, and the minimum distance between their centroids is estimated as:

$$\overline{\min_{i,j} ||\boldsymbol{\mu}_j - \boldsymbol{\mu}_i||} \simeq \frac{1}{kd} \sum_k \sum_d \min_{i,j} \left(\frac{||\mathbf{r}_{jkd} - \mathbf{r}_{ikd}||}{2\pi f_k} \right) \tag{5.57}$$

Thus, analogously to the monodimensional case, a Gaussian kernel function can be defined as

$$g_2(E_{nkd}(\mathbf{T})) = e^{-E_{nkd}(\mathbf{T})^2/(2\alpha_k^2)}, \;\; \alpha_k = h2\pi f_k \tag{5.58}$$

Although a straightforward statistical interpretation is not always available, the GSCT can be applied even when the error is defined with a different type of norm. Hence, we generalize the function in (5.36) as:

$$g_1(E_{nkd}(\mathbf{T})) = 1 - \tanh(\alpha_k E_{nkd}(\mathbf{T})), \;\; \alpha_k = \frac{1}{h\sqrt[\gamma]{P}4\pi f_k} \tag{5.59}$$

$$g_2(E_{nkd}(\mathbf{T})) = e^{-E_{nkd}(\mathbf{T})^2/(2\alpha_k^2)}, \;\; \alpha_k = h2\pi f_k \tag{5.60}$$

where h is computed as in (5.56)-(5.57) substituting $|| \cdot ||_2$ with $|| \cdot ||_\gamma$. As a particular case, if $\gamma \simeq \inf$ the above functions are equivalent to a multiplicative kernel across the dimensions P.

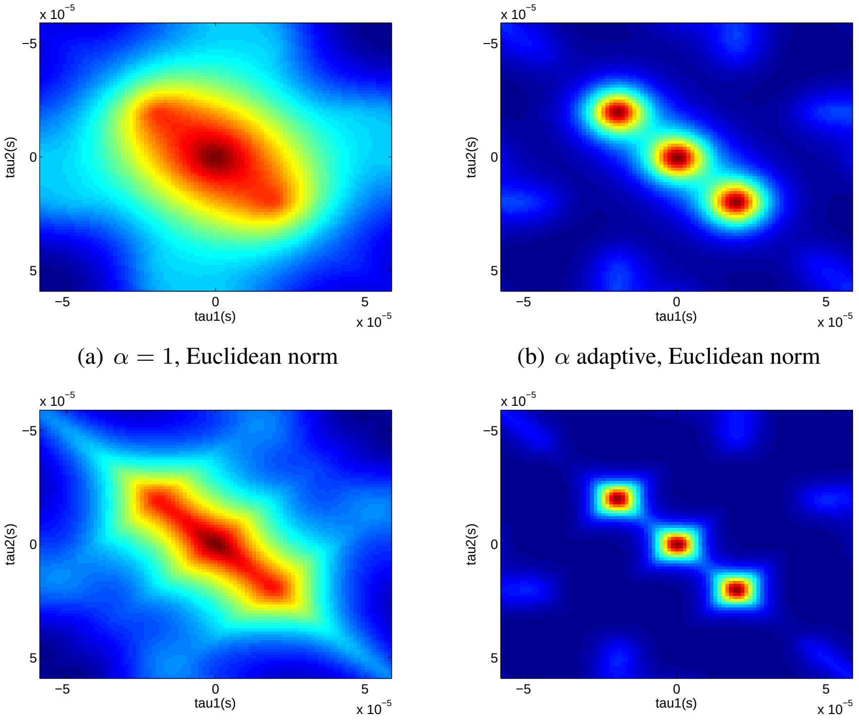
5.4.1 Numerical Evaluation

We evaluate the effect of the non-linearities in a 2-dimensional GSCT. For sake of simplicity the states in (5.51) are simulated by using a 2-D gaussian generator with a diagonal covariance matrix $\sigma\mathbf{I}$ and center $\boldsymbol{\mu}_j$ where the parameters are computed as:

$$\boldsymbol{\mu}_j = \left[\frac{v \times sin(\theta_j)}{d_{max}}, \frac{v \times sin(\phi_j)}{d_{max}} \right], \quad \sigma = \frac{\min_{i,j} ||\boldsymbol{\mu}_j - \boldsymbol{\mu}_i||}{10}, \tag{5.61}$$

assuming three sources located at $(\theta_1 = 0°; \phi_1 = 0°)$, $(\theta_2 = 0°; \phi_2 = 20°)$ and $(\theta_3 = -20°; \phi_3 = -20°)$ with $d_{max} = 0.02m$ and $d_{max} = 0.1m$. The function g_1 was adopted comparing the results for two different vector metrics: $L|| \cdot ||_2$(Euclidean norm) and $L|| \cdot ||_{\text{inf}}$(Chebyshev norm).

First of all, it should be noted that if $\alpha = 1$, which is equivalent to do not apply any non-



(a) $\alpha = 1$, Euclidean norm

(b) α adaptive, Euclidean norm

(c) $\alpha = 1$, Chebyshev distance

(d) α adaptive, Chebyshev distance

Figure 5.4: GSCT computed according to different vector metrics. Case with microphone distance $d = 0.01$

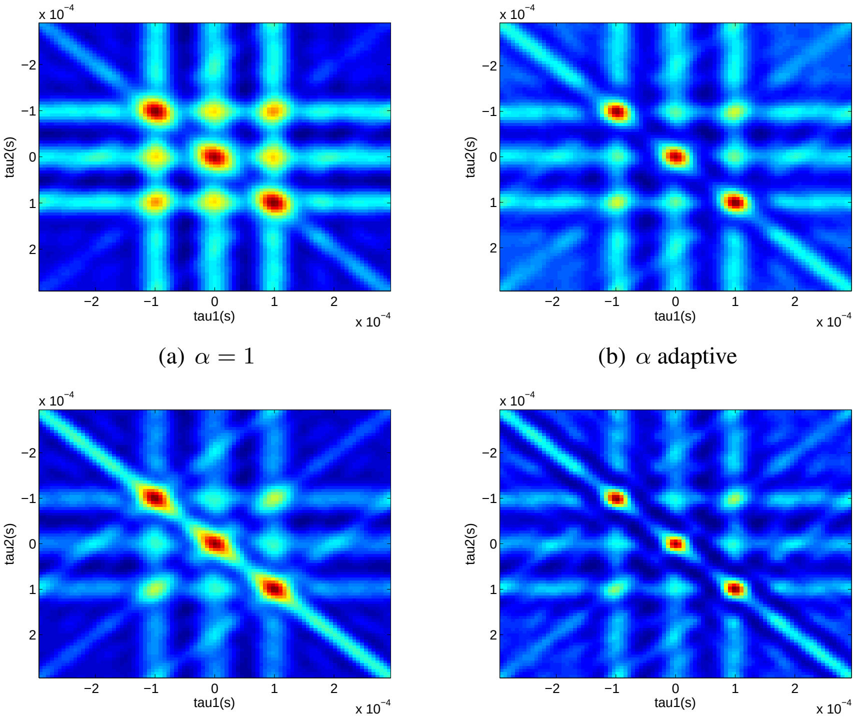
linearity, the resolution of the surface is not always sufficient to discriminate all the peaks (see figures 5.4a, 5.4c). Moreover, even when the resolution is sufficiently high, secondary local minima makes difficult the detection of the true peaks (see figure 5.5a). On the other hand with the proposed adaptive kernel, for all the cases sharp peaks arise in the corresponding center of the distributions. Further on, it can be observed that the Euclidean distance gives a smoother surface since it is less sensitive to outliers while the Chebyshev distance gives a more contrasted map since it better discriminates states belonging to different sources.

5.4.2 GSCT and Computational Issues

The GSCT is based on the direct evaluation of the state coherence in the multidimensional space. If one denotes with C the computational cost of $\mathbf{g}(E_{nkd}(\mathbf{T}))$ in (5.52) the total cost for the search of the maxima becomes:

$$\Gamma = (T \times N \times N_k \times N_d \times N_\tau)^P \tag{5.62}$$

where N is the number of the states estimated by each ICA, N_k is the number of the discrete frequencies, N_d the number of time blocks, N_τ the number of discrete time-delays according to the



(a) $\alpha = 1$

(b) α adaptive

(c) $\alpha = 1$

(d) α adaptive

Figure 5.5: GSCT computed according to different vector metrics. Case with microphone distance $d = 0.1$

desired resolution and P is the number of microphone pairs. Therefore, the computational cost is proportional to the spatial resolution and to the number of frequencies but, more important, increases exponentially with the number of dimensions P. It becomes clear that, for a general multidimensional case, the computation of the likelihood function in (5.52) becomes unfeasible as P increases, and thus an approximation is necessary to deal with a real-time implementation. A first approximation is to assume that $\mathbf{T}$ has a spherical distribution which means that the probability of a source, being located at the position $\mathbf{T} = [\tau_1, \tau_2, ...\tau_P]$, can be marginalized as:

$$p(\mathbf{T}) \simeq \prod_{p=1}^{P} p(\tau_p) \tag{5.63}$$

where $p(\tau_p)$ is the probability that in the *p-th* dimension there is a coherent source propagating along the direction τ_p. This approximation is equivalent to rewrite (5.52) as:

$$\widehat{GSCT}(\mathbf{T}) \simeq \prod_{p=1}^{P} SCT^p(\tau_p) \tag{5.64}$$

where $SCT^p(\tau_p)$ is the SCT in (5.47) evaluated at τ_p for the *p-th* microphone pair (a_p, b_p). Thus, each dimension can be treated independently in the estimation of the likelihood of the source locations. Note that with the above approximation the complexity would reduce from $O(n^P)$ to $O(n)$ but at cost of a main drawback: since the assumption in (5.63) implies that the dimensions are treated independently, maxima would arise in the $\widehat{GSCT}$ even at *ghost* locations that do not correspond to true sources. However, false locations can be rejected *a posteriori* by computing the multidimensional coherence of the observed states over selected locations.
The proposed procedure can be summarized as follows:

- the SCT^p is independently estimated and N maxima $(\mu_1^p, \mu_2^p, \cdots, \mu_N^p)$ are selected by each *p-th* SCT envelope;

- N^P multidimensional locations are generated combinatorially by means of the selected locations μ_n^p;

- the GSCT is computed only in the N^P candidate locations and the N highest peaks are selected as the most likely multidimensional source locations.

It is straightforward to observe that if $N \lll N_\tau$ the computational cost of the third step can be neglected and the computational complexity is reduced from $O(n^P)$ to $O(n)$.

5.5 Solving Real-World problems with the SCT/GSCT

The usefulness of the GSCT can be identified in three main fields:

- spatial localization of multiple sources;

- solution to the permutation problem of frequency-domain BSS;

- underdetermined source separation by time-frequency masking.

5.5.1 Spatial Localization of Multiple Sources;

In this subsection we show the experimental results for two localization tasks:

- DOA localization of 7 sources by a two channel array by an on-line SCT estimation

- 2-D localization of 7 sources by a ULA (Uniform Linear Array) array of three microphones

On-Line Two-Channel DOA Estimation

We consider the case of two microphones which record 7 sources whose activity overlaps in time. The implementation of the monodimensional SCT algorithm (i.e. one state observation for each source) has been coded both in matlab and C++ and works in real-time on a normal laptop. To perform an on-line TDOA estimation rather than applying directly the formula (5.47), we evaluate the SCT for each single block d as in (5.10) and we recursively average the envelopes over the time as follows:

$$\overline{SCT}_d(\tau) = \frac{1}{d} SCT_d(\tau) + \frac{(d-1)}{d} \overline{SCT}_{d-1}(\tau) \tag{5.65}$$

where SCT_d is the SCT is evaluated for the *d-th* time-block. Note, here we remove the superscript p since for the two-channel case only one state r_{nk} is available. We summarize in pseudo-code the main steps of the implemented algorithm:

```
apply the Short-Time Fourier Transform to the recorded signals
subdivide each time-frequency series in b_max blocks
for b=1 to d_max
   for k=maximum_frequency to 1
      compute the matrix W(k) by the RR-ICA (4) for the d-th block
      compute the normalized ratios $\overline{r}_{ik}$ as in (3.27) and (5.1)
   end
   compute the instantaneous SCT as in (5.10)
   estimate the cumulative SCT as in (5.65)
   extract the TDOAs corresponding to the peaks of the cumulative SCT envelope
end
```

In this experiment the SCT has been evaluated for the estimation of the TDOAs of 7 loudspeakers playing simultaneously sound files of about 10 seconds: 3 male utterances, 3 female utterances and 1 pop song. The loudspeakers have been spaced with an average angular distance of about $13°$ and located at an average distance of about 1 meter from the center of the two microphones. Recordings were performed in a room with $T_{60} = 700ms$ with a sampling rate of $fs = 16kHz$ and the FFT analysis was performed with an Hanning window of 2048 samples and a frame-shifting of 512 samples. The length of the time-block used for the ICA and the SCT analysis was $300ms$. The signals were recorded with two microphones spaced of $0.26m$; according to the sound speed and to the maximum admissible time-delay, the SCT was computed for 180 values of τ in the range from -13 to +13 samples.
The SCT has been compared with two other GCC-PHAT based approach:

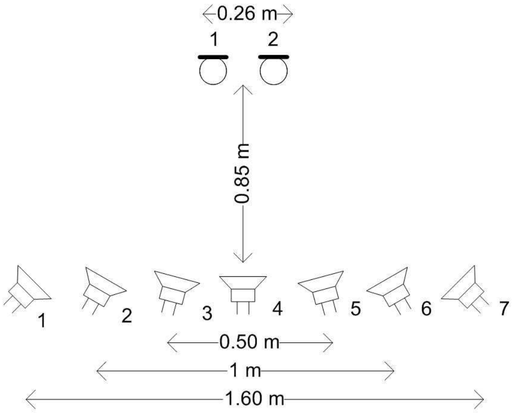


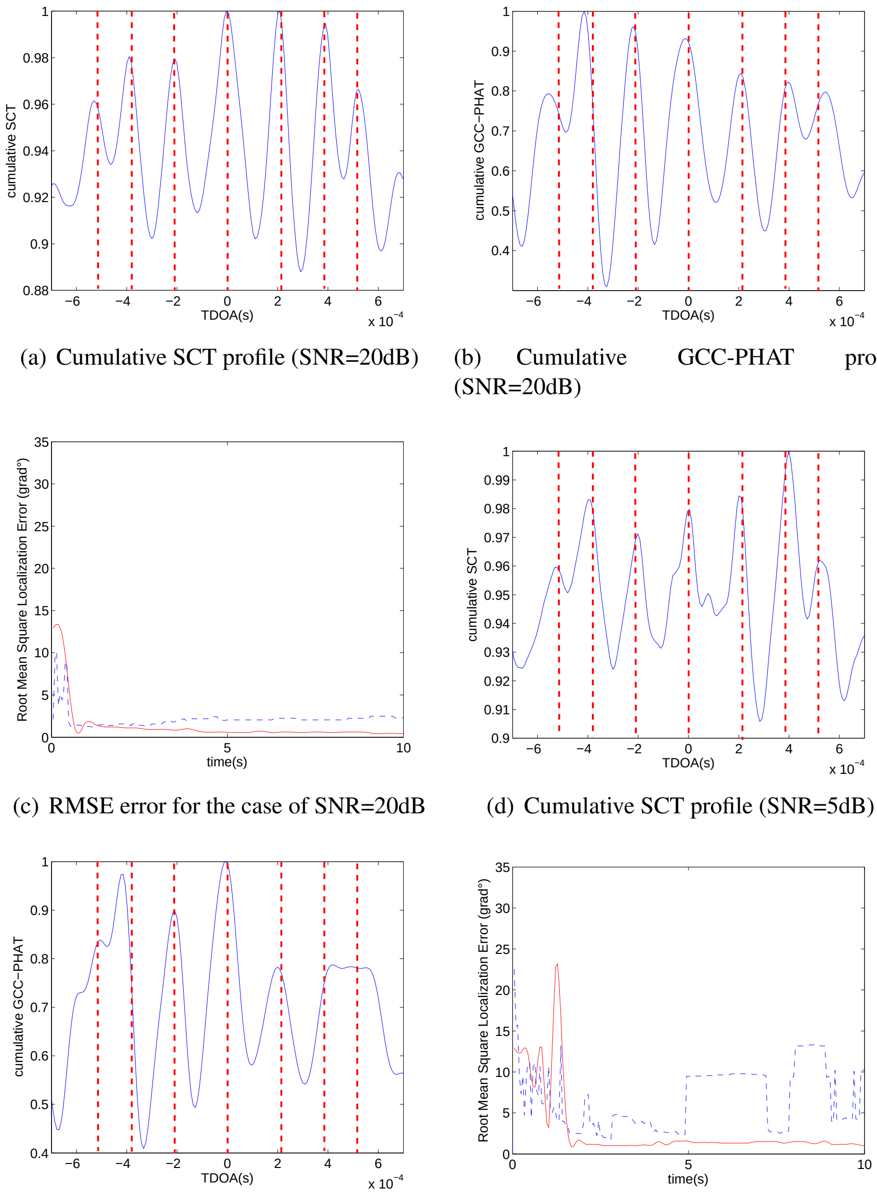
Figure 5.6: Experimental setup for the case of 7 sources.

- TDOAs selected by a cumulative GCC-PHAT
- TDOAs selected by Time-frequency histogram (TFH) [1]

For both the methods the GCC-PHAT was computed over frames of 4096 points with step of 256 points and an interpolation was applied to get the same TDOA resolution as for the SCT. The cumulative GCC-PHAT was obtained by a recursive averaging of the envelopes similarly to formula (5.65).

Figure 5.7(subpictures (a)(b)(d)(e)) shows the final envelopes associated with the cumulative SCT and with the cumulative GCC-PHAT, when the signals are affected by an Additive White Gaussian Noise (AWGN) resulting in a SNR of $20dB$ and $5dB$. For the case of lower SNR both the envelopes show clear peaks at values close to the corresponding expected TDOAs (red dotted lines). However as the noise increases the cumulative SCT clearly outperforms the GCC-PHAT. We remind that the SCT uses the demixing matrix obtained by the ICA stage and it is consequently less sensitive to the noise, while the GCC-PHAT exploits instantaneous observations of the inter-channel cross correlation. In subpictures (c) and (f) of figure 5.7 the Root Mean Square localization Error is plotted. The corresponding directions of arrival were computed according to the geometrical information and using the TDOAs estimated at each time block. For both the cases the SCT converges to a small error and just in few seconds.

In figure 5.9 we show the resulting envelopes obtained with the cumulative GCC-PHAT and the SCT, when 4 sources are recorded by two microphones spaced of $0.05m$. Since the phase difference between the propagation models of different sources is reduced, the resolution of the resulting cumulative GCC-PHAT is too low to enable the discrimination of peaks related to different sources. The advantages of the non-linear mapping of the SCT, in increasing the



(a) Cumulative SCT profile (SNR=20dB)

(b) Cumulative GCC-PHAT profile (SNR=20dB)

(c) RMSE error for the case of SNR=20dB

(d) Cumulative SCT profile (SNR=5dB)

(e) Cumulative GCC-PHAT profile (SNR=5dB)

(f) RMSE error for the case of SNR=5dB

Figure 5.7: Comparison between the cSCT and the cumulative GCC-PHAT. The red dotted lines are the true expected TDOAs. (c) and (f) show the RMSE localization error for the cumulative GCC-PHAT (blue dotted line) and the cSCT (solid red line)

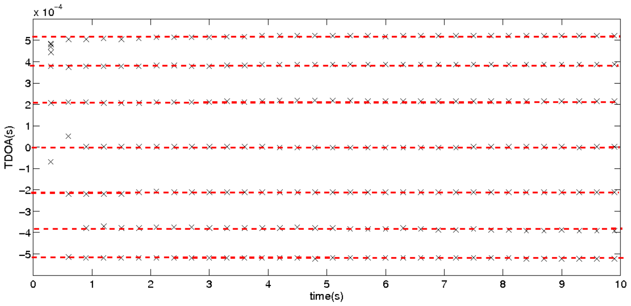
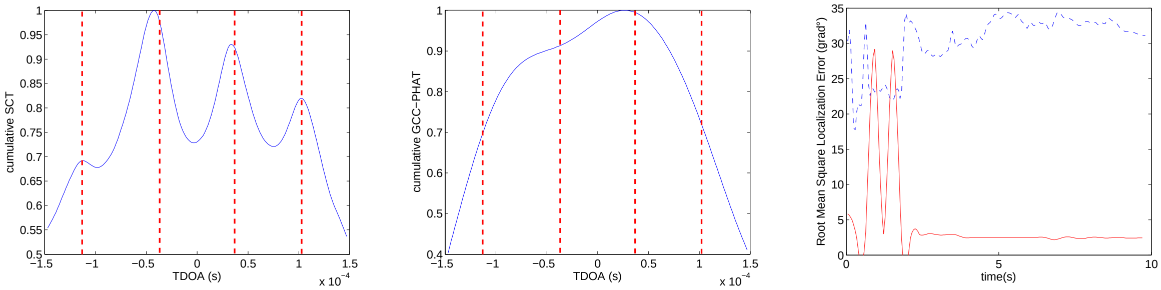


Figure 5.8: Selected TDOAs from the estimated cumulative SCT (the red dotted lines are the true expected TDOAs).



(a) Cumulative SCT profile (b) Cumulative GCC-PHAT profile (c) RMSE error

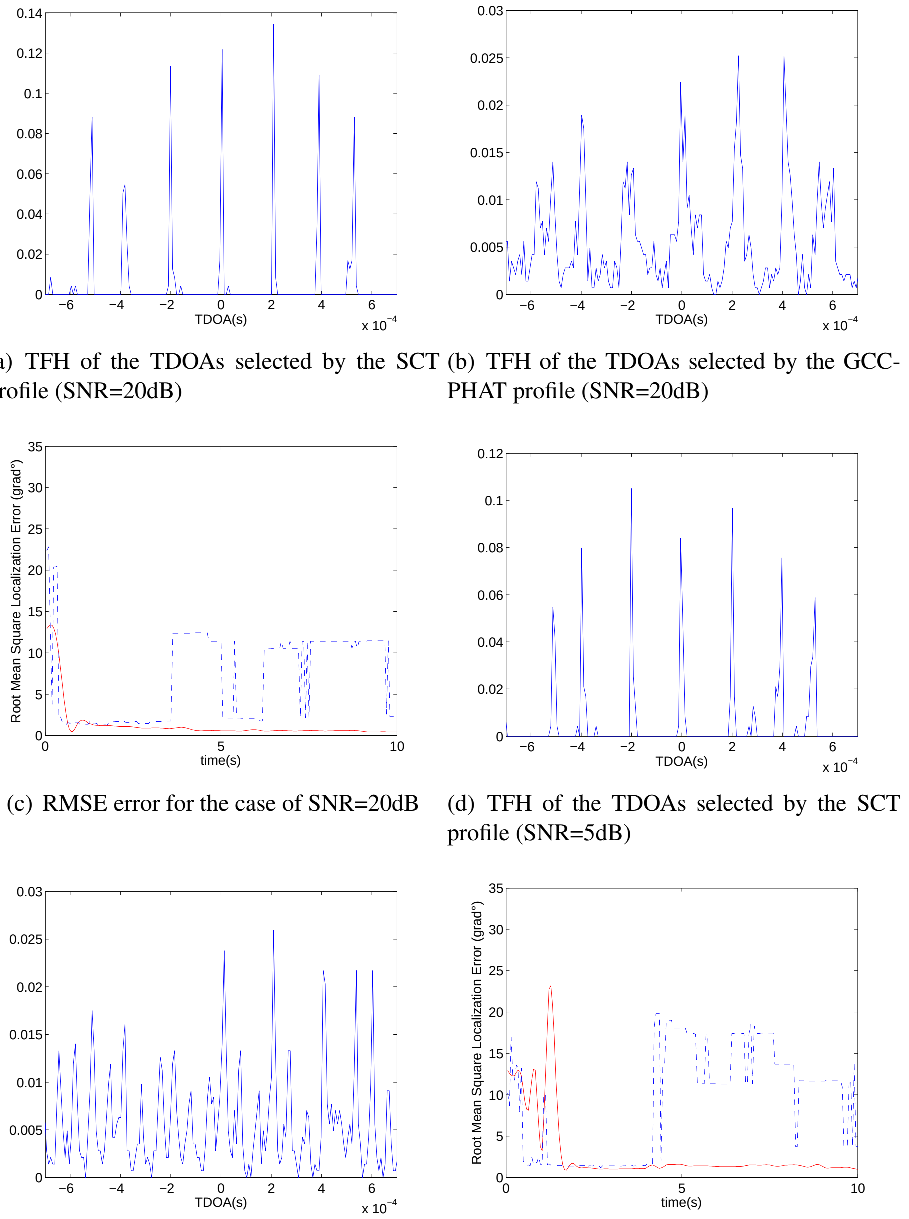
Figure 5.9: Comparison between the SCT and the cumulative GCC-PHAT for 4 sources recorded with microphones spaced of $0.05m$. In the envelopes the red dotted lines are the true expected TDOAs. In (c) is plotted the RMSE localization error for the cumulative GCC-PHAT (blue dotted line) and the SCT (solid red line)

envelope resolution, becomes clear in this case.

To complete the evaluation we provide a comparison between the SCT and a GCC-PHAT based Time-Frequency Histogram (TFH) [1]. For the TFH all the TDOAs chosen from the GCC-PHAT of each frame are pooled in a histogram where the maxima are expected to correspond to the most probable TDOAs. Additionally, we similarly computed the histogram of the TDOAs obtained at each time-block, using the peaks of the SCT. We observe from figure 5.10 that when the GCC-PHAT is used, the histogram is very noisy and do not allow a reliable estimation of the TDOA. On the other hand, even for low SNR the SCT gives much more stable results.

2-D Localization of Multiple Sources

With appropriate array geometry, the GSCT can be used to directly estimate multidimensional locations of multiple sources. Spatial locations can be mapped in the TDOA space according to the array geometry and the localization is performed by seeking maxima directly in the



(a) TFH of the TDOAs selected by the SCT profile (SNR=20dB)

(b) TFH of the TDOAs selected by the GCC-PHAT profile (SNR=20dB)

(c) RMSE error for the case of SNR=20dB

(d) TFH of the TDOAs selected by the SCT profile (SNR=5dB)

(e) TFH of the TDOAs selected by the GCC-PHAT profile (SNR=5dB)

(f) RMSE error for the case of SNR=5dB

Figure 5.10: Comparison between the cumulative SCT and the cumulative GCC-PHAT. In the envelopes the red dotted lines are the true expected TDOAs. In (c) and (f) is plotted the RMSE localization error for the cumulative GCC-PHAT (blue dotted line) and the cumulative SCT (solid red line)

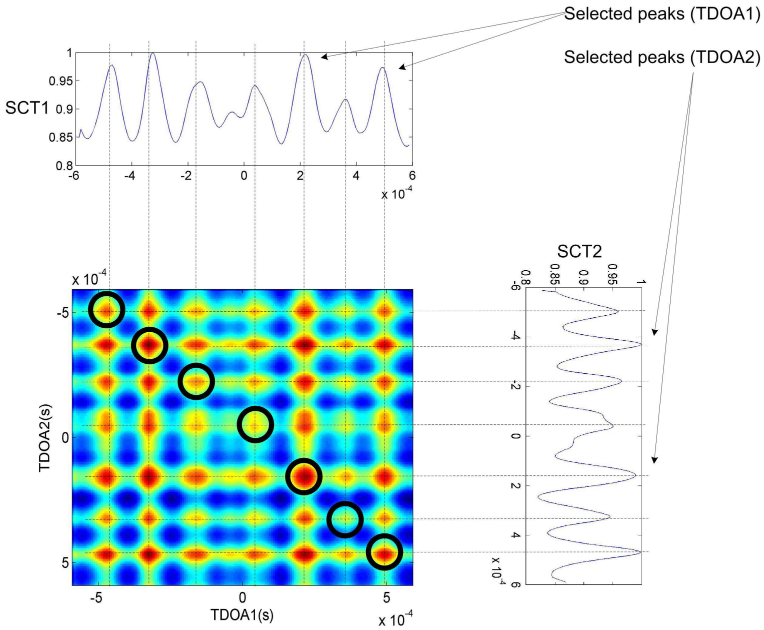


Figure 5.11: Approximated GSCT and cumulative SCT over the dimension τ_1 and τ_2

Cartesian coordinates:

$$\mathbf{T} = map(x, y, z) \tag{5.66}$$

$$\underset{x,y,z}{\operatorname{argmax}} GSCT(\mathbf{T}) \tag{5.67}$$

The spatial resolution of the resulting map would not only depend on the effectiveness of the GSCT but even on the array geometry which defines the geometrical transformation $map(x, y, z)$. Therefore, rather than computing the source location likelihood in the spatial coordinate we believe it is more effective to demonstrate here the capability of the GSCT by showing the source location likelihood in the time-delay coordinates $\mathbf{T}$.

Seven independent sources were located in a room with $T_{60} = 300\,ms$ and recorded for 6 seconds with an array of three microphones according to the configuration shown in figure 5.12. The sources were symmetrically located with respect to the microphones, in order to simplify the graphical interpretation of the results. The impulse responses were simulated by using the Lehmann's image method [46]. A sampling frequency of $f_s = 16kHz$ was used and the time-domain signals were transformed in time-frequency domain by means of STFT analysis using Hanning windows of 2048 points overlapped of 75%. The signals were partitioned in non-overlapping blocks of $1s$ and the demixing filters $\mathbf{W}(k)$, used to determine the states, were computed applying the RR-ICA (chapter 4) to each block.

Figure 5.11 shows the approximated GSCT computed by means of the projection of each monodimensional SCT across τ_1 and τ_2: the highest seven SCT peaks chosen in each dimension are projected in the map (dotted lines). According to the approximated GSCT the most likely

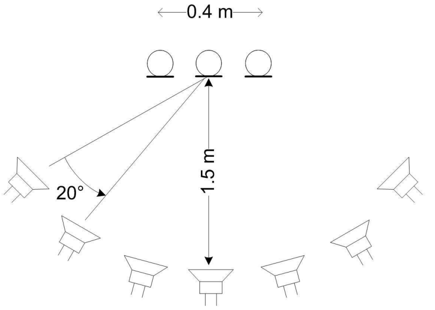


Figure 5.12: Configuration of the simulated setup ($T_{60} = 300ms$)

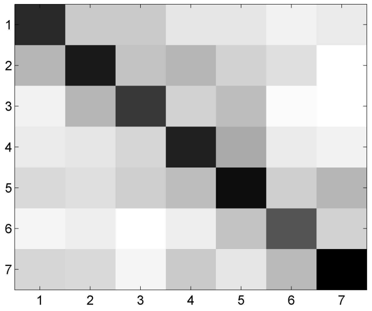
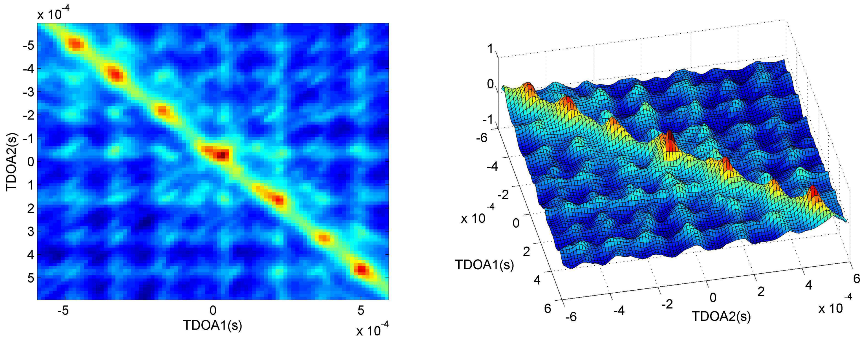


Figure 5.13: GSCT map computed over the 7 × 7 TDOA pairs

source locations are all the combinations between the selected peaks in each dimension, which are represented in figure 5.11 by the intersection points between the dotted lines. The bold circles represent the expected source locations according to the geometry of the problem. In each dimension the TDOA selected by means of the highest peaks were sorted in ascending order (i.e. index 1 corresponds to a negative TDOA). Figure 5.13 shows the map of GSCT values computed for each 2-D TDOA pair. The darkest points, which correspond to highest GSCT values, are clearly located on the main diagonal. In other terms, only the locations corresponding to the bold circles in figure 5.11 are coherent in the 2-D space. Figures 5.14(a),(b) show the true multidimensional GSCT map, which confirm that the true locations are that over the main diagonal. Thus the algorithm was able to identify all the true sources, rejecting the false locations by a *posteriori* evaluation of the 2-D coherence. Furthermore, by evaluating the GSCT only in the corresponding most-likely TDOA pairs, the problem was solved with a drastically reduced computational complexity.

5.5.2 Reduction of Permutations

For the determinated case, that is $N = M$, the estimated TDOAs can be used to reduce the permutation problem of frequency domain ICA without knowledge of the microphone array



(a) 2-D view (b) 3-D view

Figure 5.14: GSCT map

geometry. In this case we do not need to estimate the acoustic propagation at different time-blocks and we can remove the index d from the state vector in (5.51). An extensive evaluation of the performance of the SCT was presented in [47] based on a joint work with Benedikt Loesch (currently a phD student at university of Stuttgart). We report here the main results obtained in this evaluation.

Let $\Pi_k(\cdot)$ be a permutation function for frequency bin k which defines the mapping between the indices of the true sources and indices of the demixed sources. $\Pi_k(\cdot)$ is another but equivalent notation of the permutation matrix $\mathbf{\Pi}(k)$. If e.g. $\mathbf{\Pi}(k) = \mathbf{I}$, then $\Pi_k(n) = n, \ \forall n$. Given the estimated state vectors $\overline{\mathbf{r}}_{nk}$ and the TDOA estimates $\overline{\tau}_n$, we determine the permutation $\Pi_k(\cdot)$ at frequency bin k by the optimization

$$\hat{\mathbf{\Pi}}_k = \underset{\mathbf{\Pi}_k}{\operatorname{argmin}} \sum_{n=1}^{N} D[\mathbf{c}(k, \overline{\tau}_n), \overline{\mathbf{r}}_{\Pi_k(n)k}]. \tag{5.68}$$

The optimization in (5.68) aims to find the best match between the estimated state vectors $\overline{\mathbf{r}}_{nk}$ and the ideal state vectors $\mathbf{c}(k, \overline{\tau}_n)$ by finding the permutation which minimizes the sum of the distance metrics $D[\mathbf{c}(k, \overline{\tau}_n), \overline{\mathbf{r}}_{\Pi_k(n)k}]$. The number of permutation errors P_k after permutation correction in each frequency bin k can be calculated by counting the number of zero elements on the diagonal of $\mathbf{\Pi}(k)\hat{\mathbf{\Pi}}(k)^{-1}$, where $\mathbf{\Pi}(k)$ is the optimal permutation matrix at frequency bin k. The matrix $\mathbf{\Pi}(k)$ is that which maximizes the correlation between the individual source components $[\mathbf{H}_{mn}(k)\mathbf{s}_n(k, l)]_{mn}$ and the images at each microphone of the estimated source signals $x_m^{s_n}(k, l)$ estimated as in (3.18). The percentage of permutation errors is then

$$P = \frac{1}{K \cdot N} \sum_{k=1}^{K} P_k. \tag{5.69}$$

The permutation error P is expected to assume a low value if the estimated states $\overline{\mathbf{r}}_{nk}$ for different sources are sufficiently separated, which means that there is enough spatial diversity between the propagation of each source. The effectiveness of (5.68) does not only depend on the acoustic conditions but also on the locations of sources and microphones which intrinsically modify such a spatial diversity. If for example the reverberation is high or the DRR is low the state vector estimates $\overline{\mathbf{r}}_n(k)$ will have a large variance across the frequency and the optimization in (5.68) would become less effective. In fact, in such conditions, there could be some frequencies for which $D[\mathbf{c}(k, \overline{\tau}_n), \overline{\mathbf{r}}_{\Pi_k(n)k}] < D[\mathbf{c}(k, \overline{\tau}_n), \overline{\mathbf{r}}_{\Pi_k(m)k}]$ for any $n \neq m$. It means in practice that the resulting direction of the acoustic propagation, for those frequencies, is quite different from that of the direct path wavefront. In other words, the rule in (5.68) is ideally optimal only in anechoic conditions and is expected to worsen as the DRR reduces.

Evaluation method

To evaluate the robustness of the proposed approach, we have conducted extensive simulations with an L-shaped array with different microphone spacings d and three sources shown in Fig. 5.15. We fix one source at a direction-of-arrival (DOA) of $\theta_3 = 0°$ and vary the two other DOAs in the range of $[0°, 360°]$ in $10°$ steps. For each scenario, we simulate the room impulse responses assuming omni-directional sources and sensors using the ISM RoomSim toolbox [46] with a sampling frequency of f_s = 16kHz and a room size of $6.0\,\mathrm{m} \times 6.0\,\mathrm{m} \times 2.5\,\mathrm{m}$. We then generate mixtures by convolving three speech signals of length $5\,\mathrm{s}$ with the simulated impulse responses. The first three experiments consider the noiseless case, while the fourth experiment uses SNRs of $(0, 10, 20)\,\mathrm{dB}$. In the following evaluations, we use the states of the microphone pairs $(1, 2)$ and $(1, 3)$. Please note that, although we use the DOA for the source positions, the permutation correction is performed without any knowledge of the array geometry by evaluating the SCT as a function of the TDOAs.

We consider two evaluation criteria:

- The percentage P of remaining permutation errors
- The distribution of the permutation errors P_k over frequency, but averaged for all considered source positions

We plot P as a function of the two DOAs θ_1, θ_2. This plot is called a (remaining) permutation map.

Experiment1: Comparison Between Predicted and Estimated Permutation Errors

It is interesting to define a model to predict the permutation errors according to the reverberation and assuming the reverberation as a diffuse noise field. At each frequency, the transfer function between each source and each microphone depends on the wavefront time-delay which can be

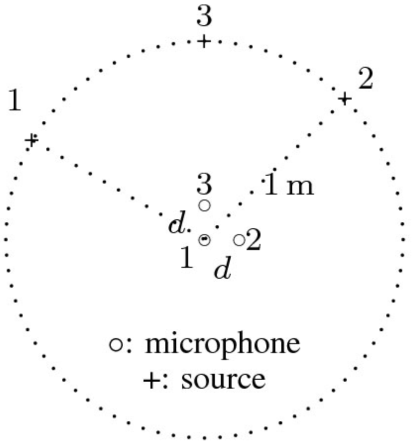


Figure 5.15: Experimental setup

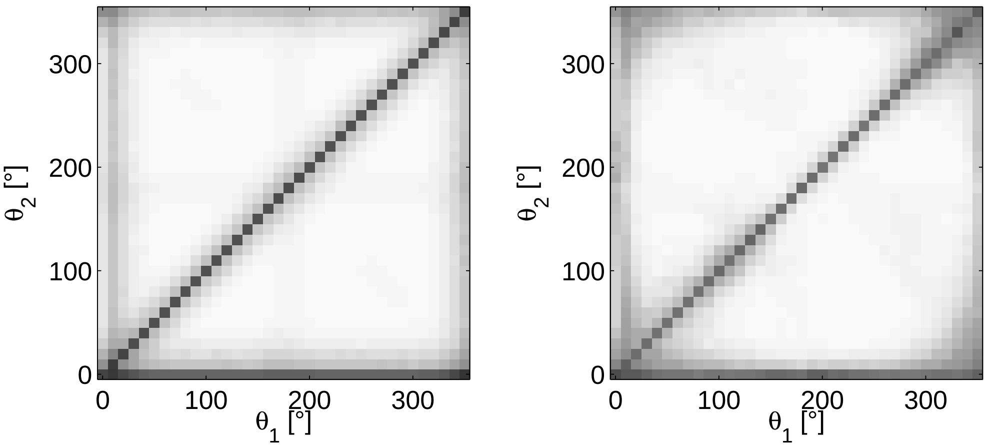
modeled with a random variable of given distribution. Consequently also the observed TDOA of each source and microphone pair can be modeled with a random variable. To simplify the analysis we model the estimated TDOA vectors $\overline{\tau}_n$ as Gaussian random vectors and we simulate the estimated states as follows:

$$\overline{\mathbf{r}}_{\Pi_k(n)k} = \mathbf{c}(k, \hat{\tau}_n) \tag{5.70}$$

where $\hat{\tau}_n \sim \mathcal{N}(\tau_n, \mathbf{C})$, τ_n is the vector of true TDOAs of source n and $\mathbf{C}$ is the covariance matrix of the noise. We can then estimate the expected value of P by Monte-Carlo simulation as follows:

- For each τ_n, generate Q realizations $\overline{\tau}_n$ according to the Gaussian distribution.
- Calculate Q corresponding state vectors according to (5.70).
- Average the percentage of permutation errors P for all Q realizations using $\mathbf{c}(k, \tau_n)$ instead of $\mathbf{c}(k, \overline{\tau}_n)$ in (5.68).

If we assume the measurements from different microphone pairs to be independent, the covariance matrix $\mathbf{C}$ is diagonal. The diagonal elements c_{nn} represents the variance in the time-delay propagation of each source which is mainly related to the DRR. The DRR depends on many parameters which determine the characteristic of the acoustic propagation. Since we assumed to be in a diffuse reverberation field it is reasonable to take on the DRR being dependent on the source-to-microphone distances. For sake of simplicity we assume such distances to be comparable and we model the vector of TDOAs $\overline{\tau}_n$ as a Gaussian random vector with covariance matrix $\mathbf{C} = \sigma^2\mathbf{I}$. Note, there is not a model to uniquely define the variance σ^2 for a given geometrical setup since the DRR in the simulated room depends on many acoustic



(a) Model

(b) ICA

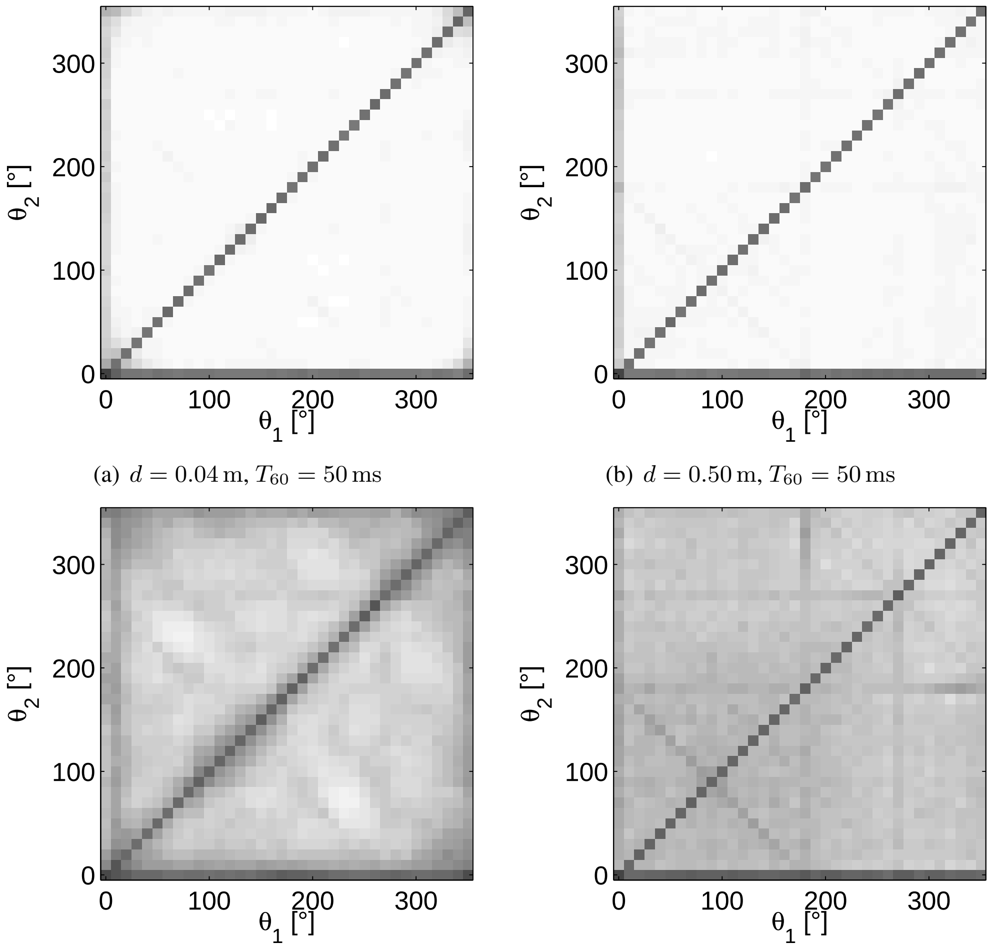
Figure 5.16: Permutation map for $T_{60} = 150\,\mathrm{ms}$, $d = 0.02\,\mathrm{m}$

parameters. Therefore, σ^2 was determined from the states obtained with the true simulated impulse responses. In Fig. 5.16, we compare the permutation map obtained using the model for the estimated states (5.70) with the permutation map obtained using ICA to estimate the states (3.27). For both cases, we use the ideal model states $\mathbf{c}(k, \overline{\tau}_n)$ for the permutation correction step (5.68). The model matches the results from ICA quite well and, as expected, when the sources approach to very close locations the error increases.

Experiments 2: Permutation reduction with true and estimated TDOA

In order to understand the intrinsic limitations of methods based on the TDOA information, regardless of the accuracy of the propagation model parameters estimated by the SCT, we first evaluate the effectiveness of the optimization rule in (5.68) with a perfect knowledge of the TDOAs. Thus, the permutation correction step (5.68) uses the ideal model states $\mathbf{c}(k, \overline{\tau}_n)$ instead of $\mathbf{c}(k, \overline{\tau}_n)$. In this experiment we evaluate the errors obtained when the data is recorded by arrays with different inter microphone distances $(0.02, 0.04, 0.1, 0.2, 0.5)\,\mathrm{m}$ and in three different reverberation conditions $T_{60} = 50\,\mathrm{ms}, 150\,\mathrm{ms}$ and $300\,\mathrm{ms}$.

Fig. 5.17 shows the permutation map for $T_{60} = 50\,\mathrm{ms}$ and $T_{60} = 300\,\mathrm{ms}$ for two different microphone spacings: $d = 0.04\,\mathrm{m}$ and $d = 0.50\,\mathrm{m}$. A quantitatively evaluation of the permutation errors is summarized in Table 5.1. The mean μ_P and the standard deviation σ_P of P over all considered source positions are computed excluding the degenerate cases (i.e. at least two sources at the same location). Furthermore histograms of P for $T_{60} = 300\,\mathrm{ms}$ are shown in Fig. 5.18. Note, in all the experiments we considered a far-field model in order to represent the location of the sources only with a single DOA. This was necessary to allow a 2D representation of the spatial source configurations. However, we want to remind that the



(a) $d = 0.04\,\mathrm{m}$, $T_{60} = 50\,\mathrm{ms}$

(b) $d = 0.50\,\mathrm{m}$, $T_{60} = 50\,\mathrm{ms}$

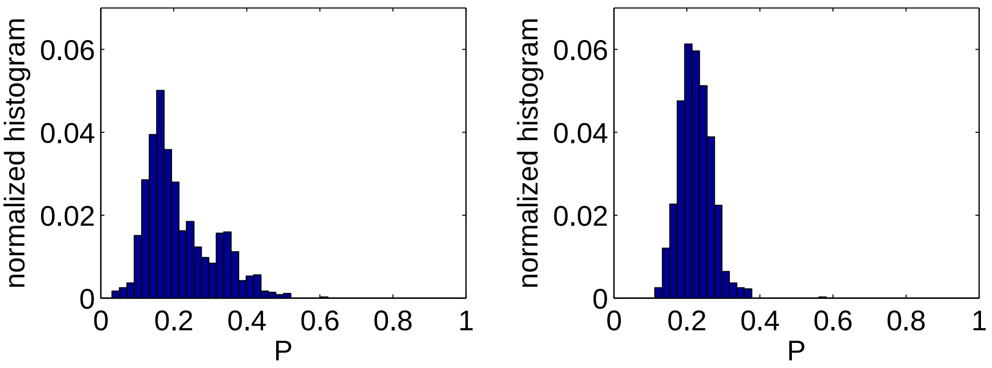
(c) $d = 0.04\,\mathrm{m}$, $T_{60} = 300\,\mathrm{ms}$

(d) $d = 0.50\,\mathrm{m}$, $T_{60} = 300\,\mathrm{ms}$

Figure 5.17: Permutation map with known TDOAs

mic spacing	$T_{60} = 50\,\mathrm{ms}$	$T_{60} = 150\,\mathrm{ms}$	$T_{60} = 300\,\mathrm{ms}$
$d = 0.02\,\mathrm{m}$	$1.62(4.79)$	$8.44(10.58)$	$20.86(13.54)$
$d = 0.04\,\mathrm{m}$	$1.31(2.78)$	$7.43(7.04)$	$21.59(9.28)$
$d = 0.10\,\mathrm{m}$	$1.16(1.83)$	$6.59(3.56)$	$20.20(4.79)$
$d = 0.20\,\mathrm{m}$	$1.32(1.68)$	$7.01(2.83)$	$21.37(3.80)$
$d = 0.50\,\mathrm{m}$	$1.43(1.69)$	$7.20(2.71)$	$22.38(4.44)$

Table 5.1: $\mu_P(\sigma_P)$ in $\%$ with known TDOAs



(a) $d = 0.04\,\mathrm{m}$

(b) $d = 0.50\,\mathrm{m}$

Figure 5.18: Histograms of P, known TDOAs, $T_{60} = 300\,\mathrm{ms}$

SCT is still able to estimate the acoustic propagation over multiple dimensions and degenerated cases can be reduced by using the 2D/3D source locations, using arrays of appropriate geometry. Looking at the figure Fig. 5.17, for $T_{60} = 50\,\mathrm{ms}$ and neglecting the degenerate case (i.e. where two of the three sources arrive from exactly the same direction) the proposed approach has very little permutation errors. As the reverberation increases for all the microphone spacing, the performance gets worse even though the error is limited to an acceptable level. We can note that when the microphone spacing is larger the variance is lower, which means that the performance is more uniform over the source locations. In fact, for closely spaced microphones the spatial diversity of the sources drastically reduces with the microphone spacing and the optimization in (5.68) becomes less effective. Furthermore a large spacing d distributes the permutation errors more equally across the frequency bins and hence has lower permutation error in the low frequency region than a small spacing d. This is due to the fact that the estimated states $\bar{\mathbf{r}}_n(k)$ can be better separated at low frequencies using a large spacing d than with small spacing d. Fig. 5.19 shows the permutation errors across the frequency range $(0, 4)\,\mathrm{kHz}$ for $d = 0.02\,\mathrm{m}, 0.1\,\mathrm{m}, 0.5\,\mathrm{m}$ smoothed with a moving average filter of order 32. This corresponds to a window size of $500\,\mathrm{Hz}$.

From table 5.1, we note that increasing too much the microphone spacing (e.g. $d = 0.5\ m$), the performance may get worse since the spatial diversity of the states also depends on the phase-wrapping due to the spatial aliasing. In the simulated scenario, a distance of $d = 0.1m$

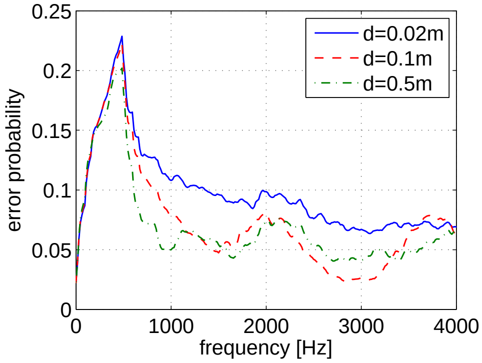


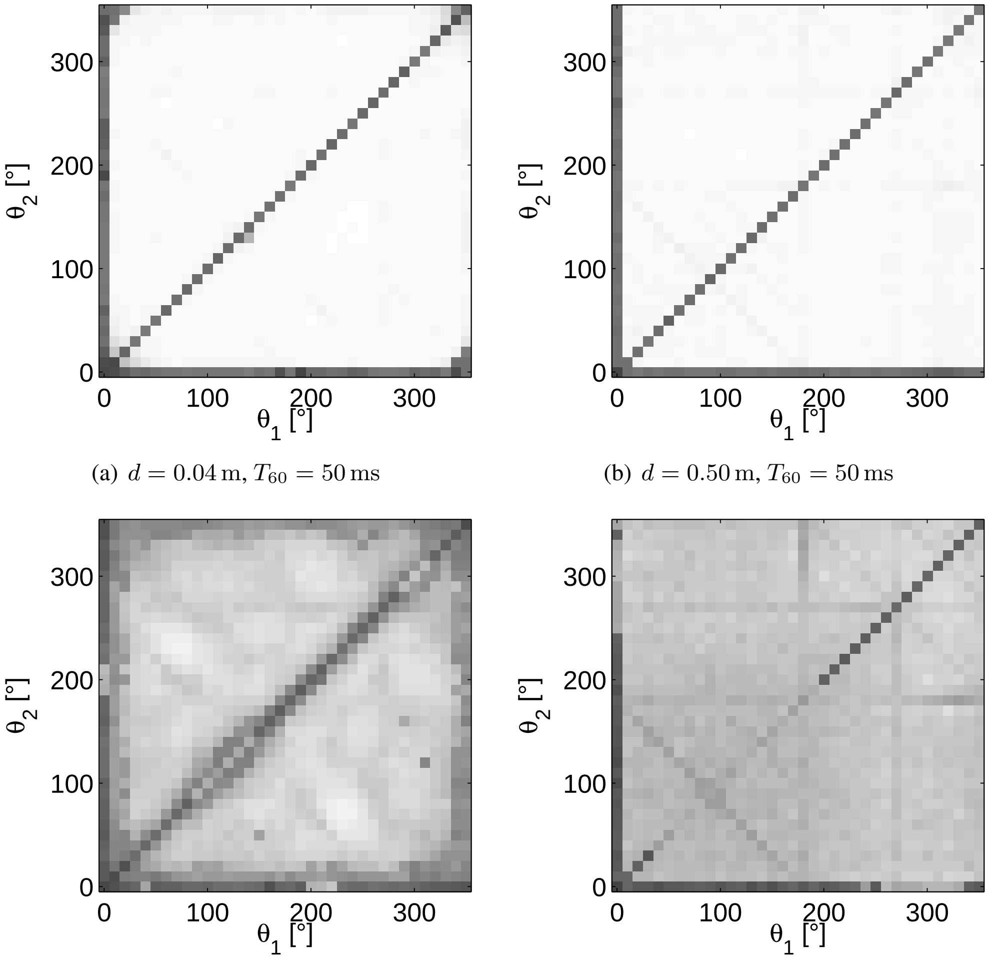
Figure 5.19: Permutation error across frequency, $T_{60} = 150\,\mathrm{ms}$

mic spacing	$T_{60} = 50\,\mathrm{ms}$	$T_{60} = 150\,\mathrm{ms}$	$T_{60} = 300\,\mathrm{ms}$
$d = 0.04\,\mathrm{m}$	$1.50(4.50)$	$9.25(11.04)$	$23.39(11.70)$
$d = 0.10\,\mathrm{m}$	$1.17(1.82)$	$6.59(3.60)$	$20.48(5.40)$
$d = 0.20\,\mathrm{m}$	$1.33(1.72)$	$6.99(2.83)$	$21.47(3.83)$
$d = 0.50\,\mathrm{m}$	$1.41(1.63)$	$7.17(2.73)$	$22.39(4.46)$

Table 5.2: $\mu_P(\sigma_P)$ in $\%$ with SCT

seems to be the best trade-off between spatial resolution and spatial-aliasing to obtain the best performance. Therefore, it may be concluded that the permutation solution based on the TDOA information is intrinsically limited by the probability of errors introduced by the optimization (5.68), which is expected to be optimal if states belonging to the propagation of different sources are disjoint in the solution space. However, due to the variance introduced by the reverberation and for the ambiguity generated by the spatial aliasing the solution space of states belonging to different sources is not linearly separable. Then, in general the phase information of the acoustic propagation can not completely solve alone the permutation problem and other techniques which use the non-stationarity of the sources must be considered as a complement (see chapter 6).

In the next experiment we compare the performance of permutation correction based on the TDOAs estimated by the SCT, i.e. permutation correction using the estimated model states $\mathbf{c}(k, \overline{\tau}_n)$, and based on the perfectly known TDOAs, i.e. permutation correction using the ideal model states $\mathbf{c}(k, \tau_n)$. In both cases, we use ICA to obtain the estimated states $\overline{\mathbf{r}}_n(k)$. Clearly, the SCT is able to estimate the TDOAs very precisely. Hence, permutation correction using the estimated DOA works almost as well as with the perfectly known TDOAs. This is reflected in Figs. 5.20 and 5.21 which show the permutation error map and the histograms of permutation error P using SCT. Comparing μ_P and σ_P for the SCT (Table 5.2) with the results for perfect knowledge of the TDOAs (Table 5.1) we note that the results match quite well.



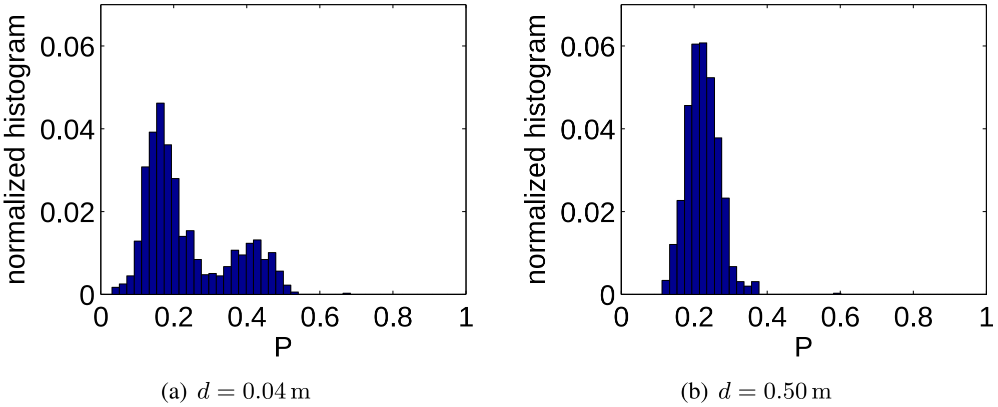
(a) $d = 0.04\,\mathrm{m}$, $T_{60} = 50\,\mathrm{ms}$

(b) $d = 0.50\,\mathrm{m}$, $T_{60} = 50\,\mathrm{ms}$

(c) $d = 0.04\,\mathrm{m}$, $T_{60} = 300\,\mathrm{ms}$

(d) $d = 0.50\,\mathrm{m}$, $T_{60} = 300\,\mathrm{ms}$

Figure 5.20: Permutation map for SCT



(a) $d = 0.04\,\mathrm{m}$

(b) $d = 0.50\,\mathrm{m}$

Figure 5.21: Histogram of P with SCT, $T_{60} = 300\,\mathrm{ms}$

SNR	known DOA	SCT
20 dB	$12.19(4.52)$	$12.27(4.98)$
10 dB	$24.79(5.01)$	$27.84(8.12)$
0 dB	$39.19(4.00)$	$51.65(6.60)$

Table 5.3: $\mu_P(\sigma_P)$ in % for $T_{60} = 150\,\text{ms}$ and $d = 0.1\,\text{m}$

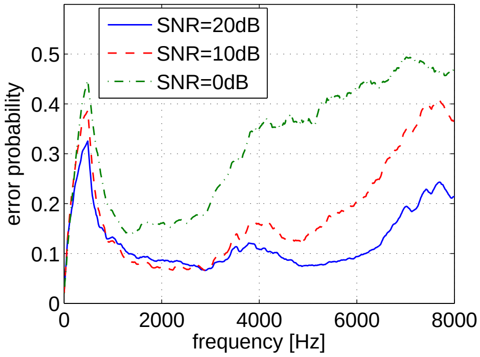


Figure 5.22: Perm. error across frequency, $T_{60} = 150\,\text{ms}$, $d = 0.1\,\text{m}$

Experiments 3: Robustness to Noise

Another peculiarity of permutation solution methods based on phase information (such as TDOA or DOA) is that they are very robust to noise compared to methods which are based on the solely non-stationarity of the sources. We consider the case of $T_{60} = 150\,\text{ms}$ and $d = 0.1\,\text{m}$. Table 5.3 summarizes the mean μ_P and standard deviation σ_P of the permutation errors P for perfect knowledge of the TDOAs and for the SCT for different noise conditions. Clearly, the number of permutation errors increases with the noise. As we note from Fig. 5.22, the error probability increases mainly at higher frequencies. This is due to the fact that speech signals have less power in the high frequency region and hence the local SNR in that region is much worse than for the lower frequencies. However, except for the unrealistic case of SNR=0dB, the SCT is able to solve the permutation with an error closer to the lower bound of TDOA-based methods.

5.5.3 Underdetermined Source Separation by Spatial Filtering

By seeking maxima on the GSCT we can identify TDOA vectors which can be used to parametrize the ideal complex-valued propagation model of each source. The GSCT can identify a number of sources greater than the number of the microphones and thus even the propagation models of secondary reflections can be estimated. The spatial diversity of the acoustic propagation depends on both source locations and array geometry. If the spatial diversity is considered sufficiently high, assuming time-frequency sparseness, the source separation is possible by computing corresponding binary masks by means of the estimated TDOA clusters. Let us define with $\mathbf{ratio}(k,l)$ the vector of the normalized ratios between all the possible microphone pairs, computed similarly to (5.51):

$$\mathbf{ratio}(k,l) = \begin{bmatrix} Re\left(\frac{x_1(k,l)}{x_2(k,l)} \cdot \frac{|x_2(k,l)|}{|x_1(k,l)|}\right) \\ Im\left(\frac{x_1(k,l)}{x_2(k,l)} \cdot \frac{|x_2(k,l)|}{|x_1(k,l)|}\right) \\ \cdot \\ \cdot \\ Re\left(\frac{x_{a_p}(k,l)}{x_{b_p}(k,l)} \cdot \frac{|x_{b_p}(k,l)|}{|x_{a_p}(k,l)|}\right) \\ Im\left(\frac{x_{a_p}(k,l)}{x_{b_p}(k,l)} \cdot \frac{|x_{b_p}(k,l)|}{|x_{a_p}(k,l)|}\right) \end{bmatrix} \tag{5.71}$$

where (a_p, b_p) are the indexes of the *p-th* microphone pair. The binary mask for the *n-th* source may be computed as:

$$mask_n(k,l) = \begin{cases} 1, & if \quad \underset{i}{argmin}\, D[\mathbf{c}(k,\boldsymbol{\mu}_i); \mathbf{ratio}(k,l)] = n \\ \\ 0, & \qquad otherwise \end{cases} \tag{5.72}$$

Then we applied such a filtering to separate 5 sources located as in figure (5.23). Note, to reduce the distortion introduced by the hard decisions in (5.72), many alternative soft-masks have been proposed [98].

A largely spaced microphone array was used to induce sufficient spatial diversity on the acoustic propagation. Two different reverberation time were simulated, $T_{60} = 150\ ms$ and $T_{60} = 300\ ms$. Note that since the filtering is based only on spatial informations, even coherent secondary reflections can be considered as sources. Then, the binary masks were obtained by considering the ideal propagation models parametrized by means of the TDOA vectors corresponding to the first 5 and 10 maxima of the GSCT. In the first case we neglect the secondary reflections and all the time-frequency points were clustered around the propagation models corresponding only to the propagation of the sources along the direct path. In the second case we generate 10 binary masks were the time-frequency points were clustered also around the propagation models of the first 5 main secondary reflections. It is important underlining that the number of the models should be defined according to the number of dominant coherent sources which can be directly evaluated by means of the magnitude of the GSCT. Nevertheless, here it was defined *a priori* to show the changes in the performance when a different number of sources is assumed. As in the above simulation, performance was evaluated by means of the BSS_EVAL toolbox decomposing signal with time-invariant filters as allowed distortions with $L = 1024$ [96]. The time-frequency representation of the signals is obtained adopting the same STFT parameters used for the localization (i.e. Hanning windows of 2048 points overlapped of

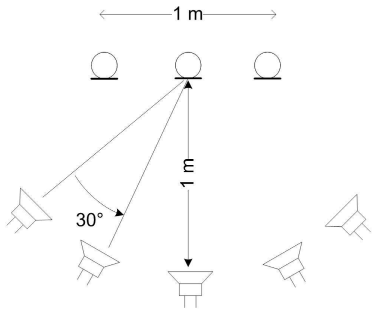
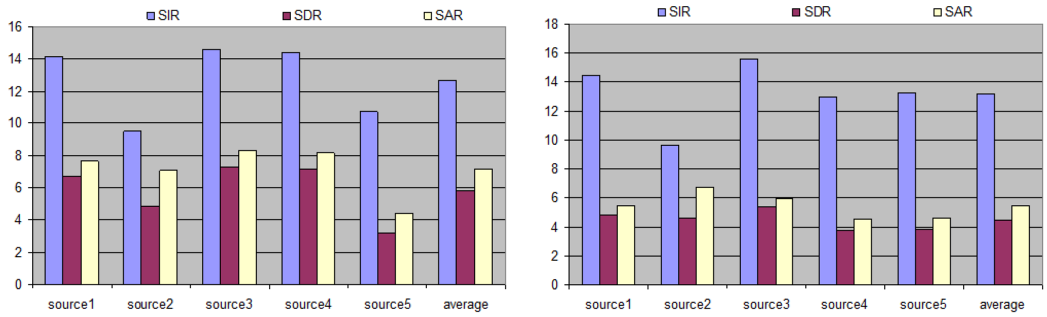


Figure 5.23: Recording setup for undetermined source separation



(a) 5 clusters

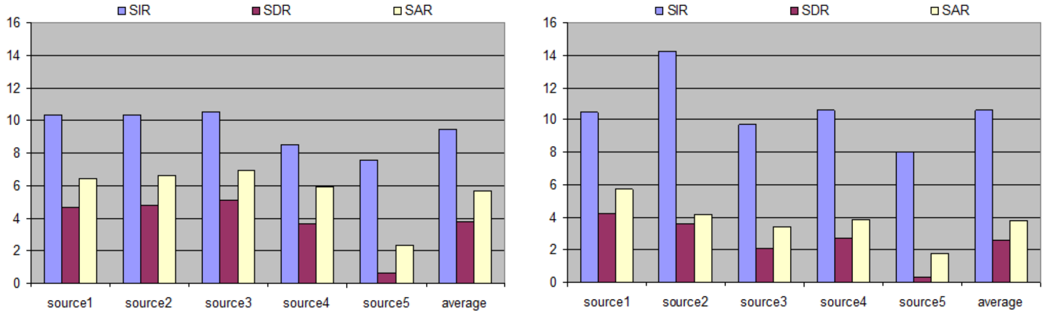
(b) 10 clusters

Figure 5.24: Performance undetermined separation $T_{60} = 150\ ms$

87.5%). Note that the mask $mask_n(k,l)$ can be used to determine the image of the *n-th* source to each microphone. In this experiment we considered the best separated sources over all the microphones images. Figures 5.24 and 5.25 show the resulting performance in SIR, SDR and Signal Artifacts Ratio (SAR). From the overall performance we can see that an effective separation is possible only by means of the estimated multidimensional propagation models even with relatively high reverberation time. We note that by considering also secondary reflections as sources, the SIR is increased but at cost of a reduced SDR and SAR. On the other hand, clustering the time-frequency points only over the direct paths acoustic propagation, artifacts and distortions reduce but with an overall lower separation.

5.6 Concluding Remarks

In this chapter we presented a novel efficient framework for estimating the parameters of the multidimensional acoustic propagation of multiple acoustic sources. The framework is based on the definition of the State Coherence Transform (SCT) and on its generalization (Generalized SCT). The GSCT was shown to be an approximated kernel density estimator of the multivariate TDOA distribution of the acoustic propagation. The estimation of such parameters is useful to



(a) 5 clusters

(b) 10 clusters

Figure 5.25: Performance undetermined separation $T_{60} = 300\ ms$

solve many problems related to acoustic applications such as localization of multiple sources, reduction of permutations and underdetermined blind source separation. Promising results over simulation and extensive real-world experiments confirm that the method is theoretically consistent and attractive for many real-world applications, due to its robustness and relatively low computation complexity.

Many future directions can be considered to further improve the good property of the GSCT. First of all, the proposed GSCT is based on batch or batch on-line adaptation. It may be interesting to derive an on-line adaptive estimator of the GSCT envelope, for both the univariate and multivariate cases. For such scenarios the impact of on-line ICA algorithm need to be investigated. Second, the computational complexity and the spatial resolution are other important issues that require particular attention to make the GSCT reliable for real-time application. Lastly, merging the GSCT with a more general Bayesian framework for localization and/or source separation (e.g. a particle filtering) can be another interesting direction of investigation.

Chapter 6

Coherent spectra estimation

This chapter describes a new method to solve the permutation problem of FD-BSS which combines the robustness of the SCT with the precision of the correlation approach. The method overcomes the limitation of permutation alignment strategies solely based on the phase coherence of the mixing system and avoids typical instabilities encountered by correlation approaches, when short signals are used. The method and the results provided in this chapter have been published in [66].

6.1 Introduction

The permutation ambiguity is an open problem of FD-BSS since a general solution is not yet available. As discussed in chapter 3, the permutation alignment methods can be grouped into two main categories. One consists of methods based on Time Difference of Arrival (TDOA) or Direction of Arrivals (DOA) (such as the method proposed in chapter 5) while the other consists of methods based on the inter-frequency magnitude correlation and the spectral continuity across frequency. We showed that methods of the first category, as those based on the SCT/GSCT, are robust to local errors but are imprecise when there is a high spatial aliasing and high reverberation. When the microphone spacing is small (resulting in small TDOA), the consistency in the phase difference across frequency related to the signal from a point source can be easily observed as a clear trajectory in a polar plot of the frequency-phase relationship (see Figure 6.1). Different point sources produce different trajectories, and the separated frequency-phase trajectories can be used to guide the permutation procedure. As discussed in the previous chapter, this is ideally possible if there is no reverberation since the error introduced by the optimization in (5.68) would be ideally absent. However, if microphones are spaced too far apart, such a separation in frequency-phase relationship becomes difficult to obtain because the reverberation and spatial aliasing introduce distortion in the ideal acoustic propagation and the rule in (5.68) may be not optimal for all the frequencies. Thus, in reverberant conditions, the knowledge of the TDOA is not sufficient for solving the permutation problem.

The second approach is generally more accurate when a sufficient number of observations are available to guarantee a reliable calibration of inter-frequency magnitude correlation. However, it suffers from instability when only a small amount of data is observed and when the noise is not negligible. In such a situation, a local error (at a frequency point) can propagate and cause wrong permutation across the rest of frequencies.

In this chapter, we propose an efficient way of combining the two aforementioned techniques to preserve both the robustness of the SCT and the precision of the correlation approach. The method is based on the main idea that the recovered sources must exhibit a global coherence in the demixing matrices phase and spectra locally smooth. Therefore at the fist stage the frequency bins are aligned according to the TDOA and the resulting aligned frequency bins are used for estimated smooth spectrum which will represent a reference base for each source. In the subsequent stages the frequency bins are aligned according to a pseudo-correlation measure between each bin and the previously estimated spectrum, which includes also the information of the estimated TDOAs. Such a measure generates permutation locally optimized but with a global phase-coherence related to the TDOA of each source. The permutation correction and the spectrum estimation is not performed within the same iteration and this prevent that a local error could propagate across frequency. Ad each stage the number of wrong permutation is reduced and the process is iterated till the number of the permutation corrections converges.

The method has been evaluated separating three sources recorded with three microphone spaced by 50 cm, in order to generate an high phase wrapping, in a simulated room with $T_{60} = 300ms$. Numerical evaluation shows that the method is able to drastically reduce the number of permutation errors of the TDOA method, still inheriting his robustness. Perceptual and numerical evaluation confirms that the achieved separation performance is comparable with the optimal one even when signals are separated in a short-time.

The chapter is organized as follows. In section 6.2, we show the limitations of the SCT for solving the permutation problem. In section 6.3, correlation-based permutation alignment approaches are discussed. In section 6.4, a new pseudo-correlation measure is introduced. In section 6.5, an algorithmic procedure for the proposed method is explained in detail. In section 6.6, experimental results are provided, followed by conclusions.

6.2 Limitation of the SCT for Solving the Permutation Problem

Again, we start from the model defined in Chapter 3 in the equations (3.23)-(3.27) and assume a number of microphones N equal to the number of the sources M. It has been shown in chapter 5 that if the source propagation is coherent in the frequency domain, which generally happens when the propagation over the direct path is predominant, it is possible to estimate the TDOA by locating the peaks of the State Coherence Transform (SCT) computed as in 5.10. Note that the "coherence" of 5.10 is an aggregate figure summed over all frequencies.

Let $\Pi_k(\cdot)$ be a permutation function for frequency bin k which defines the mapping between the indices of the true sources and indices of the demixed sources. $\Pi_k(\cdot)$ is another but equivalent notation of the permutation matrix $\mathbf{\Pi}(k)$. If e.g. $\mathbf{\Pi}(k) = \mathbf{I}$, then $\Pi_k(n) = n,\ \forall n$. For the monodimensional case (i.e. when only the propagation of the acoustic wave to a single microphone pair is considered) the frequency component can be aligned according to the permutation $\hat{\mathbf{\Pi}}_k$ that minimizes the following quantity:

$$\hat{\mathbf{\Pi}}_k = \underset{\mathbf{\Pi}_k}{\operatorname{argmin}} \sum_{i=1}^{N} |c(f_k, \tau_i) - \overline{r}_{\Pi_k(i)k}| \tag{6.1}$$

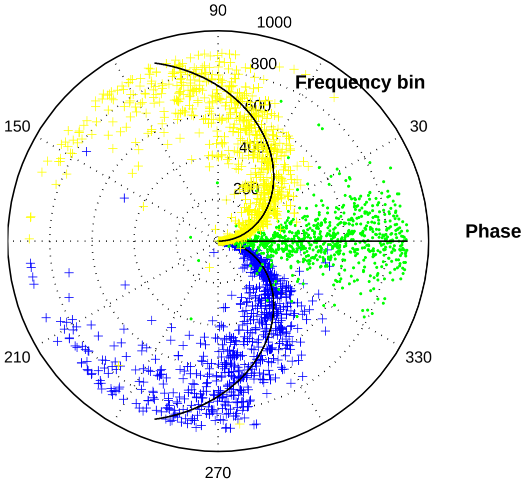
where $\overline{r}_{\Pi_k(i)k}$ is the propagation state of the i^{th} source after permuting the rows of $\mathbf{W}(k)$ according to $\Pi_k(\cdot)$.

When signals of very short duration are considered, the permutation solution based on the source localization of (6.1) is more robust than a correlation based approach [80]. This is true under the assumption that, at each frequency, the phases corresponding to different sources define separable trajectories. However, this condition is not always satisfied especially when the microphone spacing is large. To better understand this problem we look at the polar diagram in Figure 6.1 that shows the propagation model for three recorded sources in a simulated environment (with one source located at an equilateral position). In the figure, the phase of the resulting ratio $\overline{r}_{ik}$ for each source is plotted for two choices of microphone spacing: (a) d = 0.02 m and (b) d = 0.50 m. For the case of d = 0.50 m, the knowledge of the ideal propagation model (represented by the solid lines) is not useful since the phase trajectories are overlapped. Thus, the permutation correction methods based on the TDOA information are intrinsically limited and correlation methods must be considered as an alternative.

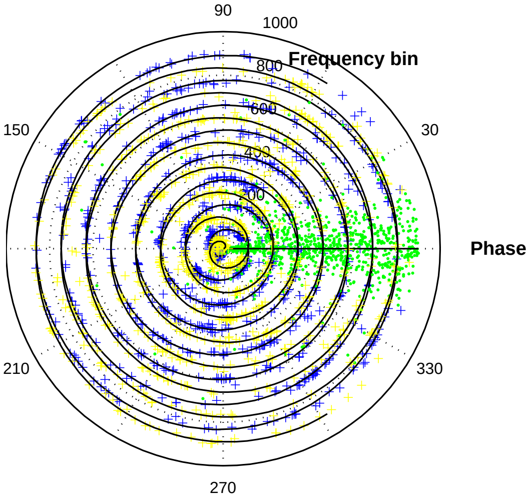
6.3 Correlation-based Approaches and Their Limitations

Correlation based approach are based on the idea that acoustic sources such as speech and music are non-stationary and have a common amplitude modulation. Thus, the energy of adjacent frequencies are highly correlated across time. A measure of correlation can be defined as:

$$corr\{e_i^k(l), e_j^{k-1}(l)\} = \frac{E[e_i^k(l)e_j^{k-1}(l)] - E[e_i^k(l)]E[e_j^{k-1}(l)]}{\sigma[e_i^k(l)]\sigma[e_j^{k-1}(l)]} \tag{6.2}$$



(a) Microphone spacing of $d = 0.02$ m.



(b) Microphone spacing of $d = 0.50$ m.

Figure 6.1: Polar diagram of the phase of $\overline{r}_{ik}$.

where k is the frequency bin index, σ the standard deviation, and $e_i^k(l)$,$e_j^k(l)$ the i^{th} and the j^{th} envelopes defined by:

$$e_i^k(l) = \phi(y_i(k,l)) \tag{6.3}$$

where ϕ is a function generally defined as $|\cdot|$ or $ln|\cdot|$ and y_i is the i^{th} element of the vector $\mathbf{y}(k,l)$ of the estimate of separated signals.

$$\mathbf{y}(k,l) = \mathbf{W}(k)\mathbf{x}(k,l) \tag{6.4}$$

Then at the frequency k, the bins can be aligned by using the permutation $\mathbf{\Pi}_k$ that maximizes the cumulative correlation between the aligned envelope at the k^{th} frequency and the original envelope at the $(k-1)^{th}$ frequency:

$$\hat{\mathbf{\Pi}}_k = \underset{\mathbf{\Pi}_k}{\operatorname{argmax}} \sum_{i=1}^{N} corr\{e_{\Pi_k(i)}^k(l), e_i^{k-1}(l)\} \tag{6.5}$$

where N is the number of sources and $e_{\Pi_k(i)}^k(l)$ is the i^{th} envelope after permuting the elements of the vector $\mathbf{e}^k(l) = [e_1^k(l), e_2^k(l), ..., e_N^k(l)]$ according to Π (which is equivalent to permuting the rows of $\mathbf{W}(k)$). Note, to apply (6.5) we need to use the MDP to reduce the scaling ambiguity.

Since the permutation decision is based on a local measure of energy correlation between the k^{th} and the $(k-1)^{th}$ frequency bins, a common method to solve the permutation for all of the frequencies is to proceed recursively from one end of the spectrum to the other. However, typical in the recursive approach, an error at one frequency can then propagate to subsequent frequencies, resulting in a severe misalignment. Figure 6.2 shows the spectrogram of a badly permuted output source. We note many discontinuities across blocks of frequencies which means that different frequency bands belong to different sources. To limit the propagation of errors, the correlation can be computed across a range of neighboring frequencies within a defined frequency band or alternatively across a subset of harmonic frequencies [84]. Another approach to further confine the local errors is to consider the permutation that maximizes the cumulative correlation between the currently aligned envelope and a smoothed version $b_i^k(l)$ of the previously aligned envelope [75], [67], i.e.,

$$\hat{\mathbf{\Pi}}_k = \underset{\mathbf{\Pi}_k}{\operatorname{argmax}} \sum_{i=1}^{N} corr\{e_{\Pi_k(i)}^k(l), b_i^{k-1}(l)\} \tag{6.6}$$

That is, $b_i^k(l)$ is considered as a “reference” that represents a smoothed spectrum of the i^{th} source at the k^{th} frequency and can be recursively estimated as:

$$b_i^k(l) = (1-\gamma)e_i^k(l) + \gamma b_i^{k-1}(l) \tag{6.7}$$

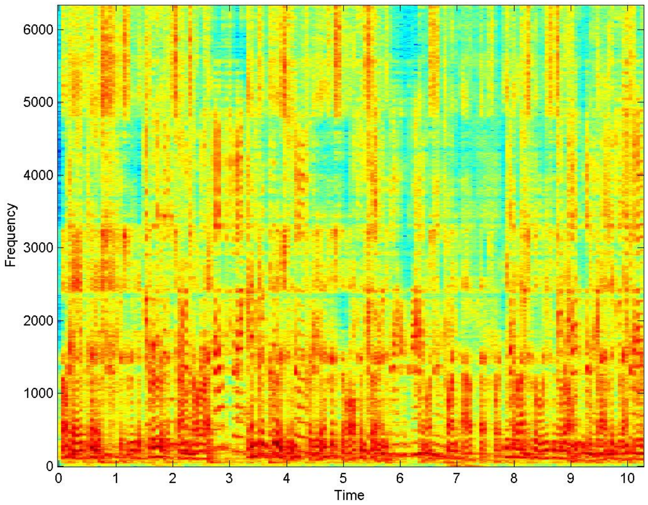


Figure 6.2: Typical spectrogram of an bad permuted output source

where γ is a first-order auto-regressive smoothing coefficient. The use of a smoothed estimate of the previously aligned envelope reduces the probability that a local error propagates to subsequent frequencies. The main issue of all these strategies is that they are highly sensitive to the correct choice of the right parameters. Furthermore, lack they lack of robustness as the correlation measure becomes not be reliable, due to bad ICA convergence.

6.4 Locally and Globally Coherent Spectrum Estimation

To require the reference envelope to have "coherent spectrum" across frequency, it is possible to use the estimated TDOAs to constrain the permutation decisions. In fact the measure of correlation is based only on local information and, if there are local errors, without any global constraint the reference $b_i^k(l)$ would not necessarily represent the same source over all the frequencies. Lets define a matrix $\mathbf{V}(k)$ of elements v_{ij}^k corresponding to the correlation between the i^{th} envelope at the k^{th} frequency (according to the permutation Π) and the j^{th} reference at the $(k-1)^{th}$ frequency:

$$v_{ij}^k = corr\{e_{\Pi_{(i)}}^k(l), b_j^{k-1}(l)\} \tag{6.8}$$

The reference $b_j^{k-1}(l)$ can be associated to a source for which the estimated TDOA is τ_j. The sorting of the envelopes depends on $\mathbf{W}(k)$ according to (6.4) at the frequency k. Then it is possible to associate a ratio $\overline{r}_{ik}$ for each $e_{\Pi_{(i)}}^k(l)$. We can define a matrix $\mathbf{T}(k)$ of elements t_{ij}^k computed as follows:

$$t_{ij}^k = 1 - \frac{|c(f_k, \tau_j) - \overline{r}_{\Pi(i)k}|}{2} \tag{6.9}$$

Finally, a pseudo-correlation matrix $\mathbf{C}(k)$ is computed as:

$$\mathbf{C}(k) = \mathbf{V}(k) \odot \mathbf{T}(k) \tag{6.10}$$

where $\odot$ is the Hadamard (i.e., element-wise) product, and the permutation at each frequency k is decided by the optimization:

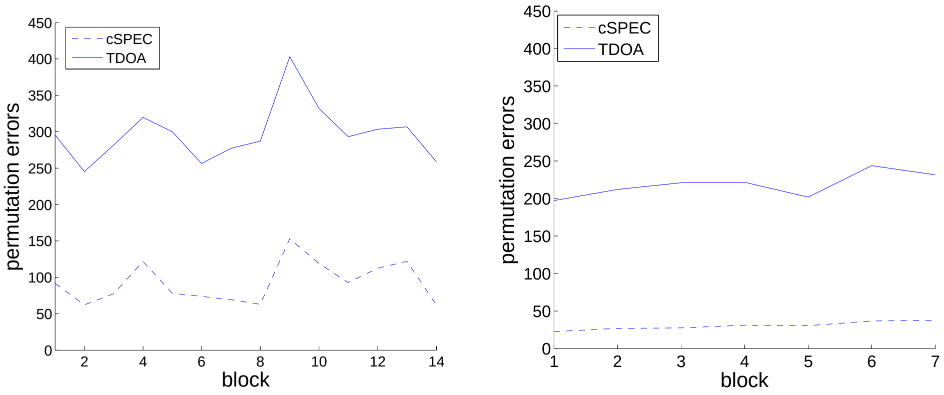
$$\hat{\mathbf{\Pi}}_k = \underset{\mathbf{\Pi}_k}{\operatorname{argmax}}\{trace[\mathbf{C}(k)]\} \tag{6.11}$$

which corresponds to the diagonalization of the matrix $\mathbf{C}(k)$. The matrix $\mathbf{C}(k)$ is jointly diagonalized by the phase-coherence of the impulse responses of the resulting $\mathbf{W}(k)$ and by the local correlation between the energy of neighboring frequencies. For example, if in some frequencies the propagation models of different sources are very similar, $\mathbf{T}(k)$ would be almost singular. Then the effect of $\mathbf{V}(k)$ in diagonalizing $\mathbf{C}(k)$ would be predominant. On the other hand, when $\mathbf{V}(k)$ is singular due to unreliable correlation measures, $\mathbf{C}(k)$ is diagonalized by $\mathbf{T}(k)$. In other words, the interleaved influence of both the matrices $\mathbf{T}(k)$ and $\mathbf{C}(k)$ guarantees a local optimization but still maintaining a global constraint of phase-coherence across frequency.

6.5 Proposed Algorithm

The following scheme summarizes how to use both the phase-coherence and local correlation to estimate the best permutation matrix:

```
for k=highest_frequency_index to 1
    Compute T as in (6.9)
    Set C=T
    Permute W(k) according to (6.11)
end
pass=1;
REPEAT (pass=pass+1)
1° Stage (spectrum estimation)
for k=highest_frequency_index to 1
    Estimate $b_i^k(l)$ as in (6.7)
end
2° Stage (permutation correction)
for k=highest_frequency_index to 1
    Compute C as in (6.8)-(6.10)
    Permute W(k) according to (6.11)
```



(a) Block length of 1 s.

(b) Block length of 2 s.

Figure 6.3: Permutation errors for TDOA and cSPEC alignment.

```
    end
    UNTIL (corrections $\le \alpha$) and (pass $\le$ MaxPass)
```

At the first iteration, the permutation is decided only by considering the TDOA information estimated by the SCT. After that, the reference envelopes are estimated by smoothing the previously aligned envelopes, which are then used to compute $\mathbf{V}(k)$ and to refine the permutation alignment. The spectrum estimation and the permutation correction procedures are iterated separately until the number of permutation corrections converges to a certain value α.

6.6 Experimental Results

The cSPEC performance is assessed as follows:

- counting the number of permutation errors after the procedure
- measuring the global separation performance (SIR and SDR)

The experiments were carried out simulating three speakers placed in a room with T_{60} = 300 ms at roughly 1 meter from three microphones spaced apart by 0.5 meter. Signals were sampled at f_s = 16 kHz, and the time-frequency representation was obtained with a short-time Fourier Transform by using Hanning windows of 2048 taps and overlap of 75%. The number of permutation errors have been evaluated as in 5.5.2. The recorded signals have been framed in short blocks of 1, 2, and 4 seconds, and the separation was independently performed in each block, where the RR-ICA was adopted at each frequency. For the reference envelope estimation in (6.7), the smoothing parameter of $\gamma = 0.9$ was chosen. The number of permutation errors in each block is plotted in Figure 6.3 for the alignment based only on the TDOA and on the coherent spectrum estimation (cSPEC). Even for the case of only 1 second of data block, we observe

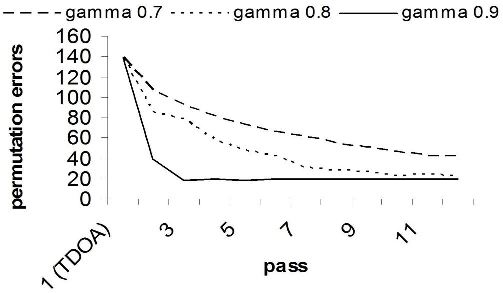


Figure 6.4: Convergence curve of the permutation errors.

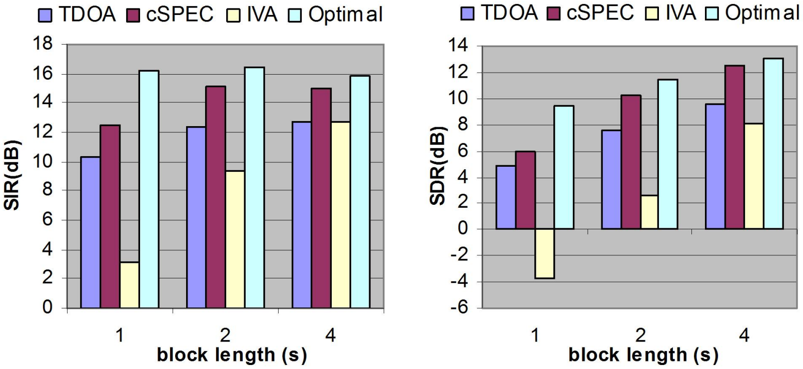


Figure 6.5: Performance evaluation.

a considerable reduction in the permutation error and we also note that the performance does not vary considerably across the blocks, which proves the robustness of the proposed strategy. Figure 6.4 shows that within a few passes the number of errors converges to a stable minimum almost independently of the parameter γ, which only conditions the convergence speed. Figure 6.5 shows the average performance in terms of the source-to-interferences ratio (SIR) and the source-to-distortion ratio (SDR) computed using the original sources and the BSS_EVAL toolbox [96]. It is important to underline the SIR and the SDR are highly influenced by the lower frequencies, which contain most of source energy. However they often do not represent the perceptual quality of the separated sources very well. Furthermore, SIR and SDR are inadequate in revealing permutation errors at high frequencies even though the distortions are clearly audible. The proposed method is compared with the Independent Vector Analysis (IVA) [45], which is a method erroneously believed to be *permutation free*. Note that the IVA approaches a performance comparable to that of TDOA and cSPEC methods when signal blocks of 4s are processed. However, the performance degrades as the size of the block is reduced since the IVA, similarly to time-domain methods, is sensitive to the presence of local minima (i.e. is still affected by permutation errors). On the other hands, the cSPEC has stable performance since it inherits the robustness of the TDOA method. Furthermore, due to the precision of the correlation measure, it approaches the optimal performance as the block size is increased.

6.7 Concluding Remarks

In this chapter, a new robust method for solving the permutation problem is presented, which combines the robustness of the SCT and the precision of the inter-frequency correlation. Although the method is based on heuristics (i.e. is sub-optimal), experimental results confirm its robustness in adverse conditions.

As a possible future direction, it may be interesting to extend the procedure to the case of the underdetermined source separation. The recent underdetermined sources separation methods for convolutive mixtures are based on approaches similar to the FD-BSS. The sources are first separated at each frequency and after that a permutation correction method is exploited to cluster the frequency components related to the same source [26][34]. Even for the underdetermined case, the phase coherence and the spectral continuity can be exploited to reduce the permutation and generate separated sources with low distortion. Finally, computational issues need to be considered since the computational complexity increases exponentially with the number of the sources.

Chapter 7

Semiblind Source Separation applied to Multichannel Acoustic Echo Canceller

In this chapter we analyze and propose a new framework which merges together the two problems of convolutive BSS and multichannel AEC (MCAEC). By exploiting the principles of source separation a new robust and efficient algorithm for MCAEC is proposed, which is robust against ill-conditioned adaptation of standard MCAEC systems and is intrinsically able to tackle the presence of multiple interfering sources, without the need of any double-talk detector. A paper derived by the following discussions has been accepted for publication in the IEEE Transactions on Audio Speech and Language processing, co-authored by Ted S. Wada and Biing-Hwang (Fred) Juang, with the name ”Batch-Online Semi-Blind Source Separation Applied to Multi-Channel Acoustic Echo Cancellation” [29].

7.1 Introduction

In the previous chapters we discussed the problem of the ”blind” source separation applied to the separation of acoustic sources. As discussed in chapter 2, the source separation can never be totally ”blind” since some assumptions on the source statistics and a model to represent the mixing system are required. However, the term ”blind” is widely used to refer to the fact that BSS techniques, compared against other methods, exploit a very small knowledge about the sources: that is, the activity and the location of the sources is generally not known in advance. A less ”blind” case of source separation is when some of the sources are already known in advance. This problem is addressed by Acoustic Echo Canceller systems. AEC systems aim to reduce the feedback coming from the response sound of the loudspeakers, recorded by the microphones. They are required in many applications such as distant-talk speech driven systems (e.g. a television or multimedia set-top-box controlled by speech commands) or robust headphone-free videoconference systems. The first systems are generally controlled by a full-duplex interaction and thus they generate interfering noise from the loudspeakers of the system. This interferer, known as echo of the system, is known by the system and can be cancelled by AEC signal processing techniques. The importance of AEC becomes even more relevant for headphone-free full-duplex videoconference applications. In this case the sound coming from the far-end side of the conference is played through the loudspeakers at the near-end side and is recorded from the microphones. If the loudspeaker echos are not properly cancelled at the near-end side, the far-end talker would receive a delayed version (according to the network propagation) of it own speech. This may also lead to an accidentally saturation of the channel because of the loop propagation of the feedback (Larsen effect). Traditional AEC systems attempt the reduction of the echos by estimating the impulse responses between the loudspeakers and the microphones. This estimation is performed by adaptive methods such as least mean square (LMS) based approaches. In spite of the popularity of such methods, two main limitations reduce the effectiveness of standard AEC systems in more complex real-world scenarios. First, since the near-end sources are not intrinsically modeled in the adaptation, the estimation must be opportunely controlled according to the presence of the near-end source. Second, the multichannel extension of these techniques suffers of ill-conditioning problems due to the intrinsic correlation between the loudspeaker signals, which often makes a unique and stable estimation of the mixing filters impossible.

In this chapter we propose and discuss a new framework of semi-blind source separation (SBSS) which better handles the aforementioned problems. The new framework shifts the conventional MCAEC paradigm to a new approach based on the Blind source separation (BSS), which is a powerful signal enhancement method for recovering a target signal from a mixture of signals when no prior information on the original source signals are available. SBSS is a direct extension of BSS when some partial knowledge of the source signals are already available (e.g., components of the reference signals). It will be shown that the SBSS is naturally suited for the MCAEC purpose in the presence of multiple interfering signals.

Many implementations of BSS are available in literature, among of them, those based in the frequency-domain are the most popular ones. FD-BSS can be implemented through a batch, i.e., offline, adaptation based on independent component analysis (ICA), which aims to maximize the statistical independence of separated signal components given that the original sources themselves are independent. In the previous chapters we referred to the natural gradient algorithm [3] which has become a standard in realizing the ICA optimization. The SBSS approach to the MCAEC problem was first proposed in [50] and was successfully implemented in [55] as a combination of multi-channel BSS and a single-channel AEC in the frequency domain. Furthermore it was shown in [97] that BSS and stereophonic AEC (SAEC) can be effectively implemented together in such a framework.

In this chapter we analyze the applicability of a general framework of SBSS to the MCAEC problem and we investigate how the performance can be improved over standard methods. First of all, a deep theoretical analysis is described which involves also the connection of the proposed framework with standard techniques of BSS and MCAEC. Second of all, many algorithmic issues are discussed in order to provide suggestions for the design of robust and practical SBSS systems. In particular, we proposed a batch and online adaptations and a proper constraint on the de-mixing matrix to achieve both high echo cancellation and relatively low misalignment without any pre-decorrelation procedure, even with a persistent activity of interfering sources.

The chapter is organized as follows:

(a) a review of the LMS algorithm is presented in Section 7.2;

(b) A deep theoretical analysis of the SBSS framework is presented in Section 7.3, which includes: presentation of the model of an SBSS system (7.3.1), discussion of the origin of the non-uniqueness problem in SBSS (7.3.2), steady-state analysis of the SBSS system (7.3.3), illustration of a connection between the MSE-based and the ICA-based approach (7.3.4), and exploration of constraints on the SBSS de-mixing filter and their effect on the ICA optimization (7.3.5).

(c) The main issues of the algorithmic SBSS design are discussed in Section 7.4, which includes: detailed outline of an online implementation of the SBSS system (7.4.1) and description of the implemented SBSS algorithm (7.4.2).

(d) Methods for evaluating the SBSS performance and simulated and real-world results are provided in (7.5.1) and (7.5.2), respectively.

(e) Concluding remarks and discussion on future direction are made in Section 7.7.

7.2 A review of the LMS algorithm

First, let us consider a single channel AEC system which is summarized in Figure 7.1. A single microphone records the acoustic echo coming from a loudspeaker together with the sound propagating from a near-end source $s(n)$. The resulting mixture is modeled as:

$$x(n) = s(n) + \mathbf{h}(n) * r(n) \tag{7.1}$$

where we indicated with $*$ the convolution operator and with n the sample index. Assuming the near-end source $s(n)$ being not active, the estimation of the mixing system $h(n)$ may be obtained by minimizing the expectation of the mean-square error $e(n)^2$. Let us indicate with L

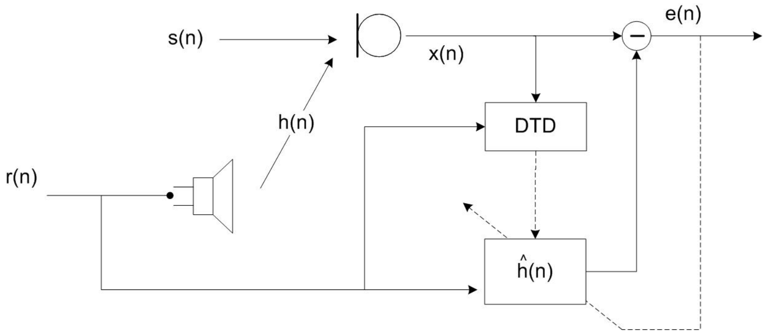


Figure 7.1: Model per a single-channel AEC system

the length of the filter to be estimated (which should be equal to the length of $\mathbf{h}(n)$) and with $\mathbf{r}(n)$ and $\mathbf{x}(n)$ the reference and signal column vectors of length L, respectively.

$$\overline{\mathbf{h}}(n) = \underset{\hat{\mathbf{h}}(n)}{\operatorname{argmin}} E[(\mathbf{x}(n) - \hat{\mathbf{h}}(n) * \mathbf{r}(n))^2] \tag{7.2}$$

Equation (7.2) can expanded as:

$$E[\mathbf{e}(n)^2]{=}E[\mathbf{s}(n)^2]{+}E[\mathbf{r}(n)^2{*}(\mathbf{h}(n){-}\hat{\mathbf{h}}(n))^2]{+}E[\mathbf{s}(n)^2]{+}2E[\mathbf{s}(n)\mathbf{r}(n){*}(\mathbf{h}(n){-}\hat{\mathbf{h}}(n))] \tag{7.3}$$

Since the near-end source is generally assumed to be not correlated with the reference signal, the third term of equation (7.3) can be neglected and the error vector is minimized when $\hat{\mathbf{h}}(n)$ approaches the true impulse response $\mathbf{h}(n)$. A preferred method to estimate the mixing system $\mathbf{h}(n)$ is the least-mean square (LMS) on-line adaptation [99] due to its computational simplicity and inherent adaptability. The algorithm is summarized by the filter coefficient update equation

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) - \mu e(n)\mathbf{r}(n), \tag{7.4}$$

where μ is the adaptation step-size, and $e(n)$ is the error between the observed and estimated actual echo

$$e(n) = x(n) - \hat{\mathbf{h}}(n) * \mathbf{r}(n) \tag{7.5}$$

LMS relies on information from the second-order statistics (SOS) for adaptation as the mean-square error (MSE) $E[e^2(n)]$ is minimized to obtain the optimal filter coefficients that best represent the true echo path. The LMS algorithm is generally used in its normalized version [100] which has a convergence rate independent on the non-stationarity in the reference signal. It is obtained by modifying the LMS as:

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) - \mu e(n)\frac{\mathbf{r}(n)}{\mathbf{r}'(n)\mathbf{r}(n)} \tag{7.6}$$

where $'$ indicates the vector transpose. The LMS algorithm suffers from a slow convergence when the reference signal is colored, e.g., a speech signal. The convergence rate can be improved by applying a decorrelation filter to the reference signal [35] or by implementing the adaptive algorithm in the frequency domain, e.g., frequency-block LMS (FBLMS) [19].

There are two other main issues and the related complications that prevent the LMS algorithm from achieving a desired echo cancellation performance. These are: the presence of local acoustic noise (or near-end speech) and the non-uniqueness problem that arise during multichannel AEC (MCAEC).

First, the LMS algorithm by itself has difficulty converging to the optimal solution when there is a local noise since each noisy signal sample directly perturbs the gradient of the MSE, $\nabla_{\mathbf{w}} E[e^2] = -2E[e\mathbf{r}]$, thereby leading to a "noisy" update of the filter coefficients. In the specific case of an NLMS adaptation and assuming the far-end and near-end signals being uncorrelated stochastic processes the expectation of the steady-state normalized error (or misalignment) between the true and the estimated mixing system can be computed as:

$$E\left[\left(\frac{||\mathbf{h}-\hat{\mathbf{h}}||}{||\mathbf{h}||}\right)^2\right] = \frac{\mu}{2 \times EBR} \tag{7.7}$$

where EBR is the echo-to-background (i.e. additive and near-end noise) power ratio [36]. It becomes evident that if the EBR is too small (e.g. smaller than 1) a small step-size is required, in order to guarantee low steady-state performance. However, this also means to lead to a very slow convergence rate in the adaptation with a consequential degradation of the overall performance. The issue is generally solved by an adaptive step-size or a regularization procedure that can be used to stabilize the adaptation process [35]. However, most of the adaptive step-size algorithms in the AEC literature are ineffective in a so-called double-talk situation that occurs when a near-end talker speaks concurrently with a far-end talker. A traditional way to prevent instabilities is to stop adaptation process when the activity of near-end sources is detected (double-talk). The inhibition of the adaptation during double-talk can be a major drawback in a non-stationary acoustic environment where an adaptive filter needs to continue learning as much as possible.

When generalized to multi-channel acoustic echo cancellation (MCAEC), traditionally, the algorithmic framework is still based on a single-channel implementation of the LMS algorithm, which relies solely on the second-order statistics (SOS). For instance, take the stereo AEC (SAEC) case as illustrated in Figure 7.2, and assume that $E[s\mathbf{r}] = \mathbf{0}$ and that the near-end impulse response vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ are of finite length L. Then the conventional approach is to

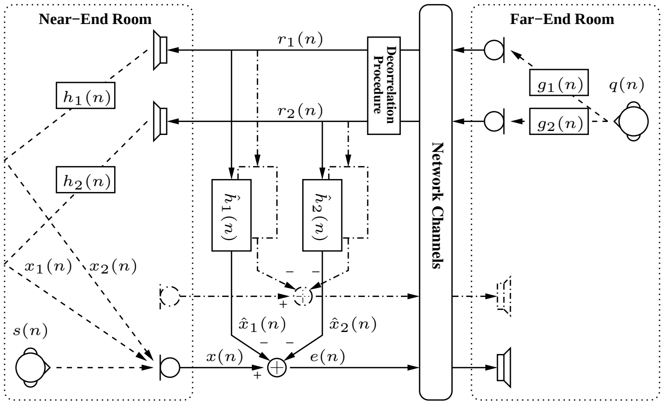


Figure 7.2: Model for stereophonic acoustic echo cancellation (SAEC).

obtain a filter coefficients vector $\hat{\mathbf{h}} = [\hat{\mathbf{h}}_1^T, \hat{\mathbf{h}}_2^T]^T$ of length $2L$ that minimizes the MSE

$$E[e^2(n)] = [\mathbf{h}(n) - \hat{\mathbf{h}}(n)]^T \mathbf{R}_{rr} [\mathbf{h}(n) - \hat{\mathbf{h}}(n)] + E[s^2(n)] \tag{7.8}$$

for each near-end microphone channel, where $\mathbf{h} = [\mathbf{h}_1^T, \mathbf{h}_2^T]^T$ corresponds the two echo paths to be estimated and $\mathbf{R}_{rr} = E[\mathbf{r}\mathbf{r}^T]$ is the $2L \times 2L$ autocorrelation matrix of $\mathbf{r} = [\mathbf{r}_1^T, \mathbf{r}_2^T]^T$. Thus the LMS-based MCAEC inherits the same problems associated with the single-channel LMS algorithm, i.e., the effect of the noise s and the additional potential ill-conditioning of $\mathbf{R}_{rr}$.

Second, the so-called non-uniqueness problem [90] occurs in an MCAEC setting when the adaptive filter length L is longer than or equal to the far-end room impulse response length M. The problem can be illustrated for the SAEC case in Figure 7.2 when the near-end source is silent, in which case (7.8) becomes

$$E[e^2(n)] = [\mathbf{h}(n) - \hat{\mathbf{h}}(n)]^T \mathbf{G}(n) \mathbf{R}_{qq} \mathbf{G}^T(n) [\mathbf{h}(n) - \hat{\mathbf{h}}(n)], \tag{7.9}$$

where $\mathbf{R}_{qq}$ is a $(M + L - 1) \times (M + L - 1)$ far-end source autocorrelation matrix and $\mathbf{G}$ is a $2L \times (M + L - 1)$ matrix with the row rank of $2L$ formed by delayed versions of the far-end impulse response vectors $\mathbf{g}_1$ and $\mathbf{g}_2$ of length M (refer to [41] for a more detailed treatment). Assuming that $\mathbf{R}_{qq}$ is fully ranked, the optimal solution for $\mathbf{h}$ is given by

$$[\mathbf{h}(n) - \hat{\mathbf{h}}(n)]^T \mathbf{G}(n) = \mathbf{0}, \tag{7.10}$$

where $\mathbf{0}$ is a zero vector of length $2L$. (7.10) is an exact or at least an over-determined problem as long as $2L \leq L + M - 1$ and thus has a unique solution if $L < M$. Therefore, the non-uniqueness problem arises when $L \geq M$ such that the far-end impulse response matrix $\mathbf{G}$ is row-rank deficient, which leads to linear dependence of the reference signals. The same analysis can be extended to the MCAEC case.

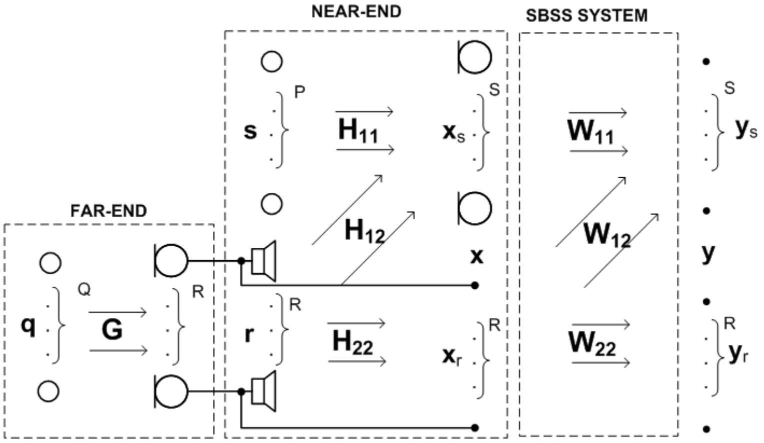

NEAR-END
SBSS SYSTEM
FAR-END
s
P
$\overrightarrow{\mathbf{H}_{11}}$
$\mathbf{x}_s$
S
$\overrightarrow{\mathbf{W}_{11}}$
S
$\mathbf{y}_s$
$\mathbf{H}_{12}$
x
$\mathbf{W}_{12}$
y
q
Q
$\overrightarrow{\mathbf{G}}$
R
r
R
$\overrightarrow{\mathbf{H}_{22}}$
$\mathbf{x}_r$
R
$\overrightarrow{\mathbf{W}_{22}}$
R
$\mathbf{y}_r$

Figure 7.3: Model of the near-end and the far-end mixing systems and the semi-blind source separation (SBSS) system.

It may be argued that the non-uniqueness condition of $L \geq M$ is actually rare since in reality the acoustic impulse response is infinite in length. However, when only one colored source is active at the far end, a severe ill-conditioning of the autocorrelation matrix occurs. In these conditions the high level of the resulting inter-channel correlation would hamper the convergence rate of an MSE-based adaptive algorithm. Furthermore, when the adaptation are formulated in the frequency-domain, due to the sparse representations of the signals and the impulse responses after taking the short-time Fourier transform (STFT), the ill-conditions of the autocorrelation matrix would be more severe. Furthermore, movement or changes in the active far-end talkers would also require an adaptive algorithm to re-converge in spite of the effect of the non-uniqueness problem. Thus some pre-processing procedure is necessary most of the time to decorrelate the reference signals before a playback at the near end, e.g., application of memory-less nonlinearities [9]. Most of the decorrelation procedures generates a degradation in the signal quality as perceived by the near-end listeners, although a perceptually motivated decorrelation technique have been recently proposed in [42].

7.3 Theoretical analysis of SBSS

7.3.1 Model

In this section we define a time-invariant mixing model in the frequency domain that is as general as possible and the notations which will be used in the subsequent sections. Moreover, we assume zero-mean random processes that generate the involved signals. A model for the near-end and the far-end mixing systems and the SBSS system is illustrated in Figure 7.3. At far-end Q sources are recorded by an array of R microphone. At near-end an array of S microphones records P near-end sources and R loudspeaker echos. A far-end source signals vector $\mathbf{q}$ is multiplied by a frequency response $R \times Q$ matrix $\mathbf{G}$, which represents the far-end mixing system, to give the reference signal vector $\mathbf{r}$:

$$\mathbf{r}(\omega, t) = [r_1(\omega, t), \cdots, r_R(\omega, t)], \quad \mathbf{q}(\omega, t) = [q_1(\omega, t), \cdots, q_Q(\omega, t)]$$
$$\mathbf{r}(\omega, t) = \mathbf{G}(\omega)\mathbf{q}(\omega, t), \tag{7.11}$$

A near-end multi-channel source-signal vector $\mathbf{s}$ is then multiplied by a frequency response $S \times P$ matrix $\mathbf{H}_{11}$, and a multi-channel reference-signal vector $\mathbf{r}$ is multiplied by a frequency response $S \times R$ matrix $\mathbf{H}_{12}$, i.e., the echo paths. The two matrices can be combined into a single $(S + R) \times (P + R)$ matrix $\mathbf{H}$ that represents the entire near-end mixing system:

$$\mathbf{s}(\omega, t) = [s_1(\omega, t), \cdots, s_P(\omega, t)],$$
$$\mathbf{x}_s(\omega, t) = [x_1(\omega, t), \cdots, x_S(\omega, t)], \;\; \mathbf{x}_r(\omega, t) = [x_{S+1}(\omega, t), \cdots, x_{S+R}(\omega, t)],$$
$$\mathbf{x}(\omega, t) = \begin{bmatrix} \mathbf{x}_s(\omega, t) \\ \mathbf{x}_r(\omega, t) \end{bmatrix} = \mathbf{H}(\omega) \begin{bmatrix} \mathbf{s}(\omega, t) \\ \mathbf{r}(\omega, t) \end{bmatrix}, \tag{7.12}$$

$$\mathbf{H}(\omega) = \begin{bmatrix} \mathbf{H}_{11}(\omega) & \mathbf{H}_{12}(\omega) \\ \mathbf{O}_{R\times P} & \mathbf{I}_R \end{bmatrix}_{(S+R)\times(P+R)}, \tag{7.13}$$

where $\mathbf{O}_{R\times P}$ is a $R \times P$ matrix with all elements equal to 0 and $\mathbf{H}_{22}$ is automatically assigned be an identity $R \times R$ matrix (i.e. $\mathbf{I}_R$). Furthermore, by substituting (7.11) into (7.12), the far-end and the near-end mixing systems can be combined into a unified mixing system represented by a $(S + R) \times (P + Q)$ matrix $\tilde{\mathbf{H}}$:

$$\mathbf{x}(\omega, t) = \tilde{\mathbf{H}}(\omega) \begin{bmatrix} \mathbf{s}(\omega, t) \\ \mathbf{q}(\omega, t) \end{bmatrix}, \tag{7.14}$$

$$\tilde{\mathbf{H}}(\omega) = \begin{bmatrix} \mathbf{H}_{11}(\omega) & \mathbf{H}_{12}(\omega)\mathbf{G}(\omega) \\ \mathbf{O}_{R\times P} & \mathbf{G}(\omega) \end{bmatrix}_{(S+R)\times(P+Q)}. \tag{7.15}$$

By now on, we assume $S = P$ and $Q = R$ which means that the matrix $\tilde{\mathbf{H}}$ is invertible. However, it we will shown that changes in the number of active sources do not introduce any problem in the adaptation, on condition that proper constrains are adopted. An SBSS system has to goal to perform the estimation of the near-end source signals by using an $(S + R) \times (S + R)$

de-mixing matrix $\mathbf{W}$ such that

$$\mathbf{y}(\omega) = \left[\begin{array}{c} \mathbf{y}_s(\omega, t) \\ \mathbf{y}_r(\omega, t) \end{array} \right] = \mathbf{W}(\omega)\mathbf{x}(\omega, t) \simeq \left[\begin{array}{c} \mathbf{s}(\omega, t) \\ \mathbf{q}(\omega, t) \end{array} \right],$$
$$\mathbf{y}_s(\omega, t) = [y_1(\omega, t), \cdots, y_S(\omega, t)], \;\; \mathbf{y}_r(\omega, t) = [y_{S+1}(\omega, t), \cdots, y_{S+R}(\omega, t)] \tag{7.16}$$

which implies that the acoustic echo cancellation is also performed to give only the source signals in the output vector $\mathbf{y}_s$, where we generalize the structure of $\mathbf{W}$ as

$$\mathbf{W}_{11} = [w_{ij}]_{1 \leq i,j \leq S}, \;\; \mathbf{W}_{12} = [w_{ij}]_{\substack{1 \leq i \leq S \\ S+1 \leq j \leq R}}, \;\; \mathbf{W}_{22} = [w_{ij}]_{S+1 \leq i,j \leq S+R}, \tag{7.17}$$

$$\mathbf{W}(\omega) = \left[\begin{array}{cc} \mathbf{W}_{11}(\omega) & \mathbf{W}_{12}(\omega) \\ \mathbf{O}_{R \times S} & \mathbf{W}_{22}(\omega) \end{array} \right]_{(S+R) \times (S+R)}. \tag{7.18}$$

From the structure of the demixing matrix in 7.17 we can discuss about two important key aspects:

- it is worth noting that $\mathbf{W}$ is a block upper triangular matrix, i.e., $\mathbf{W}_{21} = \mathbf{O}_{R \times S}$, since the reference signal channels are completely blind to the near-end source signals, i.e., $\mathbf{H}_{21} = \mathbf{O}_{R \times S}$. However, the near-end signals themselves may take a round-trip in a full-duplex teleconferencing system as illustrated in Figure 7.2, and the resulting echo may not be cancelled entirely by the far-end SBSS system such that the reference signals would contain some remnants of the near-end signals. To neglect such a circumstance we consider the systems at the steady-state. In other terms it may be assumed that the far-end system has performed a reasonable job in suppressing the echo and at convergence the near-end signals is uncorrelated with its own echo due to a sufficiently large round-trip delay. Hence by requiring $\mathbf{W}_{21} = \mathbf{O}_{R \times S}$, we avoid the possibility that $\mathbf{y}_r$ contains the near-end source signals. In other words, the echo-cancelled output components of the SBSS system in $\mathbf{y}_s$ are subject to the permutation ambiguity introduced only through the separation of the near-end sources, and the ambiguity problem then just needs to be solved for the sub-matrix $\mathbf{W}_{11}$;

- although for the demixing matrix structure $S+R$ sources would be recovered at the output, we are not interested in recovering the signals played through the loudspeakers since we already have them as the reference signals. It follows that we do not have to adapt $\mathbf{W}_{22}$ for the purpose of separating the far-end signals to obtain the original source vector $\mathbf{q}$ and $\mathbf{y}_r$ can be any linear combination of $\mathbf{r}$. Thus, the exact form of $\mathbf{W}_{22}$ may be controlled to optimize the SBSS performance appropriately. Note, the far-end sources can be still separated but the responsibility of this separation is with the far-end SBSS system and then does not have to be involved in the adaptation used to cancel the echos from the near-end side;

7.3.2 Non-Uniqueness Problem

In section 4.2.3 we discussed of the problem of the ill-conditioning of the autocorrelation matrix in a standard MCAEC algorithm which in the worst case may lead to the problem of the non-unique solution of its optimization. We show in the following discussion that this problem equivalently occurs also in the SBSS.

We see from (7.16) that the optimal solution for $\mathbf{W}$ is obtained when $\mathbf{W}\tilde{\mathbf{H}} = \mathbf{I}$, i.e.,

$$\mathbf{W}_{11}(\omega)\mathbf{H}_{11}(\omega) = \mathbf{I}_S, \tag{7.19}$$

$$\mathbf{W}_{22}(\omega)\mathbf{G}(\omega) = \mathbf{I}_R, \tag{7.20}$$

$$[\mathbf{W}_{11}(\omega)\mathbf{H}_{12}(\omega) + \mathbf{W}_{12}(\omega)]\mathbf{G}(\omega) = \mathbf{O}_{S\times R}. \tag{7.21}$$

That is, the SBSS system is able to jointly perform the separation of near-end source signals and the cancellation of acoustic echoes such that the diagonal terms become an identity while the off-diagonal terms become null. Then if $\mathbf{W}_{11}$ and $\mathbf{G}$ are not singular, $\mathbf{H}_{12}$ that corresponds to the echo paths can be uniquely obtained from (7.21) as

$$\widehat{\mathbf{H}}_{12}(\omega) = -\mathbf{W}_{11}(\omega)^{-1}\mathbf{W}_{12}(\omega). \tag{7.22}$$

When the number of the sources equals that of the microphones at the near end (i.e. $P = S$), by assuming spatial diversity and mutual independence between the sources, the physical interpretation of the frequency-domain BSS [83][61] ensures that $\mathbf{W}_{11}$ is almost always non-singular. If there are more sources than microphones, i.e., the under-determined case ($P > S$), $\mathbf{H}_{11}$ is not invertible. In this case there is no unique solution for $\mathbf{W}_{11}$ but the estimate of $\mathbf{W}_{11}$ is not necessarily singular, and its inversion is still attainable for the estimation of $\mathbf{H}_{12}$. A more serious problem is the severe ill-conditioning, or near-singularity, of $\mathbf{G}$ which would prevent from the unique identification of the echo paths by the MSE-based MCAEC. Also, (7.21) indicates the dependence of the solution for $\mathbf{H}_{12}$ on $\mathbf{G}$ just as in the MCAEC case [90]. Therefore, due to the rare occurrence of the singularity of $\mathbf{W}_{11}$, it can be stated that the non-uniqueness problem in the SBSS is equivalent to that of the traditional MCAEC system.

7.3.3 Steady-State Solution for $\widehat{\mathbf{H}}_{12}$

The global de-mixing matrix $\mathbf{W}$ can be estimated through the gradient-descent estimation procedure:

$$\mathbf{y}_n(\omega,t) = \mathbf{W}_n(\omega)\mathbf{x}(\omega,t), \tag{7.23}$$

$$\mathbf{W}_{n+1}(\omega) = \mathbf{W}_n(\omega) + \eta\boldsymbol{\Gamma}[\mathbf{W}_n(\omega),\mathbf{y}_n(\omega,t)], \tag{7.24}$$

where $\boldsymbol{\Gamma}$ is the updating term, η is the adaptation step-size, and n is the iteration index. $\boldsymbol{\Gamma}$ may take different forms according to the cost function that is to be minimized through the gradient-descent procedure. Although any gradient-descent algorithm can be used for the estimation of $\mathbf{W}$, we focus on an ICA optimization procedure based on the Natural Gradient algorithm discussed in section (2.2). Furthermore, we consider a batch adaptation in order to have a sufficient amount of observation for statistically consistent estimation. According to the natural gradient algorithm, the gradient term is given by

$$\boldsymbol{\Gamma}[\mathbf{W}_n(\omega),\mathbf{y}_n(\omega,t)] = \left\{\mathbf{I}_{S+R} - E[\Phi(\mathbf{y}_n(\omega,t))\mathbf{y}_n(\omega,t)^H]\right\}\mathbf{W}_n(\omega), \tag{7.25}$$

where $\Phi(\cdot)$ is a nonlinear function and $E[\cdot]$ is the expectation operator that can be approximated by averaging over time. Then assuming that all of the near-end sources and the echo paths are mutually independent, (7.24) converges to a solution that minimizes the Kullback-Leibler divergence (see equation (2.27)) between the separated output signals given by (7.23).

To simplify the analysis we can approximate the mutual dependence by referring to the *generalized covariance matrix* $\mathrm{E}[\Phi(\mathbf{y})\mathbf{y}^H]$ and assuming that the components in the output vector $\mathbf{y}$ are zero-mean random variables. In fact, according to the Taylor expansion of the nonlinear function $\Phi(\cdot)$:

$$E[y_a^{\mathrm{u}}(\omega)y_b^*(\omega)] = 0 \quad \forall \mathrm{u} \in \mathbb{N}, \tag{7.26}$$

the sources can be considered statistically independent when $\mathrm{E}[\Phi(\mathbf{y})\mathbf{y}^H]$ is diagonalized.

We can analyze the structure of the steady-state solution for $\mathbf{H}_{12}$ as follows. First, the input signals for (7.23) are obtained by applying (7.13) to (7.12):

$$\begin{bmatrix} \mathbf{x}_s(\omega,t) \\ \mathbf{x}_r(\omega,t) \end{bmatrix} = \begin{bmatrix} \mathbf{H}_{11}(\omega)\mathbf{s}(\omega,t) + \mathbf{H}_{12}(\omega)\mathbf{r}(\omega,t) \\ \mathbf{r}(\omega,t) \end{bmatrix}. \tag{7.27}$$

Next, the output signals used for updating $\mathbf{W}$ in (7.24) are obtained from (7.16) and (7.17):

$$\begin{bmatrix} \mathbf{y}_s(\omega,t) \\ \mathbf{y}_r(\omega,t) \end{bmatrix} = \begin{bmatrix} \mathbf{W}_{11}(\omega)\mathbf{H}_{11}(\omega)\mathbf{s}(\omega,t) + \\ \left[\mathbf{W}_{11}(\omega)\mathbf{H}_{12}(\omega) + \mathbf{W}_{12}(\omega)\right]\mathbf{r}(\omega,t) \\ \mathbf{W}_{22}(\omega)\mathbf{r}(\omega,t) \end{bmatrix}. \tag{7.28}$$

Now, let's for the moment consider statistical independence between the separated sources vector $\mathbf{y}_s$ associated with the near-end system and the separated sources vector $\mathbf{y}_r$ associated with the reference signals. The optimal solution for $\mathbf{H}_{12}$ is obtained by setting $E[\mathbf{y}_s^{\mathtt{u}}\mathbf{y}_r^H] = \mathbf{O}_{S\times R}$ $\forall \mathtt{u} \in \mathbb{N}$. Then, after omitting the frequency and time dependencies for notation convenience, we obtain the following expression by applying (7.26) to (7.28):

$$\begin{aligned} E\{\mathbf{y}_s^{\mathtt{u}}\mathbf{y}_r^H\} = E\{[\mathbf{W}_{11}\mathbf{H}_{11}\mathbf{s}+ \\ (\mathbf{W}_{11}\mathbf{H}_{12} + \mathbf{W}_{12})\mathbf{r}]^{\mathtt{u}}\mathbf{r}^H\mathbf{W}_{22}^H\} \\ = \mathbf{O}_{S\times R} \quad \forall \mathtt{u} \in \mathbb{N}, \end{aligned} \tag{7.29}$$

where $\mathbf{y}_s^{\mathtt{u}}$ indicates the raising of each component of $\mathbf{y}_s$ to the integer power $\mathtt{u}$ and $\mathbf{y}_s^H$ denotes the Hermetian (conjugate) transpose of $\mathbf{y}_s$ (i.e., the scalar sources y_a and y_b from (7.26) are simply substituted by the vectors $\mathbf{y}_s$ and $\mathbf{y}_r$). By applying the binomial expansion, we can rewrite (7.29) as:

$$\begin{aligned} E\{[\mathbf{W}_{11}\mathbf{H}_{11}\mathbf{s}]^{\mathtt{u}}\mathbf{r}^H\mathbf{W}_{22}^H\}+ \\ E\{[(\mathbf{W}_{11}\mathbf{H}_{12} + \mathbf{W}_{12})\mathbf{r}]^{\mathtt{u}}\mathbf{r}^H\mathbf{W}_{22}^H\}+ \\ E\{[\textstyle\sum_{k=1}^{\mathtt{u}-1} \frac{(\mathtt{u}-1)!}{k!(\mathtt{u}-1-k!)}(\mathbf{W}_{11}\mathbf{H}_{11}\mathbf{s})^{\mathtt{u}-1-k}\odot \\ ((\mathbf{W}_{11}\mathbf{H}_{12} + \mathbf{W}_{12})\mathbf{r})^{k}]\mathbf{r}^H\mathbf{W}_{22}^H\} \\ = \mathbf{O}_{S\times R} \quad \forall \mathtt{u} \in \mathbb{N}, \end{aligned} \tag{7.30}$$

where $\odot$ indicates is the Hadamard (element-wise) product. By using the multinomial expansion to further expand the additive terms with powers $\mathtt{u}$, $\mathtt{u} - 1 - k$, and k, it is possible to demonstrate that if $\mathbf{r}$ and $\mathbf{s}$ are statistically independent from each other, the first and the third terms in (7.30) are zero. In fact, all the matrix elements would be factorized as a sum of moments $E[s_i^{\mathtt{u}} r_j]$ that are zero for each $\mathtt{u}$ if s_i and r_j are zero-mean and mutually independent. It means the solution for $\mathbf{H}_{12}$ that satisfies (7.29) does not depend on the near-end sources, and the optimization is possible even though both the near-end and the far-end sources are active at the same time (i.e., the double-talk situation). It follows then that we can substitute (7.11) into (7.30) to obtain

$$\begin{aligned} E\{[(\mathbf{W}_{11}\mathbf{H}_{12} + \mathbf{W}_{12})\mathbf{G}\mathbf{q}]^{\mathtt{u}}\mathbf{q}^H\times \\ \mathbf{G}^H\mathbf{W}_{22}^H\} = \mathbf{O}_{S\times R} \quad \forall \mathtt{u}. \end{aligned} \tag{7.31}$$

Since the far-end sources are assumed to be statistically independent, we can rewrite (7.31) as (see Appendix 7.6)

$$\begin{aligned} [(\mathbf{W}_{11}\mathbf{H}_{12} + \mathbf{W}_{12})\mathbf{G}]^{\mathtt{u}}E[\mathbf{q}^{\mathtt{u}}\mathbf{q}^H]\times \\ \mathbf{G}^H\mathbf{W}_{22}^H = \mathbf{O}_{S\times R} \quad \forall \mathtt{u}, \end{aligned} \tag{7.32}$$

where $E[\mathbf{q}^{\mathtt{u}}\mathbf{q}^H]$ is the full-ranked generalized covariance matrix of the far-end source. If $\mathbf{W}_{22}$

and $\mathbf{G}$ are not singular, then (7.32) is satisfied when

$$\mathbf{W}_{11}\mathbf{H}_{12} + \mathbf{W}_{12} = \mathbf{O}_{S\times R}. \tag{7.33}$$

Finally, assuming that $\mathbf{W}_{11}$ is known and invertible, $\mathbf{H}_{12}$ can be estimated as

$$\widehat{\mathbf{H}}_{12} = -\mathbf{W}_{11}^{-1}\mathbf{W}_{12}, \tag{7.34}$$

which agrees with (7.22). From the above derivation some important observations can be made:

(a) we assume by (7.31) that there is always a solution $\mathbf{W}_{12} = -\mathbf{W}_{11}\mathbf{H}_{12}$ that maximizes the statistical independence of the output signals in $\mathbf{y}_s$, but the exact echo path identification is possible only if $\mathbf{W}_{11}$, $\mathbf{W}_{22}$ and $\mathbf{G}$ are fully ranked. As discussed before, the singularity of $\mathbf{W}_{11}$ and $\mathbf{G}$ is a rare occurrence and can be neglected. Furthermore, since we are not interested in separating the reference signals the structure of $\mathbf{W}_{22}$ can be properly optimized to make robust the system to ill-conditioned situations;

(b) in section 7.2 we showed that AEC is possible with traditional MSE optimization even if there is a local noise as long as the reference signals and the noise are uncorrelated with each other. However, the adaptation may be instable due to the noisy *sample-wise* estimate of the gradient of the MSE. Analogously, going from (7.30) to (7.31) was made possible by assuming independence between the reference signals in $\mathbf{r}$ and the near-end source signals in $\mathbf{s}$. In other words, the optimal solution for $\mathbf{H}_{12}$ that satisfies (7.34) can be achieved through the ICA optimization even though both the near-end and the far-end sources are active at the same time. However, like for MSE-based techniques we still must be careful how the gradient term in (7.25) is estimated in order to prevent instabilities which would prevent the system to converge to the optimal solution;

(c) the natural gradient algorithm would still converge to a solution for $\mathbf{W}_{12}$ even if the non-uniqueness problem exists. However, such solution would be dependent on the changes in the far-end mixing system and a continuous and stable sample-wise or batch-wise adaptation would not be possible. Although the ICA optimization uses more information than SOS based (e.g. MSE) adaptations and thus it should be less sensitive to the effect of the non-uniqueness problem, it was shown in [97] that a decorrelation procedure may still improve its convergence. Nevertheless, we will show in Section 7.3.5 that for a scenario of multichannel audio-conference system, the low spatial correlation between the far-end room impulse responses is sufficient for reducing the fluctuations in the estimate of $\mathbf{H}_{12}$ during the ICA optimization so that the adaptation process becomes stable even without using any decorrelation procedure;

7.3.4 Connection between MSE and ICA

It is possible to compare the MSE-based and ICA-based optimal steady-state conditions to show the relationship between the two optimization rules. According to the ICA-based SBSS, the optimal cancellation of the echo paths is obtained when is satisfied (7.31) for all integer powers u. Since the near-end sources are assumed to be inactive, (7.31) can be simplified by setting $\mathbf{W}_{11} = \mathbf{I}_S$. As discussed before, we are not interested in the separation of the reference signals at the near end and thus we can impose the constraint $\mathbf{W}_{22} = \mathbf{I}_R$. Substituting $\mathbf{r} = \mathbf{Gq}$ into the result gives

$$E\{[(\mathbf{H}_{12}(\omega) + \mathbf{W}_{12}(\omega))\mathbf{r}(\omega,t)]^{\mathrm{u}}\mathbf{r}^H(\omega,t)\} = \mathbf{O}_{S\times R} \quad \forall \mathrm{u} \in \mathbb{N}. \tag{7.35}$$

Once an ICA optimization procedure converges to the steady-state solution $\mathbf{H}_{12} = -\mathbf{W}_{11}^{-1}\mathbf{W}_{12} = -\mathbf{W}_{12}$, (7.35) can be rewritten for $\mathrm{u} = 1$ (i.e., only the SOS are considered) as

$$\begin{aligned} E\{[\mathbf{H}_{12}(\omega) + \widehat{\mathbf{W}}_{12}(\omega)]\mathbf{r}(\omega,t)\mathbf{r}^H(\omega,t)\} &= \\ E\{[\mathbf{H}_{12}(\omega) - \widehat{\mathbf{H}}_{12}(\omega)]\mathbf{r}(\omega,t)\mathbf{r}^H(\omega,t)\} &= \\ E[\mathbf{e}(\omega,t)\mathbf{r}^H(\omega,t)] &= \mathbf{O}_{S\times R}, \end{aligned} \tag{7.36}$$

where $\mathbf{e} = [e_1, \cdots, e_S]^T$ is a vector of estimation errors corresponding to S microphone signals. Let us now consider the MSE optimization. In a traditional MSE-based AEC we necessarily need to assume that there are no active near-end sources. That is, we must determine the optimal value for the estimated echos in order to impose $\mathbf{y}_s = \mathbf{0}_S$, where $\mathbf{0}_S$ is a zero vector of length S. This can be achieved by minimizing the MSE in each near-end microphone channel to obtain the optimal solution:

$$\begin{aligned} \widehat{\mathbf{h}}_i(\omega) &= \underset{\widetilde{\mathbf{h}}_j}{\operatorname{argmin}}\ E[|e_i(\omega,t)|^2] \\ &= \underset{\widetilde{\mathbf{h}}_i}{\operatorname{argmin}}\ E\{|[\mathbf{h}_i(\omega) - \widetilde{\mathbf{h}}_i(\omega)]^T\mathbf{r}(\omega,t)|^2\}, \end{aligned} \tag{7.37}$$

where $e_i(\omega,t)$ is the estimation error for the i^{th} microphone channel, $1 \leq i \leq S$, $\mathbf{r} = [r_1, \cdots, r_R]^T$ is a vector of R loudspeaker (i.e., reference) signals, and $\mathbf{h}_i = [h_{i1}, \cdots, h_{ij}, \cdots, h_{iR}]^T$ is a vector of R frequency responses corresponding to the echo paths from the j^{th} loudspeaker to the i^{th} microphone. Taking the gradient of (7.37) with respect to $\widetilde{\mathbf{h}}_j$ and setting the result to zero gives

$$E[e_i^*(\omega,t)\mathbf{r}(\omega,t)] = E[e_i(\omega,t)\mathbf{r}^*(\omega,t)] = \mathbf{0}_R, \tag{7.38}$$

Comparing (7.36) with (7.37) we can state that the ICA-based SBSS also minimizes the MSEs for all microphone channels. However, the ICA-based SBSS is capable of *jointly* minimizing not only the MSE for every near-end microphone channel but also all the higher-order cross-correlations between the reference and the microphone channels through the diagonalization of the generalized covariance matrix $E[\Phi(\mathbf{y})\mathbf{y}^H]$ such that (7.35) is satisfied for all $\mathtt{u}$. Thus the ICA-based SBSS should be able to handle multiple acoustic echoes better than the traditional MSE-based MCAEC since it exploits more information about the statistics of the involved sources.

7.3.5 Effect of Constraint on $\mathbf{W}(\omega)$

We assume here that the generalized autocovariance matrix $E[\Phi(\mathbf{r})\mathbf{r}^H]$ is fully ranked and the uniqueness condition on the far-end mixing system is satisfied. To discuss the effects of constraining the global de-mixing matrix $\mathbf{W}$ on the SBSS adaptation process, we distinguish three main cases at near end:

(A) the number of the sources is equal to the number of microphones;

(B) the number of active sources is greater than the number of microphones;

(C) the number of active sources is less than the number of microphones;

Case A

For the first case we can split the analysis considering three different mixing conditions at the far end.

1. $(\mathbf{Q} = \mathbf{R})$ *The number of active sources is equal to the number of microphones.*
In this case the reference signals in **r** are guaranteed to be linearly independent and are correlated according to the impulse responses corresponding to an individual source. Since we are not interested in the separation of the far-end source signals, the matrix $\mathbf{W}_{22}$ does not need to be estimated and, for example, may be forced to $\mathbf{W}_{22} = \mathbf{I}$. However, if $\mathbf{W}_{22}$ is not constrained we would get a full benefit of the Natural Gradient algorithm since the progressive decorrelation of the reference signals during the adaptation would stabilize its convergence behavior. We remind the reader that in the ICA based on the Natural Gradient, a stable minimum is approached when the full matrix $E[\Phi(\mathbf{y})\mathbf{y}^H]$ is diagonalized. Nevertheless, when no constraint on $\mathbf{W}_{22}$ is enforced, the estimate of the matrix $\mathbf{W}_{22}$ may accidentally approach a singularity which would make impossible the inversion of $\mathbf{W}$, needed for the application of the minimal distortion principle (MDP). However, it can be observed that $\mathbf{W}$ is a block upper triangular matrix and his inverse has a diagonal structure. Thus, the correct re-scaled demixing matrix would be

computed as:

$$\widehat{\mathbf{W}}(\omega) = \begin{bmatrix} \mathrm{diag}[\mathbf{W}_{11}^{-1}(\omega)] & \mathbf{O} \\ \mathbf{O} & \mathrm{diag}[\mathbf{W}_{22}^{-1}(\omega)] \end{bmatrix} \mathbf{W}(\omega). \tag{7.39}$$

From the above structure it can be noted that, since we are not interested in the final output components corresponding to the decorrelated reference signals, we can avoid the inversion of the entire de-mixing matrix $\mathbf{W}$, which may be not invertible if $\mathbf{W}_{22}$ is singular, and determine the scaling only for the near-end sources:

$$\widehat{\mathbf{W}}_{11}(\omega) = \mathrm{diag}[\mathbf{W}_{11}^{-1}(\omega)]\mathbf{W}_{11}(\omega). \tag{7.40}$$

In other words, the scaling ambiguity only depends on the separation of the near-end sources and not on the echo paths estimation.

2. $(\mathbf{Q} > \mathbf{R})$ *The number of active sources is greater than the number of microphones.*
In this conditions $E[\Phi(\mathbf{y}_r)\mathbf{y}_r^H]$ is fully-ranked and the optimal solution for $\mathbf{H}_{12}$ can be found. If there are no constraints in $\mathbf{W}_{22}$ the ICA adaptation would tries to diagonalize the whole generalized covariance matrix. However, $E[\Phi(\mathbf{y}_r)\mathbf{y}_r^H]$ can never be diagonalized since the number of observations are less than the number of sources and the adaptation would never converge to a stable minima. Consequently, it is advised to constrain $\mathbf{W}_{22}$ to be a fixed matrix, e.g., $\mathbf{W}_{22} = \mathbf{I}_R$, in order to improve the stability of the ICA adaptation.

3. $(\mathbf{Q} < \mathbf{R})$ *The number of active sources is equal to the number of microphones.*
Theoretically the optimal solution $\widehat{\mathbf{H}}_{12}$ may not be unique when there are fewer sources than the microphones at the far end. In a real-life scenario, $E[\Phi(\mathbf{r})\mathbf{r}^H]$ should always be fully ranked since the modeling filters are generally shorter in length than the far-end impulse responses. Also, nonlinear loudspeaker distortions and additive background noises naturally decorrelate the reference signals to help improve the conditioning of $E[\Phi(\mathbf{r})\mathbf{r}^H]$. However, the ill-conditioning of $\mathbf{G}$ in the frequency domain ultimately hampers an ICA optimization procedure from converging to the true echo paths, thus the overall echo cancellation performance becomes very sensitive to the variations in the far-end and the near-end mixing systems.

Therefore exploiting the HOS cannot solve the non-uniqueness problem since it is not related to the optimization rule used for the adaptation. However, the likelihood that the gradient of the ICA optimization cost would point towards a specific region in the solution space during a gradient-descent adaptation is strongly related to the structure of the de-mixing matrix and to the characteristics of the far-end impulse responses.

According to the statistical model discussed in 5 if the far-end microphones are sufficiently spaced apart as in a realistic situation the far-end impulse responses are already sparse in the time domain. The sparseness is not directly inherited from the time domain at each frequency in the frequency domain, but it can be shown that the cross correlation in 0 between the frequency responses decreases with the microphones spacing. This is due to the assumption that under certain conditions the reverberation can be modeled as a diffuse spatially uncorrelated noise. Then we can approximate the first factor of (7.32) as (after dropping ω for notational convenience)

$$[(\mathbf{W}_{11}\mathbf{H}_{12}+\mathbf{W}_{12})\mathbf{G}]^{\mathrm{u}} \simeq (\mathbf{W}_{11}\mathbf{H}_{12}+\mathbf{W}_{12})^{\mathrm{u}}\mathbf{G}^{\mathrm{u}}, \tag{7.41}$$

which can be derived as in Appendix 7.6 by considering $\mathbf{W}_{11}$, $\mathbf{H}_{12}$ and $\mathbf{W}_{12}$ to be constant matrices and $\mathbf{G}$ a matrix of zero-mean independent random variables, taking the expectation of $[(\mathbf{W}_{11}\mathbf{H}_{12}+\mathbf{W}_{12})\mathbf{G}]^{\mathrm{u}}$, and estimating $E[\mathbf{G}^{\mathrm{u}}]$ by $\mathbf{G}^{\mathrm{u}}$. Also, since the far-end sources are assumed to be independent, the generalized covariance matrix $E[\mathbf{q}^{\mathrm{u}}\mathbf{q}^{H}]$ is expected to be diagonal. Thus we can approximate (7.32) as

$$\mathbf{G}^{\mathrm{u}}E[\mathbf{q}^{\mathrm{u}}\mathbf{q}^{H}]\mathbf{G}^{H} \simeq \mathrm{diag}\{E[q_i^{\mathrm{u}}q_i^{*}]\sum_j g_{ij}^{\mathrm{u}}g_{ij}^{*}\} = \mathbf{D}, \tag{7.42}$$

where * denotes the complex conjugation and $\mathbf{D}$ is a diagonal matrix. Therefore, by constraining $\mathbf{W}_{22}$ to be equal to an identity matrix $\mathbf{I}$, $\mathbf{W}$ assumes the following structure:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{O} & \mathbf{I} \end{bmatrix}. \tag{7.43}$$

Assuming for simplicity $\mathbf{W}_{11} = \mathbf{I}_S$ (i.e., no near-end source separation is performed) and using the $\mathbf{W}_{22} = \mathbf{I}_R$ constraint, (7.32) reduces to

$$E[\mathbf{y}_s\mathbf{y}_r^H] \simeq (\mathbf{H}_{12}+\mathbf{W}_{12})^{\mathrm{u}}\mathbf{D}. \tag{7.44}$$

From (7.44) it is straightforward that, under the "ideal" assumption of 0 correlation of the far-end frequency responses, with the de-mixing matrix constraint shown in (7.43) the element of the matrix $\mathbf{W}_{12}$ are independently optimized. In other words, the update direction during the gradient-descent optimization procedure for the $(i,j)^{th}$ element of $\mathbf{W}_{12}$ does not depend on the other elements that are related to different echo paths, and hence the effect of the non-uniqueness problem is alleviated. Nonetheless, it should pointed out that the diagonal constraint does not mathematically solve the non-uniqueness problem since (7.42) is only an ideal approximation. However, since the contribution of the off-diagonal terms in (7.42) can be neglected for many frequencies, the ideal analysis indicates that the constraint would ensures that the global iterative solution for $\mathbf{W}_{12}$ is attracted more likely to a region very close to the point corresponding to the true echo paths. In other words, the constraint globally binds the solution space of the time-domain filters related to $\mathbf{W}_{12}$ reducing in the adaptation the fluctuations of the gradient direction, which may worse the overall misalignment.

Case B, $P < S$

If $P < S$ the mixing system is not fully ranked and it may not be possible to estimate an invertible demixing matrix corresponding to the inverse of $\mathbf{H}_{11}$. However, since in the Natural Gradient no matrix inversion is needed , the adaptation may still converge to a singular matrix which would minimize the output dependence of the estimated near-end sources $\mathbf{y}_s$ from the outputs references $\mathbf{y}_r$. In other words, the SBSS can be still applied for cancelling the echos even if is not possible to separate the near-end sources.

Case C, $P > S$

In this case the matrix $\mathbf{H}_{11}$ is not square and the Natural Gradient can not converge to any $S \times S$ demixing matrix able separate the original near-end sources. Also in this case during the adaptation $\mathbf{W}_{11}$ may accidentally approach to the singularity which would hamper the use of the MDP for the correction of the scaling ambiguity. In real-world scenarios the singularity is a very rare occurrence since the number of independent acoustic sources is in general greater than the number of the microphones (e.g. a secondary reflections can be consider as further independent sources). Furthermore, we remind that at near-end also the loudspeakers represents some independent sources. Thus, the singularity of $\mathbf{W}$ would occur only when both the near-end sources and the echos are absent, which is clearly a degenerate case. Consequently, to prevent instabilities, the global adaptation needs to be stopped when the reference signals are silent. Although in real-world the singularity of $\mathbf{W}$ is a rare occurrence, we remind that regularized variants of the Natural Gradient, that improve the stability of the adaptation, are available in literature (e.g. Flexible ICA proposed in [17]).

7.4 Algorithmic design and related issues

7.4.1 Online Implementation of SBSS

In the discussion in chapter 3 a batch implementation of the BSS system was considered. Here we extend the model for the SBSS system considering two adaptation structures: 1) on-line 2) batch on-line.

Online Adaptation

For the frequency-domain on-line estimation of $\mathbf{W}$ we consider the STFT observations of the output sources $\mathbf{y}$. By substituting the iteration index n with the current time index t, and the expectation in (7.25) with the instantaneous generalized covariance matrix $\Phi(\mathbf{y})\mathbf{y}^H$, the ICA solution is updated with the incoming data by iterating over the following formulas:

$$\mathbf{y}(\omega, t) = \mathbf{W}_t(\omega)\mathbf{x}(\omega, t), \tag{7.45}$$

$$\mathbf{W}_{t+1}(\omega) = \mathbf{W}_t(\omega) + \eta\{\mathbf{I}_{S+R} - \Phi[\mathbf{y}(\omega, t)]\mathbf{y}(\omega, t)^H\}\mathbf{W}_t(\omega), \tag{7.46}$$

where η is the adaptation step-size of the online adaptation which should be opportunely adapted to improve the steady-state performance avoiding instabilities. As for the algorithm presented in chapter 4, we propose to use the scaling normalization in [24]. The need of such a normalization becomes even more relevant when the constraint $\mathbf{W}_{22} = \mathbf{I}_R$ is enforced. In fact, the matrix $\mathbf{W}$ is re-scaled by the intrinsic normalization effect of the natural gradient algorithm that regularizes the convergence behavior and ensures the convergence of the generalized covariance matrix, i.e.,

$$\Phi[\mathbf{y}(\omega, t)]\mathbf{y}(\omega, t)^H \rightarrow \mathbf{I}_{S+R}, \tag{7.47}$$

However, if the constraint $\mathbf{W}_{22} = \mathbf{I}_R$ is enforced, the normalization effect does not apply to $\mathbf{W}_{22}$, and hence the norm of the de-mixing matrix may increase and possibly lead to divergence. According to the matrix constraints the scaling normalization is imposed as follows:

$$\mathbf{W}_{22} = \mathbf{I}_R, \;\; \Delta\mathbf{W}_{21} = \mathbf{O}_{R\times S}, \;\; \Delta\mathbf{W}_{22} = \mathbf{O}_{R\times R}, \tag{7.48}$$

where $\Delta\mathbf{W}_{21}$ and $\Delta\mathbf{W}_{22}$ are the sub-matrices relative to the gradient

$$\begin{aligned}\Delta\mathbf{W}_t(\omega) &= \left\{\mathbf{I} - \frac{1}{d(\omega, t)}\Phi[\mathbf{y}(\omega, t)]\mathbf{y}(\omega, t)^H\right\}\mathbf{W}_t(\omega) \\ &= \begin{bmatrix} \Delta\mathbf{W}_{11}(\omega) & \Delta\mathbf{W}_{12}(\omega) \\ \Delta\mathbf{W}_{21}(\omega) & \Delta\mathbf{W}_{22}(\omega) \end{bmatrix}.\end{aligned} \tag{7.49}$$

and d is an inverse scaling factor. After the calculation of (7.49) with the constraints in (7.48), the matrix $\mathbf{W}$ is updated by

$$\mathbf{W}_{t+1}(\omega) = c(\omega, t)[\mathbf{W}_t(\omega) + \eta\Delta\mathbf{W}_t(\omega)], \tag{7.50}$$

where c is the scaling normalization. The scaling c and d are computed as in [24]

$$d(\omega, t) = \frac{1}{S+R} \sum_i \sum_j |\phi y_{ij}(\omega, t)|, \quad c(\omega, t) = \frac{1}{h[d(\omega, t)]} \tag{7.51}$$

where ϕy_{ij} is the generic element of the covariance matrix $\Phi(\mathbf{y})\mathbf{y}^H$ and $h(\cdot)$ is the inverse function of the magnitude of $y^*\phi(y)$ (i.e. $h[y^*\phi(y)] = |y|$) with respect to a chosen nonlinear function $\Phi(\cdot)$.

Batch-Online Adaptation

Although an online optimization is preferred for its ability in tracking the changes over time, the statistical bias in the instantaneous covariance matrix may degrade the stability of the overall adaptation. Thus it is preferred to adopt batch-online strategies to define a better trade-off between stability and adaptiveness. In this case the higher-order correlations between the de-mixed output signal components is estimated over a certain number of time observations. Accordingly, (7.45) and (7.46) are then modified as

$$\begin{aligned} \mathbf{y}_n^{(b)}(\omega, t) &= \mathbf{W}_n(\omega)\mathbf{x}^{(b)}(\omega, t), && (7.52) \\ \mathbf{W}_{n+1}(\omega) &= \mathbf{W}_n(\omega) + \\ &\qquad \eta\{\mathbf{I}_{S+R} - E[\Phi(\mathbf{y}_n^{(b)}(\omega, t))\mathbf{y}_n^{(b)}(\omega, t)^H]\}\mathbf{W}_n(\omega) && (7.53) \end{aligned}$$

where $\mathbf{x}^{(b)}$ is a vector of input signals in the b^{th} batch and $\mathbf{y}_n^{(b)}$ is a vector of output signals obtained at the n^{th} iteration in the b^{th} batch. Then in a batch-online implementation, the solution for $\mathbf{W}$ is recursively refined for a certain number of iterations by using the expected generalized covariance matrix $E[\Phi(\mathbf{y}_n^{(b)})(\mathbf{y}_n^{(b)})^H]$.

The expectation operator $E\{\cdot\}$ may be approximated by averaging over time in a batch adaptation procedure, assuming the mixing processes and the sources to be stationary in each batch. On the other hand, it is estimated through a moving average procedure in a batch-online approach:

$$\begin{aligned} E\{\Phi[\mathbf{y}_n^{(b)}(\omega, t)\}\mathbf{y}_n^{(b)}(\omega, t)^H] = \\ &\mu E\{\Phi[\mathbf{y}^{(b-1)}(\omega, t)]\mathbf{y}^{(b-1)}(\omega, t)^H\} + \\ &(1-\mu)A\{\Phi[\mathbf{y}_n^{(b)}(\omega, t)]\mathbf{y}_n^{(b)}(\omega, t)^H\}, \qquad (7.54) \end{aligned}$$

where μ is a smoothing parameter that controls the averaging across batches, $E\{\Phi[\mathbf{y}^{(b-1)}](\mathbf{y}^{(b-1)})^H\}$ is the generalized covariance matrix estimated from the previous batch and $A\{\cdot\}$ is the time-averaging operator. However, because the full ergodicity and stationarity of the random pro-

Table 7.1: Outline of the proposed SBSS algorithm.

1. Generalized covariance matrix estimation
$\mathbf{y}(k,l) = \mathbf{W}(k)\mathbf{x}(k,l)$ $E\{\Phi[\mathbf{y}(k,l)]\mathbf{y}(k,l)^H\} = A\{\Phi[\mathbf{y}(k,l)]\mathbf{y}(k,l)^H\}$ Computation of the scaling factors $c(k)$ and $d(k)$ according to [24]
2. Adaptation with the $\mathbf{W}_{22} = \mathbf{I}$ constraint
$\Delta\mathbf{W}(k) = \{\mathbf{I}_{S+R} - \frac{1}{d(k)}E[\Phi(\mathbf{y}(k,l))\mathbf{y}(k,l)^H]\}\mathbf{W}(k)$ $\mathbf{W}_{22} = \mathbf{I}_R, \Delta\mathbf{W}_{21} = \mathbf{O}_{R\times S}, \Delta\mathbf{W}_{22} = \mathbf{O}_{R\times R}$ $\mathbf{W}(k) = c(k)[\mathbf{W}(k) + \eta\Delta\mathbf{W}(k)]$

cesses cannot be guaranteed, the higher-order correlations cannot be estimated accurately by a moving average model. To avoid this drawback we suggest to use an alternative approach proposed in [57] that consists of the approximation

$$E\{\Phi[\mathbf{y}_n^{(b)}(\omega,t)]\mathbf{y}_n^{(b)}(\omega,t)^H\} \simeq A\{\Phi[\mathbf{y}_n^{(b)}(\omega,t)]\mathbf{y}_n^{(b)}(\omega,t)^H\}, \tag{7.55}$$

where the output signal components in $\mathbf{y}_n(\omega,t)$ are obtained by initializing the iteration in (7.52) and (7.53) with the matrix obtained at the $(b-1)^{th}$ batch.

7.4.2 Proposed SBSS Algorithm

We describe here a practical implementation of the proposed batch-online SBSS algorithm described in Section 7.4.1 which is later used for the performance assessment.

The time-domain microphone and reference signals in $\mathbf{x}(t) = [\mathbf{x}_s^T(t), \mathbf{x}_r^T(t)]^T$ are transformed through the STFT

$$\mathbf{x}(k,l) = STFT[\mathbf{x}(t)], \tag{7.56}$$

where k is the frequency bin index and l is the block index in time. Signals are divided into non-overlapping batches of STFT blocks. The estimate of the de-mixing matrix $\mathbf{W}(k)$ for each frequency bin is adapted in each batch by recursing over the algorithm summarized in Table 7.1 for n_{max} iterations. The adaptation of the de-mixing matrix across batches is summarized in a pseudo-code in Table 7.2.

Table 7.2: Procedure for adaptation of the de-mixing matrix.

```
$\widehat{\mathbf{W}}(k) = \mathbf{I}_{S+R}$;
while $b$
    for k=1 to $N_k$
        $\mathbf{W}(k) = \widehat{\mathbf{W}}(k)$
        for n=1 to $n_{max}$
            refine $\mathbf{W}(k)$ as in Table 7.1
        end for
        $\widehat{\mathbf{W}}(k) = \mathbf{W}(k)$
        $\mathbf{W}_{11}(k) = \mathrm{diag}(\mathbf{W}_{11}^{-1}(k))\mathbf{W}_{11}(k)$
        Solve permutation for $\mathbf{W}_{11}(k)$
    end for
end while
```

Note, in this work we are not interested in solving the permutation ambiguity problem for $\mathbf{W}_{11}$ which still exists and needs to be solved. Techniques based on the TDOA information such as the SCT (desribed in chapter 5) would be a reasonable choice even for the SBSS framework.

Stability control for $\mathbf{W}(k)$

The above discussed scaling normalization introduces *a posteriori* unit norm constraint which makes the adaptation stable in sense of the Frobenius norm of the demixing matrix $\mathbf{W}(k)$. Furthermore, the normalization of the gradient increases the convergence rate of the adaptation which becomes less dependent on the demixing matrix initialization and on the source magnitude. However, since the normalization is applied according to the norm of the whole covariance matrix, once the adaptation converged to a stable minima, it may happen that the norm of the submatrix $\mathbf{W}_{22}(k)$ is progressively reduced approaching to the degenerate case of $\mathbf{W}_{22}(k) = \mathbf{O}_{R\times R}$. In fact, since at convergence the norm of the generalized covariance matrix becomes unitary, the scaling normalization does not longer apply. In this conditions, if there is a change in the mixing system and the magnitude of the reference signals is much smaller than the output signals, the system would converge to the new solution with a very slow convergence. To avoid this drawback we introduce a stability control based on the determinant of $\mathbf{W}_{22}(k)$. If the determinant magnitude of $\mathbf{W}_{22}$ is smaller than a certain threshold (e.g. 0.1), $\mathbf{W}$ is rescaled applying a unit norm constraint to each row of the matrix as:

$$\mathbf{w}_i = \frac{\mathbf{w}_i}{||\mathbf{w}_i||_1} \tag{7.57}$$

where $\mathbf{w}_i$ indicates the vector corresponding to the *i-th* row of the matrix.

7.5 Experimental evaluation

7.5.1 Evaluation Methods

In traditional MSE-based AEC, the identification of the echo path is performed by minimizing the energy of the estimation error in (7.5). The estimation error at time t is estimated by using the previous L taps from the reference signal r during the adaptation process. Thus $\widehat{h}$ is constrained to be causal. On the other hand, in the batch-online SBSS, the de-mixing matrix $\mathbf{W}$ is estimated by evaluating the generalized covariance of the de-mixed signals over time observations within a batch of STFT blocks. Consequently, the estimated de-mixing matrix does not generally correspond to causal filters. Therefore, for a correct measurement of the misalignment, we need to transform the obtained filters from an acausal to a causal one by a circular rotation of $L/2$ taps:

$$\widehat{\mathbf{H}}(t) = IFFT[-\mathbf{W}_{11}^{-1}(\omega)\mathbf{W}_{12}(\omega)], \tag{7.58}$$

$$\widehat{\mathbf{H}}_{causal}(t) = \text{shift}[\widehat{\mathbf{H}}(t), L/2], \tag{7.59}$$

where $\widehat{\mathbf{H}}$ represents a set of time-domain filters corresponding to all possible echo paths.

The causality of the estimated filters depend strictly on the rank of the far-end response matrix $\mathbf{G}$. If $\mathbf{G}$ has full rank, the optimal solution for the echo paths can be obtained regardless of the adaptation technique. In such a case, the time-domain filters in $\widehat{\mathbf{H}}$ are already causal. On the other hand, when the rank of $\mathbf{G}$ is not full, the ICA optimization procedure can converge to a solution for $\mathbf{W}_{12}$ whose physical interpretation does not match with the true echo paths represented by $\mathbf{H}_{12}$. In this case the corresponding time-domain filters may be not causal.

The echo cancellation performance are evaluated by two metrics:

- the misalignment, defined as:

$$\text{Misalignment}(i,j) = 10\ \log_{10} \frac{||\mathbf{h}_{ij}(t) - \widehat{\mathbf{h}}_{ij}(t)||^2}{||\mathbf{h}_{ij}(t)||^2}, \tag{7.60}$$

  where $\mathbf{h}_{ij}$ is an impulse response vector corresponding to the echo path from the j^{th} loudspeaker to the i^{th} microphone and $\widehat{\mathbf{h}}_{ij}$ is the estimated impulse response vector.

- the *true* echo return loss enhancement (tERLE) is defined as

$$\text{tERLE}(i) = 10\ \log_{10} \frac{|x_i(t) - \sum_j \tilde{\mathbf{h}}_{ij}^T(t)\mathbf{s}_j(t)|^2}{|y_i(t) - \sum_j \tilde{\mathbf{h}}_{ij}^T(t)\mathbf{s}_j(t)|^2}, \tag{7.61}$$

where $\tilde{\mathbf{h}}_{ij}$ is a vector corresponding to the impulse response from the j^{th} near-end source to the i^{th} microphone.

Note, for a fair evaluation of the SBSS performance, the above two metrics must be used with caution. The tERLE must be considered the only meaningful metric for the performance evaluation since a high misalignment does not necessarily correspond to a poor echo cancellation performance. In fact the misalignment only indicates the system identification performance but a high degree of echo cancellation can still be achieved even with non-causal filters whose physical interpretation is not related to the true system identification.

7.5.2 Experimental Results

The SBSS algorithm proposed in Section 7.4.2 was evaluated on simulated and real-word data. If not differently specified, a Hanning window of 4096 taps with 75% overlap was applied to speech signals sampled at f_s = 16 kHz before taking the STFT. The impulse responses were simulated with by the Lehmann & Johansson's image source method [46]. The far-end and the near-end impulse responses were truncated to 4096 and 3200 taps, respectively. The algorithm parameters used during the adaptation are summarized in Table 7.3.

Table 7.3: Summary of parameters used during adaptation.

Parameters
$\eta = 0.1$ $\Phi(x) = \tanh(10 \cdot \lvert x \rvert)\exp(j\phi(x))$ $1 \leq n_{max} \leq$ 20 (depending on the test situation)

For a fair performance evaluation, we considered an implementation of the batch-online adaptation without any *input-output* delay. That is, the data in the b^{th} batch are processed with a filter estimated from the previous $(b-2)^{th}$ batch. Such a method takes into account both the algorithmic and the computational delays assuming an unitary real-time factor. Thus the tERLE for first two batches at the start of the adaptation are always equal to 0 dB. The tERLE and the misalignment are averaged over all possible combinations of signals and echo paths, respectively, to obtain a single performance measure for the entire system. Furthermore for a more clear graphical representation the tERLE performance has been smoothed over time with a first-order autoregressive moving average with smoothing parameter equal to 0.8.

We defined three groups of experiments:

(a) simulation of the worst-case scenario;

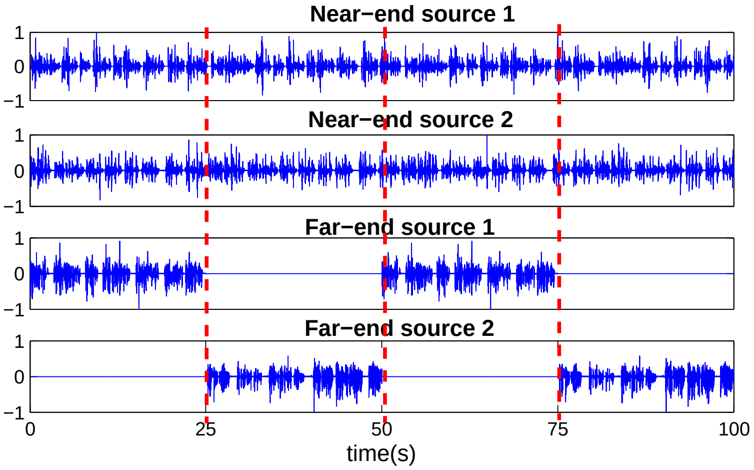


Figure 7.4: Source activities in the worst-case scenario.

(b) simulation of different far-end/near-end sources configurations;

(c) comparison between the SBSS and a standard MCAEC method on real-world data.

Simulation of the worst-case scenario

We defined a configuration with two near-end sources and microphones (i.e. P=2 and S=2) and two far-end sources and microphones (i.e. Q=2 and R=2). The theoretically worst-case scenario was generated by imposing the near-end sources to be always active and alternating the activity of two far-end sources every 25 seconds as illustrated in Figure 7.4. In this way we generate a persistent double-talk condition. According to the used STFT parameters the non-uniqueness problem is expected since $L = M = 4096$. The batch-online adaptation was evaluated with non-overlapping batches, where each batch lasted for 1 second to avoid more than one far-end source being active within a same batch and maintain the worst-case scenario. We defined the following comparisons:

- performance with and without decorrelation of the reference signals;
- performance with and without matrix constraints;
- performance with different number of ICA iterations per batch;
- performance with different values of EBR;
- performance with AWGN added at the near-end microphone signals for different values of SNR;
- performance for different FFT frame size;

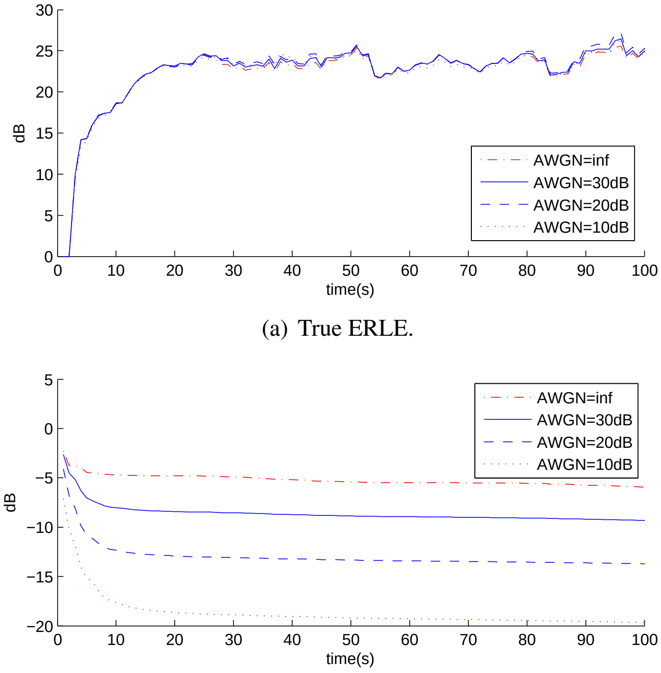
- performance for different reverberation time and fixed FFT frame size;
- performance for different values of the ICA step-size η;
- performance for different STFT frame size and iterations number.

Figure 7.5 shows the SBSS performance when additive white Gaussian noise (AWGN) is used to decorrelate the reference signals, where “AWGN=inf” corresponds to the case of no decorrelation procedure. The decorrelation level is measured in terms of the signal-to-noise ratio (SNR). As expected, the misalignment is reduced as the SNR is decreased to better decorrelate the reference signals. The misalignment converges to a relatively low value regardless of the presence of the near-end source signal. However, we note that the tERLE is almost equivalent for any value of SNR. In fact, the optimal solution obtained through SBSS is not always equivalent to the actual echo paths, and an effective echo cancellation is possible even if the reference signals are linearly dependent. Note, here we added an AWGN noise only to show that the SBSS performance are insensitive to the decorrelation of the reference signals. However, we do not suggest the AWGN as a good method for decorrelation since more recent advanced techniques, with less a audible degradation, are available in literature.

Figure 7.6 shows the performance of SBSS with and without the $\mathbf{W}_{22} = \mathbf{I}_R$ constraint. As the two far-end sources alternate in activity every 25 seconds, the unconstrained SBSS evidently displays a large degradation in the tERLE since the estimation of $\mathbf{W}_{12}$ depends on the far-end mixing condition. On the other hand the constrained SBSS is much more stable and the tERLE rapidly converge to a value of about 23 dB while the misalignment slowly decreases. A deeper understanding can be obtained observing the behavior of the unconstrained SBSS during the first 25 seconds. At beginning the misalignment is considerably high even though the echo cancellation performance is acceptable. This means that the SBSS algorithm converges to a solution that is strongly dependent on the far-end mixing system and the solution does not have to make a physical sense. However, without the constraint, the converge direction of the natural gradient is more dependent on the change of the far-end mixing system and a large degradation in tERLE is observed when the far-end source activity changes. Figure 7.7 shows the performance of SBSS with various number of iterations per batch. The convergence rate is directly proportional to the number of the iterations. However, as the number of iterations is decreased, the tERLE converge asymptotically to a higher value. Therefore a trade-off a trade-off between the convergence rate and the steady-state performance is needed.

Figure 7.8 shows the performance for different values of EBR between the loudspeakers and the near-end signals. Indicating with $\langle\cdot\rangle_t$ the average operator over t, the EBR is defined as:

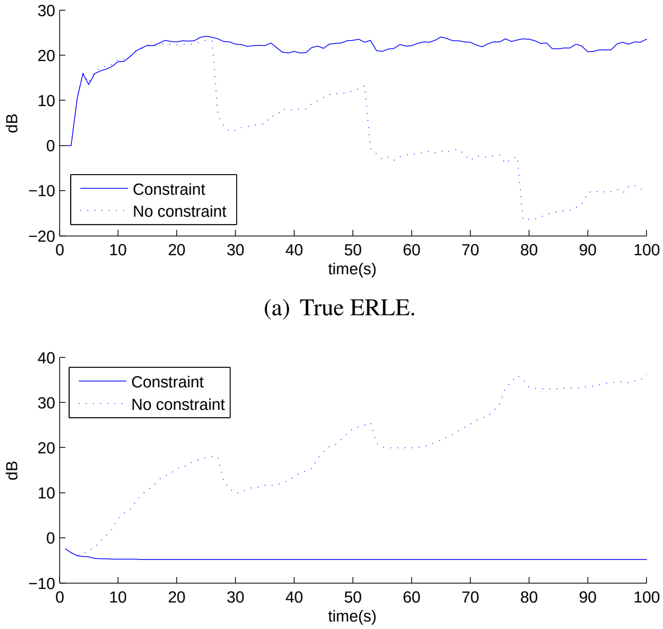
$$EBR = 10\log\frac{\langle r_i^2(t)\rangle_t}{\langle s_j^2(t)\rangle_t}, \quad \forall i, j \tag{7.62}$$



(a) True ERLE.

(b) Misalignment.

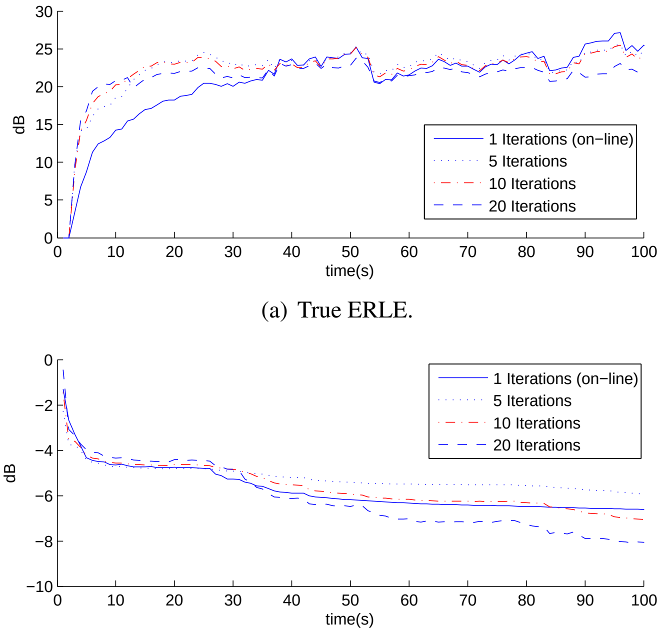
Figure 7.5: Performance of SBSS after using AWGN to decorrelate the reference signals (with de-mixing matrix constraints and 5 iterations).



(a) True ERLE.

(b) Misalignment.

Figure 7.6: Performance of constrained and unconstrained SBSS (with 5 iterations).



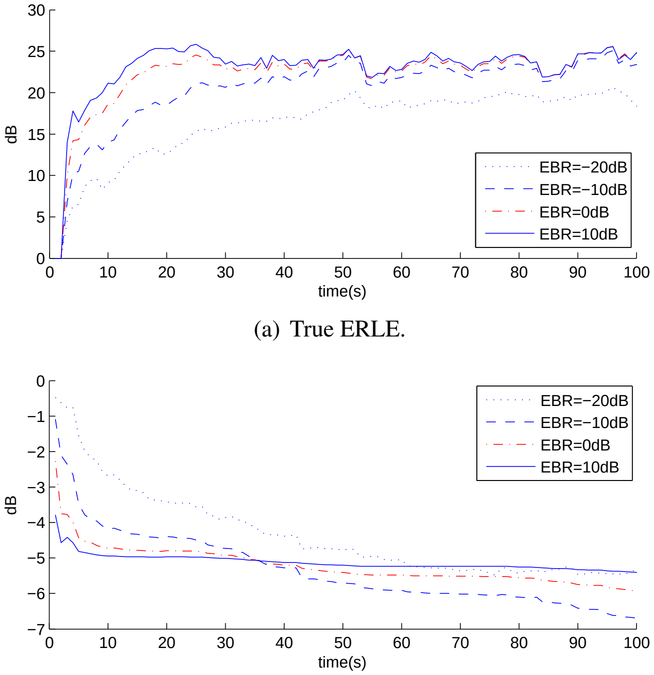
(a) True ERLE.

(b) Misalignment.

Figure 7.7: Performance of SBSS with different number of iterations (with de-mixing matrix constraint).

where r_i and s_j are the generic *i-th* and *j-th* reference and near-end signals, respectively. The high performance, even when the EBR is at -20 dB indicates the almost insensitivity of the SBSS to the presence of very loud near-end sources. We believe that the effective self-normalization introduced by the scaled Natural Gradient is the responsible of such impressive performance and worth to be better investigated in future.

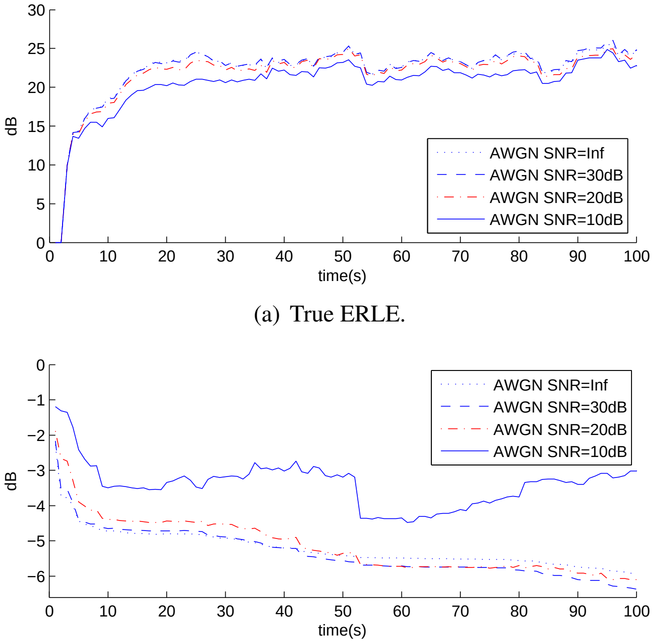
Figure 7.9 shows the SBSS performance when uncorrelated AWGN noise is added to the near-end microphones. Note, the performance does not vary considerably even with a very low SNR. It is worth noting that SBSS intrinsically consider the presence of an interfering noise which can be either a near-end source or an uncorrelated noise. As a particular case, when only 1 microphone at near-end is used the noise and all the near-end sources can be considered as a single interfering source with given pdf. Therefore, the effectiveness of the SBSS depends also on the correct choice of the ICA contrast function $\Phi(x)$. Figure 7.11 shows the SBSS performance with different FFT frame sizes. According to the property of the Fourier transform, the mixing system can be approximated at each frequency by a linear instantaneous model as long as the windows size of the STFT is sufficiently larger than the reverberation time. Therefore, the performance increases with the size of the FFT frame. Analogously, for a fixed FFT frame size (4096 points) the performance gets worse as the reverberation is increased (see Figure 7.10)



(a) True ERLE.

(b) Misalignment.

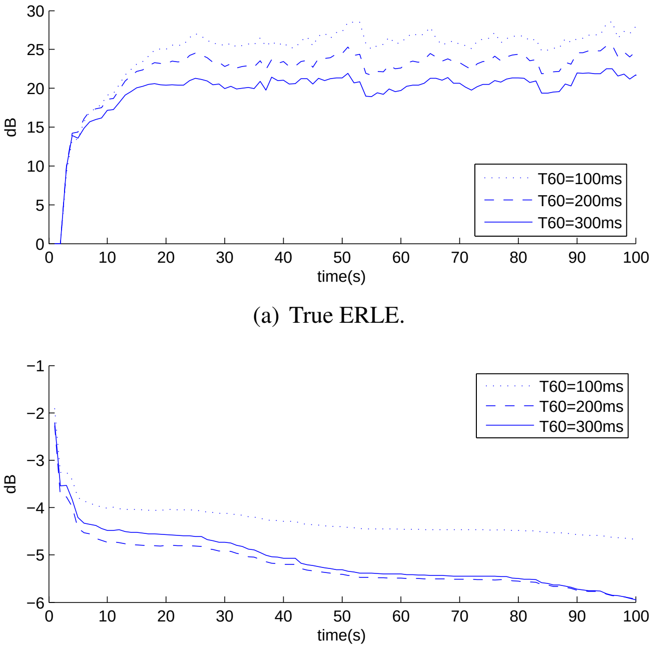
Figure 7.8: Performance of SBSS with different EBR (with de-mixing matrix constraint and 5 iterations).



(a) True ERLE.

(b) Misalignment.

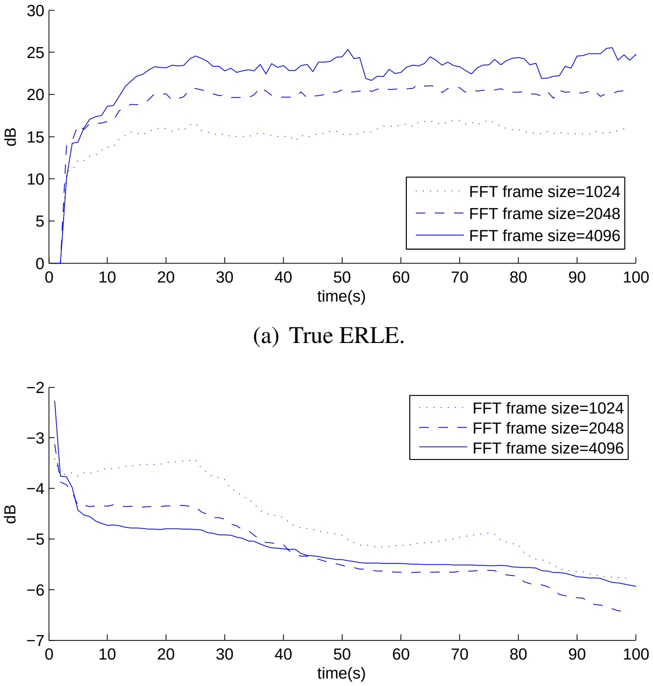
Figure 7.9: Performance of SBSS against AWGN noise at near-end microphones (with de-mixing matrix constraint and 5 iterations).



(a) True ERLE.

(b) Misalignment.

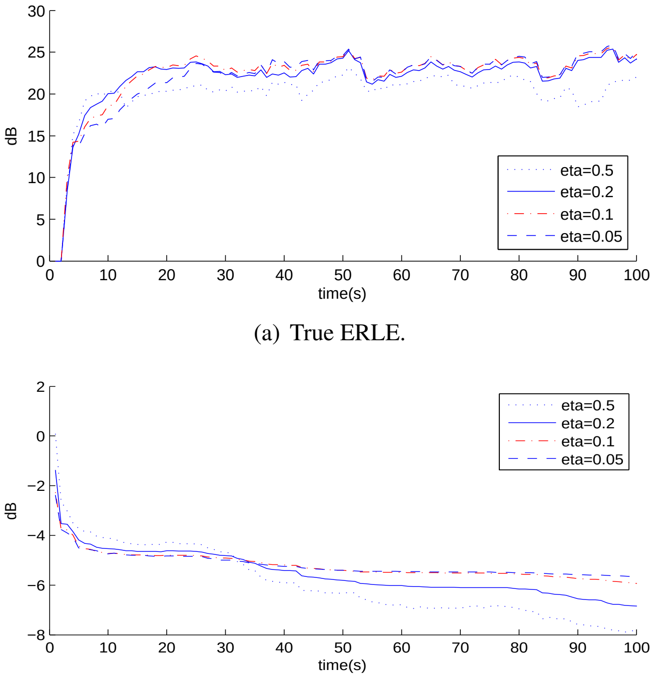
Figure 7.10: Performance of SBSS against reverberation time T_{60} (with de-mixing matrix constraint and 5 iterations).



(a) True ERLE.

(b) Misalignment.

Figure 7.11: Performance of SBSS for different FFT frame size (with de-mixing matrix constraint and 5 iterations).



(a) True ERLE.

(b) Misalignment.

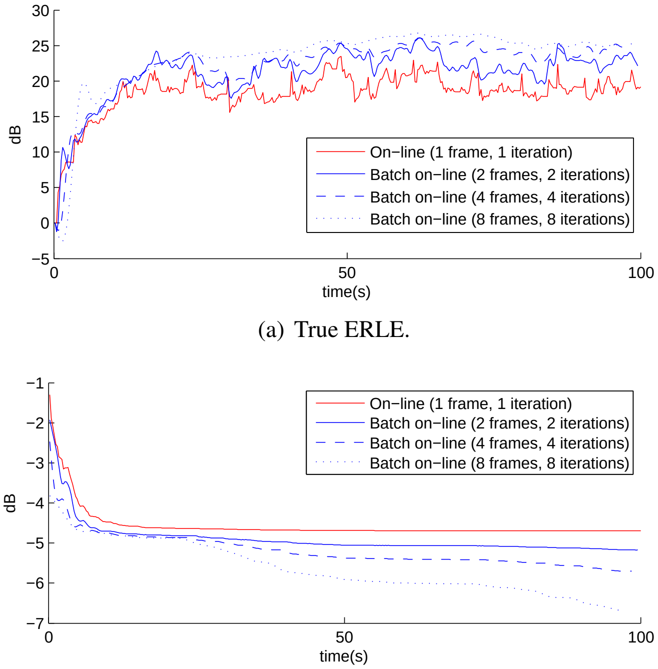
Figure 7.12: Performance of SBSS for different ICA step-size η (with de-mixing matrix constraint and 5 iterations).

Figure 7.12 shows the SBSS performance for different values of the ICA step-size η. Similarly to the influence of number of iterations, the converge rate increases with the step-size at cost of the steady-state performance.

Figure 7.13 shows the SBSS performance for different values of number of iterations and STFT frames used at each batch. We move from the on-line implementation, i.e. covariance matrix $\mathbf{\Phi}$ computed with 1 frame and adapting (7.50) for 1 iteration, to a batch on-line implementation, i.e. covariance matrix averaged over more than 1 frame and adapting (7.50) for more than 1 iteration). 7.10). Similarly to the batch on-line frequency-domain BSS in [57], it can be noted that the performance increases as we move from an on-line to a batch on-line strategy. However using batches with a too large number of frames may reduce the adaptability of the system to variation of the mixing system. Therefore, according to the application domain a trade-off between convergence rate and the steady-state performance must be defined.

Simulation of different far-end/near-end sources configurations

In this set of experiments we simulated a more realistic scenario where both the number of the near-end and far-end sources changes over time and we also analyzed the behavior of SBSS for more complex channel configurations. We simulated at near-end and far-end the presence of 3



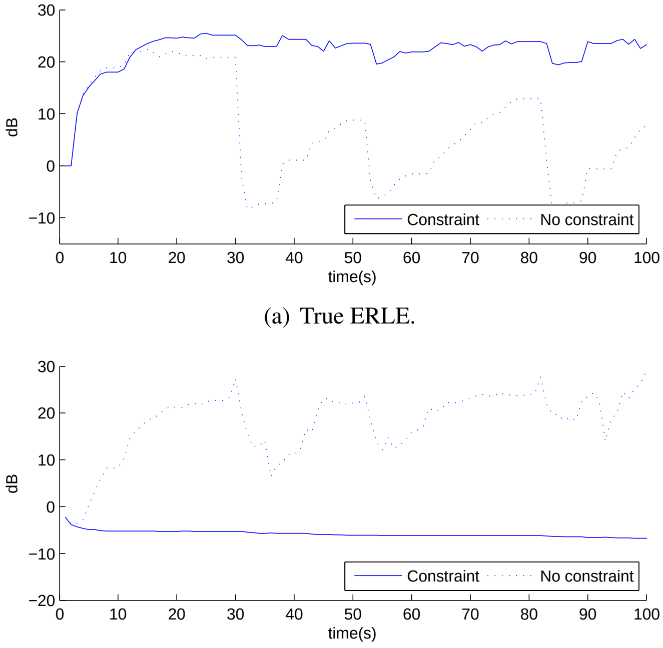
(a) True ERLE.

(b) Misalignment.

Figure 7.13: Performance of SBSS with on-line and batch on-line implementations (with de-mixing matrix constraint and 5 iterations).

sources which activity can be totally or partially overlapped (i.e. $0 \leq P \leq 2$ and $0 \leq Q \leq 3$). The number of near-end and far-end microphones are still equal to 2. In this case, varying the number of active sources at near-end and far-end we can show the behavior of SBSS for a general configuration of source and microphone numbers (see the possible cases in 4.2). At near-end an AWGN noise of SNR=80dB was added to the microphones. Figure 7.14 shows the performance of SBSS with and without the $\mathbf{W}_{22} = \mathbf{I}_R$ constraint. Still, similarly to the case discussed in Figure 7.6, with the matrix constraint the adaptation stably converges regardless the variation of the source number at near-end and far-end.

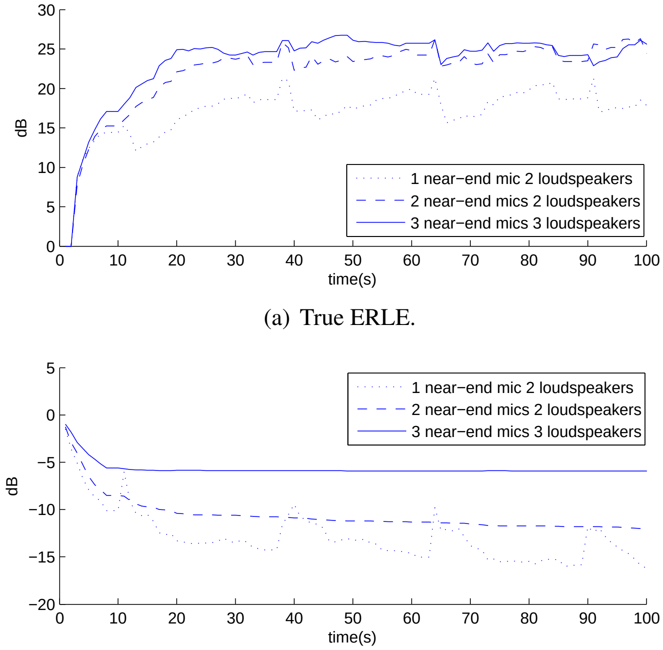
In a second test we considered three different mixing conditions: 1) 1 near-end microphone and 2 far-end microphones; 2) 2 near-end microphones and far-end microphones; 3) 3 near-end microphones and 3 far-end microphones. The sequence of the source activity (at near-end and far-end) is the same as for the previous experiment. Figure 7.15 shows the performance for tERLE and misalignment. Note, the performance are better with a higher number of near-end microphones while the misalignment is lower when a reduced number of microphones. This result confirms that for the SBSS the performance of the echo suppression does not correspond to the correct system identification. In fact, when many near-end microphones are used the full multichannel nature of the SBSS adaptation allows to exploit also spatial information by properly adapting the demixing matrices $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$. To better clarify this aspect we show



(a) True ERLE.

(b) Misalignment.

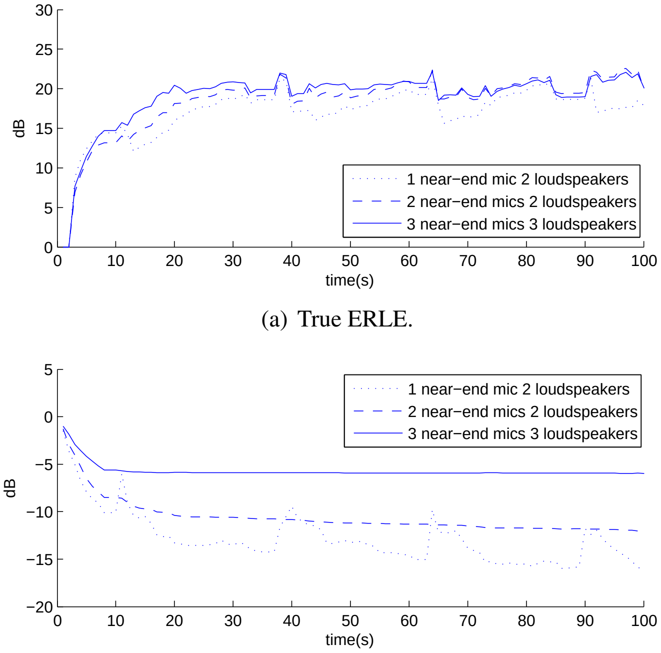
Figure 7.14: Performance of SBSS against variation of the active source number (with de-mixing matrix constraint and 5 iterations).



(a) True ERLE.

(b) Misalignment.

Figure 7.15: Performance of SBSS for different microphone configurations (with de-mixing matrix constraint and 5 iterations).



(a) True ERLE.

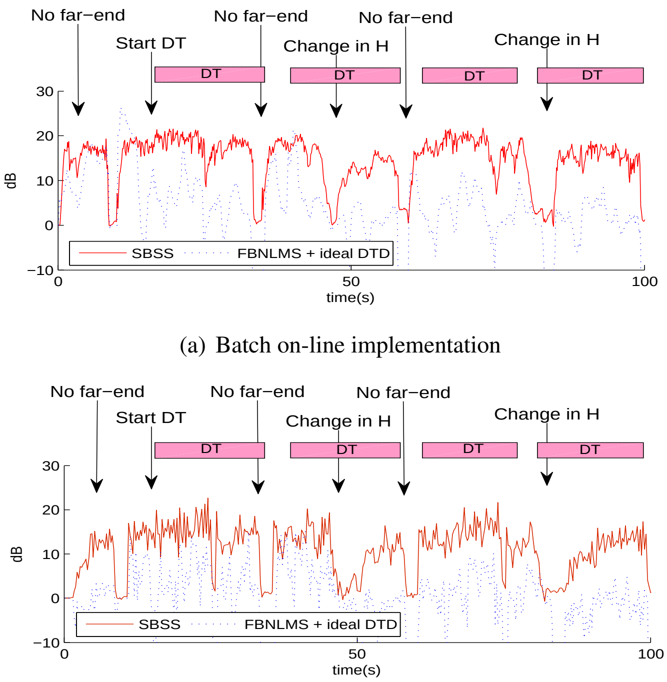
(b) Misalignment.

Figure 7.16: Performance of SBSS for different microphone configurations without applying the near-end source separation (with de-mixing matrix constraint and 5 iterations).

in Figure 7.16 the SBSS performance when the near-end source separation is not applied. The matrices $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ are multiplied by $\mathbf{W}_{11}^{-1}$ which means that the echo cancellation is obtained by means of the causal mixing filters estimated as in (7.58)-(7.58). As expected, the performance for the configurations with 2 and 3 near-end microphones becomes similar to the single microphone case. In fact, removing the effect of $\mathbf{W}_{11}$ it is almost equivalent to apply the SBSS to each near-end channel separately.

Comparison between the SBSS and a standard MCAEC method on real-world data

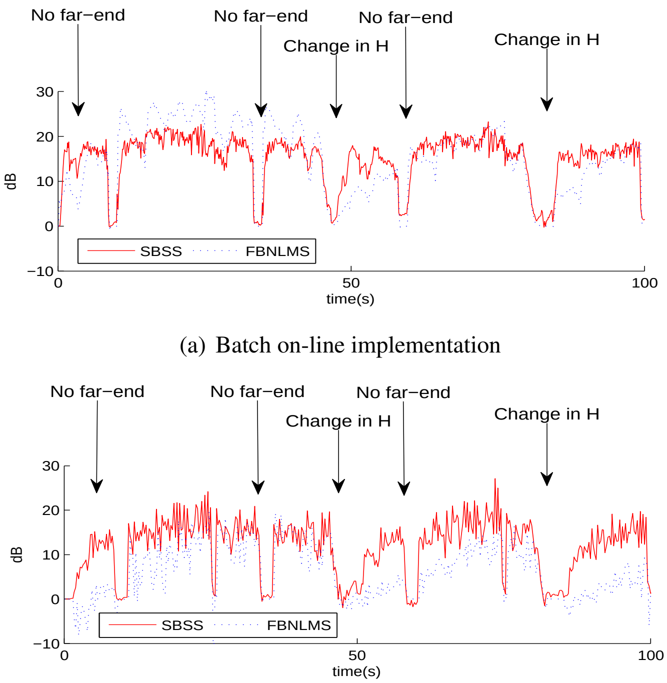
In these experiments we compare the SBSS with a conventional multichannel AEC based on the MSE. A batch implementation of a frequency-block NLMS has been implemented and the adaptation was applied only to the time intervals with no double-talk. In this experiment real-world data were recorded in a room with a reverberation time of approximatively $T_{60} = 200ms$ and the algorithms have been tested for the configuration with two near-end sources and two near-end loudspeakers (i.e. two far-end microphones). At far-end and near-end the microphones record a variable number of sources over time, $0 \leq P, Q \leq 2$. The near-end sources are active in the time interval indicated in the Figures 7.17 and 7.18 by the label ”DT”. The changes in the mixing conditions have been generated moving the near-end microphones approximatively after 50s and 80s. We compare the algorithms for two adaptation strategies: 1) a batch on-line adaptation with blocks of 1.024s, overlapped by shifting the blocks of 0.128s 2) an on-line STFT frame-by-frame adaptation. In both the adaptation strategies the signals were transformed in the frequency domain by a STFT applied to non overlapping Hanning windows of 4096 points (i.e. to update the demixing matrix at each block the batch on-line algorithms average the covariance matrix over 4 frames). In both the cases the solution for $\mathbf{W}$ was updated for 1 iteration per each block/frame. Figure 7.17(a) shows the tERLE performance for the batch on-line strategy. In this case no smoothing was applied to the tERLE. In the first 15 seconds, the FBNLMS converges quickly to a high tERLE but as soon as the near-end sources become active, the performance progressively degrades. On the other hand, the SBSS has similar performance in the first 15 seconds but clearly outperforms the FBNLMS during the double-talk. In particular, when the near-end is active and the mixing conditions changes over time, the FBNLMS is not able to correctly track the variation in the demixing filters while the SBSS stably converges to a new estimate for $\mathbf{W}$. We observe that since the FBNLMS adaptation is frozen during the double-talk, the FBNLMS performance is low even when the mixing condition at near-end does not change. In fact, since no decorrelation procedure was applied to the loudspeaker signals, the solution highly depends also on the far-end mixing conditions which changes over time according to the active speakers. A continuous adaptation then, is not useful to only track the variation of the near-end mixing system but also to track the variation on the far-end and mitigate the degradation due to the non-uniqueness problem. Figure 7.17(b) shows the tERLE performance for the batch on-line strategy. In this case the FBNLMS has a lower convergence rate due to the sample-wise noisy update of $\mathbf{W}$ even when the near-end sources are not active. The SBSS, has a slower convergence rate than the corresponding batch on-line implementation. However, thanks to the scaling normalization of the Natural Gradient, the adaptation is stable enough to ultimately lead to acceptable overall performance. We remind that the generalized covariance matrix used in the SBSS adaptation exploits high order source statistics. Similarly to the case of BSS, any measures of high-order source dependencies would be highly biased if the amount of data is limited. Therefore, to take benefit of the SBSS structure and achieve the best performance batch on-line adaptation are highly suggested. Finally we compare in Figure 7.18 the SBSS and FBNLMS performance in the ideal case when the near-end sources are not active. We observe that the FBNLMS, despite of the correlation between the loudspeakers, is able to achieve acceptable performance, on conditions that the filters are continuously updated. However, this is not always possible in presence of double-talk. On the other hand, the SBSS with and without the near-end sources being active has almost the same performance, which confirm the insensitivity of the adaptation to the presence of double-talk.



(a) Batch on-line implementation

(b) On-line implementation

Figure 7.17: Comparison between SBSS and FBNLMS for a real-world scenario



(a) Batch on-line implementation

(b) On-line implementation

Figure 7.18: Comparison between SBSS and FBNLMS for a real-world scenario in the ideal case of no near-end sources

7.6 Appendix

Assuming that $\mathbf{x} = \{x_j\}$ is a vector of zero-mean random variables of length N and that $\mathbf{A} = \{a_{ij}\}$ is an $N \times N$ matrix with constant elements, the statistical moment of order u for the i^{th} element of $\mathbf{Ax}$ is given by

$$E\left[(\sum_{j=1}^{N} a_{ij}x_j)^{\mathrm{u}}\right] \quad \forall i. \tag{7.63}$$

By using the multinomial expansion, (7.63) can be rewritten as

$$E\left[\sum_{\substack{l_1,l_2,\ldots,l_N \geq 0 \\ l_1+l_2+\ldots+l_N=\mathrm{u}}} \frac{\mathrm{u}!}{l_1!..l_N!} \prod_{j=1}^{N} (a_{ij}x_j)^{l_j}\right] \quad \forall i. \tag{7.64}$$

If the elements in $\mathbf{x}$ are mutually independent, (7.64) reduces to

$$E\left[\sum_j (a_{ij}x_j)^{\mathrm{u}}\right] = \sum_j a_{ij}^{\mathrm{u}} E[x_j^{\mathrm{u}}] \quad \forall i. \tag{7.65}$$

We can then generalize that

$$E[(\mathbf{Ax})^{\mathrm{u}}] = \mathbf{A}^{\mathrm{u}} E[\mathbf{x}^{\mathrm{u}}], \tag{7.66}$$

where $\mathbf{x}^{\mathrm{u}}$ and $\mathbf{A}^{\mathrm{u}}$ indicate the raising of each element of the vector $\mathbf{x}$ and of the matrix $\mathbf{A}$ to the power u. By the property of the covariance of linear combinations of variables, we know that if the random variables in $\mathbf{x}$ are independent, then given the $N \times N$ matrices $\mathbf{A}$ and $\mathbf{B}$, we have

$$E[\mathbf{A}\mathbf{x}\mathbf{x}^H\mathbf{B}] = \mathbf{A}E[\mathbf{x}\mathbf{x}^H]\mathbf{B}. \tag{7.67}$$

By using (7.64) and following the derivation of (7.66), it is possible to generalize (7.67) for higher-order moments as

$$E[(\mathbf{Ax})^{\mathrm{u}}\mathbf{x}^H\mathbf{B}] = \mathbf{A}^{\mathrm{u}}E[\mathbf{x}^{\mathrm{u}}\mathbf{x}^H]\mathbf{B}. \tag{7.68}$$

7.7 Concluding Remarks

In this chapter, we discussed on many issues related to the implementation of a semi-blind source separation (SBSS) system. A deep analysis of the structure of the SBSS optimization is presented and algorithmic issues are discussed in order to define a guideline for the implementation of robust SBSS systems. Promising results show that the SBSS is an attractive framework for MCAEC which is able to overcome the limitations of traditional systems, whose performance are degraded by the non-uniqueness and double-talk conditions. An implemented SBSS real-time system, described in chapter 9, shown that the framework is flexible and robust enough to be used for real-world applications even in difficult scenarios. Many future research directions can be identified:

- for multichannel video-conference purpose, the benchmarking of the algorithm with a full-duplex system may be considered;
- the variation of the performance with alternative ICA algorithms is worth to be analyzed; in particular it should be considered the case where the contrast functions are optimized according to the statistic of the observed reference signals;
- time-domain optimization may be considered to reduce the length of the required demixing filters;
- different scaling normalization or alternative adaptive step-size approach should be considered, in order to increase both the convergence rate and steady-state performance of the NG;
- performance should be evaluated complementing the algorithm with non-linear post processing based on the output energy correlation.

Chapter 8

Towards Distributed Blind Source Separation

A general limitation of standard methods of BSS is that the signals that are processed by the system are recorded with a single array with a relatively small aperture. If one want to increase the spatial coverage in a large environment, the centralized approach is not longer appropriate and distributed strategies must be adopted. In this chapter we present a preliminary analysis of a possible extension of the BSS applied to distributed arrays. The work has been published in [64].

8.1 Introduction

In traditional FD-BSS the estimation of the demixing filters, which maximize the statistical independence of the output signals, is performed considering all the mixtures at the same time. The mixtures are generally recorded by microphone arrays which typically have a small aperture (e.g. microphone spaced of 10-20cm). As shown in chapter 2 and 4 a FD-BSS is equivalent to a set of adaptively optimized null-beamformers. Thus, for each frequency the FD-BSS generates a beampattern which steers a spatial null in a given direction, with the effect of removing the interfering source. Large aperture arrays are preferred to separate the sources with a higher spatial resolution. Nevertheless, two major drawbacks limit the use of such arrays for BSS system:

- high spatial aliasing complicates the estimation of the source location which is useful to solve the well-known permutation problem;
- in presence of diffuse reverberation, the spatial coherence between acoustic waves of each source recorded at different microphones, decreases with the microphone spacing [27].

Hence, the capability of ICA to estimate reliable demixing filters reduces with the microphone distance.

Similarly, such drawbacks are present also in system for source localization. To overcome the limitations of large aperture arrays, distributed arrays have been exploited in the localization community [10]. Distributed arrays are formed by many small-aperture sub-arrays disseminated in all the environment. The main idea behind the distributed-array paradigm is that spatial signal processing techniques (e.g. localization based on GCC-PHAT) are applied only to the data recorded by the same sub-array and the information provided by different sub-arrays is merged a posteriori in a different domain.

For the localization of acoustic sources distributed arrays have been widely used in order to improve the localization accuracy. Resolution and spatial coverage can be increased by exploiting appropriate array geometries, without the needs of increasing the inter-microphone distances. However, to our knowledge no work has been done to bring this principle into the problem of the acoustic blind source separation (BSS). In this chapter we propose a new BSS scheme, which has been called Cooperative Wiener ICA (CW-ICA), able to perform the BSS with distributed arrays. We propose a new structure where different ICA adaptations are applied to the signals recorded by each array but are interconnected by a collaborative scheme. An ICA algorithm based on the Natural Gradient [3] and Kullback-Liebler divergence is modified in order to introduce in the adaptation *a priori* knowledge regarding to the time-activity of the sources, obtained according to Bayesian estimation theory. Such knowledge is estimated and propagated across the network in order to constrain the global adaptation to converge to a physically coherent solution, without the need to use any geometrical knowledge on the recording setup.

The chapter is organized as follows. In section 8.2, we recall the basis of frequency-domain BSS and the physical interpretation of the estimated mixing matrix and of the demixed output signals. In section 8.3 we recall the Wiener regularization presented in chapter 4 and we describe the network topology used to interconnect all the ICA blocks. In section 8.4, experimental results on simulated data confirm the validity of the theoretical reasoning. Conclusions follows at end of the chapter.

8.2 BSS as Soft-Masks Estimation

According to the definition in (3.27) (chapter 3) the ratios r_{nk}^{p} between the a_p-*th* and b_p-*th* elements of the *n-th* column of $\mathbf{W}(k)^{-1}$ are scaling invariant and, assuming the permutation problem is solved, they represent the observations of the acoustic propagation from the *n-th* source to the microphone pair (a_p, b_p).

For the sake of simplicity, let us first consider the case of two sources and two microphones. The ratios related to the propagation of each source can be derived as:

$$r_{nk} = -\frac{w_{n2}(k)}{w_{n1}(k)}, \quad r_{1k} = \frac{h_{12}(k)}{h_{22}(k)}, \quad r_{2k} = \frac{h_{11}(k)}{h_{21}(k)}$$

Then the first source estimate $y_1(k,l)$ is obtained as:

$$\begin{array}{rcl} y_1(k,l) & = & w_{11}(k)x_1(k,l) + w_{12}(k)x_2(k,l) \\ & = & w_{11}(k)x_1(k,l)\left(1 - \frac{r_1(k)}{q(k,l)}\right) \end{array} \tag{8.1}$$

where $q(k,l) = \frac{x_1(k,l)}{x_2(k,l)}$. Note that $w_{11}(k)$ can assume any value for the scaling indeterminacy of ICA. Neglecting the phase of $y_1(k,l)$ its magnitude can be expressed as:

$$\begin{array}{rcl} |y_1(k,l)| & = & |w_{11}(k)||x_1(k,l)|\left(\frac{|q(k,l)-r_1(k)|}{|q(k,l)|}\right) \\ & = & |w_{11}(k)||x_1(k,l)|D[q(k,l), r_{1k}] \end{array} \tag{8.2}$$

where $D[\cdot,\cdot]$ is a normalized Euclidean distance between the estimated propagation model and the observed output ratios. For the time-frequency point for which the input ratio is very close to the estimated propagation model, the function $D[\cdot,\cdot]$ becomes zero. Therefore, neglecting the phase, in a frequency-domain BSS each demixing matrix defines a soft mask which has the effect to extract the energy of the source of interest, removing the contribute of the interfering source identified by r_{1k}. Motivated by the intrinsic sparseness of the time-frequency representation of acoustic sources, a binary mask can be used alternatively where at each time-frequency point the non-dominant sources are floored to 0:

$$mask_n(k,l) = \begin{cases} 1, & if \quad \underset{i}{argmin}\; D[q(k,l); r_{ik}] = n \\ \\ 0, & \qquad otherwise \end{cases} \tag{8.3}$$

We showed in section 5.5.3 that for the generic case of a greater number of microphones, the mask can be defined as in (5.72). Therefore, according to (5.72) estimating the acoustic propagation of each source in multiple TDOA dimension (i.e. with respect to different microphone pairs) is sufficient to define a time-frequency mask that extracts the image of each source at each microphone without the need of a global demixing matrix. Furthermore, the separation is possible even if the models r_{nk}^p (or subsets of them) are estimated by independent ICA adaptations, assuming that for each source they can be grouped to define a global model $\mathbf{r}_{nk}$. Nevertheless, this cannot be guaranteed due to the permutation ambiguity of ICA, that is: the rows of the demixing matrix may be differently permuted and the estimated *n-th* model r_{nk}^p would not always represent the acoustic propagation of the same source. As discussed in section 5 this problem is related to the localization of multiple sources in multiple dimensions. The ambiguity does not exist if the vector of the propagation models r_{nk}^{p} is estimated by applying the same ICA adaptation to all the signals $\mathbf{x}(k,l)$ recorded by the microphones of all the P pairs. However, in a distributed case we use different ICA adaptations and a method to remove such ambiguity becomes necessary.

8.3 Cooperative Wiener-ICA

In chapter 4 it was shown that in a batch on-line adaptation the ICA estimation can be improved by a proper weighting of the instantaneous gradient. The weights in (4.7) are based on the estimation of energy of the sources. It was experimentally shown that by estimating the time activity of the sources across the frequencies the accuracy of ICA is improved since the weighting reduces the statistical bias due the intrinsic dependencies between sources, observed in a short time. Besides the frequency redundancies the Wiener weights can be estimated also by spatial redundancies. According to the definition in (4.7) each weight $mask_n(k,l)$ can be interpreted as the probability that the *n-th* source is dominant in the time-frequency point (k,l). The sequence of such probabilities is related to the activity of the sources which, for the sparseness of acoustic signals, is expected to be sparse in time-frequency domain. Therefore, it can be assumed that the time activity of the sources estimated by the Wiener weights does not depend on the selected microphone pairs but only on the probability that the sources are present at certain time-frequency point. Hence, similarly to the interconnection between ICA of different frequencies proposed in Chapter 4, the weights can be used to interconnect ICA adaptations regarding different arrays. In this case the described weighting procedure has two important effects. First of all, if the weight are sufficiently accurate the *a priori* knowledge of the source activity would improve the estimation of the gradient by a reduced statistical bias and reduced sensitivity to errors localized at particular sensors. Second of all, the *n-th* column of the gradient, which updates the matrix element related to the *n-th* source, is constrained to converge in the direction of the source identified by $\psi_n(k,l)$. Therefore, if the ICA adaptation of each array is constrained by the same weighting matrix $\mathbf{\Psi}(k,l)$, each estimated mixing matrix is likely to be affected by the same permutation. Hence, the multidimensional propagation model $\mathbf{r}_{nk}$ of each source can be estimated without any ambiguity.

The interconnection of each ICA adaptation by means of the Wiener-like regularization can be performed through different networking strategies. In this work, we propose a cooperative scheme similar to a *token ring* network, where the weighting matrix $\mathbf{\Psi}(k,l)$ is circularly propagated across the ICA stages. Figure 8.1 describes how to interconnect each block according to the proposed topology.

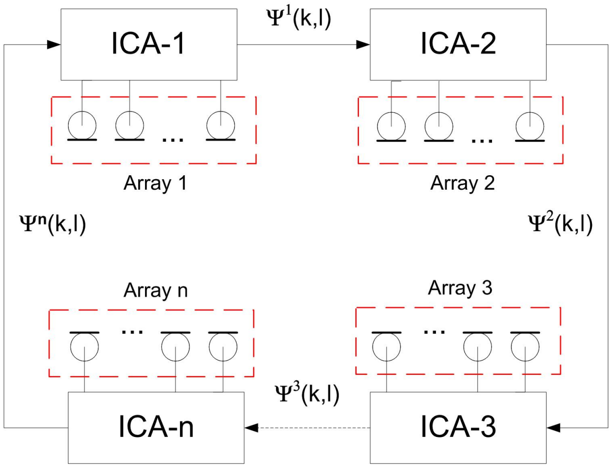


Figure 8.1: Interconnection of the network by a token ring topology.

```
for k=1 to $N_k$
    $\boldsymbol{\Psi}(k,l) = 1, \ \forall l$
    for i=1 to $N_i$
        for b=1 to $N_b$
            Compute $\mathbf{y}^b_{(i)}(k,l)$ and $\mathbf{H}^b_{(i)}(k)$
            as in (4.1)-(4.3)
            Compute $\boldsymbol{\Psi}^b(k,l)$ as in (4.5)
            $\boldsymbol{\Psi}(k,l) = \boldsymbol{\Psi}^b(k,l)$
        end
    end
end
```

Table 8.1: Pseudo code description of the Cooperative Wiener ICA.

A pseudo code of the algorithm is summarized in Table 8.1. In the given algorithm, N_k, N_i, and N_b denote the number of frequency bins, the maximum number of the ICA iterations, and the number of nodes (i.e. the number of arrays in the network), respectively. Furthermore, $\mathbf{y}^b_{(i)}(k,l)$, $\mathbf{H}^b_{(i)}(k)$ and $\boldsymbol{\Psi}^b(k,l)$ indicate the output signals, the estimated mixing matrix, and the weighting matrix related to the *b-th* node, respectively. At the first iteration, the ICA running at the first node (i.e. $b = 1$) estimates the output signals without any prior knowledge of the source activity since all the elements of the weighting matrix are set to 1. After that, the resulting output signals are used to compute the weighting matrix $\boldsymbol{\Psi}^{\mathbf{1}}(k,l)$ which is adopted to initialize the matrix $\boldsymbol{\Psi}(k,l)$ for the subsequent node. The second node uses the previously estimated weights by performing the ICA adaptation with the data provided by the second array; then, it refines the estimation of $\boldsymbol{\Psi}(k,l)$ which is propagated to the next node. The iterative process continues across all the nodes, circularly propagating and constraining the ICA adaptation of each node of the network.

The propagation of the weights in the network is expected to be effective when source power

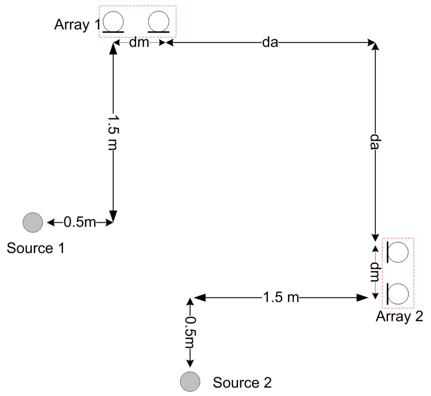
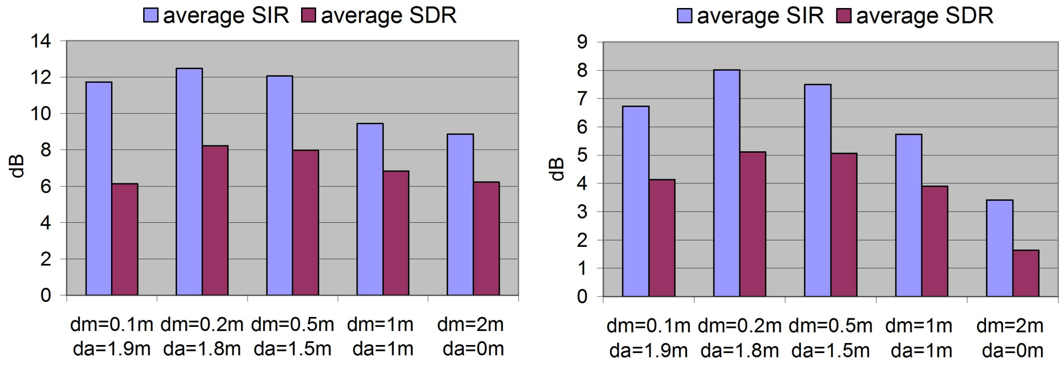


Figure 8.2: Experimental setup.



(a) $T_{60} = 150ms$

(b) $T_{60} = 300ms$

Figure 8.3: Performance comparison for different inter-microphones and inter-pair spacing.

contributions at different arrays are comparable. In some real-world scenarios, this assumption may be unrealistic and thus other network topologies deserve to be investigated in the future. In general, due to the sparseness of the acoustic sources the weights assume values close to 1 or to 0. Moreover, they represent the time-activity of the sources, which must be coherent at different arrays even when power contributions differ from one array to another.

8.4 Experimental Results

We evaluated the performance of CW-ICA according to the network topology in Figure 8.1. The impulse responses were simulated arranging sources and microphones in a room with T_{60}=300ms as shown in figure 8.2. The distances *dm* and *da* were changed in order to evaluate their impact on the global performance. Signals were sampled at f_s = 16kHz, and the time-frequency representation was obtained with a short-time Fourier transform by using Hanning windows of 2048 taps and an overlap factor of 87.5%. The permutation problem was solved with the GSCT as in section 5.5.2.

	WC-ICA	Traditional ICA
SIR	8.01	1.34
SDR	5.1	-0.25

Table 8.2: Average performance for $T_{60} = 300ms$.

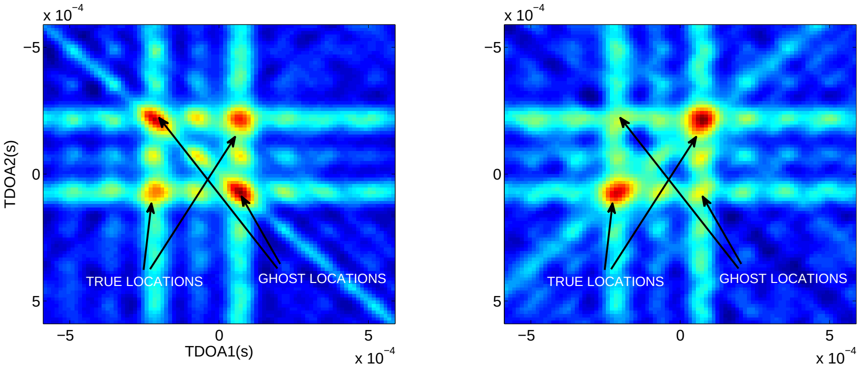


Figure 8.4: Source location likelihood obtained by means of the GSCT computed with the states $\mathbf{r}_{nk}$; states obtained by independent unconstrained ICA adaptations (case a), and states obtained by CW-ICA adaptations (case b).

.

Figure 8.3 shows the performance obtained for different microphone setups. It is interesting to observe that the maximum performance is obtained for the case da=1.8 m, dm=0.2 m. Such a configuration can be retained as a good trade-off between the spatial resolution and the resulting coherence. In fact, despite of an increased spatial resolution, increasing the inter-microphone distance (i.e. da=1 m, dm=1 m) deteriorates the global performance, due to the poor ICA accuracy. On the other hand, if the microphone distance is reduced (da=1.9 m, dm=0.1 m), the performance is also degraded due to a lower spatial resolution.

Table 8.2 compares the performance obtained when the ICA adaptations are interleaved, as described in section 8.3, and when they are performed independently. As expected, if the ICA adaptations are not connected the permutation ambiguity reduces the effectiveness of the estimated state vectors $\mathbf{r}_{nk}$ which are used to determine the binary masks. A direct evaluation of such ambiguity can also be shown by plotting the spatial likelihood of the source location by means of GSCT. As shown in Figure 8.4, the correct multidimensional spatial location cannot be determined if the ICA adaptations are not constrained by the Wiener-like regularization. In the latter case, the likehood of $ghost$ sources would increase at wrong locations (as shown in the case (a) of Figure 4).

8.5 Concluding Remarks

In this chapter, the problem of source separation was extended to the case of distributed microphone arrays. A cooperative scheme was proposed to interconnect ICA adaptations regarding different arrays of the given network. The Wiener-like regularization introduced in chapter 4 was exploited to estimate and propagate the information of the source activity across the network. Simulation experiments provided very promising results, showing that both separation and localization can be achieved without ambiguity in the proposed framework. This work was a preliminary analysis of the distributed paradigm applied to the problem of BSS which may open many new future research directions. First, the applicability to real-world scenarios must be investigated, for example in presence of sources sparsely located in large environments. Second, different network topologies can be analyzed as well as different networking strategies. On this regards, a convergence analysis of the global adaptation for different network topologies is of high interest.

Chapter 9

Real-time BSS/SBSS system implementation

9.1 Introduction

We describe here the implementation of two real-time systems for BSS and SBSS which use the techniques proposed in the previous chapters. We focused on a system design suitable for home applications such as multi-speaker videoconference and automatically speech controlled systems. Particular attention has been given to robustness and computational issues. Furthermore, a system design with a reduced cost of the hardware has been considered. The systems, developed to work on a standard laptop, have been successfully demonstrated in international conferences such as HSCMA 2008, IWAENC 2008, ICASSP 2009, ASRU 2009, ICASSP 2010. We considered two different system implementations:

- Real-time BSS: a two channel real-time BSS system.
- Real-time SBSS: a four channel SBSS system for stereo echo suppression, separation of two near-end sources and automatic recognition.

9.2 Architecture of the BSS system

Figure 9.1 shows the logical architecture of the implemented real-time BSS system. The corresponding hardware architecture is shown in Figure 9.2. Two hypercardioid microphones AKG C 400 BL are used to record two sources which are represented by two loudspeakers connected to an mp3 player. The datasheet of the used microphones are reported in Figure 9.3. Note, the intrinsic directionality of the microphones does not affect the performance of the implemented system since we assumed that the target sources are frontally located to the array. This

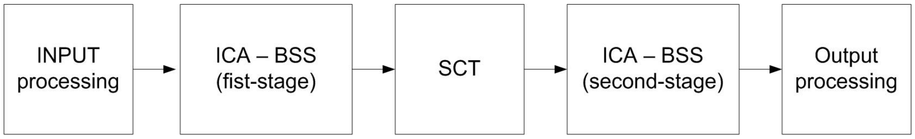


Figure 9.1: Logical architecture of the developed BSS system.

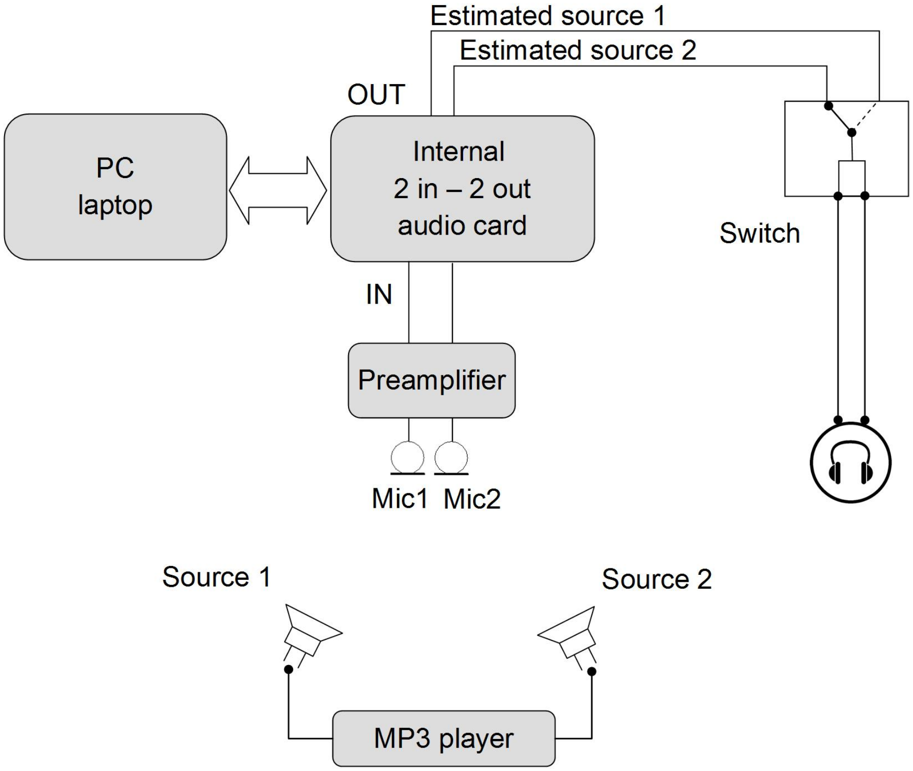


Figure 9.2: Hardware architecture of the developed BSS system.

restriction is essential since symmetric locations cannot be distinguished by a two-channel microphone array. The signals are preamplified with a SHURE FP24 (the internal audio card of laptops does not provide any preamplification on the ”line-in” input channel). The preamplified signals are recorded by the internal audio card and sampled at Fs=16kHz and resolution of 16 bits. The system runs on Linux and is coded in C++. The data is acquired by the audio card by using the Alsa audio driver interface. No sub-sampling is applied to the sampled data since the internal card natively support the sample frequency of 16kHz. The separated output signals are sent to a stereo output channel connected to a stereo-mono switch. By means of the switch, each channel is duplicated to the two channels of the headphone in order to allow the user listening each source separately.

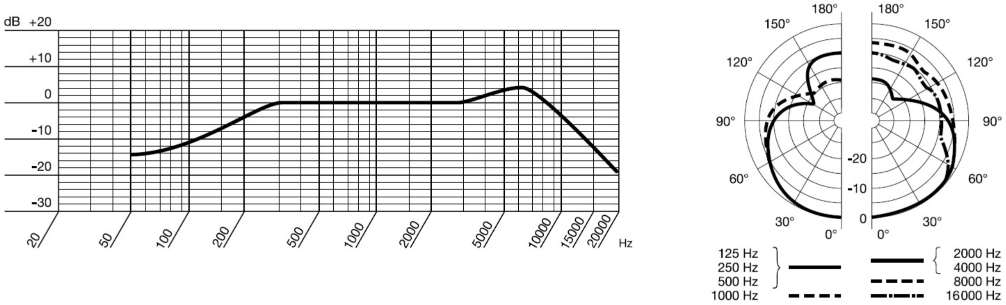


Figure 9.3: Frequency response and directivity pattern of the AKG C400 BL (from the original specification datasheet).

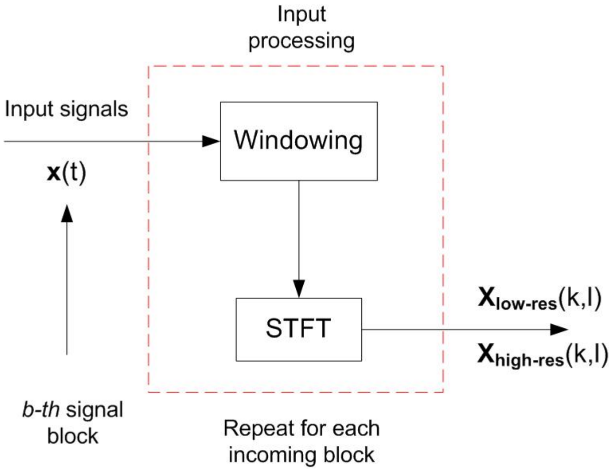


Figure 9.4: Description of the input processing stage.

9.2.1 Input processing

The input processing stage is described in Figure 9.4. The sampled data is split into different blocks of specified size (generally a multiple of the STFT frame-size). Each block is framed according to the parameter of the STFT analysis, windowed with an Hanning window and converted into the frequency-domain with a FFT. Two different STFT analysis are applied to obtain two different resolutions of the corresponding time-frequency representation of the signals. Note, each signal block is separated independently and no link is provided across time (i.e. over different blocks). It is reasonable to assume for acoustic sources that the mixing conditions are *quasi-stationary*, which means that the sources may change locations rapidly but do not move during their activity. On conditions that the length of the block is small enough and the source separation strategy is of sufficient accuracy, this *block off-line* strategy allows one to achieve high tracking capability of the mixing conditions even when the sources change quickly their locations.

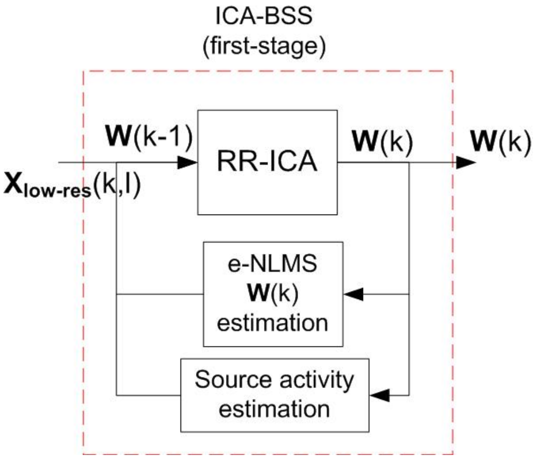


Figure 9.5: Description of the first ICA stage for the BSS.

9.2.2 ICA-BSS (first-stage)

The first ICA stage is described in Figure 9.5. The frequency components $\mathbf{x}(k,l)$ is the input of the system. The RR-ICA described in the chapter 4 is implemented at this stage. The ICA is recursively applied from the highest to the lowest frequency bins. At each frequency a smooth estimate of $\mathbf{W}(k)$ and a weighting matrix $\mathbf{\Psi}(k,l)$ is recursively computed and used to regularize the ICA of the next frequency. The system is designed to work also with largely spaced microphones. Therefore, the frequency linking provided by RR-ICA is not sufficient to avoid the permutation. For this reason the permutations are corrected *a posteriori* by the SCT in the next stage. However, the RR-ICA is essential to improve the ICA solution when the size of the block is limited (e.g. b=0.5-1s) and to speed-up the convergence rate.

9.2.3 SCT

This stage, described in Figure 9.6, aims to reduce the permutation problem by means of the ideal propagation model estimated by the SCT. First of all, a monodimensional SCT is implemented according to the theory in Chapter 5. The SCT is used to detect the maximum-likelihood TDOA of the sources. The peaks of its envelope are detected according to the SCT values and the first-derivative. The ideal propagation models parameterized with the estimated TDOA are used to align the demixing matrix $\mathbf{W}(k)$ according to the optimization in (5.68).

9.2.4 ICA-BSS (second stage)

This stage, described in Figure 9.7, is used to further improve the accuracy of the estimated demixing filters by a higher frequency resolution refinement of the demixing matrices estimated at the first ICA stage. In fact, despite of the good property of RR-ICA in improving the ICA

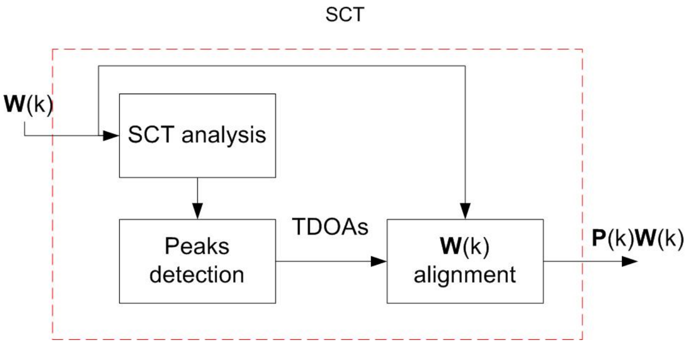


Figure 9.6: Description of the permutation correction stage based on the SCT

.

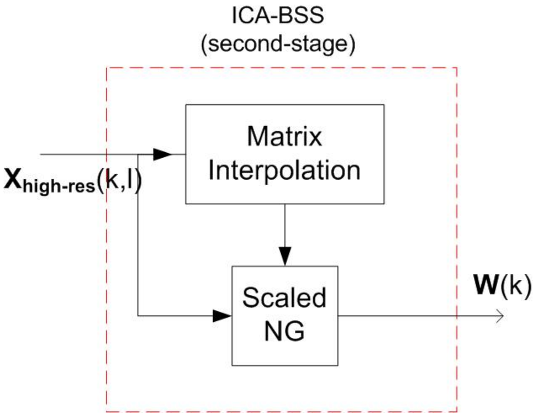


Figure 9.7: Description of the second ICA stage.

accuracy when a small amount of data is used, STFT with a resolution higher than 4096 bins may generate too much variance in the estimated demixing matrices $\mathbf{W}(k)$ and lead to audible distortions in the recovered output sources. Therefore, we found convenient to apply the first-stage to data converted into the frequency-domain by a lower STFT resolution (e.g. 2048-4096 points) and implement a further re-estimation of ICA with demixing matrices interpolated to a higher resolution. This strategy can be viewed as a simplified multi-resolution approach to the frequency-domain BSS. The matrices are interpolated from the lower to the higher resolution with a simple zero-order hold (ZOH) to avoid that higher order interpolation may accidentally generate singular demixing matrices which can not be used in the initialization of the ICA adaptation.

9.2.5 Output processing

The output processing stage implements the scaling normalization of the estimated demixing matrices (according to the MDP[53]) and the smoothing to avoid spikes due to the circularity effect of the FFT (see Chapter 3). The resulting demixing matrices are converted to time-

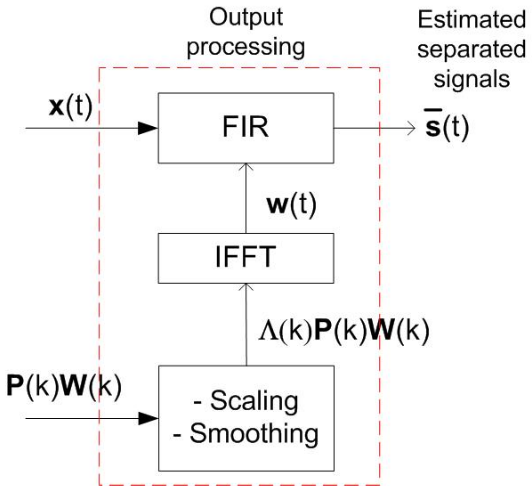


Figure 9.8: Description of the output processing stage for BSS.

domain filters by the inverse FFT and circular shifting of half-length in order to convert the filters from acausal to causal. This operation is essential if we want recover the output sources by continuously filtering the signals $\mathbf{x}(t)$ with FIR filters in time-domain.

9.3 Evaluation of the BSS algorithm

The implemented algorithm has been tested by generating dynamic mixtures of two sources, that is the case of sources changing over time their activity and their locations. The model used for the generation of the data have been used to define two tasks for the second Signal Separation Evaluation Campaign (SISEC2010). The dataset generated and evaluated in this Chapter has been submitted in the audio source separation task under the name of ”Determined convolutive mixtures under dynamic conditions”[1] and is currently involved for the evaluation of other algorithms proposed by the BSS research community.

We consider the case when maximum 2 sources are active at the same time and are recorded by a stereo microphone. The source mixtures are obtained by summing the individual source components recorded by each microphone. The components are generated by convolving random utterances with measured impulse responses and contaminated with an additive white, Gaussian noise (AWGN) according to an SNR of 40dB. The impulse responses between the microphones and different source locations (corresponding to different angular directions) are measured in a real room with a high reverberation time (T60 around 700-800ms). The distance between the sources and the microphones is about 1.1m.

The impulse responses have been measured by generating the chirp signal proposed in [91] and computing the correlation between the original signal and those recorded by the micro-

[1] *http://sisec.wiki.irisa.fr/*

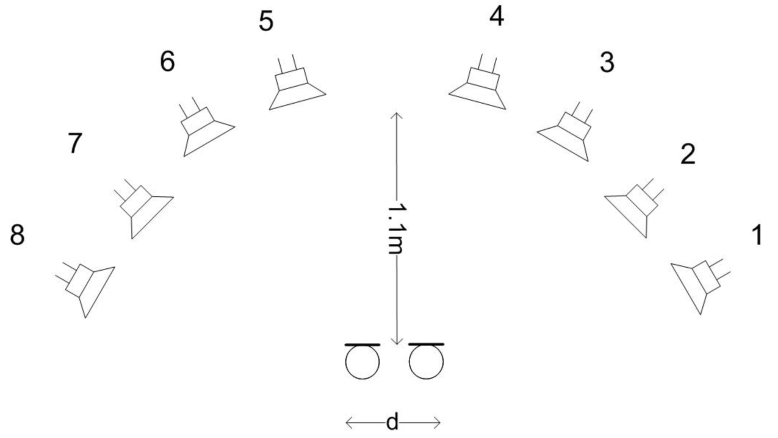


Figure 9.9: Setup of the source locations.

phones. The chirp signal is a time-stretched pulse signal having a flat overall power spectrum and autocorrelation equal to a pulse that enables a very accurate measurement of the acoustic impulse response. The pulse is defined on the discrete frequency-domain as the N-point sequence:

$$Pulse(n) = \begin{cases} e^{j2m\pi n^2/N^2}, & 0 \leq n \leq \frac{N}{2} \\ Pulse^*(N-n), & \frac{N}{2} \leq n \leq N \end{cases} \tag{9.1}$$

where * indicates the complex-conjugate operator, N is the FFT size and m defines the length of the generated impulse. A CHIRP of 4 seconds has been generated and the impulse responses have been measured for sources located approximatively as shown in Figure 9.9. Three different microphone spacing have been considered: d=0.02 m, d=0.06 m, d=0.10 m. Then, a total of 48 impulse responses have been measured. For each different array configuration, the quasi-stationary dynamic mixtures have been generated by first convolving the source signals for each impulse response. After that, a sequence of concatenated random sources located at random locations have been generated. Utterances of 8 speakers (4 male + 4 female), drawn from the Lite TIMIT database [1], have been used as source signals. For each speaker three utterances of different length between 1s and 4s have been considered. Two different tasks have been defined:

- Task1: the activity of the sources and their length is known. That is, we generate separated audio files for each mixture of two random sources (and utterances) in random locations. Thus, each mixture can be separated independently. This case is explained in Figure 9.10, where each color indicates a different speaker.

- Task2: the activity of the sources and their length is not known. That is, we generate a continuous sequence where at maximum two sources are active at the same time, still

[1] *http://web.mit.edu/course/6/6.863/share/nltk_lite/timit/*

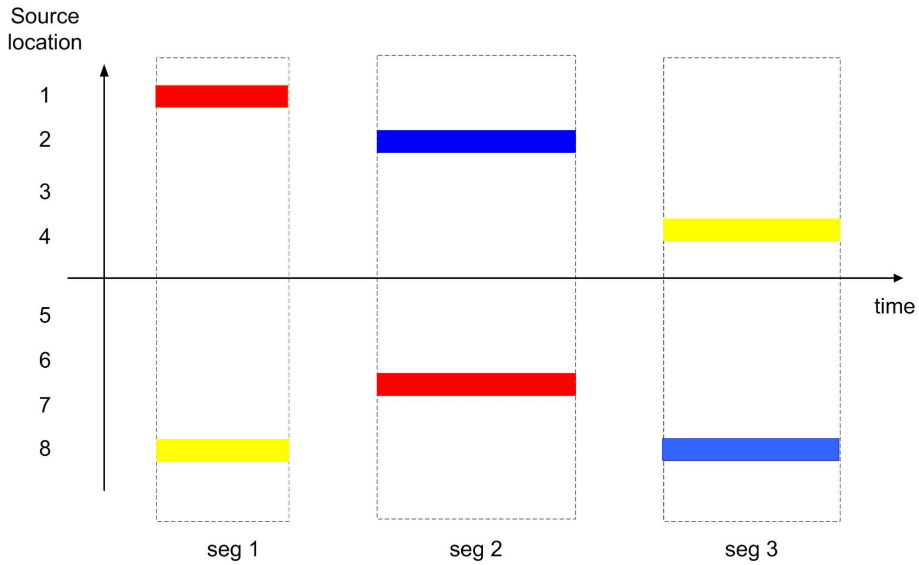


Figure 9.10: Explanation of the dynamic conditions generated in Task1.

located in random locations. This case is explained in Figure 9.11, where each color indicates a different speaker.

Task 2 is obviously more realistic than Task 1 but also more challenging since the start/end point of each utterance is not known in advance. In this case the BSS algorithm needs to adapt the separation across the sequence according to the change of the conditions. In Task1 the performance can be evaluated with the entire output signals since each mixture is independently generated. The evaluation of Task2 is more complex. For a fair evaluation we need to consider the separation performance with time segments where only two sources are simultaneously active (see Figure 9.11). Therefore, the true start and end marks of each segment have been determined during the simulation of the sequence by scanning the time activity of the sources and using the finite state-machine described in Figure 9.12.

The performance is evaluated by the criteria defined in the BSS_EVAL toolbox [96]. The BSS_EVAL toolbox decomposes the estimated sources in a sum of components corresponding to: 1) a deformation of the original source 2) a deformation of the sources accounting for the presence of unwanted interfering sources 3) artifacts introduced by the BSS procedure 4) a deformation of a perturbating noise. According to this decomposition, different performance measures are defined. In this task only the SIR and SDR are evaluated, considering the best input/output ordering which leads to the best SIR result. In order to give measures which are close to the human perception, we also evaluate the SIR/SDR by filtering the components estimated by the BSS_EVAL with an A-weighting filter (its frequency response is shown in Figure 9.13). The filtering is introduced to emphasize frequencies around 3-6 kHz where the human ear is most sensitive, while attenuating higher and lower frequencies. The aim of this filter-

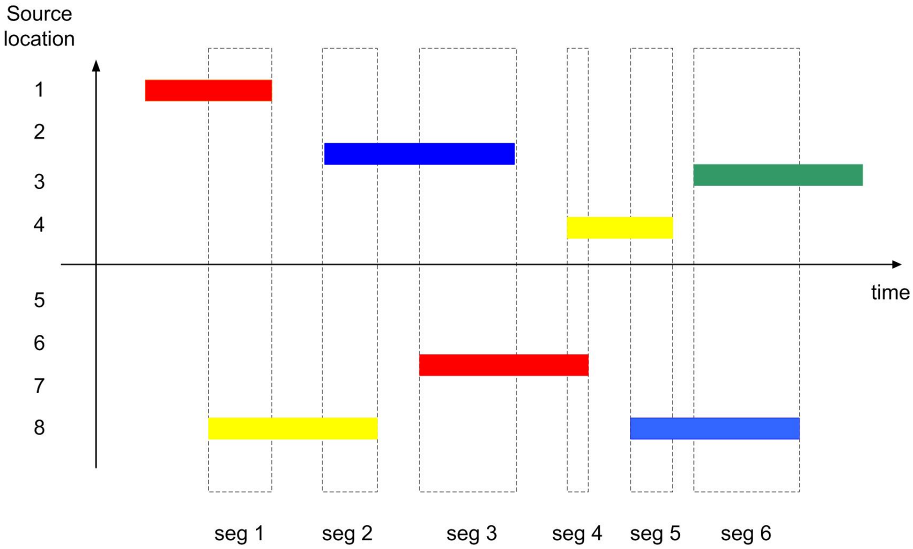


Figure 9.11: Explanation of the dynamic conditions generated in Task2.

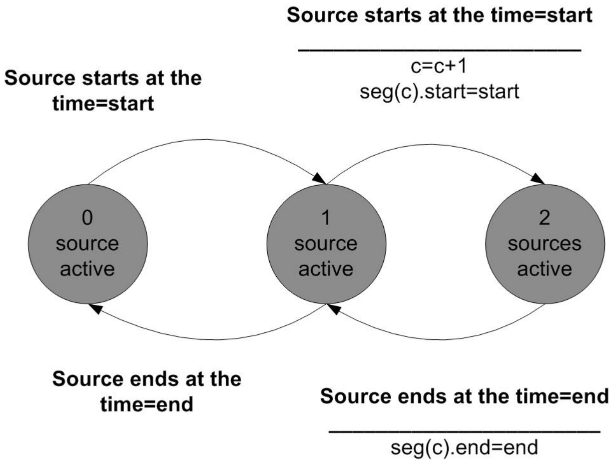

Source starts at the time=start
c=c+1
seg(c).start=start
Source starts at the time=start
0 source active
1 source active
2 sources active
Source ends at the time=end
Source ends at the time=end
seg(c).end=end

Figure 9.12: State-machine for the detection of the number of active sources

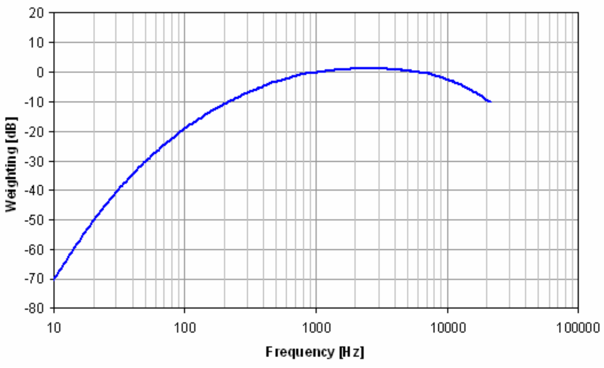


Figure 9.13: Frequency response of the A-weighting filter.

ing is to ensure that the measured separation performance better matches with the objective perceived separation. Note, this is only an approximation of the subjective separation performance perceived by a human listener since it only takes into account the perceived loudness. In general, modeling the human perception of the sound separation is still an open issue since it involves many complex phenomena such as harmonic relationships between sources, spatial cues, masking effects and so on.

Task1 has been evaluated generating a random dataset of 28 mixture combinations for each array configuration. Each mixture have been independently separated with all the available data (i.e. the block length is equivalent to the mixture length). Task2 has been evaluated generating a random sequence of about three minutes for each array configuration. In this task the BSS algorithm was parameterized with the following parameters: blocks of 1s, FFT size (lower resolution)=2048 points, FFT size (higher resolution)=8192 points, frame overlap=75%, maxIter=20, L=40, $\mu = 0.01$, $\eta = 0.1$.

Figures 9.14, 9.15, 9.16 and 9.17 show the distribution of the SIR and SDR performance obtained for Task1, with and without applying the A-weighting filter to the estimated source components. The distributions have been computed by a kernel density approximation of the performance obtained in each test (using a Gaussian kernel). We note that the average performance is satisfactory despite of the difficult mixing conditions and, more important, the system exhibits a clear robustness since there are no outliers in the distributions which could be approximated by unimodal distribution models. This observation becomes more consistent looking at the distributions when the performance is computed with the A-weighting filtering. Also, from a listening test, the filtering better approximates the computed performance to the perceptual separation. In other terms, the algorithm produced output signals with very slight fluctuations in the perceptual separation evaluated by a human listener. Looking at all the figures one could also note that the performance is better for closely-spaced microphones. This also confirms the motivation at the basis of the distributed BSS method proposed in Chapter 8.

The same analysis has been performed for Task2 and the results are shown in Figures 9.18,

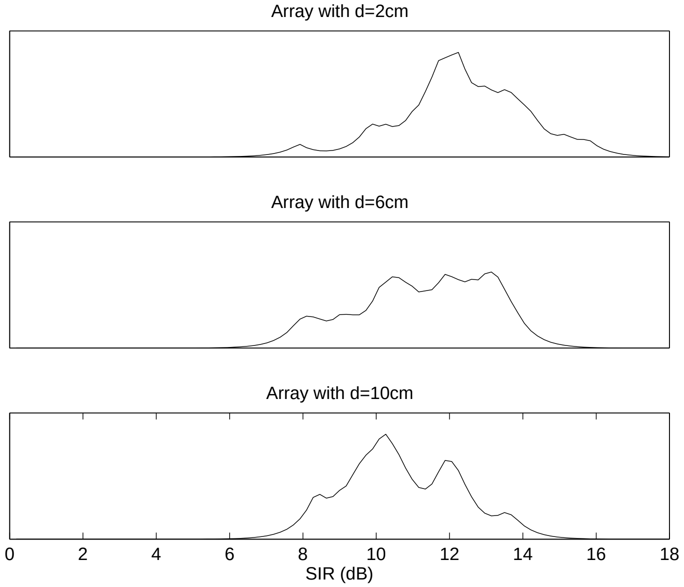


Figure 9.14: Task1, pdf of the SIR performance computed without A-weighting filter.

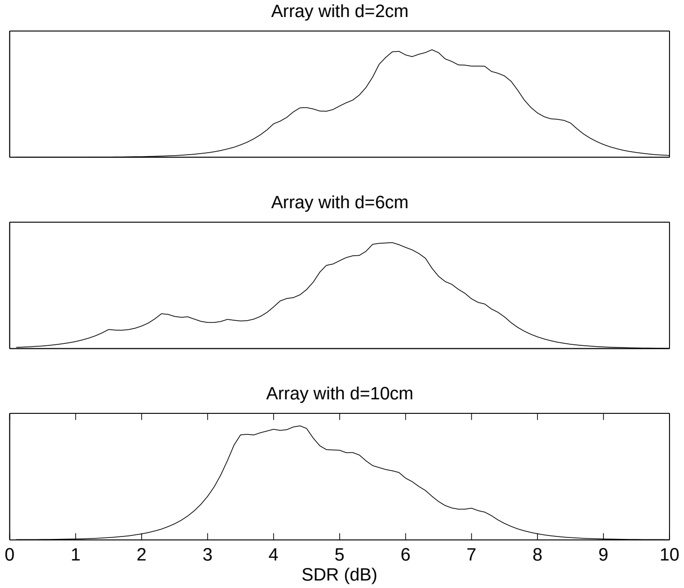


Figure 9.15: Task1, pdf of the SDR performance computed without A-weighting filter.

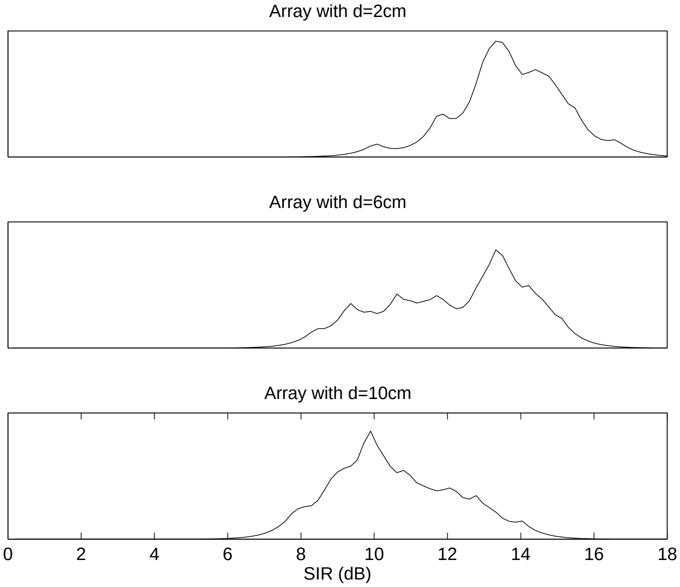


Figure 9.16: Task1, pdf of the SIR performance computed with A-weighting filter.

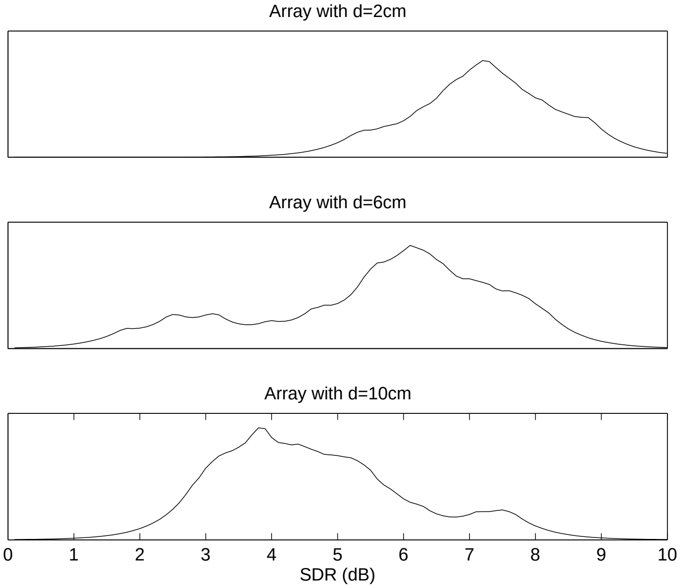


Figure 9.17: Task1, pdf of the SDR performance computed with A-weighting filter.

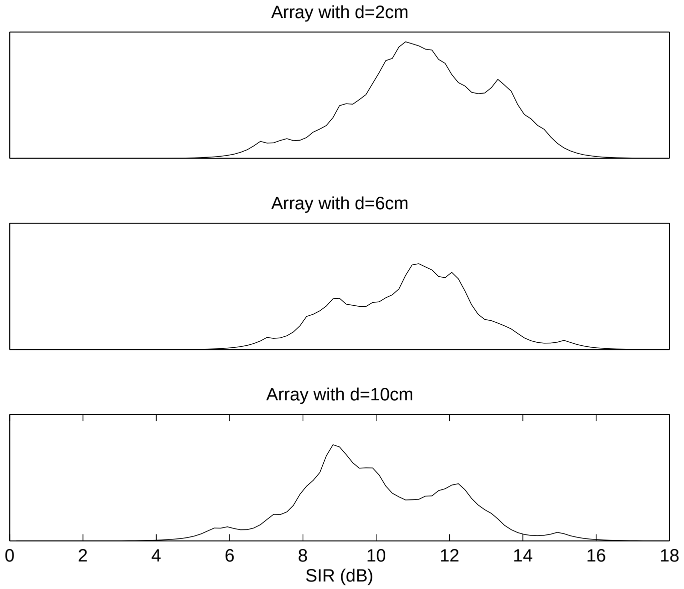


Figure 9.18: Task2, pdf of the SIR performance computed without A-weighting filter.

9.19, 9.20 and 9.21. In this task, the distributions have been computed by a kernel density approximation of the performance obtained in each segment of the sequence where two sources are active at the same time. Figures 9.22, 9.23, 9.24 and 9.25 show the measured performance of the whole sequence. Even in this case, the performance is stable which means that, under the assumption of quasi-stationary mixing conditions, the source separation can be effectively achieved separating short blocks of data with the proposed BSS algorithm.

9.4 Architecture of the real-time SBSS

The SBSS system slightly differs from the BSS for a further module used to suppress the echos and for a module of Automatic Speech Recognition (ASR). Figure 9.26 shows the logical architecture of the system. Differently from what is discussed in Chapter 7 the source separation and the echo removal is not performed in the same ICA adaptation. The reason of this architecture is discussed in the next subsections.

Figure 9.27 shows the hardware architecture of the system. Two hypercardioid microphones AKG C 400 BL are used to record the acoustic scene premplified with a SHURE FP24. The preamplified signals are sent to the first two input channels of an external low cost 4-in 4-out USB audio card *ESI-Maya 44USB*. A stereo signal (e.g. a stereo TV audio signal) is played through the loudspeaker using two output channels of the USB audio card. The signals are duplicated and sent back to the last two input channels of the audio card. These signals correspond to the references $\mathbf{r}(t)$ used in the SBSS adaptation. All the signals are sampled at Fs=16kHz and resolution of 16 bits. The system runs on Linux and is coded in C++. The data are acquired

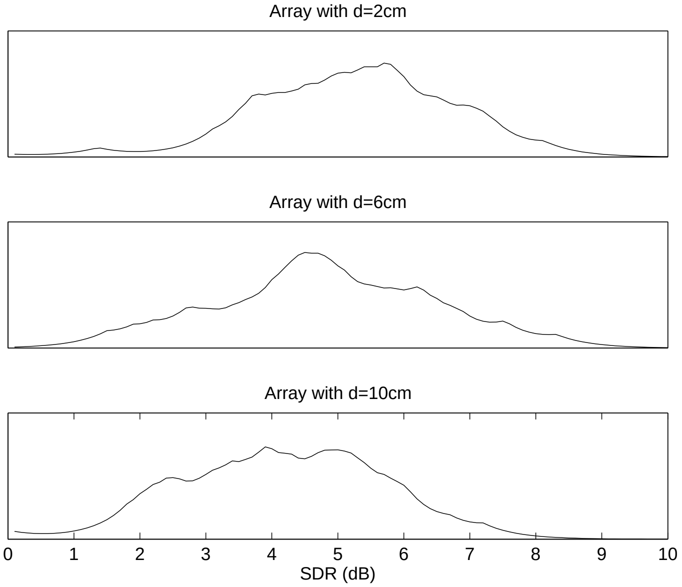


Figure 9.19: Task2, pdf of the SDR performance computed without A-weighting filter.

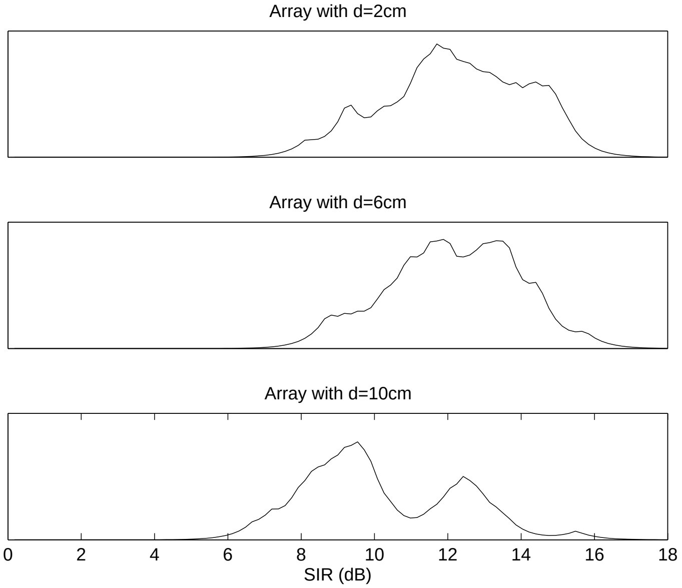


Figure 9.20: Task2, pdf of the SIR performance computed with A-weighting filter.

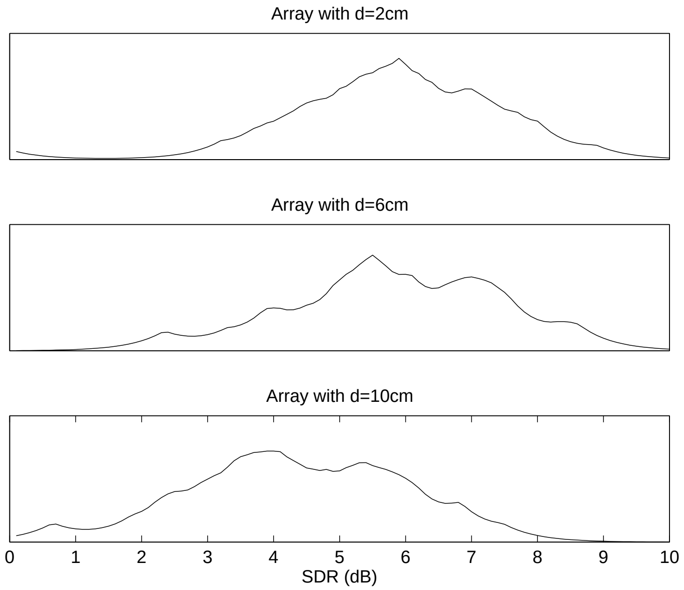


Figure 9.21: Task2, pdf of the SDR performance computed with A-weighting filter.

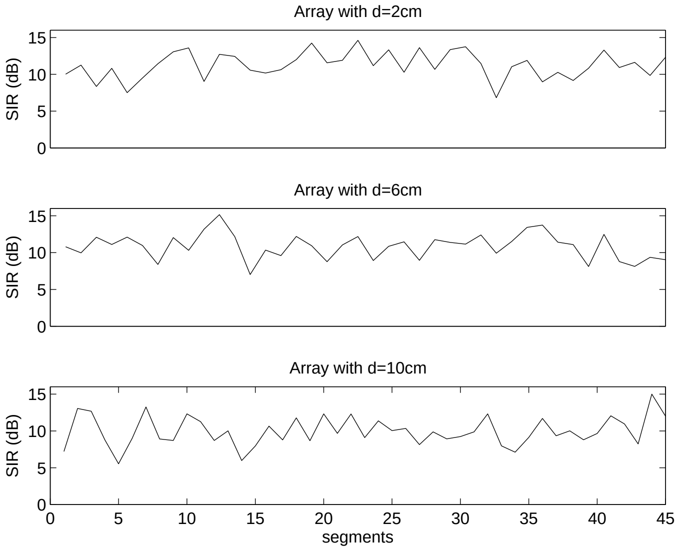


Figure 9.22: Task2, SIR performance envelope computed without A-weighting filter.

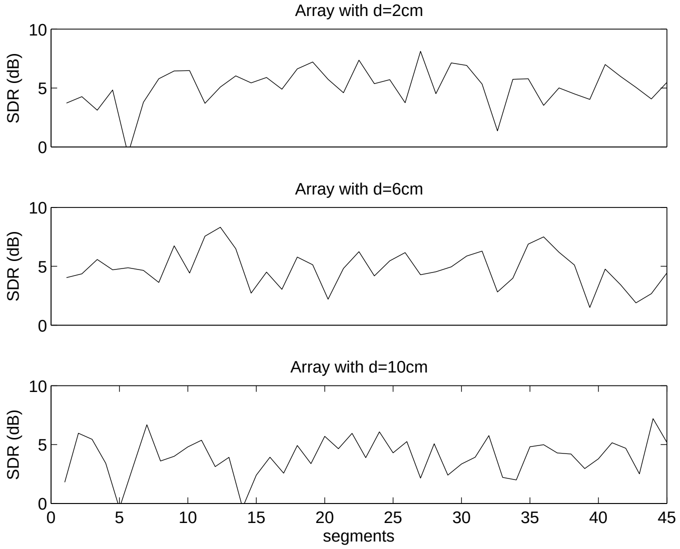


Figure 9.23: Task2, SDR performance envelope computed without A-weighting filter.

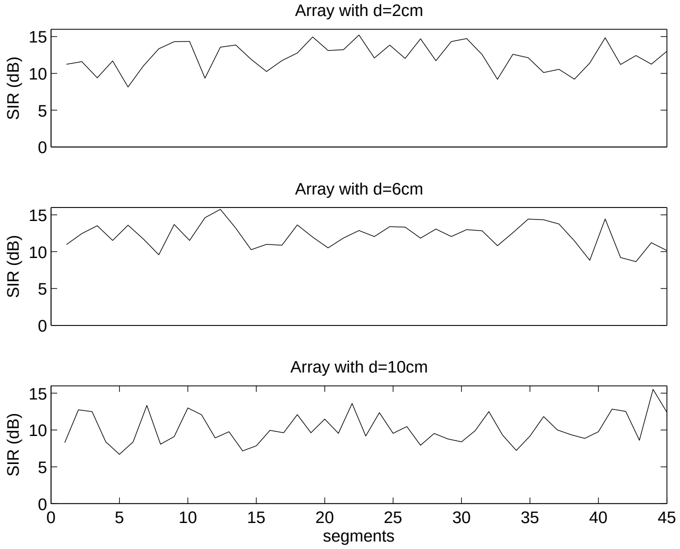


Figure 9.24: Task2, SIR performance envelope computed with A-weighting filter.

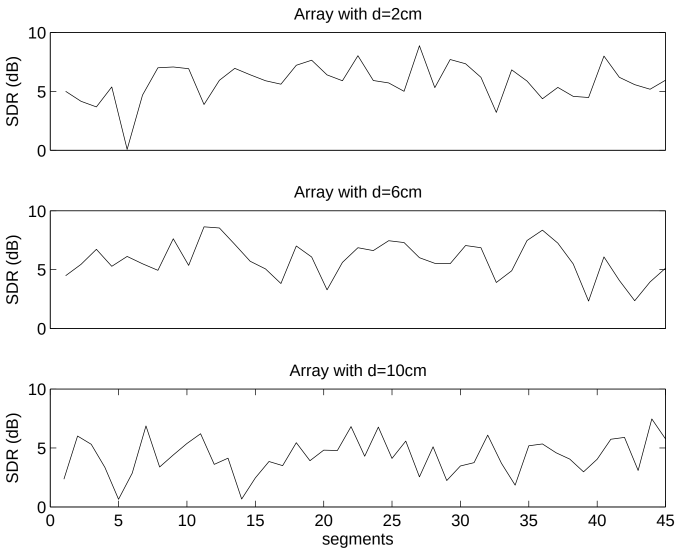


Figure 9.25: Task2, SDR performance envelope computed with A-weighting filter.

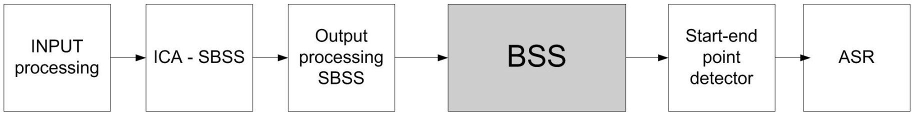


Figure 9.26: Logical architecture of the developed SBSS system.

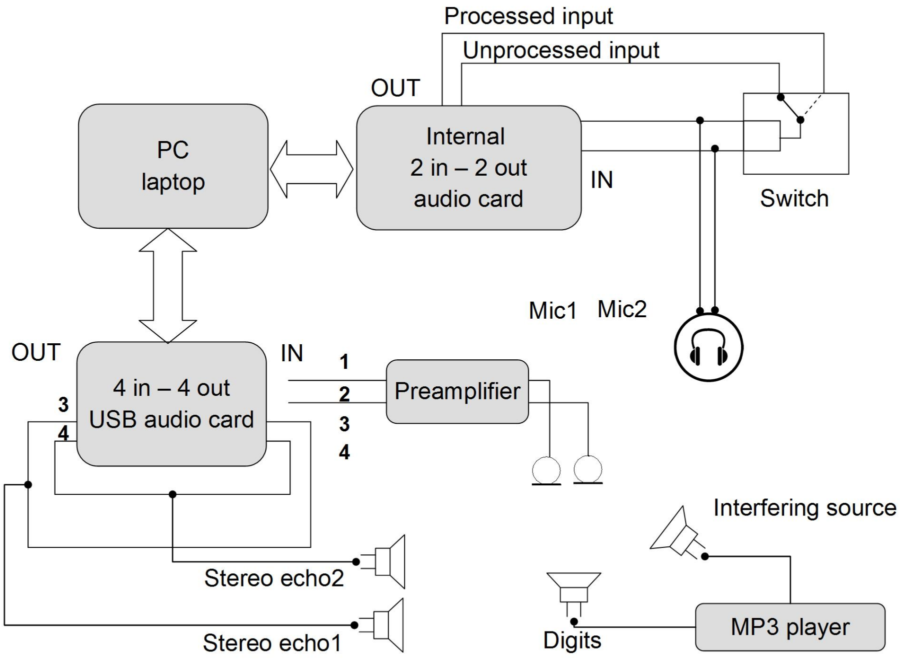


Figure 9.27: Hardware architecture of the developed SBSS system.

by the audio card by means of the Alsa audio driver interface. From the output of the internal audio card the system gives a stereo signal. The first channel represents the unprocessed signal recorded at input 1 and the other channel represents the recovered target source. The stereo output channel is connected to a stereo-mono switch. By means of the switch, each channel is forwarded to the headphone in order to allow the listening of each channel separately and to verify the effect of the global source enhancement. Furthermore, the selected output is forwarded to the input of the internal audio card in order to be recognized by an ASR module.

In this system implementation we consider the recognition of a sequence of digits. Therefore the near-end sources are represented by two loudspeakers connected to an mp3 player, one playing a digits sequence while the other is playing an audio book. These sources are completely unknown by the system (i.e. are not physically connected to the system). For the order source ambiguity of any BSS system (often known as *external permutation ambiguity*) we can not know in advance which is the channel number of the target source (i.e. the sequence of digits) and the ordering may change over time. Assuming that the digits are played through the loudspeaker more centrally located to the array, a solution is to estimate the TDOA of the sources (e.g. by means of the SCT) and opportunely ordering the output signals in order to forward to the ASR modules the source corresponding to the smallest TDOA. Alternatively, one could also implement two parallel ASR modules and decode both the sources discharging *a posteriori* the recognized world with the lowest likelihood. We remind that a general solution to the problem of the external permutation is not available and strategies to mitigate the effect of such ambiguity should be defined according to the target application.

9.4.1 Input processing

The input processing stage is described in Figure 9.28 and is equivalent to the case of the BSS configuration. The sampled data is split into different blocks of specified size (in general equal to a multiple of the STFT frame-size). Each block is framed according to the parameter of the STFT analysis, windowed with an Hanning window and converted into the frequency-domain by means of the FFT.

9.4.2 ICA-SBSS

The ICA for the SBSS stage is described in 9.29. The frequency components $\mathbf{x}(k,l)$ and the reference signals $\mathbf{r}(k,l)$ are the inputs of the system. The constrained SBSS described in chapter 7 is implemented in this stage. Note that, differently from the first ICA stage for the BSS task described in the previous section, the adaptation is not applied recursively across the frequencies. In this stage the ICA is applied independently at each frequency implementing the full block on-line linking strategy described in Chapter 7. The ICA adaptation of SBSS, which

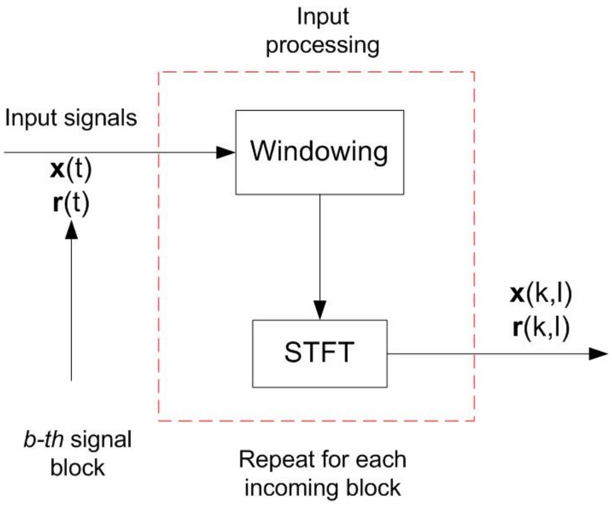


Figure 9.28: Description of the input processing stage for the SBSS.

removes the echos from the output signals, has a higher accuracy and convergence rate than the ICA adaptation applied to the blind separation case. Therefore, the frequency recursion is not essential to improve the ICA solution in each batch and, under stationary mixing conditions (i.e. loudspeakers and microphones do not change their location), on-line adaptations over time would guarantee higher steady state performance.

9.4.3 Output processing

The output processing stage described in Figure 9.30 is equivalent to that in Figure 9.8 with the difference that no permutation correction is applied to $\mathbf{W}(k)$ before the computation of the demixing filters. In our system implementation we can avoid this operation since we are interested also in the near-end source separation which is applied in the next stage. In general, we remind that if we are only interested in removing the echos without separating the near-end sources the normalizations $\mathbf{W}_{12} = \mathbf{W}_{11}^{-1}\mathbf{W}_{12}$ and $\mathbf{W}_{11} = \mathbf{I}$ must be applied before the computation of the time-domain demixing filters.

9.4.4 BSS

This stage corresponds to the whole BSS system implemented in previous section with the slight difference that the unprocessed signals $\mathbf{x}(t)$ at the input of the stage of Figure 9.28 is bypassed directly through the outputs.

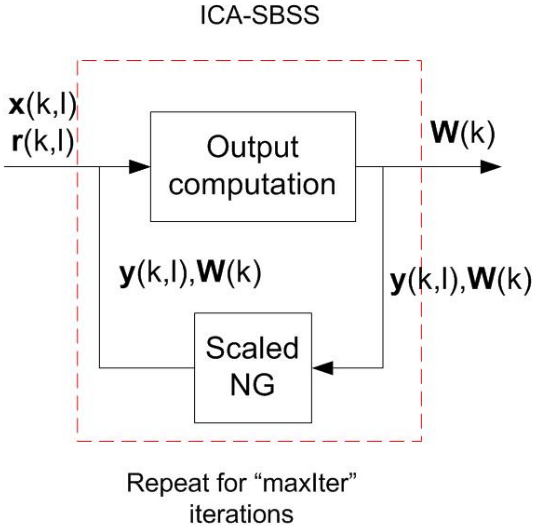


Figure 9.29: Description of the ICA stage for the SBSS.

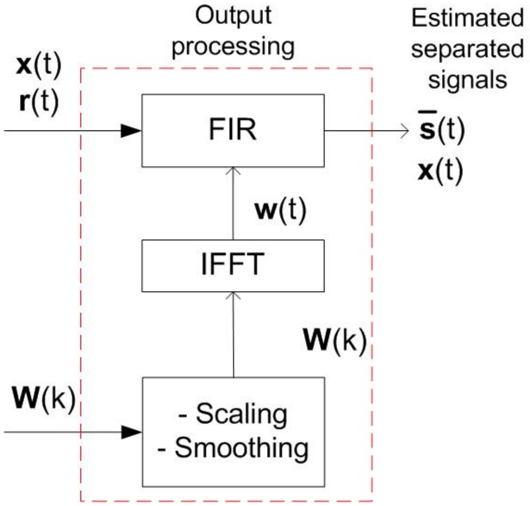


Figure 9.30: Description of the output processing stage for the SBSS.

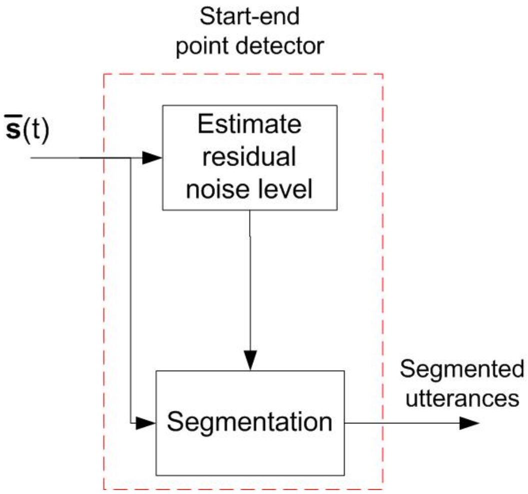


Figure 9.31: Block description of the start-end point detector.

9.4.5 Start-end point detector

This stage implements a simple energy-based voice activity detector to reject segments where no speech activity in the desired waveform is detected. The energy of the signals is adaptively estimated over time as:

$$e(t) = \gamma * e(t-1) + (1-\gamma)|\overline{s}_1(t)|^2 \tag{9.2}$$

where $|\overline{s}_1(t)|^2$ is the estimated power of the target source and γ is a smoothing factor $\gamma < 1$. The activity of the source is dynamically detected when the energy is over that of the residual noise, which is adaptively estimated as:

$$e_{noise}(t) = \gamma * e_{noise}(t) + (1-\gamma)[1-a(t)]|\overline{s}_1(t)|^2 \tag{9.3}$$

where $a(t)$ is a function indicating the presence of the source:

$$a(t) = \begin{cases} 1, & |\overline{s}_1(t)| > |\overline{s}_2(t)| \\ 0, & otherwise \end{cases} \tag{9.4}$$

and $\overline{s}_2(t)$ indicates the near-end interfering source. Furthermore, to make the start-end point detector more stable some further constraints on the minimum length of the detected segments are imposed.

9.4.6 ASR

The target signal extracted by the SBSS system is fed into a standard HMM recognizer. The signals are framed with windows of 20ms. 13 standard Mel frequency cepstral coefficients

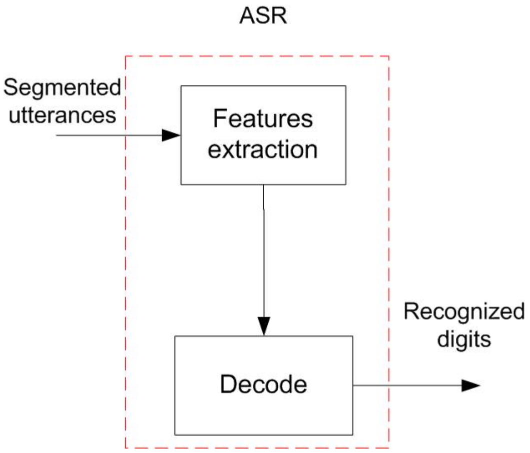


Figure 9.32: Block description of the ASR module.

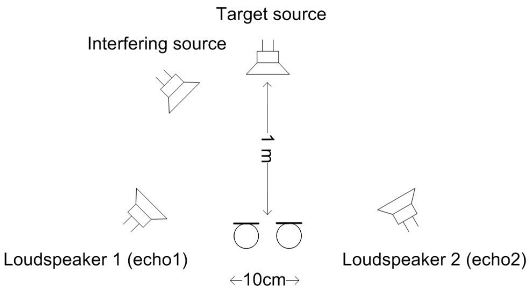


Figure 9.33: Setup for the SBSS evaluation

(MFCC) are extracted for each window and the first and second derivatives are computed. After that, mean and variance normalization is applied to all the 39 features coefficients. The recognition engine is based on the FBK engine used in [52].

The adopted acoustic models of the ASR module are trained in matched conditions: 1000 connected sentences from TIDIGITS database have been played back in the room and acquired by the acoustic front-end composed by the microphone array and the SBSS processing when both the echos and the near-end interfering sources are silent.

9.5 Evaluation of the SBSS algorithm

For this system we evaluate the global performance from an application point of view using a simple speech recognition task as benchmark.

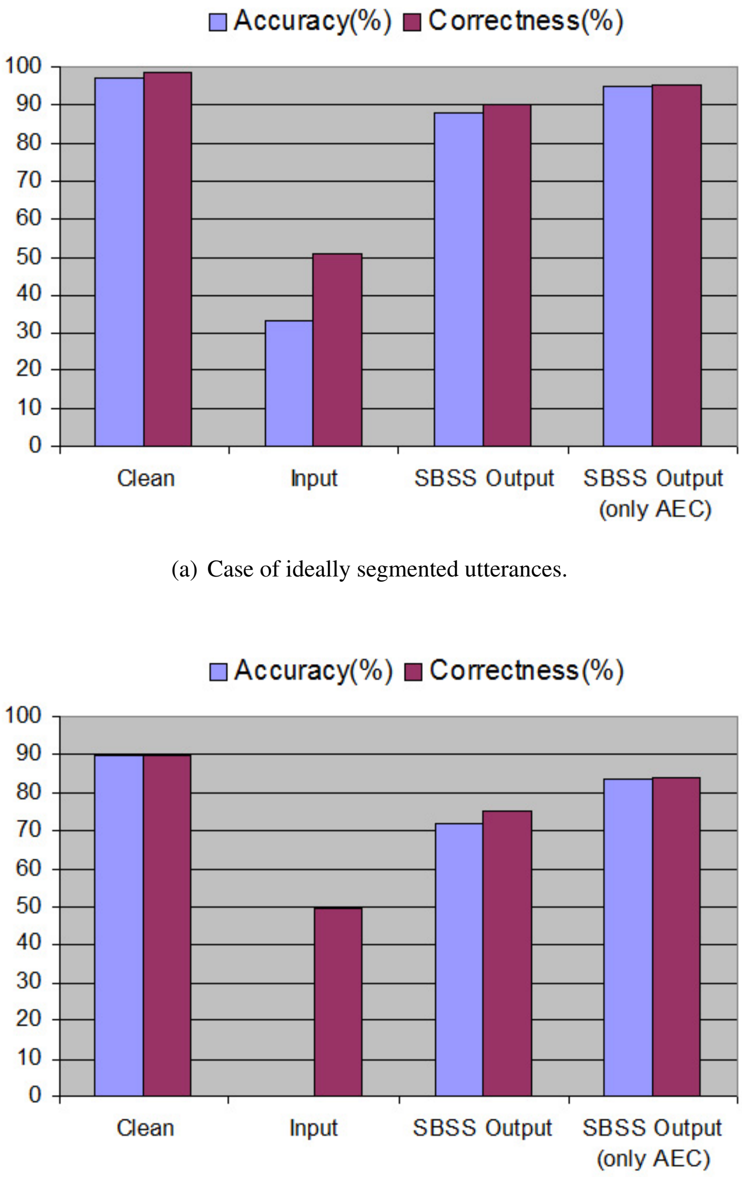
Hence, a recognition system has been setup in order to measure the performance resulting from a connected digits recognition task. The sources are located as in the configuration shown in Figure 9.33 in the same room used for the evaluation of the BSS system. The BSS block of the system was setup with the same parameters used for the previous evaluation. For the

SBSS part the following parameters were used: sampling frequency of 16kHz, signals framed with Hanning windows of 4096 points overlapped of 75%, block length of 256ms, maxIter=5, $\eta = 0.1$.

The resulting system has been tested on a short signal composed by the concatenation of several TIDIGITS sequences (450 digits). The signal have been generated introducing high variation in the dynamic among different utterances, in order to test the robustness of the start-end point detector. The performance is measured by means of the accuracy and correctness. The accuracy measures the number of correct digits taking into account of insertions, cancellations and substitutions in the recognized sequence. The correctness measures the number of correct digits taking into account of cancellations and substitutions but does not account for the insertions. The performance has been computed for the following signals:

- the clean speech reproduced by the mp3 player used as ground truth for the comparison;
- the input source mixture recorded by the microphones;
- the output of the SBSS system, which is the estimated target without the echos and the near-end interfering source;
- the output of the SBSS system when the near-end interfering source is silent and only the MCAEC is applied.

The performance is evaluated applying the separation to the ideally segmented utterances (according to the true segmentation) and to the segments obtained by the start-end point detector. Figure 9.34(a) shows the performance for the ideally segmented utterances. The performance of the noisy mixtures is clearly low, especially in terms of accuracy. In fact, the decoder of the ASR module inserts many wrong hypothesis in the pause between the digits of each sequence. On the other hand, the output of the SBSS is accurate enough and both accuracy and correctness approach a value of about 90%. The performance is even higher when the near-end interferer is silent and only the AEC is applied, which clearly shows the effectiveness of the SBSS for MCAEC purpose. We remind that, even for this simple task, the distant-talk performance is lowered by the signal coloration introduced by the reverberation which is not negligible for the room considered in the test (a reverberation time of about T_{60}=700-800ms). Figure 9.34(b) shows the performance for the automatically segmented utterances. For the noisy mixtures, the word correctness equals that obtained with the ideal segmented files while the accuracy has a very low (negative) value. The reason of this mismatch relies on the impossibility for the automatic start-end point detector to provide a correct segmentation of the sequence. In fact, the automatic speech detector is not able to converge to a reliable noise level estimate since the disturbances (i.e. the near-end interfering signals and the echos) have magnitude similar to that of the target source. Since the signal picked up at the microphone is composed by a



(a) Case of ideally segmented utterances.

(b) Case of automatically segmented utterances.

Figure 9.34: Recognition performance for the SBSS system

mixture of the digit sequence and the disturbances, without the application of any separation processing the recognizer is prone to provide hypotheses with a large number of spurious digits inserted during the undetected speech pauses. Thus, it becomes clear that a BSS/SBSS method becomes essential to enable the recognition in such a challenging scenario. Unfortunately, also the performance with the SBSS processing is lower than the case of ideal segmentation because the start-end point detector is not sufficiently robust against big variations in the dynamic of the target source. This result indicates that further investigation needs to be done in this direction.

Chapter 10

Conclusions

In this thesis, the problem of the multichannel source separation based on the Independent Component Analysis (ICA) is studied. After the analysis of the blind case (BSS) in the frequency-domain and its related physical interpretation the focus is on the problem of the separation of short utterances, where the ill-conditioning of the ICA and the high reverberation time hamper acceptable performance of traditional methods. A frequency-recursive regularized ICA (RR-ICA) is then presented to improve the estimation of the mixing system and therefore the separation performance.

By means of the physical interpretation of the ICA solutions an efficient transform (GSCT) of the estimated demixing matrices is proposed, which allows to estimate the location of multiple sources even in multiple dimensions. The source locations estimated by such a transform are combined with the spectral continuity of the source energy in order to define a joint metric of global spectral coherence (cSPEC). Such a metric is able to define a robust and precise solution to the permutation problem even when short signals are recorded under the condition of high spatial aliasing.

The paradigm of ”blind” source separation is extended to the semi-blind case (SBSS) where a partial knowledge of some source signals and/or mixing system is exploited. This new approach is particularly efficient in limiting the drawbacks of traditional acoustic echo cancellation (AEC) systems and has the advantage of merging the blind and semi-blind case into a common framework. After a deep theoretical analysis, many issues for the design of practical implementable SBSS system are discussed. Furthermore, the thesis discusses and proposes a new technique to extend the source separation to the case of a distributed array network. Preliminary results shows that the distributed paradigm could be a promising approach to increase the robustness of source separation to an acceptable level even in adverse conditions.

Finally, the algorithmic implementations of a real-time two-channel BSS system and a four-channel SBSS system is described. A detailed discussion about conclusions and future directions for each proposed technique is available in the last section of each chapter.

Bibliography

[1] P. Aarabi and S. Mavandadi. Multi-source time delays of arrival estimation using conditional time-frequency histograms. *Information Fusion*, 4(2):111–122, June 2003.

[2] Robert Aichner, Herbert Buchner, Fei Yan, and Walter Kellermann. A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments. *Signal Process.*, 86(6):1260–1277, 2006.

[3] Shunichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[4] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari. Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming for convolutive mixtures. In *EURASIP Journal on Applied Signal Processing, vol. 2003*, pages 1157–1166, November 2003.

[5] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari. Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2737–2740, 2001.

[6] Shoko Araki, Tomohiro Nakatani, Hiroshi Sawada, and Shoji Makino. Stereo source separation and source counting with map estimation with dirichlet prior considering spatial aliasing problem. In *ICA '09: Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*, pages 742–750, Berlin, Heidelberg, 2009. Springer-Verlag.

[7] S. Uhlich B. Loesch and B. Yang. Multidimensional localization of multiple sound sources using frequency-domain ica and an extended state coherence transform. In *Proceedings of SSP*, Cardiff,UK, September 2009.

[8] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. In *Neural Computation*, volume 7(6), pages 1004–1034, 1995.

[9] J. Benesty, D. R. Morgan, and M. M. Sondhi. A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation. *IEEE Transaction on Speech and Audio Processing*, 6(2):156–165, March 1998.

[10] A. Brutti. *Distributed Microphone Networks for Sound source localization in smart rooms*. PhD thesis, DIT - University of Trento, 2007.

[11] Alessio Brutti, Maurizio Omologo, and Piergiorgio Svaizer. Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays. In *Proceedings of Interspeech*, pages 2337–2340, 2005.

[12] Herbert Buchner, Robert Aichner, and Walter Kellermann. TRINICON: A versatile framework for multichannel blind signal processing. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 889–892, Montreal, Canada, May 17-21 2004.

[13] J.-F. Cardoso. Infomax and maximum likelihood for blind source separation. *Signal Processing Letters, IEEE*, 4(4):112–114, Apr 1997.

[14] Jean-Franois Cardoso and Shun ichi Amari. Maximum likelihood source separation: Equivariance and adaptivity. In *in Proc. of SYSID97, 11th IFAC symposium on system identification*, pages 1063–1068, 1997.

[15] M. A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. *In Proceedings of International Computer Music Conference*, 2000.

[16] Y.S. Choi, H.C. Shin, and W.J. Song. Robust regularization for normalized LMS algorithms. *IEEE Transaction on circuits and system*, 53(8):627–631, August 2006.

[17] A. Cichocki, I. Sabla, and S. Amari. Intelligent neural networks for blind signal separation with unknown number of sources. In *In Proc. of Conference Engineering of Intelligent Systems*, page 148154, 1998.

[18] Andrzej Cichocki and Shun-Ichi Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, Inc., New York, NY, USA, 2002.

[19] G. A. Clark, S. R. Parker, and S. K. Mitra. A unified approach to time- and frequency-domain realization of FIR adaptive digital filters. *IEEE Transactions on Audio, Speech, and Language Processing*, ASSP-31(5):1073–1083, October 1983.

[20] Pham D.-T., Garrat P., and Jutten C. Separation of a mixture of independent sources through a maximum likelihood approach. In *proceedings of EUSIPCO*, page 77177, 1992.

[21] Nathalie Delfosse and Philippe Loubaton. Adaptive blind separation of independent sources: a deflation approach. *Signal Process.*, 45(1):59–83, 1995.

[22] E. J. Diethorn. Foundations of spectral-gain formulae for speech noise reduction. In *Proceedings of IWAENC*, Eindhoven, The Netherlands, September 2005.

[23] S. Ding, A. Cichocki, J. Huang, and D. Wei. Blind source separation of acoustic signals in realistic environments based on ICA in the timefrequency domain. *Journal of pervasive computing and communications*, 1(2):88–89, 2005.

[24] S.C. Douglas and M. Gupta. Scaled natural gradient algorithms for instantaneous and convolutive blind source separation. In *Proceedings of ICASSP*, volume II, pages 637–640, April 2007.

[25] Shlomo Dubnov, Joseph Tabrikian, and Miki Arnon-Targan. Speech source separation in convolutive environments using space-time-frequency analysis. *EURASIP J. Appl. Signal Process.*, 2006(1):1–11, January 2006.

[26] Zaher El Chami, Dinh-Tuan Pham, Christine Servière, and Alexandre Guérin. A new model-based underdetermined speech separation. In *11th International Workshop on Acoustic Echo and Noise Control, IWAENC2008*, pages 1–4, Seatle, Washington, Etats-Unis, September 2008.

[27] G. W. Elko. *Spatial Coherence Functions for Differential Microphones in Isotropic Noise Fields*. Springer-Verlag, 2001.

[28] J. Eriksson and V. Koivunen. Identifiability and separability of linear ica models revisited. *Proceedings of ICA*, pages 23–27, 2003.

[29] F., T. S. Wada, and B.H. Juang. Batch-online semi-blind source separation applied to multi-channel acoustic echo cancellation. Accepted in IEEE Transactions on Audio, Speech, and Language Processing, 2010.

[30] III Frost, O.L. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8):926–935, Aug. 1972.

[31] Maria Funaro, Erkki Oja, and Harri Valpola. Independent component analysis for artefact separation in astrophysical images. *Neural Netw.*, 16(3-4):469–478, 2003.

[32] Patrick D. Gerard and William R. Schucany. Local bandwidth selection for kernel estimation of population densities with line transect sampling. *Biometrics*, 55(3):769–773, 1999.

[33] T. Gustaffson, B. D. Rao, and M. Trivedi. Source localization in reverberant environments: Modeling and statistical analysis. *IEEE Transactions on Speech and Audio Processing*, 11(6):791–803, November 2003.

[34] Sawada H., Araki S., and Making S. A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures. In *Proceedings of WASSPAA*, pages 139–142, New Paltz, October 2007.

[35] E. Hänsler and G. U. Schmidt. *Acoustic Echo and Noise Control: A Practical Approach*. John Wiley & Sons, 2004.

[36] Yiteng Huang and Jacob Benesty. *Audio Signal Processing for Next-Generation Multimedia Communication Systems*. Kluwer Academic Publishers, Norwell, MA, USA, 2004.

[37] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley, New York, NY, USA, 2001.

[38] A. Hyvarinen and E. Oja. Independent component analysis. *Neural Networks*, 13(4-5):411–430, 2000.

[39] Aapo Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 273–279, Cambridge, MA, USA, 1998. MIT Press.

[40] A. Hyvrinen and E. Oja. A fast fixed-point algorithm for independent component analysis. In *Neural Computation*, volume 9(7), pages 1483–1492, 1997.

[41] Kazushi Ikeda and Ryohei Sakamoto. Convergence analyses of stereo acoustic echo cancelers with preprocessing. *IEEE Transactions on Signal Processing*, 51(5):1324–1334, May 2003.

[42] W. Kellermann J. Herre, H. Buchner. Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement. In *Proceedings of ICASSP*, volume I, pages 17–20, April 2007.

[43] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 24, pages 320–327, 1976.

[44] Z. Koldovskỳ and P. Tichavskỳ. Time-domain blind audio source separation using advanced component clustering and reconstruction. In *Proceedings of HSCMA*, Trento, Italy, May 2008.

[45] Intae Lee, Taesu Kim, and Te-Won Lee. Independent vector analysis for convolutive blind speech separation. In *Blind Speech Separation*. Springer, September 2007.

[46] E. Lehmann and A. Johansson. Prediction of energy decay in room impulse responses simulated with an image-source model. *Journal of the Acoustical Society of America*, 124(1):269–277, July 2008.

[47] B. Loesch, Nesta F., and B. Yang. On the robustness of the multidimensional state coherence transform for solving the permutation problem of frequency-domain ICA. *Proceedings of ICASSP*, March 2010.

[48] B. Loesch, S. Uhlich, and B. Yang. Multidimensional localization of multiple sound sources using frequency domain ICA and an extended state coherence transform. *Proc. IEEE Workshop on Statistical Signal Processing (SSP)*, September 2009.

[49] A. Lombard, H. Buchner, and W. Kellerman. Multidimensional localization of multiple sound sources using blind adaptive mimo system identification. In *IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Heidelberg, Germany, 2006.

[50] H. Mathis M. Joho and G.S. Moschytz. Combined blind/nonblind source separation based on the natural gradient. *IEEE Signal Processing Letters*, 8(8):236–238, 2001.

[51] D. Maino, A. Farusi, C. Baccigalupi, F. Perrotta, A. J. Banday, L. Bedini, C. Burigana, G. De Zotti, K. M. Gorski, and E. Salerno. All-sky astrophysical component separation with fast independent component analysis (fastica). *Monthly Notices of the Royal Astronomical Society*, 2001.

[52] Marco Matassoni, Maurizio Omologo, Diego Giuliani, and Piergiorgio Svaizer. Hmm training with contaminated speech material for distant-talking speech recognition. *Computer Speech and Language*, pages 205–223, 2002.

[53] K. Matsuoka and S. Nakashima. Minimal distortion principle for blind source separation. In *Proceedings of International Symposium on ICA and Blind Signal Separation*, San Diego, CA, USA, December 2001.

[54] Tom Melia and Scott Rickard. Underdetermined blind source separation in echoic environments using desprit. *EURASIP Journal on Advances in Signal Processing, Article ID 86484, 19 pages*, 2006.

[55] S. Miyabe, T. Takatani, H. Saruwatari, K. Shikano, and Y. Tatekura. Barge-in and noise-free spoken dialogue interface based on sound field control and semi-blind source separation. In *Proceedings of the European Signal Processing Conference*, pages 232–236, Florence, Italy, September 2007.

[56] McKeown MU, Makeig S, Brown GG, Jung PT, Kindemann SS, Bell AJ, and Sejnowski TJ. Analysis of fmri data by blind separation into independent component analysis. *Human Brain Mappping*, 1:160–188, 1998.

[57] Ryo Mukai, Hiroshi Sawada, Shoko Araki, and Shoji Makino. Real-time blind source separation an DOA estimation using small 3-D microphone array. In *Proceedings of International Workshop on Acoustic Echo and Noise Control*, pages 45–48, Eindhoven, The Netherlands, September 12-15 2005.

[58] Ryo Mukai, Hiroshi Sawada, Shoko Araki, and Shoji Makino. Frequency-domain blind source separation of many speech signals using near-field and far-field models. *EURASIP J. Appl. Signal Process.*, 2006:200–200, uary.

[59] F. Nesta and M. Omologo. Generalized state coherence transform for multidimensional localization of multiple sources. In *Accepted for publication in Proc. WASPAA*, October 2009.

[60] F. Nesta, M. Omologo, and P. Svaizer. Multiple TDOA estimation by using a state coherence transform for solving the permutation problem in frequency-domain bss. In *Proceedings of MLSP*, Cancun, Mexico, October 2008.

[61] F. Nesta, M. Omologo, and P. Svaizer. A novel robust solution to the permutation problem based on a joint multiple TDOA estimation. In *Proceedings of IWAENC*, Seattle, USA, September 2008.

[62] F. Nesta, P. Svaizer, and M. Omologo. A bss method for short utterances by a recursive solution to the permutation problem. In *Proceedings of SAM*, Darmstadt, Germany, July 2008.

[63] F. Nesta, P. Svaizer, and M. Omologo. Cumulative state coherence transform for a robust two-channel multiple source localization. In Tlay Adali, Christian Jutten, Joo Marcos Travassos Romano, and Allan Kardec Barros, editors, *ICA*, volume 5441 of *Lecture Notes in Computer Science*, pages 290–297. Springer, 2009.

[64] F. Nesta, P. Svaizer, and M. Omologo. Robust two-channel tdoa estimation for multiple speaker localization by using recursive ica and a state coherence transform. In *Proceedings of ICASSP*, Taipei, Taiwan, 2009.

[65] F. Nesta, P. Svaizer, and M. Omologo. Convolutive bss of short mixtures by ica recursively regularized across frequencies. Accepted in IEEE Transactions on Audio, Speech, and Language Processing, 2010.

[66] F. Nesta, T. S. Wada, and B.-H. Juang. Coherent spectra estimation for a robust solution to the permutation problem. In *proceeding of WASPAA*, October 2009.

[67] N.Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, 2001.

[68] M. Omologo and P. Svaizer. Use of the crosspower-spectrum phase in acoustic event location. In *IEEE Transactions on Speech and Audio Processing*, volume 5, pages 288–292, May 1997.

[69] Sea-Nae Park, Dong-Gyu Sim, Seoung-Jun Oh, Chang Beom Ahn, Yung Lyul Lee, Hochong Park, Chae-Bong Sohn, and Jeongil Seo. Residual signal compression based on the blind signal decomposition for video coding. In *ICUCT*, pages 11–19, 2006.

[70] L. Parra and C. Alvino. Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Transaction on Speech and Audio Processing*, 10(6):352–362, September 2002.

[71] L. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Trans. on Speech and Audio Processing*, pages 320–327, May 2000.

[72] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra. A survey of convolutive blind source separation methods. In *Springer Handbook of Speech*, November 2007.

[73] M. S. Pedersen, D. Wang, J. Larsen, and U. Kjems. Two-microphone separation of speech mixtures. *IEEE Transactions on Neural Networks*, 19(3):475–492, mar 2008. DOI: 10.1109/TNN.2007.911740.

[74] D.-T. Pham, Ch. Servière, and H. Boumaraf. Blind separation of convolutive audio mixtures using nonstationarity. In *Proceedings of ICA 2003 Conference*, Nara, Japan, Apr. 2003.

[75] D.-T. Pham, Ch. Servière, and H. Boumaraf. Blind separation of speech mixtures based on nonstationarity. In *Proceedings of ISSPA*, Paris, France, July 2003.

[76] Dinh-Tuan Pham, Zaher El-Chami, Alexandre Guérin, and Christine Servière. Modeling the short time fourier transform ratio and application to underdetermined audio source separation. In Tülay Adali, Christian Jutten, Jo ao Marcos Travassos Romano, and Allan Kardec Barros, editors, *Independent Component Analysis and Signal Separation*, pages 98–105. Springer, 2009.

[77] Andr Puga and Artur Pimenta Alves. The independent component analysis approach to interframe video coding. In *proceedings of International Conference on Telecommunications, ICT96*, pages 620–623, Istanbul, Turkey, 1996.

[78] E. Robledo-Arnuncio, H. Sawada, and S. Makino. Frequency domain blind source separation of a reduced amount of data using frequency normalization. In *Proceedings on ICASSP*, volume 5, May 2006.

[79] Makeig S, Bell AJ, Jung T-P, and Sejnowski TJ. Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, 1996.

[80] Hiroshi Saruwatari, Satoshi Kurita, Kazuya Takeda, Fumitada Itakura, Tsuyoki Nishikawa, and Kiyohiro Shikano. Blind source separation combining independent component analysis and beamforming. *EURASIP J. Appl. Signal Process.*, 2003(1):1135–1146, 2003.

[81] H. Sawada, S. Araki, and S. Makino. Frequency-domain blind source separation. In *Blind Speech Separation*. Springer, September 2007.

[82] H. Sawada, S. Araki, R. Mukai, and S. Makino. Solving the permutation problem of frequency-domain bss when spatial aliasing occurs with wide sensor spacing. In *Proceedings of ICASSP*, volume 5, Toulouse, France, May 2006.

[83] H. Sawada, S. Araki, R. Mukai, and S. Makino. Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1592–1604, July 2007.

[84] 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 Transactions on Speech and Audio Processing*, 12:530–538, September 2004.

[85] H. Sawada, R. Mukai, S. de la Kethulle de Ryhove, S. Araki, and S. Makino. Spectral smoothing for frequency-domain blind source separation. In *Proceedings of IWAENC*, Kyoto, Japan, September 2003.

[86] Hiroshi Sawada, Stefan Winter, Ryo Mukai, , Shoko Araki, and Shoji Makino. Estimating the number of sources for frequency-domain blind source separation. In *Proceedings of international conference on independent component analysis and blind signal separation*, pages 610–617, Granada, Spain, September 22-24 2004.

[87] Friedrich Schmid and Mark Trede. Simple tests for peakedness, fat tails and leptokurtosis based on quantiles. *Comput. Stat. Data Anal.*, 43(1):1–12, 2003.

[88] Ch. Servière and D. T. Pham. Permutation correction in the frequency domain in blind separation of speech mixtures. *EURASIP J. Appl. Signal Process.*, 2006(1):1–16, January 2006.

[89] P. Smaragdis and P. Boufounos. Learning source trajectories using wrapped-phase hidden markov models. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pages 114–117, Oct. 2005.

[90] M. M. Sondhi, D. R. Morgan, and J. L. Hall. Stereophonic acoustic echo cancellation - an overview of the fundamental problem. *IEEE Signal Processing Letters*, 2(8):148–151, August 1995.

[91] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone. An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses. *JASA*, 97(2):1119–1123, 1995.

[92] R. Taylor and G. Dailey. The super-directional acoustic sensor. *Proceedings of OCEANS*, 1:386–391, 1002.

[93] T.Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *In Proceedings of Neural Information Processing System*, pages 556–562, 2001.

[94] Keisuke Toyama and Mark D. Plumbley. Estimating phase linearity in the frequency-domain ica demixing matrix. In *ICA*, volume 5441 of *Lecture Notes in Computer Science*, pages 362–370. Springer, 2009.

[95] E. Vincent, S. Araki, and P. Bofill. The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation. In *ICA '09: Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*, pages 734–741, Berlin, Heidelberg, 2009. Springer-Verlag.

[96] E. Vincent, C. Fèvotte, and R. Gribonval. Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech and Language Processing*, 14(4):1462–1469, 2006.

[97] T. S. Wada, S. Miyabe, and B.-H. Juang. Use of decorrelation procedure for source and echo suppression. In *Proceedings of IWAENC*, Seattle, USA, September 2008.

[98] DeLiang Wang. Timefrequency masking for speech separation and its potential for hearing aid design. *Trends Amplif*, 12(4):332–53, 2008.

[99] B. Widrow and M. E. Hoff, Jr. Adaptive switching circuits. *IRE Wescon Convention Record*, pages 96–104, 1960, Part 4.

[100] B. Widrow and S. D. Stearns. *Adaptive Signal Processing*. Prentice Hall, 1985.

[101] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *Signal Processing, IEEE Transactions on*, 52(7):1830–1847, July 2004.