UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
**ICT International Doctoral School**

# EXPLOITING SPATIAL AND SPECTRAL INFORMATION FOR AUDIO SOURCE SEPARATION AND SPEAKER DIARIZATION

Mahmoud Fakhry

Advisor

Maurizio Omologo

Fondazione Bruno Kessler

Co-Advisor

Piergiorgio Svaizer

Fondazione Bruno Kessler

December 2016

# Acknowledgements

My gratitude goes to my supervisor, Maurizio Omologo, who gave me the opportunity to do a PhD. I highly appreciate his patience and wisdom. His guidance throughout this period helped me to get to this point.

I am very thankful to my co-supervisor, Piergiorgio Svaizer, for sharing his valuable time and for his gracious recommendations.

I would like to thank the members of the SHINE research unit for the enjoyable conversations at lunch and coffee time. Thanks also to the people of the ICT doctoral school,

I would like to thank the members of the NTT Communication Science Laboratories, Kyoto, Japan, who helped me to enrich the content of the thesis by accepting me to spend six months as a paid intern.

I give my thanks with a grateful heart to my family and friends.

# Abstract

*The goal of multichannel audio source separation is to produce high quality separated audio signals, observing mixtures of these signals. The difficulty of tackling the problem comes from not only the source propagation through noisy and echoing environments, but also overlapped source signals. Among the different research directions pursued around this problem, the adoption of probabilistic and advanced modeling aims at exploiting the diversity of multichannel propagation, and the redundancy of source signals. Moreover, prior information about the environments or the signals is helpful to improve the quality and to accelerate the separation.*

*In this thesis, we propose methods to increase the effectiveness of model-based audio source separation methods by exploiting prior information applying spectral and sparse modeling theories. The work is divided into two main parts.*

*In the first part, spectral modeling based on Nonnegative Matrix Factorization is adopted to represent the source signals. The parameters of Gaussian model-based source separation are estimated in sense of Maximum-Likelihood using a Generalized Expectation-Maximization algorithm by applying supervised Nonnegative Matrix and Tensor Factorization, given spectral descriptions of the source signals. Three modalities of making the descriptions available are addressed, i.e. the descriptions are on-line trained during the separation, pre-trained and made directly available, or pre-trained and made indirectly available. In the latter, a detection method is proposed in order to identify the descriptions best representing the signals in the mixtures.*

*In the second part, sparse modeling is adopted to represent the propagation environments. Spatial descriptions of the environments, either deterministic or probabilistic, are pre-trained and made indirectly available. A detection method is proposed in order to identify the deterministic descriptions best representing the environments. The detected descriptions are then used to perform source separation by minimizing a non-convex $l_0$-norm function. For speaker diarization where the task is to determine "who spoke when" in real meetings, a Watson mixture model is optimized using an Expectation-Maximization algorithm in order to detect the probabilistic descriptions, best representing the environments, and to estimate the temporal activity of each source.*

*The performance of the proposed methods is experimentally evaluated using different datasets, between simulated and live-recorded. The elaborated results show the superiority of the proposed methods over recently developed methods used as baselines.*

**Keywords** Source separation, speaker diarization, spectral information, spatial information, probabilistic modeling, and spectral and spatial modeling.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

You are at a crowded party, where the music is loud, people are laughing, and different conversations are running all around you. However, you are able to focus on the specific voice you want to hear and you can also temporarily alternate between different voices. The issue of focusing on a single speaker is called the "cocktail party" problem. The ears receive stereo mixtures of overlapped sounds. To hear what a certain person is saying, the brain is able to extract the individual sound of interest from those mixtures. Multichannel audio processing for source separation aims at reproducing, by means of an automatic system, this functionality that we achieve so effectively by our ears and brain.

The addressed scenario involves the use of multiple microphones to capture, at different points in space, mixtures of individual signals emitted by multiple audio sources. The individual signals are mixed through a Multiple-Input Multiple Output (MIMO) mixing system, representing sound propagation from each source to each microphone (a mixing process). Each microphone receives overlapped and modified copies of the individual signals, plus the additive background noise of the environment. Based on the information contained in the mixture signals, it is required to design a separation system, termed a reconstruction system or an in-

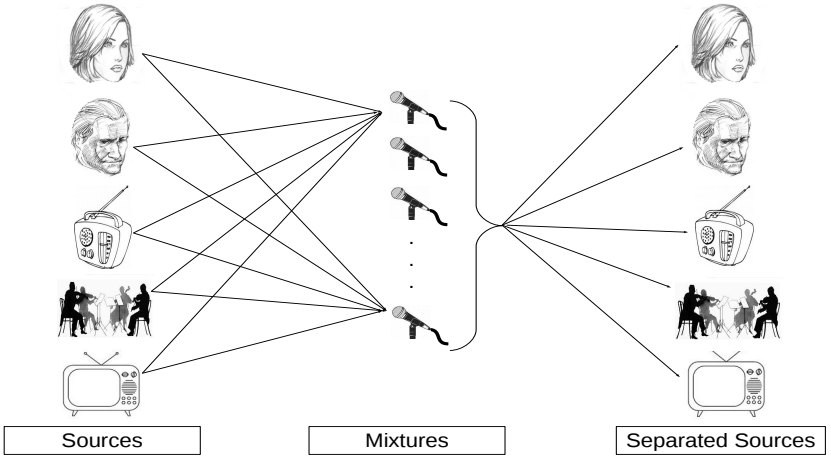


Figure 1.1: A multichannel mixing/separation process.

verse system, in order to retrieve the individual audio signals (a separation process). Figure 1.1 shows an example of the mixing/separation process.

Because the output of the source separation system consists of audio signals that may be further processed or listened to, the topic has attracted extensive research work. The importance of tackling the problem for speech processing comes from its possible applicability including:

- Speech separation to segregate simultaneous signals in echoing and noisy environments for audio communication (e.g. teleconferencing).
- Speech enhancement to extract a signal of interest in the presence of background noise when the rest are considered to be nuisance signals, for audio surveillance.
- Speaker diarization to determine "who spoke when" in real meetings, for voice activity detection.
- A front-end processing step for automatic speech recognition of multiple speakers.
- Automatic indexing of audio databases.

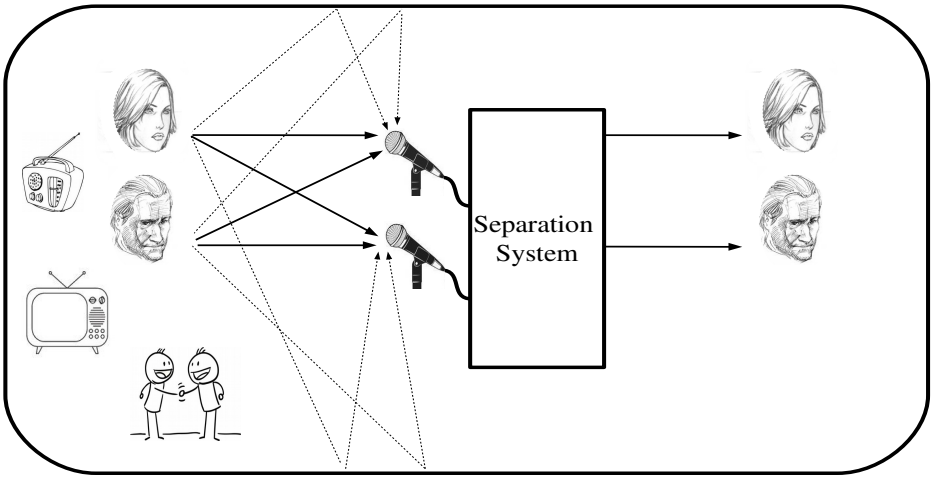


Figure 1.2: A multichannel mixing process in an echoing and noisy environment.

## 1.1 Problem statement

Starting from multiple observations of mixed audio signals generated by different sources, a successful separation system extracts each of the original source signals. We consider the case where many people are simultaneously talking inside an echoing room, each person is considered as a source of an audio signal. Furthermore, if an array of microphones is installed inside the room, each microphone picks up a combination of reproduced audio signals (see Figure 1.2). These multiple combinations are defined as **observed mixtures** of the audio signals. The reproduced audio signal is a modified copy of the original one, as the original audio signal bounces off walls and objects inside the room. The set of all the paths that an audio signal takes to arrive at a microphone is known as a propagation channel. The multiple propagation channels that the multiple audio signals broadcast through are defined as a **mixing system**, which is described by a set of mixing parameters. Starting from such observed mixtures where the sources interfere each other and the signals are distorted by the propagation through the noisy environment, it is indeed a challenging task to obtain appropriate estimation of each original audio signal.

In the search for a solutions that achieves a good performance, numerous efforts have been undertaken.

- A first trend is to simulate the human auditory system source formation process [84], assuming that the sources are instantaneously not overlapped. Exploiting this sparseness assumption, using auditory characteristics-based classification techniques [6, 48, 76], the observed mixtures are classified into clusters, each one belonging to a source.

- A second trend is to separately tackle the problem in short sub-bands. Inside each sub-band, segments of source signals are obtained by estimating separation filters applying Independent Component Analysis (ICA) [46, 59]. Later, the segments belonging to each source are grouped using source-based or spatial-based information.

- An alternative trend is to build generative models that integrate knowledge about the source production process [3]. Model-based methods exploit all the available knowledge at once and can consider multiple overlapped sources [8, 30, 31, 58]. The source signals are obtained by first estimating model hyper-parameters using for example Expectation-Maximization (EM) [27] or convex/non-convex optimization [68], then applying a suitable filtering process. In order to improve the performance, along with model-based methods, advanced modeling theories such as the spectral modeling theory [25] and the sparse modeling theory [21] are involved in the models in [10, 41, 70, 71].

- Better performance and faster convergence could be achieved when the separation system is fed by prior information on a particular mixing process. Nowadays, a new trend has grown up to take advantage of prior information on either source signals [50, 51, 74, 85] or propagation channels [11, 31, 56, 60].

## 1.2 Objectives of the thesis

In line with the recent trends, we propose to exploit prior information applying advanced modeling theories to improve the effectiveness of model-based methods. The main focus of the thesis is on multichannel audio processing for source separation, specifically stereo source separation, i.e. where the number of microphones is two. Furthermore, we exploit the deep study of source separation to investigate speaker diarization. Considering the multichannel audio processing for source separation and speaker diarization, the main objectives of the thesis are:

- To explore source separation aiming at introducing a clear explanation of its mathematical formulation and probabilistic modeling.
- To present a general review of the-state-of-the-art methods and to figure out their limitations.
- To introduce advanced modeling theories that can be applied to improve the existing methods.
- To discuss and analyse in details proposed methods for improving the overall performance.
- To experimentally evaluate the performance of the proposed methods in several conditions, including simulated and real environments.
- To compare the performance of the proposed methods to recently developed methods.
- To explore speaker diarization aiming to present a short review of the-state-of-the-art methods, and to discuss a proposed solution.

## 1.3 Original contribution

In case of a noisy environment, the elements of multichannel audio precessing are mostly defined as a) audio signals, b) propagation channels, and c) background noise. For the tasks under investigation, the audio signals are represented by their spectral descriptions, and the propagation channels are represented by their spatial descriptions. Building on this, the original contribution of this thesis can be divided into two main parts (see Figures 1.3 and 1.4), namely:

- Exploiting spectral information about audio signals in observed mixtures, for source separation, by applying the spectral modeling theory using Nonnegative Matrix and Tensor Factorization (NMF/NTF).
- Exploiting spatial information about propagation channels used to generate observed mixtures, for source separation and speaker diarization, by applying the sparse modeling theory.

### 1.3.1 Exploiting spectral information for source separation

In this part of the work, we propose to combine spectral modeling using NMF/NTF [25] with one of the recently grown up model-based audio source separation algorithms [30, 31, 39, 70]. Observed mixtures of audio signals are probabilistically described by a multivariate complex Gaussian model parametrized by spectral and spatial parameters, i.e. spectral variances describing the audio signals, and spatial covariance matrices describing the propagation channels. The parameters are estimated applying a Generalized Expectation-Maximization (GEM) algorithm [27]. In this sense, we propose to reduce the estimation dependency of the parameters, and to exploit spectral descriptions of sources. Our contribution in this direction is listed as follows:

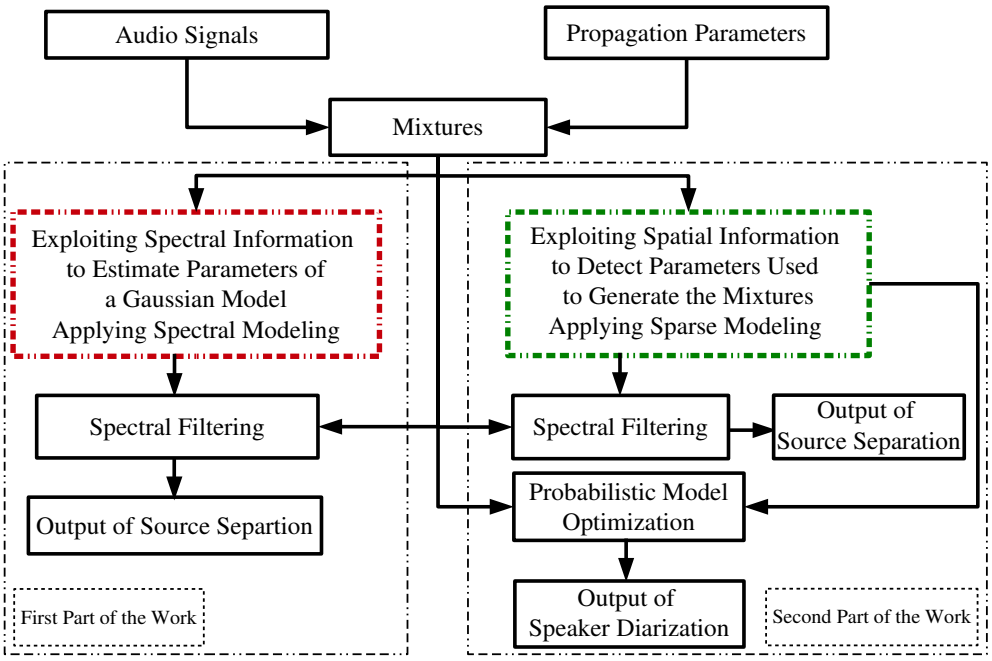


Figure 1.3: A description of the proposed work. Highlighted blocks refer to novel contributions.

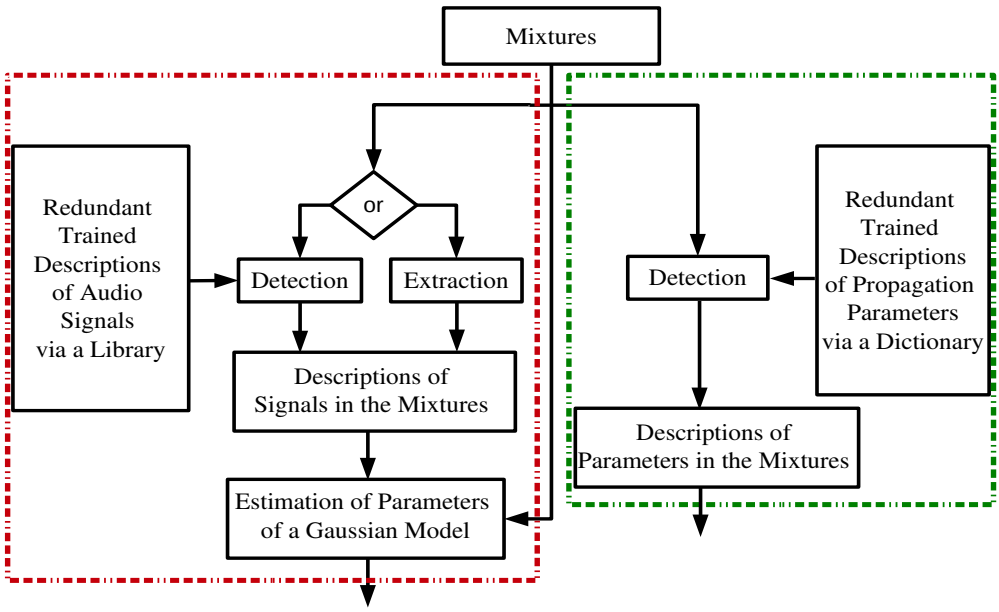


Figure 1.4: A detailed description of the thesis contribution.

- Blindly estimating the spectral parameters regardless of the spatial ones. Moreover, the spatial parameters are estimated by exploiting spectral descriptions of the estimated spectral parameters applying spectral modeling using supervised NMF/NTF. We propose that the descriptions are either trained on-line, or pre-trained in advance using a training dataset.

- Jointly estimating the parameters applying spectral modeling using supervised NMF/NTF. Spectral descriptions of the spectral parameters are made available to perform the supervised factorization. We propose that the descriptions are either extracted on-line, or made indirectly available through a redundant library containing trained spectral descriptions of many sources. In the latter case, a detection step is proposed in order to identify the descriptions that best represent source signals in observed mixtures.

### 1.3.2 Exploiting spatial information for source separation and speaker diarization

Motivated by the idea of adopting a collection of patterns (over-complete dictionaries) for sparse modeling, in this part of the work, we propose to build dictionaries composed of spatial descriptions representing propagation channels. For possible positions of a source (a finite set of spatial positions), the spatial description of propagation channels representing a certain position forms a subset (a column) of the dictionary. Observing mixtures of audio signals, given the dictionary, the spectral descriptions that best match descriptions of propagation channels used to generate the mixtures are detected, in order to perform source separation and speaker diarization. Our contribution in this direction is listed as follows:

- The description of a column is represented by trained parameters

of propagation channels for source separation. Furthermore, in case that there is mismatch between the parameters of the dictionary and the parameters used to generate observed mixtures, an unsupervised dictionary adaptation step using weighted Independent Component Analysis (wICA) is proposed to reduce such mismatch [67].

- The description of a column is represented by trained hyper-parameters of a probabilistic model describing propagation channels for speaker diarization. Moreover, in the presence of background noise and interferences, a Laplace distribution [4] is proposed to model the accompanying corruption that is generated by the noise and the interferences.

## 1.4 Dissemination of the thesis

The dissemination of the proposed work is listed as follows:

- M. Fakhry, P. Svaizer, and M. Omologo. *Audio source separation in reverberant environments using $\beta$-divergence based nonnegative factorization.* IEEE Transactions on Audio, Speech, and Language Processing (Passed the first phase of revision with two major and one minor revisions).

- M. Fakhry, N. Ito, S. Araki, and T. Nakatani. *Modeling audio directional statistics using a probabilistic dictionary for speaker diarization in real meetings.* In Proceedings of IWAENC, 2016.

- M. Fakhry, P. Svaizer, and M. Omologo. *Estimation of the spatial information in Gaussian model based audio source separation using weighted spectral bases.* In Proceedings of EUSIPCO, 2016.

- M. Fakhry, P. Svaizer, and M. Omologo. *Audio source separation using a redundant library of source spectral bases for nonnegative tensor*

*factorization.* In Proceedings of ICASSP, 2015.

- M. Fakhry, P. Svaizer, and M. Omologo. *Reverberant audio source separation using partially pre-trained nonnegative matrix factorization.* In Proceedings of IWAENC, 2014.

- F. Nesta and M. Fakhry. *Unsupervised spatial dictionary learning for sparse underdetermined multichannel source separation.* In Proceedings of ICASSP, 2013.

- M. Fakhry and F. Nesta. *Underdetermined source detection and separation using a normalized multichannel spatial dictionary.* In Proceedings of IWAENC, 2012.

## 1.5 Organization of the thesis

The rest of the thesis is organised as follows. Chapter 2 provides a general review of different mechanisms and techniques commonly applied to tackle audio source separation. Chapter 3 introduces the mathematical formulation and probabilistic modeling of the problem. Moreover, we give a brief introduction of the statistical estimation of the model parameters using the Generalised Expectation-Maximization (GEM) algorithm, and of spectral and sparse modeling theories.

A first proposed method to estimate the parameters of the Gaussian model-based audio source separation system using spectral modeling based on NMF/NTF is presented in Chapter 4. In this method, we reduce the estimation dependency of the parameters, while the separation system is informed by pre-trained spectral descriptions of audio signals in observed mixtures. Chapter 5 presents a modification of the estimation method proposed in the previous chapter, so that the performance is improved and

the whole separation system can work with pre-trained or on-line trained spectral descriptions.

A method for stable estimation of the model parameters using spectral modeling based on NMF/NTF is detailed in Chapter 6. In this method, we also reduce the estimation dependency by jointly updating the parameters. The separation system can work with pre-trained or on-line trained spectral descriptions. Furthermore, we assume that the spectral descriptions are indirectly available through a redundant library, and we detect the descriptions that best represent audio signals in observed mixtures.

Chapter 7 presents an audio source separation method based on the sparse modeling theory. In this method, we build a spatial dictionary containing trained parameters describing a finite set of propagation channels. Applying the sparse modeling theory, we detect the parameters that match the parameters used to generate observed mixtures of audio signals. In case there is mismatch between the two sets of parameters, an adaptation step is proposed to reduce such mismatch.

Chapter 8 provides a general review of speaker diarization. Moreover, the chapter presents a speaker diarization method using a spatial dictionary composing of trained parameters of probabilistic models describing a finite set of propagation channels. Applying the sparse modeling theory, we detect the parameters that match the parameters used to generate observed mixtures of audio signals by optimizing a probabilistic model. Furthermore, the temporal activity of each source is estimated for speaker diarization. Finally, the conclusion together with a perspective on future work are drawn in Chapter 9.

# Chapter 2

# Review

Multichannel audio source separation deals with the output of a Multiple-Input Multiple-Output (MIMO) mixing system. Based on the general mechanism adopted to solve the problem, approximated versions of the original source signals are directly obtained, or a separation system is estimated, then an optimization problem is adapted to extract approximated versions of the original source signals. According to the available prior information exploited to perform source separation, the-state-of-the-art is classified into three main categories of methods, namely:

- Blind Source Separation (BSS) where the problem is tackled without any prior information both on audio signals and on propagation channels.
- Source-based informed separation where prior information about audio signals is assumed to be available in different forms.
- Mixing system-based informed separation where the separation system is fed by information about propagation channels.

Several approaches tackling the problem have appeared in the literature during the last two decades [23, 46, 59]. They essentially differ in the pre-assumptions about sources:

- Some of them assume that the sources are statistically independent without temporal structure. Higher order statistics, such as mutual information, entropy, and non-Gaussianity, are mainly used to solve the problem assuming that only one source has a normal distribution.

- Other algorithms suggest less restrictive conditions than independence. They assume that each source has non-vanishing temporal correlation. In this case, second order statistics such as cross-correlation are sufficient to extract the sources if they have non-identical power spectrum shapes.

- Non-stationarity of the sources is exploited as a pre-assumption, where the source variances do not vary in time. Accordingly, second order statistics are able to estimate the sources if they are with non-identical non-stationarity properties.

- Various diversities of the sources, typically, time, frequency, and time-frequency are used as pre-assumptions. In this case, the sources are interpreted as localized, sparse or structured signals. Source signals are then obtained by masking or filtering the observed mixtures in the diversity domain.

On the other side, the configuration of the mixing process can also be constrained by assuming a specific geometrical setup identifying the spatial positions of sources and microphones, i.e. the distance between the microphones, the distance between the microphones and the sources, the angular distance between the sources, etc.

In this multichannel multiple-sources scenario, the spatial diversity between multiple signal observations at multiple microphones is generated as a result of multichannel propagation and different source spatial positions. The spatial diversity, combined with one or more of the above

source signal pre-assumptions, can be exploited to perform multichannel audio processing for source separation.

## 2.1 Blind source separation

Time-frequency representations of observed mixtures of audio signals are obtained through the discrete Short Time Fourier Transform (STFT), which is also known as the most common spectral-temporal representation of a signal. The discrete STFT represents a signal as a matrix of complex components, i.e. time-frequency points. The index of each point is defined by two variables, i.e. a frequency bin and a time frame. It has been proven that the time-frequency representation is the most suitable domain to process audio signals for the tasks under investigation.

Figure 2.1 illustrates the general mechanisms used to extract the contribution of each original source signal from the observed mixtures, for the case that two observed mixtures are generated by mixing two original source signals. Time-frequency representations of the original source signals can be estimated by clustering time-frequency representations of the observed mixtures point-by-point applying binary masking [90] or soft masking [44, 47, 77]. Full-band of frequency bins of the observed mixtures can also be binary clustered as in [48, 76]. Spectral multichannel Wiener filtering is applied to extract the contribution of each original source signal by estimating time-frequency multichannel filter gains [30, 31, 70, 71]. Time-invariant multichannel spectral filters can also be estimated inside each frequency bin over all the time frames of the observed mixtures [46, 59].

Without prior information on either the source signals or the propagation environments, Blind Source Separation (BSS) methods broadly use prior information on the source production and/or propagation processes. The BSS methods can be classified into three main groups, based on the

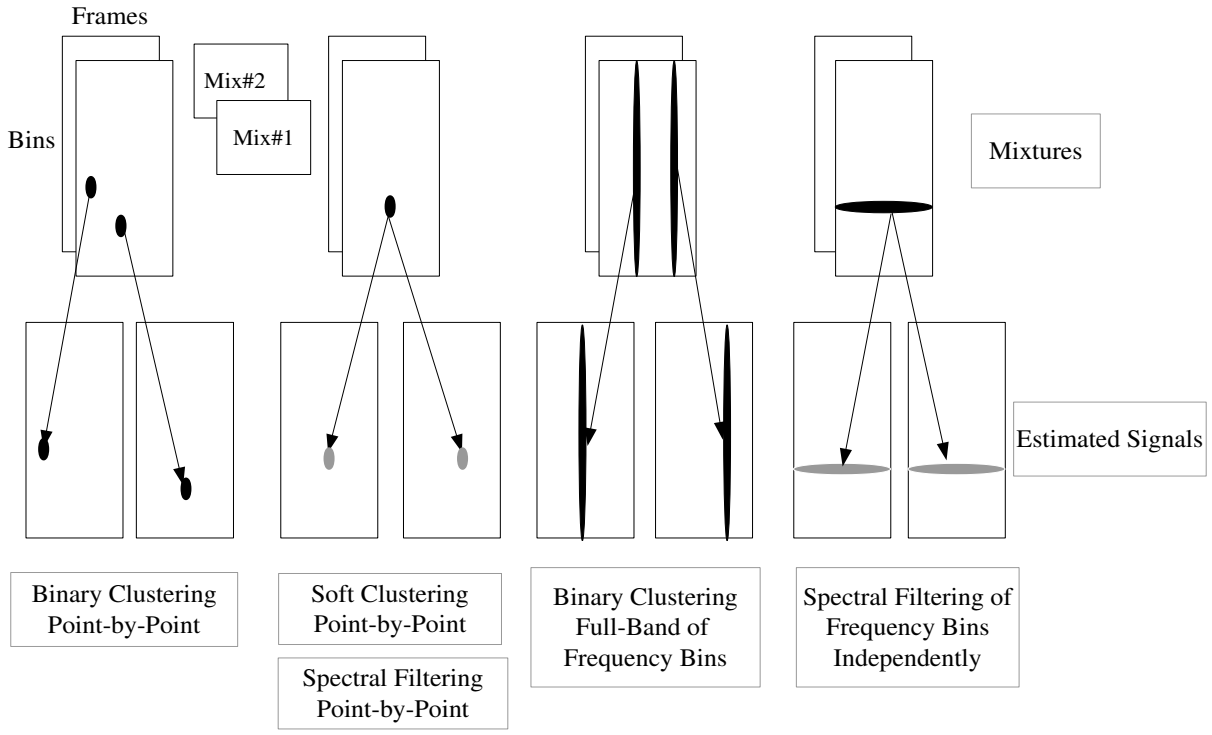


Figure 2.1: Different mechanisms to extract each signal from mixtures of signals.

general methodology and technique used to retrieve the source signals, i.e. clustering-based separation, Independent Component Analysis (ICA-based separation), and model-based separation.

### 2.1.1 Clustering-based separation

A first trend to source separation is to simulate the human auditory system source formation process [84]. In the time-frequency domain, the observed mixtures are represented as the product of time-frequency representations of audio signals and complex-valued mixing vectors. Assuming that each time-frequency point of the mixture signals is dominated by one source, the clustering is performed point-by-point. The mixture signals are classified into small clusters using auditory characteristics-based classification

techniques. These auditory characteristics state for example that the time-frequency points are clustered together when they have harmonic frequencies, smooth spectral envelopes, correlated amplitude and frequency modulations, or similar inter-channel time and intensity differences. In order to obtain a single cluster per source, the resulting clusters are processed by exploiting source-based prior knowledge such as the timber of a known speaker.

To preserve the continuity of source spectral structures and to avoid musical noise, the clustering is performed at full-band of frequency bins, assuming that one source dominates at each time frame. Most of the existing techniques rely on estimating the Time-Difference-Of-Arrival (TDOA) [6, 76] of each source at multiple microphones, or Interaural Time and Level Differences (ITD/ILD) [48] for a two-microphone stereo case. Such techniques work well in propagation environments with low reverberation. However, in propagation environments with high reverberation, the above clustering techniques are suggested as a first stage of two stages-based separation. In the second stage, source activities that are represented by amplitude envelopes of the sources [87] or clustering posterior probabilities [44] are used to regroup the time-frequency points belonging to the same source. A weighted Minimum Mean Square Error (MMSE) [32] solution is proposed in [78]. The known MMSE estimator is weighted by source activity posterior probability that is estimated using extracted spatial and spectral source cues. Furthermore, the time-frequency points are clustered in the direction that optimizes the weighted MMSE, and then regrouped as in [44].

Since these clustering-based source separation methods assume that one source is approximately dominating at each time-frequency point or at each time frame, they fail when the source signals do not fulfill the assumption of time-frequency sparseness.

### 2.1.2 ICA-based separation

Assuming that more than one source is active at each frequency bin over all the time frames, the problem is separately solved at each bin using Independent Component Analysis (ICA) [46, 59]. ICA looks for estimating linear spectral mixing/separation filters that minimize the statistical dependence between their output signals. Corresponding to the representation in the clustering-based audio source separation, the observed mixtures are represented as the product of time-frequency representations of audio signal and complex-valued mixing vectors. With a less restrictive assumption than independence in ICA, Principal Component Analysis (PCA) tackles the source separation problem assuming that the signals are statistically uncorrelated [46, 59]. Source separation based on PCA is performed by estimating the complex mixing vectors as the principal vectors of covariance matrices of the observed mixtures, applying the singular value decomposition (SVD). Although PCA can approximately estimate uncorrelated signals, it is not enough to separate the signals, especially when the probability distributions of the sources are not Gaussian. In some source separation methods, PCA is used as a pre-processing step to ICA, acting as a whitening step.

On the other side, as a result of solving the problem at each frequency bin independently, a permutation ambiguity is generated between frequency bins of the separated source signals. The permutation should be aligned so that a separated signal contains frequency bins of the same source. The permutation problem is solved exploiting spatial information of sources in [77] and spatial-temporal information in [67]. Source spectral information is also exploited to avoid the permutation ambiguity in [80].

Given estimation of the mixing vectors, in the under-determined mixing model (i.e. when the number of observed mixtures is less than the number

of original source signals), exploiting the sparsity of audio signals [16, 55], the source signals are extracted using, for example, $l_0$-norm minimization [80] or soft masking [17, 44, 77].

Since ICA-based source separation algorithms rely on higher order statistics (independence), which require much amount of data to be processed, they fail to separate short-length signals.

Regrouping in clustering-based algorithms, and permutation alignment in ICA-algorithms rely on source activity estimation, hence this group of source separation methods has a problem of convergence when they are used to separate synchronized sources such as instrumental components in mixed music signals.

### 2.1.3 Model-based separation

An alternative trend to audio source separation is to build generative models that integrate knowledge about the source production process. The Bayesian theory [3] provides an appropriate framework to exploit such models. The probabilistic distribution of observed mixtures of audio signals is specified by a set of hidden variables, including **audio signals**, **mixing parameters**, and **conditional distributions** between these variables. Given mixtures of source signals, the model variables are estimated by standard algorithms such as Expectation-Maximization (EM) [27] or convex/non-convex optimization [68]. It seems very likely that the knowledge about the sources used by probabilistic model-based source separation methods turns out to be similar to that used by human auditory-based source separation methods. However, model-based separation methods exploit all the available knowledge at once and can consider multiple active sources at each time-frequency point.

Some of model-based algorithms separate the source signals using spectral filters, exploiting simplified spatial models for the propagation channels

[8, 30, 58]. A probabilistic model of Interaural Time and Level Differences (ITD/ILD) related to each source is proposed in [62], later an EM algorithm is applied to assign time-frequency points of observed mixtures to each source. A mapping between source positions and the Interaural Level Difference (ILD) is proposed in [26]. Furthermore, a Bayesian inference using variational EM is applied to estimate the source signals.

Recently, a Gaussian framework [30, 39, 70] has grown up as model-based audio source separation. Audio signals are locally modeled by multivariate complex Gaussian distributions. The covariance matrix of each probabilistic distribution is parametrized by spatial and spectral parameters. Assuming that the audio signals are statistically independent, the likelihood function of the observed mixtures is a complex multivariate Gaussian distribution. The parameters are estimated by maximizing the likelihood function applying an EM algorithm. Given estimation of the model parameters, the source signals are obtained by means of multichannel Wiener filtering. Other algorithms exploit spectral and temporal redundancies [70, 71] to estimate the parameters of the Gaussian model by representing audio signals using Nonnegative Matrix Factorization (NMF) [25, 53]. Furthermore, a combination of spatial modeling and NMF-based spectral modeling is proposed in [9] to estimate the model parameters.

Model-based source separation systems have achieved good performance, however, most of the systems require prior knowledge about the source production process and/or the source propagation process and they fail to separate audio signals mixed in difficult conditions.

Although BSS techniques are able to perform the separation task without specific prior information on either the source signals or the mixing environments, their robustness is still limited by low convergence, by high estimation variances, and by signal conditions not well fitting the general separation hypotheses.

For a given problem, prior information on the audio signals and/or on the mixing environments can be helpful to perform source separation with improved performance. Taking advantage of prior knowledge has recently raised as a new trend to increase the performance of BSS.

## 2.2 Source-based informed separation

Nesting prior knowledge about original source signals of a particular problem as side information promises to enhance the quality and accelerate the separation process [51]. Different prior information about the temporal activities of original sources in observed mixtures have been used to help the separation system to achieve its task in [42, 57, 74]. Recently, spectral modeling based on Nonnegative Matrix Factorization (NMF) [25] has grown up to solve the problem by decomposing the observed mixtures in a supervised scenario, i.e. the separation system is informed by spectral descriptions of original source signals.

### 2.2.1 NMF-based separation

Spectral modeling based on Nonnegative Matrix Factorization (NMF) represents the nonnegative magnitude or power spectrum of a signal as the product of two nonnegative matrices, i.e. a **spectral basis matrix** and an **activation coefficient matrix** [25, 53]. In an informed scenario, the separation system is fed by trained spectral basis matrices representing source signals in observed mixtures. Given the trained matrices and observed mixtures, the initial task of the separation system is to estimate appropriate activation coefficient matrices of the sources. Moreover, source separation is performed by first reconstructing either the magnitude or power spectra of all the source signals in the observed mixtures, later building clustering masks or applying Wiener filtering.

Spectral modeling based on NMF was applied for speech enhancement purposes in [79], where a universal dictionary of spectral bases is built in advance, using a training dataset of multiple speech signals. The selection of optimal spectral bases and the estimation of activation coefficient matrices, best representing source signals in observed mixtures, are done using block sparsity constraints on top of the NMF objective. For the same purpose of speech enhancement, a set of local dictionaries of spectral bases best modeling a source signal, e.g., speech and noise, is proposed in [50]. The activation coefficient matrices are obtained applying block regularized NMF so that only a small number of blocks are active at a time-instant.

To take advantage of more information about the signal structure, e.g., speech and noise, in [86] the authors propose to train prior models for activation coefficient matrices of sources during the training of spectral basis matrices. The speech signal is extracted in the presence of nonstationary noise by applying a regularized NMF algorithm using the trained basis matrices and exploiting the trained priors. In [85] the authors propose to train discriminative spectral basis matrices to be used for the estimation of activation coefficient matrices in the testing phase. Furthermore, the estimated coefficient matrices are used to optimize the trained discriminative spectral basis matrices.

## 2.3 Mixing system-based informed separation

To solve the source separation problem by exploiting prior information, an alternative way is to have knowledge about the mixing parameters. One trend is to adopt the multichannel sparse modeling theory. Sparse modeling assumes that by using a pre-specified redundant dictionary including a proper definition of a set of basis called atoms, a signal can be uniquely represented as a sparse vector in term of these atoms [21]. In the multi-

channel situation, i.e. when the source signals are recorded by an array of microphones, the atoms of the dictionary are defined through descriptions of possible source mixing parameters. Then, the problem of estimating the mixing system is relaxed to be a problem of detecting the system best matching a real system that generated the observed mixtures.

In moderately reverberant environments, this description can be approximated by modeling the atoms as the phase differences of an anechoic model [58]. Then a distance metric is used to match the observed mixtures and the modeled atoms. Moreover, a blind beamformer is applied to estimate the original source signals.

In more echoic environments, parameters describing propagation channels can be trained from the data off-line [60] to build a set of cancellation filters, when only one source at a time is active. Then Independent Component Analysis (ICA) is applied to detect the appropriate cancellation filter and to extract the original source signals. In [10, 11], a model-based spatial dictionary of propagation channels between the microphones and a set of points is built by estimating the parameters of the model when only one source at a time instant is active. Then a sparse modeling algorithm is applied to estimate the original source signals.

An alternative way to the spatial dictionary is to train probabilistic models for fixed source spatial positions. By adopting spatial priors, the authors in [31] propose to train the hyper-parameters of inverse-Wishart distributions for fixed source positions. Given the trained priors, the parameters of the Gaussian model are estimated in sense of Maximum-A-Posterior (MAP) using a Generalized Expectation-Maximization (GEM) algorithm. The source signals are then obtained by means of multichannel Wiener filtering.

# Chapter 3

# Formulation

The mixing and filtering process of the audio source signals can be represented by different possible mathematical or physical models:

- The **instantaneous model** is the simplest one to describe the process. The mixing system contains only constant attenuation factors over time, since propagation effects are not taken into account.
- To model the propagation delays of audio signals from their positions to the locations of microphones, an **anechoic model** is applied. The elements of the mixing system are fixed attenuation and propagation time delay factors.
- To simulate the realistic situations in a closed room, a **reverberated echoic model** is necessary. The mixing system is a set of filters fully representing the propagation channels from the positions of audio sources to the locations of microphones.

The third mixing model is the most challenging one and due to the filtering and mixing processes applied to the audio source signals, it is called a **convolutive mixing model**. On the other side, if the number of audio sources is equal to the number of microphones, the mixing process is referred to as **determined** mixing model. However, **over-determined** and

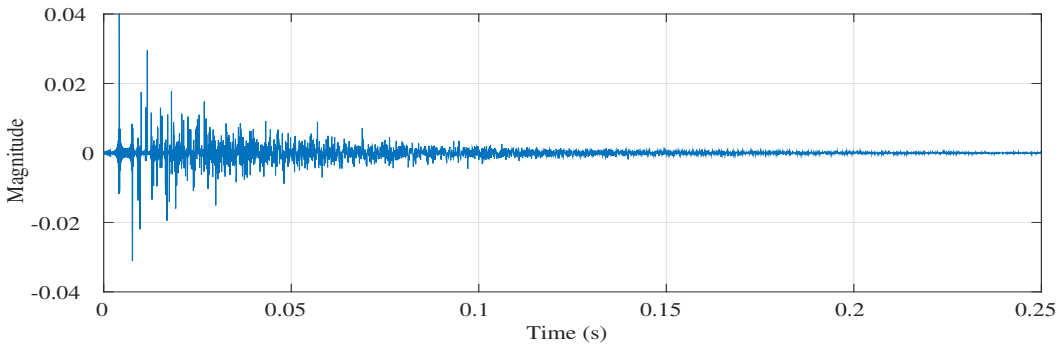


Figure 3.1: A synthetic model of a propagation channel (RIR) sampled at 16 kHz.

**under-determined** models have also been suggested to identify the cases when the number of audio sources is larger or smaller, than the number of microphones, respectively.

## 3.1 RIR model

The propagation channel from the position of an audio source to the location of a microphone is characterized by its corresponding Room Impulse Response (RIR) which provides a model indicating the effects of all possible propagation paths (see Figure 3.1). Each sample of the model represents a corresponding attenuation and phase shift. The attenuation factor defines the amount of signal amplitude decay, and the phase factor expresses the amount of time delay due to the propagation. The model of the RIR consists of a direct path, early reflections, and a reverberation tail. When an audio signal is produced in a real environment, the direct path conveys the signal directly to the microphone without interfering with any objects. The early reflections describe the reflections from the side walls, the ceiling, and the floor within a time period approximately $50 - 100$ ms after the direct path. The reverberation tail characterizes the high density reflections occurring after the first $50 - 100$ ms and it is often defined by an exponentially decaying envelope curve.

## 3.2 Mathematical formulation

A challenging situation arises when the number of microphones is less than the number of source signals (*under-determined mixing model*), and the surrounding mixing environment is reverberant (*convolutive mixing model*). To mathematically formulate this situation, we assume that $N$ sources are observed by an array of $M$ microphones, where $M < N$. The vector of observed mixtures $\mathbf{x}(t)$ received at the microphones at the time-instant $t$ is represented as

$$\mathbf{x}(t) = \sum_{n=1}^{N} \mathbf{c}_n(t), \tag{3.1}$$

where $\mathbf{c}_n(t)$ is a vector of source spatial images of the $n$-th source signal $s_n(t)$. The $m$-th component of the vector $\mathbf{c}_n(t)$ is represented as a function of the propagation channel $h_{nm}(t)$ from the $n$-th source position to the $m$-th microphone location as follows

$$c_{nm}(t) = \sum_{\tau=0}^{N_L} h_{nm}(\tau) s_n(t-\tau), \quad m = 1, .., M \tag{3.2}$$

where $N_L$ is the length of $h_{nm}(t)$. Using the discrete Short Time Fourier Transform (STFT), each point at the frequency bin $\omega$ and the time frame $l$, out of the total number of frequency bins $\Omega$ and time frames $L$, of the observed mixtures is represented by a $M \times 1$ vector of complex coefficients $\mathbf{x}(\omega, l)$. For the linear property of the STFT, the vector can be represented as the combination of $N$ source spatial images $\mathbf{c}_n(\omega, l)$ as follows

$$\mathbf{x}(\omega, l) = \sum_{n=1}^{N} \mathbf{c}_n(\omega, l). \tag{3.3}$$

At the microphones, $\mathbf{c}_n(\omega, l)$ is expressed as a function of the source signal $s_n(\omega, l)$ in the time-frequency domain and the vector of time-invariant frequency responses of $M$ propagation channels $\mathbf{h}_n(\omega)$ at the frequency $\omega$ as

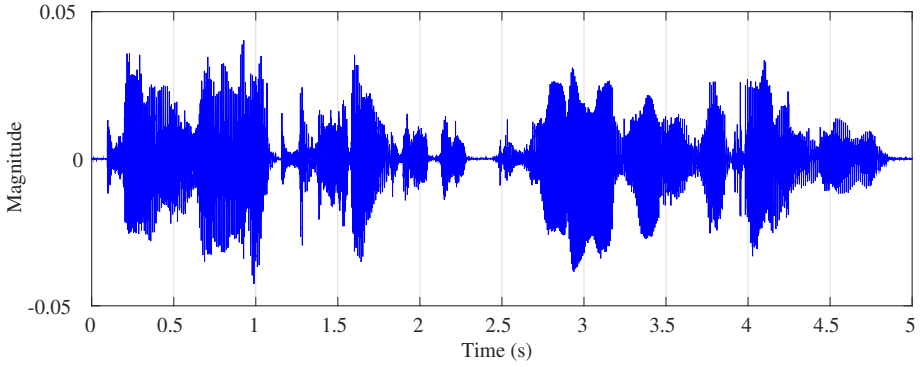


Figure 3.2: A speech signal waveform in the time domain sampled at 16 kHz.

$$\mathbf{c}_n(\omega, l) = \mathbf{h}_n(\omega) s_n(\omega, l). \tag{3.4}$$

Observing the signal mixtures $\mathbf{x}(\omega, l)$, source separation aims at finding corresponding estimate $\tilde{s}_n(\omega, l)$ of the original signal $s_n(\omega, l)$. Then the inverse short-time Fourier transform (ISTFT) is applied in order to represent back the estimated source signal $\tilde{s}_n(t)$ in the time domain.

## 3.3 Example of mixing process

As an example of the multichannel mixing process in a reverberant environment, we consider the case where two speakers at two different spatial positions are simultaneously talking in a closed room. Figure 3.2 shows an example of a speech signal waveform in the time domain sampled at 16 kHz. Figures 3.3 and 3.4 show the power spectrograms of the speech signal and its reverberated version as received at a distance microphone, i.e. the square values of $s_n(\omega, l)$ and $c_{nm}(\omega, l)$. Mixtures of two speech signals uttered by the two speakers are picked up by two spatially separated microphones which are placed far from the speaker positions. Figure 3.5 shows the power spectrograms of the two speech signals, and the power spectrograms of their two mixture signals received at the two microphones.

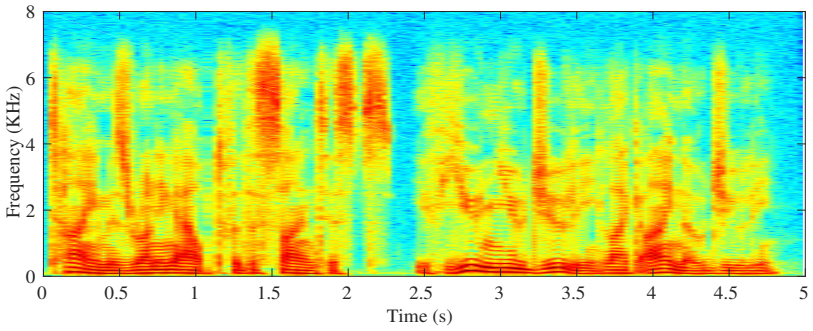


Figure 3.3: Power spectrogram of the speech signal represented by Figure 3.2.

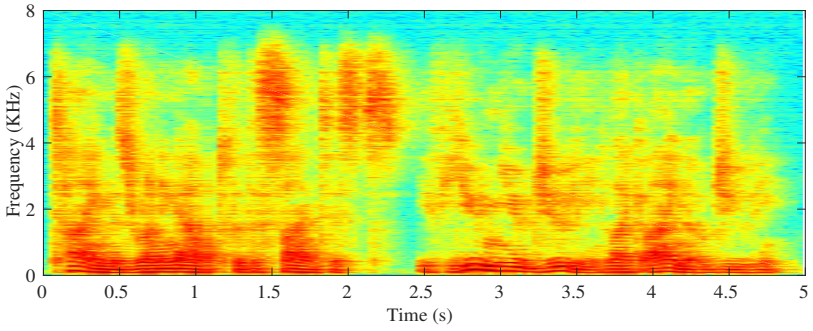


Figure 3.4: Power spectrogram of the speech signal represented by Figure 3.2 as received at a distant microphone in a reverberant environment.

## 3.4 Local Gaussian modeling

Over all the time-frequency points, the vectors of source spatial images $\mathbf{c}_n(\omega, l)$ are assumed to be independent, and probabilistically modeled by a zero-mean multivariate complex Gaussian distribution with a $M \times M$ covariance matrix $\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)$

$$\mathbf{c}_n(\omega, l) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)), \tag{3.5}$$

where $\mathbf{0}$ is a $M \times 1$ vector of zeros. Under the assumption that the source images are independent, the observed mixtures $\mathbf{x}(\omega, l)$ are also modeled by

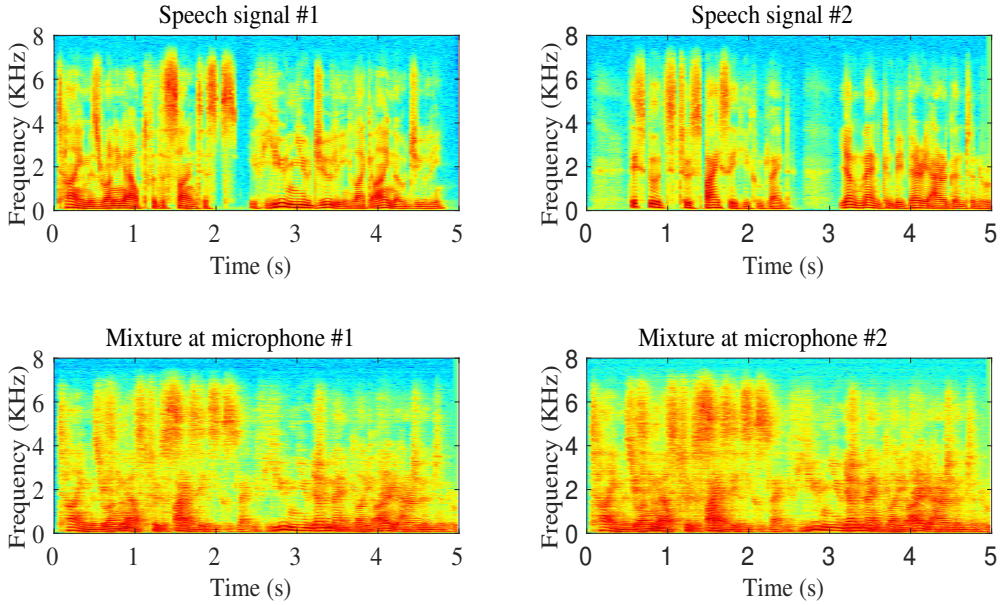


Figure 3.5: Power spectrograms of two speech signals (first row) and their generated mixtures (second row) at two spatially separated microphones in a reverberant environment.

a zero-mean multivariate complex Gaussian distribution with a covariance matrix obtained as

$$\mathbf{\Sigma}_{\mathbf{x}}(\omega, l) = \sum_{n=1}^{N} \mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l). \tag{3.6}$$

In this case, the multivariate complex Gaussian likelihood function of the observed mixtures $\mathbf{x}(\omega, l)$ is parametrised by the set of the covariance matrices of source spatial images, i.e. $\theta = \{\mathbf{\Sigma}_{\mathbf{c}_1}(\omega, l), ..., \mathbf{\Sigma}_{\mathbf{c}_N}(\omega, l)\}_{\omega,l}$. Over all the time-frequency points, Maximum-Likelihood (ML) estimation of $\theta$ is shown to result from the minimization of the minus log-likelihood function as [70]

$$\xi(\theta) = \sum_{\omega,l} \mathrm{tr}(\mathbf{\Sigma}_{\mathbf{x}}^{-1}(\omega, l)\tilde{\mathbf{R}}_{\mathbf{x}}(\omega, l)) + \log|\pi\mathbf{\Sigma}_{\mathbf{x}}(\omega, l)|, \tag{3.7}$$

where $|.|$ denotes the determinant of a square matrix, $\mathrm{tr}(.)$ indicates the trace of a matrix, and $\tilde{\mathbf{R}}_{\mathbf{x}}(\omega, l)$ is an empirical covariance matrix of the

observed mixtures that can be defined by a rank-1 model in a linear form as [71]

$$\tilde{\mathbf{R}}_{\mathbf{x}}(\omega, l) = \mathbf{x}(\omega, l)\mathbf{x}^{H}(\omega, l), \tag{3.8}$$

where $(.)^{H}$ indicates the conjugate transposition. In a quadratic form [70, 71], the empirical covariance matrix is obtained by local averaging over the neighborhood of each time-frequency point as

$$\tilde{\mathbf{R}}_{\mathbf{x}}(\omega, l) = \frac{\sum_{\tilde{\omega},\tilde{l}} \gamma(\tilde{\omega} - \omega, \tilde{l} - l)\mathbf{x}(\tilde{\omega}, \tilde{l})\mathbf{x}^{H}(\tilde{\omega}, \tilde{l})}{\sum_{\tilde{\omega},\tilde{l}} \gamma(\tilde{\omega} - \omega, \tilde{l} - l)}, \tag{3.9}$$

where $\gamma$ is a bi-dimensional window describing the shape of the neighborhood. As it is clear, the quadratic form to compute the empirical covariance matrix $\tilde{\mathbf{R}}_{\mathbf{x}}(\omega, l)$ includes additional information about the local correlation between propagation channels which often increases the accuracy of estimation.

Source separation is performed by first estimating the set $\theta$ in the sense of ML. The source spatial images are then obtained in the sense of the Minimum Mean Square Error (MMSE) by applying multichannel Wiener filtering as follows

$$\tilde{\mathbf{c}}_{n}(\omega, l) = \mathbf{G}_{n}(\omega, l)\mathbf{x}(\omega, l). \tag{3.10}$$

The smoothing filter gain $\mathbf{G}_{n}(\omega, l)$ is computed as

$$\mathbf{G}_{n}(\omega, l) = \mathbf{\Sigma}_{\mathbf{c}_{n}}(\omega, l)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(\omega, l). \tag{3.11}$$

### 3.4.1 Smooth Wiener filtering

The conventional multichannel Wiener filter expressed in (3.11) minimizes the mean-square error between the filter input and output signals. Applying extra constraints, such as smoothness or sparseness, on the minimization problem can improve the results. Multichannel spatial smoothing techniques have been proposed in order to widen the spatial response of

the filter, so as to reduce artifacts assumed to be spatially close to the target source direction. In this work, spatial smoothing [28] inspired by a weighted likelihood model is used in which the filter is expressed as

$$\mathbf{G}_n^s(\omega, l) = \mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)[(1-\mu)\mathbf{\Sigma}_{\mathbf{x}}(\omega, l) + \mu\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)]^{-1}. \tag{3.12}$$

The smoothness of the resulting filter $\mathbf{G}_n^s(\omega, l)$ increases with $\mu$, so that it is equal to the conventional Wiener filter $\mathbf{G}_n(\omega, l)$ for $\mu = 0$ and to the identity filter for $\mu = 1$.

### 3.4.2 Spatial covariance decomposition

In the spatial covariance decomposition, the covariance matrix of the $n$-th source spatial images $\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)$ is modeled as the product of a scalar spectral variance $v_n(\omega, l)$ encoding the power spectrum of the $n$-th source signal at the frequency bin $\omega$ and time frame $l$, and a $M \times M$ time-invariant spatial covariance matrix $\mathbf{R}_n(\omega)$ encoding the spatial information associated with the propagation of the $n$-th source signal at each frequency $\omega$

$$\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l) = v_n(\omega, l)\mathbf{R}_n(\omega). \tag{3.13}$$

Following such decomposition, the set of the model parameters to estimate is updated to be

$$\theta = \{\{v_1(\omega, l), ..., v_N(\omega, l)\}_l, \mathbf{R}_1(\omega), ..., \mathbf{R}_N(\omega)\}_\omega. \tag{3.14}$$

1. *Spatial parameters*: The covariance matrix models the spatial characteristics, such as intensity and phase differences between propagation channels. The matrix can be represented by a rank-1 model as $\mathbf{R}_n(\omega) = \mathbf{h}_n(\omega)\mathbf{h}_n^H(\omega)$, but it can be also described by an unconstrained model, which is considered in this work. As the focus of the work is on spectral modeling, these aspects are not further detailed.

2. *Spectral parameters*: The power spectrum of the source signal $s_n(\omega, l)$ denoted as $\mathbf{V}_n = [\{v_n(\omega, l)\}_{\omega,l}]_{\Omega \times L}$ can be represented as the product of two nonnegative matrices using NMF [25, 53]. Accordingly, the nonnegative source variance $v_n(\omega, l)$ can be represented as the multiplication of two vectors, each one with nonnegative entries

$$v_n(\omega, l) = \mathbf{u}_n^T(\omega)\mathbf{w}_n(l), \tag{3.15}$$

where $\mathbf{u}_n(\omega)$ is a spectral basis column vector of $K$ latent coefficients of a spectral basis matrix $\mathbf{U}_n = [\{\mathbf{u}_n^T(\omega)\}_\omega]_{\Omega \times K}$, and $\mathbf{w}_n(l)$ is a column vector of $K$ latent coefficients of a time-varying coefficient matrix $\mathbf{W}_n = [\{\mathbf{w}_n(l)\}_l]_{K \times L}$.

### 3.4.3 Estimation of the model parameters

Since the observed data $\mathbf{X} = \{\mathbf{x}(\omega, l)\}_{\omega,l}$ is fully expressed by the unobserved data $\mathbf{C} = \{\mathbf{c}_n(\omega, l)\}_{n,\omega,l}$ as it is mathematically modeled in (3.3), the set of complete data is defined as $\{\mathbf{X}, \mathbf{C}\}$. The natural statistics are defined as the covariance matrix of the conditional probability of the source spatial images $\mathbf{c}_n(\omega, l)$ [30]. The set $\theta$ is estimated by minimizing the criterion in (3.7) using a Generalized Expectation Maximization (GEM) algorithm [27] that consists in alternating the following two steps:

1. *E step*: given the observed mixtures $\mathbf{x}(\omega, l)$ and the current estimation of the set $\theta$, the conditional expectation of the natural statistics is computed.

2. *M step*: given the conditional expectation of the natural statistics, the set $\theta$ is updated so as to increase the conditional expectation of the likelihood of the complete data.

Applying the GEM, the two estimation steps are detailed as follows:

1. *E step*: given the current estimation of the source spatial images $\tilde{\mathbf{c}}_n(\omega, l)$ and the set of model parameters $\theta$, the natural statistics is computed, for example, as [30]

$$\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) = \tilde{\mathbf{c}}_n(\omega, l)\tilde{\mathbf{c}}_n^H(\omega, l) + (\mathbf{I} - \mathbf{G}_n(\omega, l))\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l), \tag{3.16}$$

where $\mathbf{I}$ is an $M \times M$ identity matrix and $\tilde{\mathbf{c}}_n(\omega, l)\tilde{\mathbf{c}}_n^H(\omega, l)$ is a rank-1 empirical covariance matrix of $\tilde{\mathbf{c}}_n(\omega, l)$ in the linear form.

2. *M step*: given $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, the set $\theta_n = \{\{v_n(\omega, l)\}_l, \mathbf{R}_n(\omega)\}_\omega$ belonging to the $n$-th source, is updated according to the minimization of

$$\tilde{\theta}_n = \arg\min_{\theta_n} \sum_{\omega, l} \mathrm{tr}(\mathbf{\Sigma}_{\mathbf{c}_n}^{-1}(\omega, l)\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)) + \log|\pi\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)|, \tag{3.17}$$

## 3.5 Spectral modeling using NMF

The source variance $v_n(\omega, l)$ can be modeled using spectral modeling based on Nonnegative Matrix Factorization (NMF) as the product of two nonnegative vectors as in (3.15). In other words, the source power spectrum $\mathbf{V}_n = [\{v_n(\omega, l)\}_{\omega,l}]_{\Omega \times L}$ is decomposed into two nonnegative matrices: a spectral basis matrix $\mathbf{U}_n$ containing constitutive parts of the power spectrum, and an activation coefficient matrix $\mathbf{w}_n(l)$ containing time-varying weights. Figure 3.6 shows an example of decomposing the source power spectrum using NMF. An extension to NMF has been considered by arranging multiple signal observations in a tensor form, where the observations form slices of 3-D tensor [25]. In Nonnegative Tensor Factorization (NTF), the tensor is decomposed into the multiplication of matrices and tensors. The redundancy among the original tensor slices is described by the matrices, while the diversity is represented by the decomposed tensors. The factorization is achieved by minimizing a cost function, which is an error measurement function.

Divergences are widely used as cost functions to measure the similarity between source signals. For instance, the Kullback-Leibler (KL) divergence [53] is used to compare two probability distributions. The Itakura-Saito (IS) divergence [38] is used as a measure of the perceptual differences between spectra. The generalized $\beta$-divergence [13], used as a cost function for NMF in [40, 52], encompasses the KL and IS divergences. The $\beta$-divergence $d_\beta(a/bc)$ between the elements $a$ and its element decomposition $b$ and $c$ is expressed as [40]

$$d_\beta(a/bc) = \begin{cases} \frac{a}{bc} - \log(\frac{a}{bc}) - 1, & \beta = 0 \\ a\log(\frac{a}{bc}) + bc - a, & \beta = 1 \\ \frac{a^\beta + (\beta-1)(bc)^\beta - \beta a(bc)^{\beta-1}}{\beta(\beta-1)}, & otherwise \end{cases} \tag{3.18}$$

For $\beta = 1$, $d_\beta(a/bc)$ is the Kullback-Leibler (KL) divergence [53], while it is the Itakura-Saito (IS) divergence for $\beta = 0$ [38]. Applying the commonly used multiplicative update (MU) rules, the factorization is accomplished by minimizing the $\beta$-divergence. Appendix A presents the MU rule to minimize the $\beta$-divergence for matrix and tensor factorization.

## 3.6 Sparse modeling

Sparse representations of signals have received a lot of attention to tackle the problem of audio source detection and separation. The solution of this problem basically exploits a prior assumption that the speech signals are sparse in their nature in a known domain such as the time-frequency domain. Indeed, it is a realistic assumption that, in a given basis, the audio sources have an energy compaction property and most of their coefficient values are very small. Moreover, a clear outcome of sparsity is the low probability for several sources to be simultaneously active at every time-frequency point.

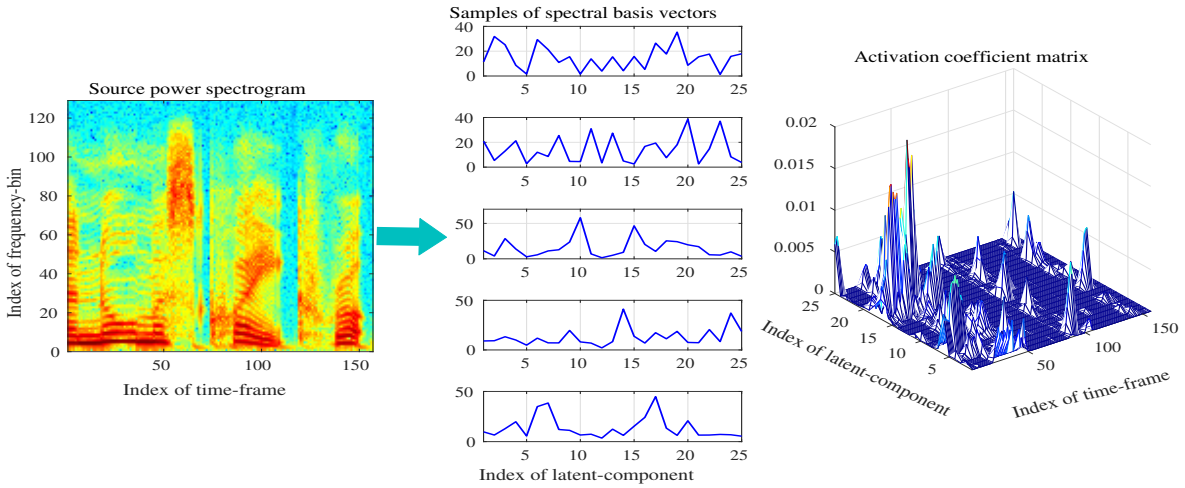


Figure 3.6: Nonnegative matrix factorization (NMF): the first figure on the left shows the power spectrogram of a speech signal, the second one shows samples of spectral basis vectors $\mathbf{u}_n(\omega)$ at different frequency bins, and the third one shows the complete activation coefficient matrix $\mathbf{W}_n$.

Sparse modeling assumes an ability to describe a signal by a small number of values using a pre-defined dictionary. The basic idea of sparse modeling is that traditional orthonormal bases are replaced by atoms (columns) in an over-complete dictionary. Given the dictionary, a signal can be represented as a linear combination of few atoms [21]. This modeling requires a proper definition of the atoms of the dictionary. As such, the choice of the dictionary is important for the success of this modeling. A very simple measure of sparsity of a vector $\mathbf{a}$ involves the number of nonzero entries; the vector is sparse if there are few non-zeros among its entries. It is more convenient to introduce the $l_0$ norm as follows

$$||\mathbf{a}||_0 = \#\{i : a_i \neq 0\}. \tag{3.19}$$

Given a vector of observations $\mathbf{y}$ and a redundant dictionary $\mathbf{D} = [\mathbf{d}_1\ \mathbf{d}_2\ ...]$ (see Figure 3.7), the sparsity optimization problem is explicitly formulated

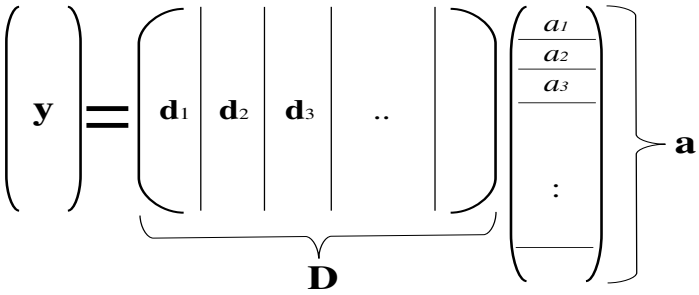


Figure 3.7: Sparse modeling.

as

$$\min_{\mathbf{a}} ||\mathbf{a}||_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{D}\mathbf{a}. \tag{3.20}$$

This optimization is a NP-hard problem. However, it is relatively easy to approximate using various techniques, including matching pursuit (MP) [61], orthogonal matching pursuit (OMP) [75] and basis pursuit (BP) [29]. By matching the $o$-th atom $\mathbf{d}_o$ of the dictionary $\mathbf{D}$ and the observations $\mathbf{y}$, an MP-based algorithm iteratively selects the active atom $\mathrm{o}^{\mathrm{match}}$ from the produced dictionary and removes its effect from the observations, such that the measurement of the error is decreased at each iteration, till a certain level of sparsity $G$. Indicating with $\mathbf{z}_i$ the residual error after the $i$-th iteration, a simple MP algorithm can be described as

---

$\mathbf{z}_0 = \mathbf{y}$
For $i = 1$; $i = i$+1; till (i == $G$),
$\mathrm{o}^{\mathrm{match}} = \arg\max_o |(\mathbf{d}_o)^H \ \mathbf{z}_{i-1}|$
$\mathbf{z}_i = \mathbf{z}_{i-1} - (\mathbf{d}_{\mathrm{o}^{\mathrm{match}}})^H \ \mathbf{z}_{i-1}\mathbf{d}^{\mathrm{o}^{\mathrm{match}}}$
Return

---

Many algorithms have followed a trend to exploit the sparse nature of the audio source signals [15]. In the multichannel scenario, i.e. when audio signals are recorded by an array of microphones, efficient sparse modeling can be obtained through a model-based definition of the mixing parameters. In mixing environments with low reverberation, anechoic models are often used to approximate the parameters of the mixing environments [58]. In more reverberant environments, a dictionary of the mixing parameters can be trained from the data itself [60].

## 3.7 Evaluation metrics

The separation performance is evaluated via signal-to-distortion ratio (SDR), source image-to-spatial distortion ratio (ISR), signal-to-interference ratio (SIR) and source-to-artifact ratio (SAR) expressed in decibels (dBs) [81], which respectively account for overall distortion, target distortion, residual crosstalk, and musical noise. The $n$-th estimated source spatial image component $\tilde{c}_{nm}(t)$ at the $m$-th microphone is represented in terms of the true spatial image component $c_{nm}(t)$, as follows

$$\tilde{c}_{nm}(t) = c_{nm}(t) + e_{nm}^{spat}(t) + e_{nm}^{interf}(t) + e_{nm}^{artif}(t), \tag{3.21}$$

where $e_{nm}^{spat}(t)$, $e_{nm}^{interf}(t)$ and $e_{nm}^{artif}(t)$ are distinct error components respectively representing spatial distortion, interference and artifacts. The above performance measurements of each source are mathematically defined as

$$\mathrm{ISR}_n = 10 \log \frac{\sum_{m,t} c_{nm}(t)^2}{\sum_{m,t} e_{nm}^{spat}(t)^2} \tag{3.22}$$

$$\mathrm{SIR}_n = 10 \log \frac{\sum_{m,t} (c_{nm}(t) + e_{nm}^{spat}(t))^2}{\sum_{m,t} e_{nm}^{interf}(t)^2} \tag{3.23}$$

$$\mathrm{SAR}_n = 10 \log \frac{\sum_{m,t} (c_{nm}(t) + e_{nm}^{spat}(t) + e_{nm}^{interf}(t))^2}{\sum_{m,t} e_{nm}^{artif}(t)^2} \tag{3.24}$$

$$\mathrm{SDR}_n = 10 \log \frac{\sum_{m,t} c_{nm}(t)^2}{\sum_{m,t} (e_{nm}^{spat}(t) + e_{nm}^{interf}(t) + e_{nm}^{artif}(t))^2} \tag{3.25}$$

# FIRST PART

Exploiting spectral information about audio signals in observed mixtures, for source separation, by applying the spectral modeling theory using Nonnegative Matrix and Tensor Factorization (NMF/NTF).

# Chapter 4

# Nonnegative Decomposition I

## Trained spectral bases

As reported in the previous chapter, the mixing process is probabilistically modeled by a multivariate complex Gaussian likelihood function. The function is described by spectral and spatial parameters, i.e. spectral source variances ($v_n(\omega, l)$) and spatial covariance matrices ($\mathbf{R}_n(\omega)$). Source separation is performed by estimating the parameters, then applying multichannel spectral Wiener filtering. The parameters are estimated in sense of Maximum-Likelihood (ML) by applying a Generalized Expectation-Maximization (GEM) algorithm. The GEM algorithm consists in alternating two steps, i.e. an Expectation step and a Maximization step. In the Expectation step, the natural statistics $\{\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)\}_{\omega,l,n}$ are computed. In the Maximization step, the set $\theta = \{\{v_n(\omega, l)\}_l, \mathbf{R}_n(\omega)\}_{n,\omega}$ is updated. In this chapter, we propose to (see Figure 4.1):

- Modify the computation method of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, in the Expectation step, in order to exploit the redundancy between multiple observations.

- Reduce the estimation dependency between the model parameters $v_n(\omega, l)$ and $\mathbf{R}_n(\omega)$, in the Maximization step, in order to reduce the accumulated estimation error.

- Exploit pre-trained spectral basis vectors $\mathbf{u}_n(\omega)$, in the Maximization step, in order to propose accurate estimation of the parameters.

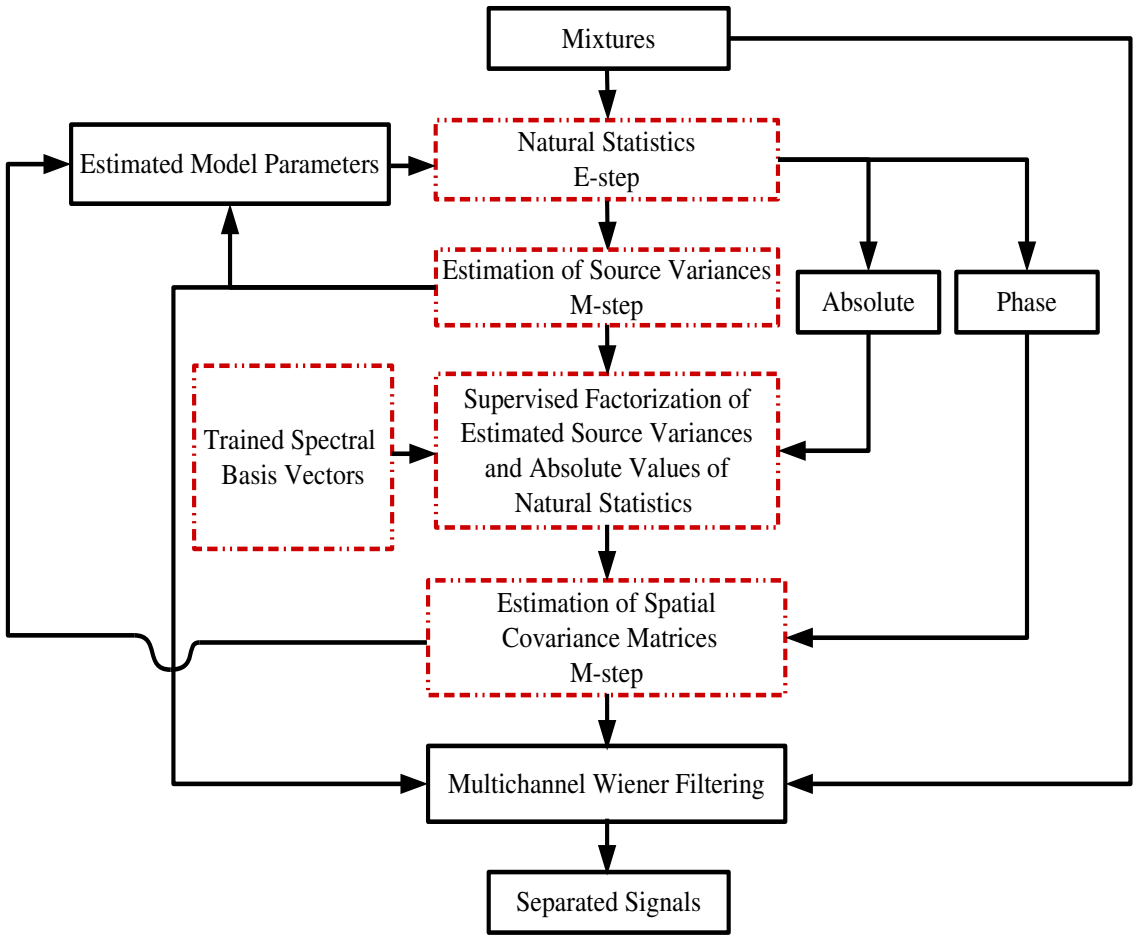


Figure 4.1: A flowchart of the proposed method. Highlighted blocks refer to novel contributions.

## 4.1 Method

In the Expectation step of the GEM to estimate the set $\theta$ in Section 3.4.3, the computation of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ in (3.16) can be modified in order to include more information about the coherence between propagation channels, and the temporal-spectral redundancy between time-frequency points of the

obtained source spatial images $\tilde{\mathbf{c}}_n(\omega, l)$. Using this empirical computation often increases the accuracy of estimation. The matrix computation in (3.16) can then be modified to be represented as follows

$$\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) = \hat{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) + (\mathbf{I} - \mathbf{G}_n(\omega, l))\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l), \tag{4.1}$$

where $\hat{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ is an empirical covariance matrix of the estimated source spatial images $\tilde{\mathbf{c}}_n(\omega, l)$, and it can be obtained in a quadratic form as in (3.9) as follows

$$\hat{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) = \frac{\sum_{\tilde{\omega},\tilde{l}} \gamma(\tilde{\omega} - \omega, \tilde{l} - l)\tilde{\mathbf{c}}_n(\tilde{\omega}, \tilde{l})\tilde{\mathbf{c}}_n^H(\tilde{\omega}, \tilde{l})}{\sum_{\tilde{\omega},\tilde{l}} \gamma(\tilde{\omega} - \omega, \tilde{l} - l)}. \tag{4.2}$$

In the Maximization step of the GEM, estimating the set of parameters $\theta = \{\{v_n(\omega, l)\}_l, \mathbf{R}_n(\omega)\}_{n,\omega}$ of the model in Section 3.4 is performed, as for example in [31], by updating one parameter, then the second parameter is updated using the first estimated one. The main limitation of this method is that the estimation error is accumulated from the first parameter to the second, and from one iteration to the next. We propose to reduce the estimation dependency of the parameters, so that the estimation in the current iteration does not depend on the previous one. This is done, here, by reducing the dependency between the parameters by estimating the source variance $v_n(\omega, l)$ regardless of the spatial covariance matrix $\mathbf{R}_n(\omega)$.

Assuming that spectral basis vectors $\mathbf{u}_n(\omega)$ trained using variances of source signals in observed mixtures are available, in this chapter, we propose methods to modify and refine the estimation of the parameters $\theta$. One way to involve the trained basis vectors $\mathbf{u}_n(\omega)$ in the optimization function requires, in practice, to slightly modify the minimization function in (3.17). To do this, the problem is broken down into the sum of sub-problems so that spectral modeling using Nonnegative Matrix/Tensor Factorization (NMF/NTF) is employed to represent the source variance $v_n(\omega, l)$ and the matrix of multiple observations $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ [35, 36].

Up to a constant, the minimization function in (3.17) can be expressed in terms of the parameters of the spatial covariance decomposition in (3.13) as follows

$$\xi(\theta) = \sum_{\omega,l,n} \mathrm{tr}(v_n^{-1}(\omega,l)\mathbf{R}_n^{-1}(\omega)\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)). \tag{4.3}$$

We recall the set of model parameters as defined in (3.14)

$$\theta = \{\{v_n(\omega,l)\}_l, \mathbf{R}_n(\omega)\}_{\omega,n}. \tag{4.4}$$

The set $\theta$ is estimated so that the function in (4.3) is minimized with respect to each parameter. As proposed in this chapter, the source spectral variance $v_n(\omega,l)$ is blindly estimated observing the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)$, whereas the spatial covariance matrix $\mathbf{R}_n(\omega)$ is estimated in an informed scenario, given the estimation of spectral variances of sources $v_n(\omega,l)$ and trained spectral basis vectors of sources $\mathbf{u}_n(\omega)$ .

### 4.1.1 Training of source-based prior information

In spectral modeling using NMF, as recently proposed, trained source spectral basis matrices can be identified as source-based prior information. For clean training audio signals of the $n$-th source in the observed mixtures $\mathbf{x}(\omega,l)$, the nonnegative power spectra of the training signals are concatenated in a matrix $\mathbf{V}_n^t$. Applying NMF, a spectral basis matrix $\mathbf{U}_n$ is extracted by decomposing $\mathbf{V}_n^t$ using the Multiplicative Update (MU) rules in order to minimize the Kullback-Leibler (KL) divergence in (3.18) [25]. The matrix $\mathbf{U}_n$ is obtained as follows (appendix A)

$$\mathbf{U}_n \leftarrow \mathbf{U}_n \circ \frac{[\mathbf{V}_n^t \circ (\mathbf{U}_n\mathbf{W}_n^t)^{-1}](\mathbf{W}_n^t)^T}{\mathbf{1}(\mathbf{W}_n^t)^T}, \tag{4.5}$$

where the division is point-wise, $\circ$ indicates point-wise multiplication, and $\mathbf{1}$ is a matrix of ones. The matrix $\mathbf{U}_n$ contains spectral basis vectors $\mathbf{u}_n(\omega)$,

each one of length $K$, i.e. $\mathbf{U}_n = [\{\mathbf{u}_n^T(\omega)\}_\omega]_{\Omega\times K}$. $\mathbf{W}_n^t$ is a time-varying activation coefficient matrix obtained as

$$\mathbf{W}_n^t \leftarrow \mathbf{W}_n^t \circ \frac{\mathbf{U}_n^T[\mathbf{V}_n^t \circ (\mathbf{U}_n\mathbf{W}_n^t)^{-1}]}{\mathbf{U}_n^T\mathbf{1}}. \tag{4.6}$$

The factorization is performed by alternating (4.5) and (4.6) till a certain point of convergence defined by a number of iterations or a value of an error measurement. The matrix $\mathbf{W}_n^t$ is not needed in the estimation step.

### 4.1.2 Estimation of $v_n(\omega, l)$ using SVD

$\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ is a matrix with a high condition number (the ratio of the largest singular value to the smallest one). In fact, the value of the number depends on many factors, one of them is the reverberation. In mixing environments with low reverberation, the number is high, however, it decreases as the environment becomes more reverberant. This means that in environments with low reverberation, the largest singular value can well describe the magnitude information of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, however, an approximate description is accomplished in environments with high reverberation. The largest singular value $\sigma_n(\omega, l)$ is derived by computing the singular value decomposition (SVD) of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$. The minimization function in (4.3) can then be described by including $\sigma_n(\omega, l)$ as follows

$$\xi(\theta) = \sum_{\omega,l,n} \mathrm{tr}(v_n^{-1}(\omega, l)\mathbf{R}_n^{-1}(\omega)\sigma_n(\omega, l)\frac{\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)}{\sigma_n(\omega, l)}). \tag{4.7}$$

In sense of Maximum-Likelihood (ML), if the source spectral variance is estimated as [35]

$$v_n(\omega, l) = \sigma_n(\omega, l), \tag{4.8}$$

the matrix $\mathbf{R}_n(\omega)$ is estimated so that its largest singular value equals one. *In fact, $\sigma_n(\omega, l)$ conveys information not only about the true spectral variance of the n-th source but also about the propagation channels that are*

*associated with the source*. For this reason, the estimated $v_n(\omega, l)$ in (4.8) can be considered as the true source variance weighted by average channel intensities, i.e. the source variance at the microphone locations. Therefore, the spatial information involved in the estimated source variance $v_n(\omega, l)$ can be helpful for the estimation of the spatial covariance matrix $\mathbf{R}_n(\omega)$.

### 4.1.3 Estimation of $\mathbf{R}_n(\omega)$ using trained basis vectors

The $n$-th estimated source power spectrum $\mathbf{V}_n = [\{v_n(\omega, l)\}_{\omega,l}]_{\Omega \times L}$ in (4.8) can be decomposed applying supervised NMF, given the trained spectral basis matrix $\mathbf{U}_n$, to compute an activation coefficient matrix $\mathbf{W}_n = [\{\mathbf{w}_n(l)\}_l]_{K \times L}$ that contains time-varying weight vectors $\mathbf{w}_n(l)$, each one of length $K$. Accordingly, given $\mathbf{u}_n(\omega)$, the estimated source variance $v_n(\omega, l)$ in (4.8) is represented in the factorization domain as follows

$$v_n(\omega, l) = \sum_k u_n(\omega, k) w_n(k, l), \tag{4.9}$$

where $u_n(\omega, k)$ and $w_n(k, l)$ are the $k$-th coefficients of the vectors $\mathbf{u}_n(\omega)$ and $\mathbf{w}_n(l)$, respectively.

On the other side, the absolute values of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ can be decomposed applying supervised NMF, given the spectral basis vectors $\mathbf{u}_n(\omega)$. Accordingly, the matrix is represented in the factorization domain as follows

$$\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) = \sum_k u_n(\omega, k) \mathbf{W}_{\mathbf{c}_n}(k, l) \angle \tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l), \tag{4.10}$$

where $\angle \tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ indicates the phase information of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$. $\mathbf{W}_{\mathbf{c}_n}(k, l)$ is a time-varying activation coefficient matrix of the absolute values of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, the matrix is represented as

$$\mathbf{W}_{\mathbf{c}_n}(k, l) = \begin{bmatrix} w_{\mathbf{c}_n}^{11}(k, l) & \cdots & w_{\mathbf{c}_n}^{1M}(k, l) \\ \vdots & \ddots & \vdots \\ w_{\mathbf{c}_n}^{M1}(k, l) & \cdots & w_{\mathbf{c}_n}^{MM}(k, l) \end{bmatrix}, \tag{4.11}$$

where $w_{\mathbf{c}_n}^{m_1 m_2}(k,l)$, $m_1, m_2 = 1, .., M$, is a weight coefficient of a time-varying vector that is obtained by factorizing the ($m_1$,$m_2$) vector of the absolute values of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$. For better understanding the building of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, we want to emphasize that its $m$-th diagonal observation $\tilde{r}_{\mathbf{c}_n}^{mm}(\omega, l)$ is a real coefficient, which is expressed in the factorization domain as follows

$$\tilde{r}_{\mathbf{c}_n}^{mm}(\omega, l) = \sum_k u_n(\omega, k) w_{\mathbf{c}_n}^{mm}(k, l), \tag{4.12}$$

and its ($m_1$,$m_2$) off-diagonal observation is a complex coefficient that is represented as

$$\tilde{r}_{\mathbf{c}_n}^{m_1 m_2}(\omega, l) = \sum_k u_n(\omega, k) w_{\mathbf{c}_n}^{m_1 m_2}(k, l) \angle \tilde{r}_{\mathbf{c}_n}^{m_1 m_2}(\omega, l), \tag{4.13}$$

where $\angle \tilde{r}_{\mathbf{c}_n}^{m_1 m_2}(\omega, l)$ indicates the phase information of the ($m_1$,$m_2$) coefficient $\tilde{r}_{\mathbf{c}_n}^{m_1 m_2}(\omega, l)$.

To estimate the spatial covariance matrix $\mathbf{R}_n(\omega)$, the minimization function in (4.3) can be expanded in order to exploit the above two factorization steps, taking advantages of the trained spectral basis vectors $\mathbf{u}_n(\omega)$. The function can be approximately described in terms of the above two factorization steps in (4.9) and (4.10) as follows

$$\xi(\theta) \approx \sum_{\omega,l,n,k} \mathrm{tr}((u_n(\omega, k) w_n(k, l))^{-1} \mathbf{R}_n^{-1}(\omega) u_n(\omega, k) \mathbf{W}_{\mathbf{c}_n}(k, l) \angle \tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)). \tag{4.14}$$

In this formulation, the optimization function in (4.3) is modified in order to perform the minimization in a high resolution domain, i.e. the factorization domain, where the problem is represented as the sum of $K$ sub-problems. The disadvantage of this formulation is that the trained spectral basis vectors $\mathbf{u}_n(\omega, k)$ are eliminated. Accordingly, the absolute values in the minimization function are frequency-independent. As a result, the changes in the absolute values of the matrix $\mathbf{R}_n(\omega)$ are not well

followed, from one frequency to another. The function in (4.14) is minimized in sense of Maximum-Likelihood (ML) with respect to the spatial covariance matrix $\mathbf{R}_n(\omega)$, in which the matrix is estimated as follows

$$\mathbf{R}_n(\omega) = \frac{1}{L} \sum_{l,k} \frac{\mathbf{W}_{\mathbf{c}_n}(k,l)}{w_n(k,l)} \angle \tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l). \tag{4.15}$$

The estimated matrix $\mathbf{R}_n(\omega)$ is then normalized using its largest singular value. As it is noted, the matrix is estimated using absolute information of the factorization, i.e. time-varying activation weights, and phase information of the time-frequency domain.

### 4.1.4 Refining the estimation of $v_n(\omega, l)$ using NTF

Since the estimation of the source variance $v_n(\omega, l)$ in (4.8) is point-wise, by locally observing each time-frequency point of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, spectral-temporal redundancy between the points is not well exploited. We propose a post-processing step to refine the estimation of the source variance $v_n(\omega, l)$ using the trained spectral basis vectors $\mathbf{u}_n(\omega)$ and applying Nonnegative Tensor Factorization (NTF).

As previously introduced, NTF is a parallel decomposition, where an original tensor is decomposed into the product of matrices representing the redundancy between slices of the original tensor, and decomposed tensors describing the diversity. In this step of refining the estimation of $v_n(\omega, l)$, not only spectral-temporal redundancy between time-frequency points of one signal observation is exploited, but also spatial redundancy between multiple signal observations.

The diagonal coefficients of the matrix $\frac{1}{L} \sum_l \frac{\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)}{\sigma_n(\omega,l)}$ can be approximately seen to describe time-invariant inter-channel intensities, and the off-diagonal coefficients to represent time-invariant cross-channel intensities and phase differences. We arrange the estimated source variance $v_n(\omega, l)$

times the diagonal coefficients of $\frac{1}{L}\sum_l \frac{\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)}{\sigma_n(\omega,l)}$ in a tensor $\tilde{\mathbf{V}}^M_{\mathbf{c}_n}(\omega,l)$ of size $\Omega \times L \times M$, where $M$ is the number of propagation channels. Each $(\omega,l)$ coefficient of the $m$-th slice $\tilde{\mathbf{V}}^m_{\mathbf{c}_n}(\omega,l)$ of the tensor is represented by $v_n(\omega,l)$ weighted by the diagonal $(m,m)$ coefficient of $\frac{1}{L}\sum_l \frac{\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)}{\sigma_n(\omega,l)}$. The $n$-th tensor is decomposed, fixing the $n$-th trained spectral basis matrix $\mathbf{U}_n$, so that the $m$-th slice of the tensor is represented as

$$\tilde{\mathbf{V}}^m_{\mathbf{c}_n}(\omega,l) = \mathbf{U}_n \tilde{\mathbf{D}}^m_n \tilde{\mathbf{W}}_n, \tag{4.16}$$

where $\tilde{\mathbf{W}}_n$ is a time-varying activation coefficient matrix, and $\tilde{\mathbf{D}}^m_n$ is a diagonal matrix of size $K \times K$. Each $(k,k)$ coefficient of the matrix $\tilde{\mathbf{D}}^m_n$ encodes the contribution of each spectral basis vector of $\mathbf{U}_n$ in the $m$-th tensor slice $\tilde{\mathbf{V}}^m_{\mathbf{c}_n}(\omega,l)$ of the $m$-th channel observation. For the $(k,k)$ vector of length $M$ in the $K \times K \times M$ tensor $\tilde{\mathbf{D}}^M_n = [\{\tilde{\mathbf{D}}^m_n\}_m]_{K\times K\times M}$, we propose to select the $m$-th channel index that maximizes the contribution of each basis vector of $\mathbf{U}_n$ in $\tilde{\mathbf{V}}^M_{\mathbf{c}_n}(\omega,l)$. The optimal channel index is selected as

$$m^*(k) = \arg\max_m \tilde{d}^m_n(k,k), \qquad m = 1,2,...,M. \tag{4.17}$$

where $\tilde{d}^m_n(k,k)$ is the $k$-th diagonal coefficient of the matrix $\tilde{\mathbf{D}}^{m^*}_n$. The source variance is then updated as follows

$$v_n(\omega,l) = \sum_k u_n(\omega,k)\tilde{d}^{m^*(k)}_n(k,k)\tilde{w}_n(k,l), \tag{4.18}$$

$\tilde{w}_n(k,l)$ is the $k$-th coefficient of the matrix $\tilde{\mathbf{W}}_n$ at the time frame $l$.

### 4.1.5 Estimation of $\mathbf{R}_n(\omega)$ using NTF

Over all the time-frequency points and coefficients, a tensor of observations of size $F \times L \times M^2$ can be built from the absolute values of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)$. The tensor is represented in the factorization domain as the multiplication of the trained matrix $\mathbf{U}_n$, a time-varying coefficient matrix, and

a tensor representing the spatial diversity between coefficients of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$. The full representation of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ is then defined as

$$\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) = \sum_k u_n(\omega, k) w_{\mathbf{c}_n}(k, l) \tilde{\mathbf{D}}_{\mathbf{c}_n}(k, k) \angle \tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l). \tag{4.19}$$

In this representation, at each frequency bin $\omega$ and time frame $l$, the redundancy between coefficients of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ is represented by $u_n(\omega, k)$ and $w_{\mathbf{c}_n}(k, l)$, and the diversity is described by the matrix $\tilde{\mathbf{D}}_{\mathbf{c}_n}(k, k)$ that is represented as follows

$$\tilde{\mathbf{D}}_{\mathbf{c}_n}(k, k) = \begin{bmatrix} \tilde{d}^{11}_{\mathbf{c}_n}(k, k) & \cdots & \tilde{d}^{1M}_{\mathbf{c}_n}(k, k) \\ \vdots & \ddots & \vdots \\ \tilde{d}^{M1}_{\mathbf{c}_n}(k, k) & \cdots & \tilde{d}^{MM}_{\mathbf{c}_n}(k, k) \end{bmatrix}. \tag{4.20}$$

The minimization function in (4.3) can be described again by involving the factorization of $v_n(\omega, l)$ in (4.18) and $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ in (4.19), as follows

$$\xi(\theta) \approx \sum_{\omega, l, n, k} \operatorname{tr}[(u_n(\omega, k) \tilde{d}_n^{m^*(k)}(k, k) w_n(k, l))^{-1} u_n(\omega, k) w_{\mathbf{c}_n}(k, l) \mathbf{R}_n^{-1}(\omega) \tilde{\mathbf{D}}_{\mathbf{c}_n}(k, k) \angle \tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)]. \tag{4.21}$$

As stated before about the formulation of the minimization function in (4.14), the formulation of the minimization function in (4.21) also suffers from that the trained spectral basis vectors $\mathbf{u}_n(\omega)$ are eliminated. Accordingly, the absolute values in the minimization function are frequency-independent. As a result, the changes in the absolute values of the matrix $\mathbf{R}_n(\omega)$ are not well followed, from one frequency to another. The function in (4.21) is minimized in sense of Maximum-Likelihood (ML) with respect to the spatial covariance matrix $\mathbf{R}_n(\omega)$, in which the matrix is estimated as

$$\mathbf{R}_n(\omega) = \frac{1}{L} \sum_{l,k} \frac{w_{\mathbf{c}_n}(k, l)}{\tilde{d}_n^{m^*}(k, k) w_n(k, l)} \tilde{\mathbf{D}}_{\mathbf{c}_n}(k, k) \angle \tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) \tag{4.22}$$

The matrix $\mathbf{R}_n(\omega)$ is then normalized using its largest singular value.

### 4.1.6 Matrix/tensor representation of multiple observations

To perform supervised NMF of multiple observations in (4.10), the absolute values of the $(m_1, m_2)$ coefficients of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, over all the frequency bins and time frames, are arranged side-by-side in a matrix $\mathbf{V}_{\mathbf{c}_n}^{m_1 m_2} = [\{|\tilde{r}_{\mathbf{c}_n}^{m_1 m_2}(\omega, l)|\}_{\omega,l}]_{\Omega \times L}$. Given the $n$-th trained source spectral basis matrix $\mathbf{U}_n$, the $(m_1, m_2)$ matrix of observations is factorized applying the MU rule as follows (appendix A)

$$\mathbf{V}_{\mathbf{c}_n}^{m_1 m_2} = \mathbf{U}_n \mathbf{W}_{\mathbf{c}_n}^{m_1 m_2}. \tag{4.23}$$

Supervised NTF of multiple observations in (4.19) is performed by arranging the matrices $\mathbf{V}_{\mathbf{c}_n}^{m_1 m_2}, m_1, m_2 = 1, .., M$, in a tensor of multiple observations, i.e. $\mathbf{V}_{\mathbf{c}_n}^{M^2} = [\{\mathbf{V}_{\mathbf{c}_n}^{m_1 m_2}\}_{m_1,m_2}]_{\Omega \times L \times M^2}$. Given the $n$-th trained source spectral basis matrix $\mathbf{U}_n$, the $m$-th slice of the tensor is factorized applying the MU rule as follows (appendix A)

$$\mathbf{V}_{\mathbf{c}_n}^{m} = \mathbf{U}_n \tilde{\mathbf{D}}_{\mathbf{c}_n}^{m} \mathbf{W}_{\mathbf{c}_n}. \tag{4.24}$$

### 4.1.7 Initialization

The source spatial images $\tilde{\mathbf{c}}_n(\omega, l)$ needed for the initial computation of the matrix $\hat{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ in (4.2) are initialized by binary clustering the time-frequency points of the observed mixtures $\mathbf{x}(\omega, l)$. The time difference of arrival (TDOA) of each source is estimated by GCC-PHAT (Generalized Cross-Correlation with PHAse Transform) as described in [69]. Given the estimated TDOAs, the time-frequency points of $\mathbf{x}(\omega, l)$ are classified into multiple clusters, each one corresponding to a source. The clustering is performed by minimizing the error between steering vectors of the estimated TDOAs and phase differences of the time-frequency points of $\mathbf{x}(\omega, l)$.

## 4.2 Full description

The full method is summarized as follows:

---

**Training**: $\mathbf{U}_n,\ n = 1, ..., N$ as in Section 4.1.1
**Input**: $\mathbf{x}(\omega, l)$
**Initialize**: $\tilde{\mathbf{c}}_n(\omega, l)$ as in Section 4.1.7, $\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l) = \mathbf{I}$
**Iterate**: *till convergence*
Compute $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ as in (4.1)
**Estimation of** $\theta = \{\{v_n(\omega, l)\}_l, \mathbf{R}_n(\omega)\}_{\omega,n}$:
NMF: Estimate $v_n(\omega, l)$ and $\mathbf{R}_n(\omega)$ as in (4.8) and (4.15)
NTF: Estimate $v_n(\omega, l)$ and $\mathbf{R}_n(\omega)$ as in (4.18) and (4.22)
**Separation**:

$$\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l) = v_n(\omega, l)\mathbf{R}_n(\omega)$$
$$\mathbf{G}_n(\omega, l) = \mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(\omega, l)$$
$$\tilde{\mathbf{c}}_n(\omega, l) = \mathbf{G}_n(\omega, l)\mathbf{x}(\omega, l)$$

**Return**
**Output**: $\tilde{\mathbf{c}}_n(\omega, l)$

---

## 4.3 Experiments

A room with size $4.45 \times 3.35 \times 2.5$ meters and an array of 2 omnidirectional microphones spaced 0.2 m are considered in a simulated scenario. The microphones are located in the middle of the room and they are at the same height (i.e., 1.4 m) of three given audio sources ($N = 3$). The distance from each source to the central point between the two microphones is either 0.5 or 1 m. The direction of arrivals of the source signals are 35, 90, and 145 angular degrees. Synthetic room impulse responses (RIRs) are simulated

through the Image Source Method (ISM) [54] with a sampling frequency of 16 kHz for three values of the reverberation time: $T_{60} = 200,\ 350$ or 500 ms. Six Italian speakers (3 males and 3 females) are considered as audio sources. Each speaker uttered 20 sentences, of average length 8.75 s. The clean speech signals are divided into 5 signals to test the method, and 15 to train the matrices $\mathbf{U}_n$, $n = 1, 2, 3$. The matrices are extracted by decomposing the power spectra of the training signals applying the multiplicative update (MU) rule [25, 53] to minimize the Kullback-Leibler (KL) divergence as in Section 4.1.1,

Four male-female mixture combinations (i.e. 3 males, 3 females, 2 females and 1 male, and 2 males and 1 female) were generated by individually convolving the full length of the simulated RIRs with the original signals and adding the source image contributions at each microphone. This resulted in a total of 20 test mixtures for each $T_{60}$. The discrete time-frequency representation of the mixtures $\mathbf{x}(\omega, l)$ is obtained through STFT with a Hanning analysis window of length 128 ms and shift 64 ms. The window $\gamma$ of computing the empirical covariance matrix of the source spatial images in (4.2) is a Hanning window of size $3 \times 3$. The separation performance is measured using the evaluation metrics detailed in Section 3.7.

As shown in Figure 4.2, the separation performance of source separation using NMF (SS-NMF) initially keeps improving as $K$ increases, and a peak is detected when $K$ is around 15. Further increase of $K$ slowly degrades the performance, as a result of overestimation of the model parameters $\theta$.

Fixing $K$ at 15, Table 4.1 shows the separation performance of SS-NMF as a function of the type of speech signals in the mixtures. It can be noted that depending on the type of speech signals, the performance slightly changes. *To achieve good performance, the trained spectral basis vectors* $\mathbf{u}_n(\omega)$ *must have high source reconstruction and discrimination properties,*

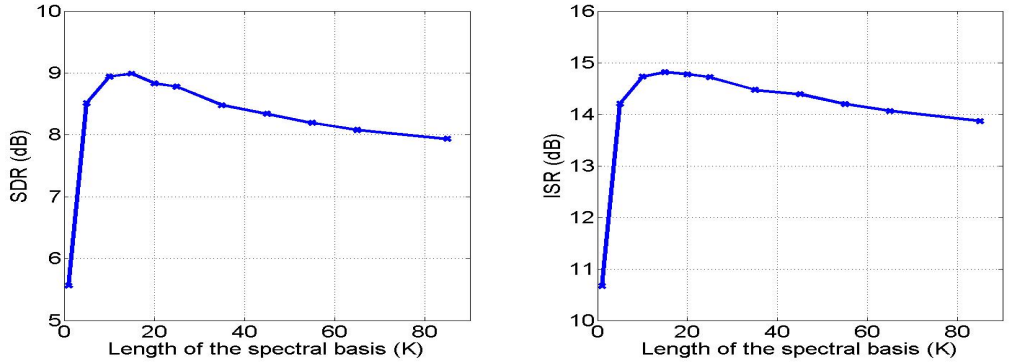


Figure 4.2: Average separation performance as a function of $K$.

Table 4.1: Separation performance as a function of the type of speech signals in mixtures. Source-to-microphone distance is 0.5 m and $K = 15$.

| $T_{60}$ | **3 males** | | **2 males&1 female** | | **2 females&1 male** | | **3 females** | |
|---|---|---|---|---|---|---|---|---|
| (ms) | SDR | ISR | SDR | ISR | SDR | ISR | SDR | ISR |
| 200 | 9.17 | 15.38 | 10.51 | 16.70 | 10.36 | 16.30 | 10.23 | 16.08 |
| 350 | 7.10 | 12.48 | 8.04 | 13.48 | 8.23 | 13.27 | 8.50 | 13.70 |
| 500 | 5.52 | 10.51 | 6.53 | 11.55 | 6.68 | 11.33 | 6.99 | 11.79 |

*and the overlap of the spectral-temporal representations of speech signals must be low (high signal sparseness).*

For mixtures of male speech signals, it is expected that there is a high overlap in the time-frequency domain, especially in low frequency bands. Even if the signals are well represented by the trained basis matrices, as a result of that overlap, low separation performance is achieved. On the other side, female speech signals are less overlapped as demonstrated by the higher separation performance obtained for mixtures of female signals.

The separation performance of SS-NMF was evaluated in terms of the source-to-microphone distances and the reverberation times as shown in Figure 4.3. The highest performance is obtained when the source spatial positions are close to the microphones (0.5 m), and the mixing environment is less reverberant ($T_{60} = 100$ ms). Moreover, the proposed algorithm

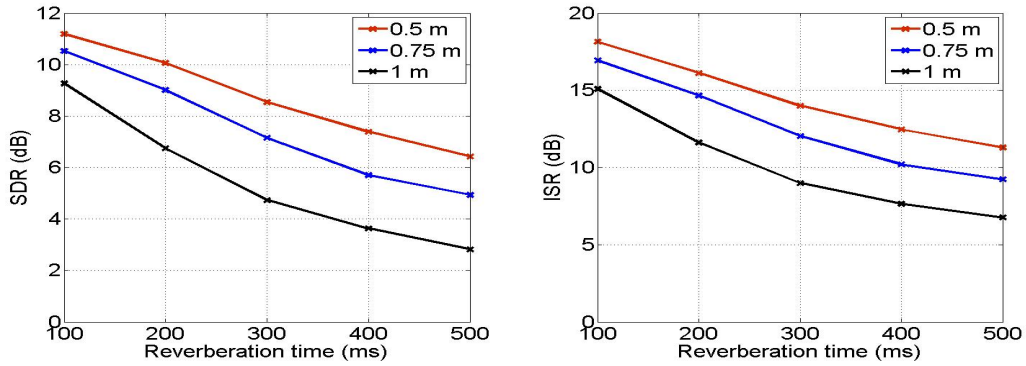


Figure 4.3: Average separation performance in terms of reverberation times and source-to-microphone distances.

achieves a high separation performance (SDR = 6.43 dB and ISR = 11.3 dB) in a reverberant mixing environment ($T_{60}$ = 500 ms). However, as expected, the performance degrades as either the source-to-microphone distance or the reverberation time increases.

### 4.3.1 Performance comparison

We compared the performance of the SS-NMF algorithm to three other source separation algorithms, i.e. the binary masking (BM) [90], the $l_0$-norm minimization [80], and the original Gaussian model-based source separation (ML-Blind) [30]. Both binary masking (BM) and $l_0$-norm minimization are fully informed by the mixing parameters in an oracle form. In practice, the binary masking and the $l_0$-norm minimization in the oracle form are used to better understand the upper bound limits of the separation performance. From Figure 4.4, in mixing environments with low reverberation ($T_{60}$ = 200 ms), the proposed algorithm outperforms the fully informed algorithms (i.e. exploiting the true values of the mixing parameters), from SDR point of view, which proves that prior information about the sources in mixing environments with low reverberation is more helpful than prior information about the mixing parameters. However,

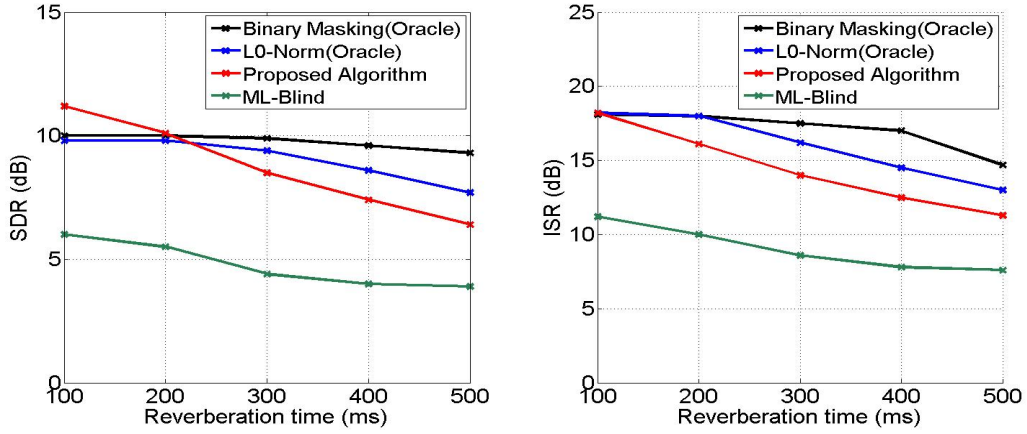


Figure 4.4: Comparison of separation performance, the source-to-microphone distance is 0.5 m.

prior information about the mixing parameters is very helpful in mixing environments with high reverberation. On the other hand, the proposed algorithm achieves high separation performance with respect to the ML-Blind algorithm; the performance has been improved by around 4 dB of SDR and more than 5 dB of ISR, averaging over all the reverberation times.

The proposed method (SS-NMF) outperforms two other methods as reported in Table 4.2. The first compared method is the original Gaussian model-based audio source separation proposed in [30], and the second one is based on clustering the time-frequency points of the observed mixtures using estimated source Time Difference-of-Arrivals (TDOAs) [69]. Furthermore, by exploiting spectral-temporal redundancy and applying parallel processing, source separation using NTF (SS-NTF) outperforms SS-NMF.

## 4.4 Conclusion

This chapter presented a method to estimate the parameters of the Gaussian model-based audio source separation. The model is parametrized by variances of audio sources in observed mixtures and corresponding spatial

Table 4.2: Comparison of separation performance, the source-to-microphone distance is 1 m.

| dB | BM Ideal | $l_0$norm Ideal | SS-NTF Inf. | SS-NMF Inf. | ML Blind | TDOAs Blind |
|---|---|---|---|---|---|---|
| SDR | 10.53 | 10.12 | 7.11 | 6.53 | 4.62 | 4.90 |
| ISR | 19.44 | 17.56 | 12.58 | 12.11 | 9.06 | 11.12 |
| $T_{60} = 200$ ms | | | | | | |
| SDR | 10.02 | 7.80 | 5.20 | 4.46 | 3.56 | 3.01 |
| ISR | 18.70 | 13.63 | 9.90 | 9.23 | 7.30 | 8.50 |
| $T_{60} = 350$ ms | | | | | | |
| SDR | 9.57 | 6.30 | 4.11 | 3.55 | 2.48 | 2.30 |
| ISR | 18.08 | 11.57 | 8.47 | 8.04 | 5.90 | 7.51 |
| $T_{60} = 500$ ms | | | | | | |

covariance matrices. We aim at reducing the estimation dependency of the parameters and exploiting trained source-based prior information to improve the separation performance. Spectral basis matrices trained using a set of power spectra of sources in observed mixtures are assumed to be available. The matrices are obtained by factorizing the power spectra using Nonnegative Matrix Factorization (NMF) applying the Multiplicative Update (MU) rules to minimize Kullback-Leibler (KL) divergence. The variances of the sources are blindly estimated by applying singular value decomposition of matrices of multiple observations. Furthermore, the spatial covariance matrices are estimated applying supervised NMF given the trained source spectral basis matrices. As an extension to the supervised NMF, we presented a supervised NTF method to refine the estimation of the variances of the sources and to estimate the spatial covariance matrices. The proposed methods were compared to the original Gaussian model-based audio source separation and it provided better performance in several mixing conditions.

# Chapter 5

# Nonnegative Decomposition II

## Weighted spectral bases

Although the method proposed in the previous chapter works well with good efficiency, the estimation does not exactly follow the changes in the absolute values of the spatial covariance matrix $\mathbf{R}_n(\omega)$, from one frequency to another, as a result of eliminating the frequency-dependent trained basis vectors $\mathbf{u}_n(\omega)$ from the optimization function in (4.14) and (4.21). To overcome this, we propose another method to involve source spectral modeling using NMF in the optimization function in (4.3), in which the amplitude values of the matrix $\mathbf{R}_n(\omega)$ are well followed and tracked [37].

The proposed method can work with good performance in either **blind** or **trained** scenarios applying either **unsupervised** or **supervised** NMF, respectively. We propose to exploit the spectral modeling using NMF of the estimated source variance $v_n(\omega, l)$ in (4.8) to obtain weighted basis vectors, which are used to factorize the matrix of multiple observations $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ (see Figure 5.1). In this context, we look for compact descriptions to the multiple observations in terms of the factorization of $v_n(\omega, l)$, which considerably improves the separation performance. Later, the factorization output of both the estimated source variance $v_n(\omega, l)$ and the matrix of multiple observations $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ are used to estimate the matrix $\mathbf{R}_n(\omega)$.

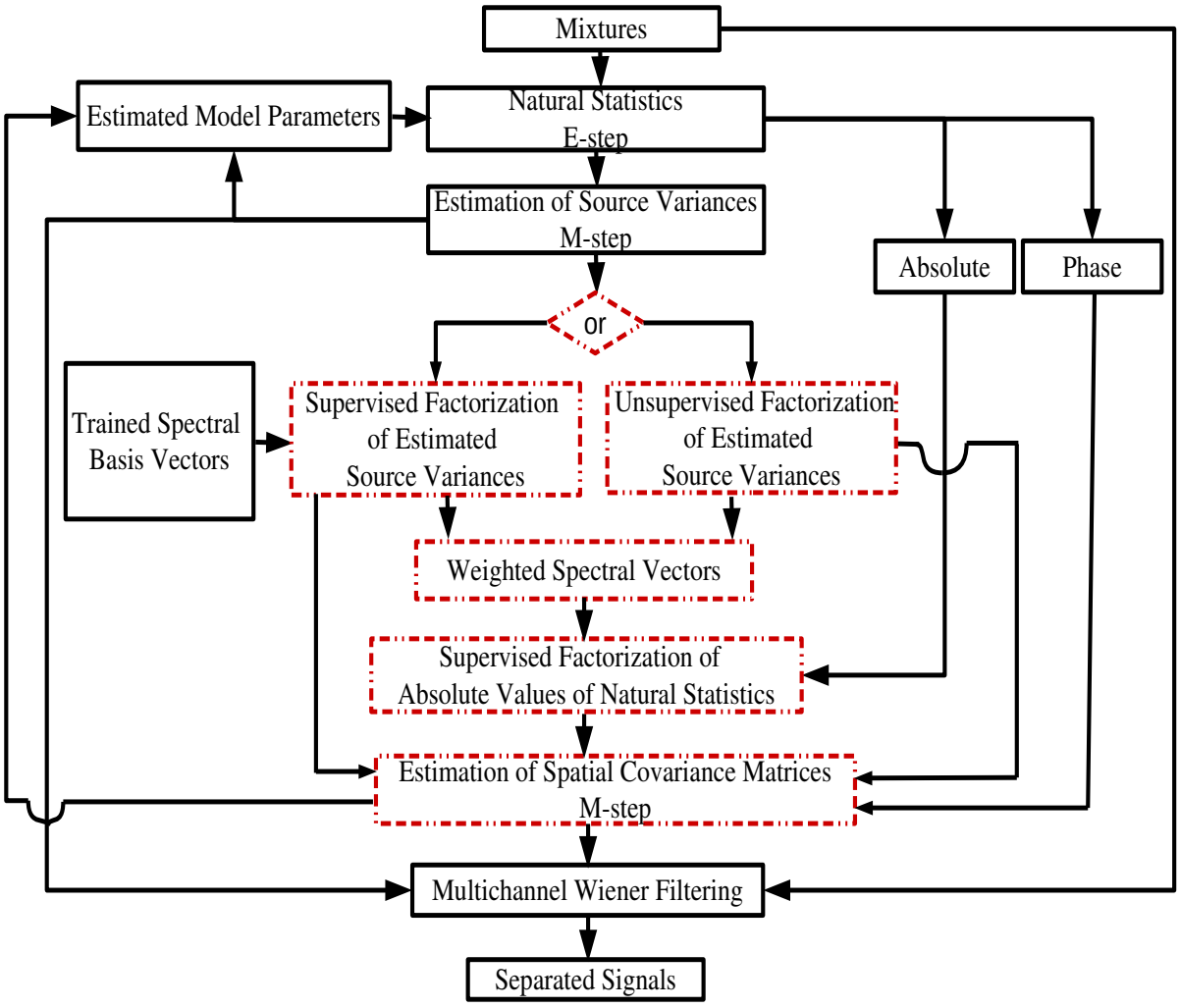


Figure 5.1: A flowchart of the proposed method. Highlighted blocks refer to novel contributions.

## 5.1 Method

As earlier stated, we aim at estimating the set of Gaussian model parameters $\theta = \{\{v_n(\omega,l)\}_l, \mathbf{R}_n(\omega)\}_{n,\omega}$. As proposed in Section 4.1.2, the source spectral variance $v_n(\omega,l)$ is estimated as the largest singular value of the matrix of multiple observations $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)$. Furthermore, applying **unsupervised** or **supervised** NMF, the estimated source variance $v_n(\omega,l)$ is approximately represented as the multiplication of a frequency-dependent spectral basis vector $\mathbf{u}_n(\omega)$ and a time-varying activation coefficient vector $\mathbf{w}_n(l)$ as in (3.15)

$$v_n(\omega,l) = \mathbf{u}_n^T(\omega)\mathbf{w}_n(l). \tag{5.1}$$

Accordingly, the estimated source power spectrum is represented in the factorization domain as follows

$$\mathbf{V}_n = [\{v_n(\omega,l)\}_{\omega,l}]_{\Omega\times L} = \mathbf{U}_n\mathbf{W}_n. \tag{5.2}$$

To estimate the matrix $\mathbf{R}_n(\omega)$ that is modeled as time-invariant, the main idea is to exploit the factorization of $v_n(\omega,l)$ in (5.1) to find compact time-invariant representations of the time-varying absolute values of the multiple observations $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)$. To do this, both the basis and coefficient vectors resulting from factorization of $v_n(\omega,l)$ are used to build frequency-dependent time-varying matrices, which are employed to decompose the absolute values of the multiple observations $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)$ applying **supervised** NMF. The factorization output consists of time-invariant compact matrices encoding the spatial diversity among coefficients of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)$ over all the time frames. The spatial covariance matrix $\mathbf{R}_n(\omega)$ is then estimated in sense of Maximum-Likelihood (ML) using the time-varying coefficient vectors $\mathbf{w}_n(l)$ of the factorized spectral variance $v_n(\omega,l)$, the time-invariant matrices of the factorized absolute values of the multiple observations $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)$, and the phase information of the multiple observations $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l)$.

Both the first and second factorization steps are carried out by minimizing the $\beta$-divergence applying the Multiplicative Update (MU) rules [40]. Furthermore, we tested several values of $\beta$ in order to identify the best performing ones from source separation point of view. Intended mismatch in selecting values of $\beta$ for the first and second factorization steps is applied, which leads in some cases to better performance.

### 5.1.1 Estimation of $\mathbf{R}_n(\omega)$

In a supervised factorization scenario, we propose to decompose the absolute values of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ using a time-varying frequency-dependent vector $\mathbf{q}_n(\omega, l)$. The matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ is represented in the factorization domain as follows

$$\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l) = \sum_k q_n(\omega, k, l)\mathbf{W}_{\mathbf{c}_n}(\omega, k) \angle \tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l), \tag{5.3}$$

where $q_n(\omega, k, l)$ is the $k$-th coefficient of the vector $\mathbf{q}_n(\omega, l)$. Unlike the formulations in (4.10) and (4.19), used to represent the matrix of multiple observations $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ in the factorization domain, here, both $q_n(\omega, k, l)$ and $\mathbf{W}_{\mathbf{c}_n}(\omega, k)$ are granted to be frequency-dependent. This proposed formulation allows to follow the amplitude values in the estimation of the matrix $\mathbf{R}_n(\omega)$, from one frequency to another. At each frequency bin $\omega$, the time indeterminacy of the absolute values of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, from one time-frame to another, is described by $q_n(\omega, k, l)$. Moreover, the spatial indeterminacy of the absolute values of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, from one coefficient to another, is specified by the time-invariant compact matrix

$$\mathbf{W}_{\mathbf{c}_n}(\omega, k) = \begin{bmatrix} w_{\mathbf{c}_n}^{11}(\omega, k) & \cdots & w_{\mathbf{c}_n}^{1M}(\omega, k) \\ \vdots & \ddots & \vdots \\ w_{\mathbf{c}_n}^{M1}(\omega, k) & \cdots & w_{\mathbf{c}_n}^{MM}(\omega, k) \end{bmatrix}, \tag{5.4}$$

where $w_{\mathbf{c}_n}^{m_1 m_2}(\omega, k)$ indicates the $(m_1, m_2)$ coefficient of the matrix $\mathbf{W}_{\mathbf{c}_n}(\omega, k)$, that corresponds to the $(m_1, m_2)$ coefficient of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$.

For better understanding the building of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ in the factorization domain, we want to confirm that its $m$-th diagonal observation $\tilde{r}^{mm}_{\mathbf{c}_n}(\omega, l)$ is a real coefficient, which is expressed in the factorization domain as follows

$$\tilde{r}^{mm}_{\mathbf{c}_n}(\omega, l) = \sum_k q_n(\omega, k, l) w^{mm}_{\mathbf{c}_n}(\omega, k), \tag{5.5}$$

and its ($m_1$,$m_2$) off-diagonal observation is a complex coefficient that is represented as

$$\tilde{r}^{m_1 m_2}_{\mathbf{c}_n}(\omega, l) = \sum_k q_n(\omega, k, l) w^{m_1 m_2}_{\mathbf{c}_n}(\omega, k) \angle \tilde{r}^{m_1 m_2}_{\mathbf{c}_n}(\omega, l), \tag{5.6}$$

where $\angle \tilde{r}^{m_1 m_2}_{\mathbf{c}_n}(\omega, l)$ indicates the phase information of the ($m_1$,$m_2$) coefficient $\tilde{r}^{m_1 m_2}_{\mathbf{c}_n}(\omega, l)$. The minimization function in (4.3) is expressed, by substituting the factorization of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ in (5.3), as

$$\xi(\theta) = \sum_{\omega, l, n} \mathrm{tr}(v_n^{-1}(\omega, l) \mathbf{R}_n^{-1}(\omega) \sum_k q_n(\omega, k, l) \mathbf{W}_{\mathbf{c}_n}(\omega, k) \angle \tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)). \tag{5.7}$$

Let's initially propose that the component $q_n(\omega, k, l)$ be defined as the product of two other scalar components as follows

$$q_n(\omega, k, l) = v_n(\omega, l) w_n(k, l), \tag{5.8}$$

where $w_n(k, l)$ is the $k$-th coefficient of the vector $\mathbf{w}_n(l)$. Keep in mind that $v_n(\omega, l)$ itself can be also represented as the product of two vectors as in (5.1). The minimization function in (5.7) is represented again by substituting the factorization of $q_n(\omega, k, l)$ in (5.8) and eliminating $v_n(\omega, l)$, as follows

$$\xi(\theta) = \sum_{\omega, l, n} \mathrm{tr}(\mathbf{R}_n^{-1}(\omega) \sum_k w_n(k, l) \mathbf{W}_{\mathbf{c}_n}(\omega, k) \angle \tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)). \tag{5.9}$$

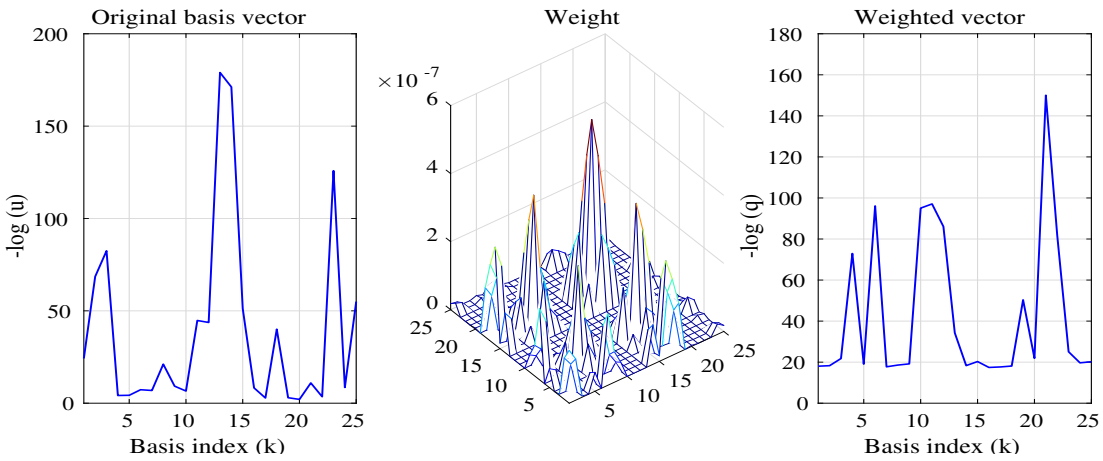


Figure 5.2: A trained basis vector and its generated weighted copy which is used to factorize a particular time-frame of multiple observations.

The spatial covariance matrix $\mathbf{R}_n(\omega)$ is then estimated in sense of Maximum-Likelihood (ML) by minimizing the optimization function in (5.9) and averaging over all the time-frames, as follows

$$\mathbf{R}_n(\omega) = \frac{1}{L} \sum_{l,k} w_n(k,l) \mathbf{W}_{\mathbf{c}_n}(\omega,k) \angle \tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega,l). \tag{5.10}$$

The matrix $\mathbf{R}_n(\omega)$ is then normalized using its largest singular value. On the other side, regarding the factorization of the estimated source variance $v_n(\omega,l)$ in (5.1) and the component $q_n(\omega,k,l)$ in (5.8), the vector $\mathbf{q}_n(\omega,l)$ is represented as a weighted copy of the spectral basis vector $\mathbf{u}_n(\omega)$ as

$$\mathbf{q}_n^T(\omega,l) = \mathbf{u}_n^T(\omega)[\mathbf{w}_n(l)\mathbf{w}_n^T(l)], \tag{5.11}$$

where the weight $[\mathbf{w}_n(l)\mathbf{w}_n^T(l)]$ is the outer-product of the vector $\mathbf{w}_n(l)$ and its transposition.

Figure 5.2 shows an example of an original spectral basis vector $\mathbf{u}_n(\omega)$ of length $K = 25$ at the frequency $\omega$, a weight matrix $[\mathbf{w}_n(l)\mathbf{w}_n^T(l)]$ of size $25 \times 25$ obtained at a specific time-frame $l$, and a generated weighted vector $\mathbf{q}_n(\omega,l)$ of length $K = 25$.

As it is clear, much trust is given to the estimation of the source variance $v_n(\omega, l)$, where a combination of its factorization ($\mathbf{q}_n(\omega, l)$ as in (5.11)) weighted by the time-invariant matrix $\mathbf{W}_{\mathbf{c}_n}(\omega, k)$, is employed to describe the temporal activity of source spatial images $\mathbf{c}_n(\omega, l)$ in the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$. As a consequence, robust estimation and efficient factorization of $v_n(\omega, l)$ are strict requirements. The estimation step is accomplished as proposed in Section 4.1.2. *The factorization can be performed in* ***supervised*** *or* ***unsupervised*** *scenarios. In case of supervised factorization, the spectral basis vectors* $\mathbf{u}_n(\omega)$ *are pre-trained in advance using a set of training data, and kept fixed during the factorization of* $v_n(\omega, l)$ *in (5.1). However, the separation can be achieved in a blind scenario by on-line training of the vectors* $\mathbf{u}_n(\omega)$ *applying unsupervised factorization of* $v_n(\omega, l)$ *in (5.1).*

### 5.1.2 Matrix representation of multiple observations

To perform the factorization of multiple observations in (5.3), a frequency-dependent matrix composed of the vectors $\mathbf{q}_n(\omega, l)$ over all the time-frames is built, i.e. $\mathbf{Q}_n(\omega) = [\{\mathbf{q}_n^T(\omega, l)\}_l]_{L\times K}$. The absolute values of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ over all the time frames are also arranged side-by-side in a matrix of observations $\tilde{\mathbf{V}}_{\mathbf{c}_n}(\omega) = [\{|\tilde{r}_{\mathbf{c}_n}^{m_1 m_2}(\omega, l)|\}_{m_1 m_2, l}]_{L\times M^2}$. At the frequency bin $\omega$, the temporal diversity between columns of the matrix $\tilde{\mathbf{V}}_{\mathbf{c}_n}(\omega)$ is modeled to be constant, and it is represented by the matrix $\mathbf{Q}_n(\omega)$. The spatial diversity between rows of the matrix $\mathbf{V}_{\mathbf{c}_n}(\omega)$ is not constant, as it depends on propagation inter and cross channel intensities, and it is encoded at each $k$-th coefficient by the matrix $\mathbf{W}_{\mathbf{c}_n}(\omega, k)$. To perform the factorization, coefficients of the matrix $\mathbf{W}_{\mathbf{c}_n}(\omega, k)$ are arranged side-by-side in a matrix of size $K \times M^2$, i.e. $\mathbf{H}_{\mathbf{c}_n}(\omega) = [\{w_{\mathbf{c}_n}^{m_1 m_2}(\omega, k)\}_{m_1 m_2, k}]_{K\times M^2}$. Accordingly, the matrix of observations is represented in the factorization domain as

$$\tilde{\mathbf{V}}_{\mathbf{c}_n}(\omega) = \mathbf{Q}_n(\omega)\mathbf{H}_{\mathbf{c}_n}(\omega). \tag{5.12}$$

## 5.2 Full description

The method is summarized as follows:

---

**Training**: $\mathbf{U}_n,\ n = 1, ..., N$ as in Section 4.1.1
**Input**: $\mathbf{x}(\omega, l)$
**Initialize**: $\tilde{\mathbf{c}}_n(\omega, l)$ as in Section 4.1.7, $\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l) = \mathbf{I}$
**Iterate**: *till convergence*
Compute $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ as in (4.1)
Estimate $v_n(\omega, l)$ as in Section 4.1.2
Arrange absolute values of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ in $\tilde{\mathbf{V}}_{\mathbf{c}_n}(\omega)$ as in Section 5.1.2
**Factorization**: as in Section 5.3
(1) Supervised or Unsupervised factorization of $\mathbf{V}_n$ in (5.2) using $\beta^1$
(2) Supervised Factorization of $\tilde{\mathbf{V}}_{\mathbf{c}_n}(\omega)$ in (5.12) using $\beta^2$
Estimate $\mathbf{R}_n(\omega)$ as in Section 5.1.1
**Separation**:

$$\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l) = v_n(\omega, l)\mathbf{R}_n(\omega)$$
$$\mathbf{G}_n(\omega, l) = \mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)\mathbf{\Sigma}_{\mathbf{x}}^{-1}(\omega, l)$$
$$\tilde{\mathbf{c}}_n(\omega, l) = \mathbf{G}_n(\omega, l)\mathbf{x}(\omega, l)$$

**Return**
**Output**: $\tilde{\mathbf{c}}_n(\omega, l)$

---

## 5.3 Supervised NMF using $\beta$-divergence

The Multiplicative Update (MU) rule to minimize the $\beta$-divergence between a matrix $\mathbf{A}$ and its matrix factorization $\mathbf{BC}$, is applied as follows (appendix A)

$$\mathbf{B} \leftarrow \mathbf{B} \circ \frac{[\mathbf{A} \circ (\mathbf{BC})^{\beta-2}]\mathbf{C}^T}{(\mathbf{BC})^{\beta-1}\mathbf{C}^T}, \tag{5.13}$$

$$\mathbf{C} \leftarrow \mathbf{C} \circ \frac{\mathbf{B}^T[\mathbf{A} \circ (\mathbf{BC})^{\beta-2}]}{\mathbf{B}^T(\mathbf{BC})^{\beta-1}}, \tag{5.14}$$

where $\circ$ indicates the point-wise multiplication, and the division is point-wise. Matrix factorization is performed by alternating (5.13) and (5.14) till the convergence is achieved. Supervised factorization is carried out by fixing one matrix, either $\mathbf{B}$ or $\mathbf{C}$, and updating only the other one. The factorization of the power spectrum in (5.2) and the multiple observations in (5.12) are accomplished by respectively replacing $\mathbf{A}$ by either $\mathbf{V}_n$ or $\mathbf{V}_{\mathbf{c}_n}(\omega)$, $\mathbf{B}$ by either $\mathbf{U}_n$ or $\mathbf{Q}_n(\omega)$, and $\mathbf{C}$ by either $\mathbf{W}_n$ or $\mathbf{H}_{\mathbf{c}_n}(\omega)$.

### 5.3.1 Analysis of semi-supervised factorization for single channel source extraction

The current work of single or multiple channel source separation based on supervised NMF relies on feeding the separation system by trained spectral basis matrices of all sources in observed mixtures, and on fixing the value of the divergence factor $\beta$ for both training and supervised reconstruction, for any value of latent coefficients $K$. In this section we experimentally study the influence of selecting values of $K$ and $\beta$ for NMF-based training and semi-supervised reconstruction, for semi-supervised single channel source extraction. In the training phase, spectral basis vectors $\mathbf{u}_t(\omega)$ of a target source $s_t(t)$ are trained using the source power spectrum, with a training divergence factor $\beta^t$. In the reconstruction phase, observing a mixture of the target source signal and an interfering signal, and given the trained vectors $\mathbf{u}_t(\omega)$, we want to estimate activation coefficient vectors $\mathbf{w}_t(l)$ best representing the target source $s_t(t)$, by testing several values of a reconstruction divergence factor $\beta^s$. To extract the target source, we built a soft clustering-based single channel source extraction system. The extraction performance is evaluated using the Source-to-Distortion Ratio (SDR).

The experimental steps are summarized as follows:

- Train $\mathbf{u}_t(\omega)$ using the power spectrum of a target speech signal $s_t(t)$ applying unsupervised NMF, with $K$ and $\beta^t$.
- Generate a mixture by adding an interfering speech signal $s_{intf}(t)$ to the target one, i.e. $y(t) = s_t(t) + s_{intf}(t)$.
- Given $\mathbf{u}_t(\omega)$, estimate $\mathbf{w}_t(l)$ by factorizing the power spectrum $p_y(\omega, l)$ of $y(t)$ applying semi-supervised NMF, with $K$ and $\beta^s$.
- Compute the Signal-to-Interference Ratio as
$$\mathrm{SIR}(\omega, l) = \frac{\mathbf{u}_t^T(\omega)\mathbf{w}_t(l)}{|p_y(\omega,l) - \mathbf{u}_t^T(\omega)\mathbf{w}_t(l)|}$$
- Avoid the outliers resulting of the division by comparing them to a threshold, and build a soft clustering mask as
$$\mathrm{SM}(\omega, l) = \log(\mathrm{SIR}(\omega, l))$$
- Set the negative values corresponding to low SIR to a small positive threshold, and apply spectral-temporal smoothing as
$$\mathrm{SM}^s(\omega, l) = \frac{\sum_{\tilde{\omega},\tilde{l}} \gamma(\tilde{\omega}-\omega,\tilde{l}-l)\mathrm{SM}(\tilde{\omega},\tilde{l})}{\sum_{\tilde{\omega},\tilde{l}} \gamma(\tilde{\omega}-\omega,\tilde{l}-l)}$$
- Obtain estimation of the source signal as
$$\tilde{s}_t(\omega, l) = \mathrm{SM}^s(\omega, l) y(\omega, l)$$
- Compute the extraction performance (SDR) using $s_t(t)$ and $\tilde{s}_t(t)$.

The experiment was carried out using eight speech signals, i.e. four males and four females (in both cases two English and two Japanese speech signals). The dataset is an excerpt from the SISEC evaluation campaign [1]. The mixtures were generated with lengths of 10 s. The discrete time-frequency representation of each mixture $y(t)$ is obtained through STFT with a Hanning analysis window of length 128 ms (or 2048 samples) and shift of 64 ms. The training divergence factor ($\beta^t$) to train the vectors

$\mathbf{u}_t(\omega)$ was either 0.5 or 0.9. The number of the trained spectral basis vectors ($K$) was 15, 25, 35 or 50. The value of the reconstruction divergence factor ($\beta^s$) was selected to span the interval between 0.1 and 1.9 by a step of 0.2. For each value of $\beta^t$, $K$ and $\beta^s$, the number of mixtures under test was 56 speech signals. The input SDR of each one of the mixtures measured between the signal $s_t(t)$ and the mixture $y(t)$ varies between small negative and positive values depending on the signal level in the mixture.

**Analysis and results**

In the reconstruction phase, the power spectrum of the mixture $y(t)$ can be approximately represented in the factorization domain as follows

$$p_y(\omega, l) \approx \mathbf{u}_t^T(\omega)\mathbf{w}_t(l) + \epsilon(\omega, l), \tag{5.15}$$

where $\epsilon(\omega, l)$ is the factorization error. For fixed value of $\beta^t$, we can find an optimal value of $\beta^s$ that best reduces the impact of the interference by constraining the sparsity of the semi-supervised reconstruction. On the other side, for any value of $\beta^s$, even if it equals $\beta^t$, the error $\epsilon(\omega, l)$ has a considerable value. This means, in some cases, we can obtain good interference reduction when $\beta^t = \beta^s$. In fact, after an extensive experimental study, we can state that the optimal value of $\beta^s$ is data-dependent and there is no optimal choice valid in general. However, the extraction performance is obtained by computing the average over all the dataset under test.

Figures 5.3 and 5.4 show the average extraction performance (SDR) in dBs. As it is expected, a large number ($K$) of trained spectral basis vectors $\mathbf{u}_t(\omega)$ benefits the performance. Large values of $\beta^s$ perform slightly better than smaller ones when $K$ equals 15 or 25 in both cases of training, $\beta^t$ equals either 0.5 or 0.9. In case that $\beta^t = 0.5$, small values of $\beta^s$ perform better than large ones when $K$ equals either 35 and 50. However, better performance is achieved when $\beta^s$ is around 0.9 when $\beta^t = 0.9$.

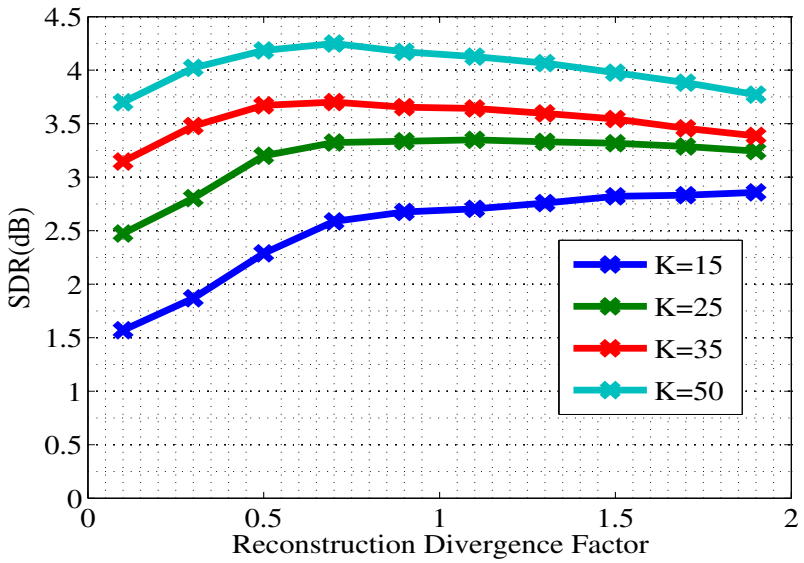


Figure 5.3: Source extraction performance, the training divergence factor = 0.9 and the average input SDR = 0.04 dB.

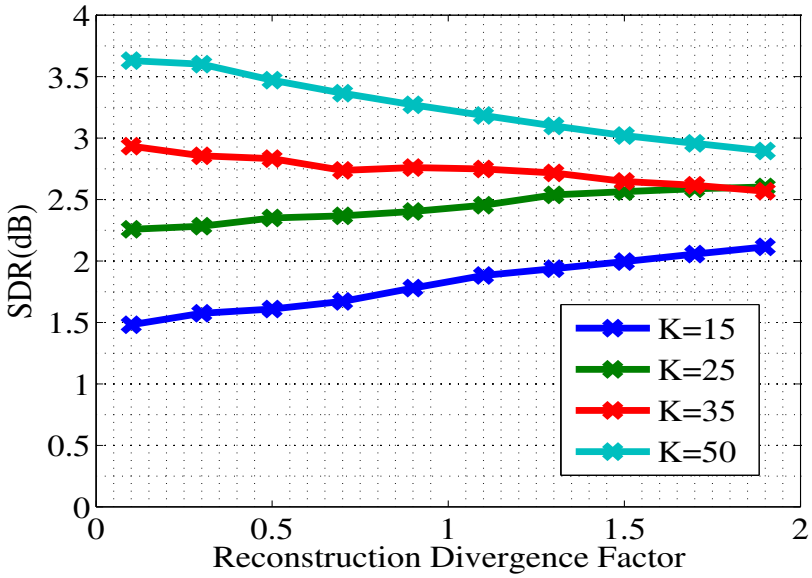


Figure 5.4: Source extraction performance, the training divergence factor = 0.5 and the average input SDR = 0.04 dB.

### 5.3.2 Analysis of semi-supervised factorization for stereo source separation

As an extension to the above single channel source extraction, we analysed the proposed multichannel source separation method. The training step of $\mathbf{u}_t(\omega)$ in the above single channel source extraction method corresponds to the training of $\mathbf{u}_n(\omega), n = 1, .., N$, using a training divergence factor $\beta^t$. The factorization of the estimated source variance $v_n(\omega, l)$ in (5.1), i.e. the estimated source power spectrum $\mathbf{V}_n$ in (5.2), is performed applying **supervised** NMF using a factorization divergence factor $\beta^1$, given the trained spectral basis vectors $\mathbf{u}_n(\omega), n = 1, .., N$. The spatial covariance matrix $\mathbf{R}_n(\omega)$ is estimated as in Section 5.1.1 using a factorization divergence factor $\beta^2$. The estimation step of the model parameters $v_n(\omega, l)$ and $\mathbf{R}_n(\omega)$ corresponds to the estimation step of $\mathbf{w}_t(l)$ in the above single channel source extraction method. Moreover, the multichannel Wiener filtering in the proposed method corresponds to the single channel soft masking.

The live-recorded dataset *dev*1 of the SISEC evaluation campaign was used for this analysis. The dataset consists of 4 live-recorded stereo mixtures of 3 Japanese and English speech signals. All the mixtures are 10 s long sampled at 16 kHz. The experimental setup consists of 2 omnidirectional microphones placed 1 m apart in a room of dimension $4.45 \times 3.35 \times 2.5$ m with reverberation times 130 and 250 ms. The distance between source positions and the central point between the microphones varies between 0.8 and 1.2 m. For all mixtures the DOAs vary between 60 and 300 angular degrees, with a minimal spacing of 15 degrees. The discrete time-frequency representation of the mixtures $\mathbf{x}(\omega, l)$ is obtained through STFT with a Hanning analysis window of length 128 ms (or 2048 samples), with shift of 64 ms. The window $\gamma$ for the computation of the empirical covariance matrix of the source images in (4.2) is a Hanning window of size $3 \times 3$.

Table 5.1: Average SDR (dB) of informed separation of live-recorded mixtures from SISEC.

| $K$ | 15 | | 25 | | 50 | |
|---|---|---|---|---|---|---|
| $\beta^2$ | 0.5 | 0.9 | 0.5 | 0.9 | 0.5 | 0.9 |
| $\beta^1 = 0.5$ | 6.29 | 5.64 | 6.25 | 6.32 | 7.88 | 7.73 |
| $\beta^1 = 0.9$ | 7.36 | 7.09 | 7.68 | 7.80 | 7.88 | 7.68 |
| $\beta^t = 0.9$ | | | | | | |
| $\beta^1 = 0.5$ | 5.77 | 5.82 | 7.82 | 7.86 | 8.31 | 8.27 |
| $\beta^1 = 0.9$ | 7.18 | 7.28 | 7.58 | 7.43 | 7.78 | 7.73 |
| $\beta^t = 0.5$ | | | | | | |

Table 5.1 reports the average performance as a function of values of $\beta^t$, $\beta^1$, $\beta^2$, and $K$. As it is noted, the performance follow the same trend observed in the above single channel case. Supervised factorization of $v_n(\omega, l)$ using $\beta^1 = 0.9$ provides better performance than using $\beta^1 = 0.5$ when $K = 15$, in both cases of training, i.e. $\beta^t$ equals either 0.5 or 0.9. Almost the same performance is achieved when $\beta^t = 0.9$ and $K = 50$, in both cases of estimation, i.e. $\beta^1$ equals either 0.9 or 0.5. Corresponding to the trend observed in the above single channel case, $\beta^1 = 0.5$ performs better than $\beta^1 = 0.9$ when $\beta^t = 0.5$ and $K = 50$. In order to apply two different sparsity constraints, we tested generated mismatch between the selected values of $\beta^1$ and $\beta^2$, which improves the performance in some cases.

## 5.4 Experiments

In this section, we evaluate the proposed method in a blind scenario. The factorization of the estimated source variance $v_n(\omega, l)$ in (5.1), i.e. the estimated power spectrum $\mathbf{V}_n$ in (5.2), is performed applying **unsupervised** NMF using a factorization divergence factor $\beta^1$, i.e. preliminar training of the spectral basis vectors $\mathbf{u}_n(\omega)$ is not needed. The spatial covariance matrix $\mathbf{R}_n(\omega)$ is estimated as in Section 5.1.1 using a factorization divergence

factor $\beta^2$. The experimental evaluation was carried out as a function of $K$, $\beta^1$, and $\beta^2$. The selected values of $K$ were 15, 25, or 50, while each of $\beta^1$ and $\beta^2$ were assigned values of 0.5 and 0.9. Two different datasets were used for this evaluation, including synthetic and live-recorded data.

### 5.4.1 Synthetic simulated dataset

A room with size $4.45 \times 3.35 \times 2.5$ m and an array of 2 omnidirectional microphones spaced of 0.2 m are considered. The microphones are located in the middle of the room and they are at the same height (i.e., 1.4 m) of three given sources. The distance from each source to the mid point between the two microphones is 1 m. The direction of arrivals of the sources are 40, 85, and 130 degrees. Synthetic room impulse responses (RIRs) are simulated through ISM [54] with a sampling frequency of 16 kHz for two reverberation times: $T_{60} = 130$ or 250 ms. Six Italian speakers (3 males and 3 females) are considered as audio sources. Six mixture combinations of male-female speech signals were generated for each case of the reverberation times. The discrete time-frequency representation of the mixtures is obtained through STFT with a Hanning analysis window of length 128 ms (or 2048 samples), with shift of 64 ms ($L = 137$). The window $\gamma$ for the computation of the empirical covariance matrix of the source images in (4.2) is a Hanning window of size $3 \times 3$.

Table 5.2 shows the average performance as a function of values of $K$, $\beta^1$, and $\beta^2$. As previously observed, large values of $K$ benefit the performance. On the average, mismatch between values of $\beta^1$ and $\beta^2$ does not influence much the performance. For comparison, the proposed method (SS-WSB) outperforms two other methods as reported in Table 5.3. The first method is the original Gaussian model-based source separation proposed in [30], and the second one is based on clustering time-frequency points of the mixtures using estimated Time Difference-of-Arrivals (TDOAs) [69].

Table 5.2: Average performance of blind separation of synthetic stereo mixtures.

| $K$ | | 15 | | 25 | | 50 | |
|---|---|---|---|---|---|---|---|
| $\beta^2$ | | 0.5 | 0.9 | 0.5 | 0.9 | 0.5 | 0.9 |
| $\beta^1 = 0.5$ | SDR | 7.32 | 7.46 | 7.69 | 7.76 | 8.13 | 8.22 |
| | ISR | 12.49 | 12.78 | 13.05 | 13.07 | 13.59 | 13.76 |
| | SIR | 12.47 | 12.76 | 13.06 | 13.16 | 13.68 | 13.76 |
| | SAR | 9.39 | 9.92 | 9.67 | 9.91 | 10.20 | 10.23 |
| $\beta^1 = 0.9$ | SDR | 7.03 | 7.31 | 7.80 | 7.99 | 8.19 | 8.33 |
| | ISR | 11.95 | 12.33 | 13.15 | 13.33 | 13.75 | 13.81 |
| | SIR | 11.98 | 12.26 | 13.16 | 13.41 | 13.82 | 13.84 |
| | SAR | 9.51 | 9.78 | 9.79 | 10.19 | 10.18 | 10.42 |

Table 5.3: Comparison of blind separation performance.

| dB | SS-WSB | ML-blind | TDOAs |
|---|---|---|---|
| SDR | 8.33 | 5.72 | 5.88 |
| ISR | 13.81 | 10.46 | 12.23 |
| SIR | 13.84 | 8.38 | 11.53 |
| SAR | 10.42 | 10.46 | 8.72 |

### 5.4.2 Live-recorded dataset of SISEC

The SISEC dataset explained in Section 5.3.2 is used to compare the separation performance in a blind scenario. In this case the performance of the proposed method is compared to two recently developed methods in [22, 67]. The separation results of these two methods are published on the SISEC website. The separation performance of the proposed method was evaluated in details as a function of values of $K$, $\beta^1$ and $\beta^2$. Table 5.4 reports the average separation performance of the four live-recorded stereo mixtures generated using two different reverberation times, i.e. $T_{60} = 130$ ms and 250 ms. On the average, the proposed method outperforms the one developed in [67] and achieves separation performance comparable to the one developed in [22]. Better performance is obtained

Table 5.4: Average SDR (dB) of blind separation of 4 live-recorded stereo mixtures of three male and three female speech signals from SISEC, $T_{60} = 130$ ms and 250 ms.

| Proposed | | | | | | | | Baseline | |
|---|---|---|---|---|---|---|---|---|---|
| $K$ | | 15 | | 25 | | 50 | | Nesta | Cho |
| $\beta^2$ | | 0.5 | 0.9 | 0.5 | 0.9 | 0.5 | 0.9 | [67] | [22] |
| $\beta^1 = 0.5$ | SDR | 5.30 | 6.35 | 6.01 | 6.08 | 6.26 | 6.61 | 6.35 | 6.75 |
| $\beta^1 = 0.9$ | SDR | 4.67 | 5.20 | 5.34 | 5.85 | 6.04 | 5.94 | | |

Table 5.5: Detailed performance of blind separation of live-recorded stereo mixtures of three female speech signals from SISEC, $T_{60} = 130$ ms.

| Proposed | | | | | | | | Baseline | |
|---|---|---|---|---|---|---|---|---|---|
| $K$ | | 15 | | 25 | | 50 | | Nesta | Cho |
| $\beta^2$ | | 0.5 | 0.9 | 0.5 | 0.9 | 0.5 | 0.9 | [67] | [22] |
| $\beta^1 = 0.5$ | SDR | 8.62 | 9.34 | 9.67 | 9.34 | 8.92 | 9.24 | 7.70 | 8.40 |
| | ISR | 13.98 | 14.56 | 14.87 | 14.29 | 14.19 | 14.36 | 10.50 | 13.00 |
| | SIR | 14.56 | 15.42 | 15.74 | 14.93 | 14.36 | 14.41 | 13.30 | 12.60 |
| | SAR | 12.22 | 12.10 | 12.60 | 13.00 | 11.54 | 12.39 | 11.80 | 12.10 |
| $\beta^1 = 0.9$ | SDR | 6.80 | 7.78 | 7.44 | 8.73 | 8.67 | 8.63 | | |
| | ISR | 12.15 | 12.62 | 12.77 | 13.56 | 13.33 | 13.14 | | |
| | SIR | 12.37 | 12.32 | 12.42 | 13.50 | 13.50 | 13.26 | | |
| | SAR | 10.30 | 11.61 | 10.86 | 12.24 | 12.18 | 12.41 | | |

if $\beta^1$ is assigned a value of 0.5. In this case, if $\beta^2$ is assigned a value of 0.9, improved performance is achieved.

Observing the separation performance of each one of the stereo mixtures independently as reported in Tables 5.5, 5.6, 5.7, it is noted that values of $K$, $\beta^1$ and $\beta^2$ can be tuned for each one of the mixtures, in order to get the best performance. All in all, large values of $K$ mostly benefit the separation performance, moreover, mismatch between values of $\beta^1$ and $\beta^2$ plays a good role in improving the performance.

Table 5.6: Detailed performance of blind separation of live-recorded stereo mixtures of three male speech signals from SISEC, $T_{60} = 130$ ms.

| | | Proposed | | | | | | Baseline | |
|---|---|---|---|---|---|---|---|---|---|
| $K$ | | 15 | | 25 | | 50 | | Nesta | Cho |
| $\beta^2$ | | 0.5 | 0.9 | 0.5 | 0.9 | 0.5 | 0.9 | [67] | [22] |
| $\beta^1 = 0.5$ | SDR | 4.41 | 4.90 | 5.25 | 5.46 | 6.83 | 6.68 | 6.50 | 6.50 |
| | ISR | 8.82 | 9.56 | 9.90 | 10.23 | 11.90 | 11.58 | 9.30 | 11.40 |
| | SIR | 7.36 | 8.60 | 9.08 | 9.38 | 11.82 | 11.55 | 10.90 | 10.00 |
| | SAR | 7.68 | 7.62 | 7.81 | 8.40 | 9.00 | 8.97 | 9.60 | 10.50 |
| $\beta^1 = 0.9$ | SDR | 4.58 | 5.32 | 5.89 | 6.42 | 6.95 | 6.40 | | |
| | ISR | 9.24 | 10.10 | 10.91 | 11.55 | 12.05 | 11.32 | | |
| | SIR | 8.29 | 9.73 | 10.65 | 11.39 | 11.98 | 10.96 | | |
| | SAR | 7.21 | 7.74 | 8.00 | 8.44 | 8.99 | 8.70 | | |

Table 5.7: Detailed performance of blind separation of live-recorded stereo mixtures of three female speech signals from SISEC, $T_{60} = 250$ ms.

| | | Proposed | | | | | | Baseline | |
|---|---|---|---|---|---|---|---|---|---|
| $K$ | | 15 | | 25 | | 50 | | Nesta | Cho |
| $\beta^2$ | | 0.5 | 0.9 | 0.5 | 0.9 | 0.5 | 0.9 | [67] | [22] |
| $\beta^1 = 0.5$ | SDR | 4.71 | 6.82 | 4.74 | 4.63 | 4.77 | 4.70 | 6.00 | 6.10 |
| | ISR | 9.61 | 11.02 | 9.74 | 9.37 | 10.29 | 10.22 | 8.90 | 10.90 |
| | SIR | 9.31 | 10.84 | 8.28 | 7.45 | 8.66 | 8.45 | 10.60 | 9.00 |
| | SAR | 8.41 | 10.84 | 10.16 | 10.41 | 9.91 | 10.11 | 8.70 | 10.00 |
| $\beta^1 = 0.9$ | SDR | 3.76 | 4.00 | 4.40 | 4.32 | 4.51 | 4.43 | | |
| | ISR | 8.36 | 8.66 | 9.25 | 9.03 | 9.22 | 8.84 | | |
| | SIR | 7.07 | 7.30 | 7.21 | 6.86 | 7.33 | 7.11 | | |
| | SAR | 8.67 | 9.03 | 9.68 | 9.86 | 9.82 | 10.00 | | |

Table 5.8: Detailed performance of blind separation of live-recorded stereo mixtures of three male speech signals from SISEC, $T_{60} = 250$ ms.

| | | Proposed | | | | | | Baseline | |
|---|---|---|---|---|---|---|---|---|---|
| $K$ | | 15 | | 25 | | 50 | | Nesta | Cho |
| $\beta^2$ | | 0.5 | 0.9 | 0.5 | 0.9 | 0.5 | 0.9 | [67] | [22] |
| $\beta^1 = 0.5$ | SDR | 3.44 | 4.32 | 4.38 | 4.88 | 4.50 | 5.81 | 5.20 | 6.00 |
| | ISR | 7.82 | 8.59 | 9.14 | 9.37 | 8.87 | 10.54 | 8.30 | 10.50 |
| | SIR | 7.11 | 8.32 | 8.70 | 8.85 | 8.07 | 10.38 | 9.00 | 9.10 |
| | SAR | 5.56 | 6.87 | 6.69 | 7.19 | 7.23 | 7.78 | 8.00 | 9.20 |
| $\beta^1 = 0.9$ | SDR | 3.55 | 3.68 | 3.64 | 3.94 | 4.16 | 4.31 | | |
| | ISR | 7.91 | 7.84 | 8.07 | 8.10 | 8.50 | 8.49 | | |
| | SIR | 6.55 | 6.36 | 6.93 | 6.92 | 7.59 | 7.74 | | |
| | SAR | 6.13 | 6.62 | 6.65 | 7.28 | 7.01 | 7.39 | | |

## 5.5 Conclusion

In this chapter we introduced a method to estimate the parameters of the Gaussian model-based audio source separation. We aimed at mitigating a weakness point in estimation method proposed in the previous chapter. The model is parametrized by variances of audio sources in observed mixtures and corresponding spatial covariance matrices. We exploited unsupervised or supervised factorization of the blindly estimated source variances to build weighted basis metrics, which are used to factorize matrices of multiple observations in order to estimate the spatial covariance matrices. The proposed method can work in either informed or blind scenarios. In both cases, source spectral basis matrices are used to estimate the spatial covariance matrices applying supervised Nonnegative Matrix Factorization (NMF). In case of informed scenario the basis matrices are trained in advance and the estimated source variances are decomposed applying supervised NMF. However, in the other case the basis matrices are trained online by decomposing the estimated source variances applying unsuper-

vised NMF. The above factorization steps are performed by minimizing the $\beta$-divergence using the Multiplicative Update (MU) rules. The separation performance of the proposed method was evaluated as a function of the size of the spectral basis matrices, and the values of $\beta$ for each task of training and estimation. According to the detailed results, we found that both the size of the basis matrices and the values of $\beta$ can be tuned for each mixing condition in order to obtain the best performance. The performance of the proposed method was compared to two recently developed source separation methods. On the average, the proposed method outperforms one of the compared methods and achieves comparable performance to the other one. However, it outperforms both the methods when the comparison is done for each on of the mixtures independently, by tuning the above stated variables.

# Chapter 6

# Nonnegative Decomposition III

## Extracted and trained spectral bases

In line with the existing and proposed methods, the source spectral variance $v_n(\omega, l)$ is modeled using Nonnegative Matrix Factorization (NMF), i.e. $v_n(\omega, l) = \mathbf{u}_n^T(\omega)\mathbf{w}_n(l)$. Given spectral basis vectors $\mathbf{u}_n(\omega)$, here, multichannel decomposition using supervised Nonnegative Matrix/Tensor Factorization (NMF/NTF) applying the $\beta$-divergence is adopted to estimate the set $\theta = \{\{v_n(\omega, l)\}_l, \mathbf{R}_n(\omega)\}_{n,\omega}$. The parameters are jointly updated at one step, which increases the estimation robustness and stability. Using supervised factorization has the advantage that not only the parameters are jointly estimated by factorizing the matrix of multiple observations $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, but also unwanted artifacts are avoided. Due to the implicit requirements of nonnegativity, we propose to split the parameters of the model $\theta$ into two subsets: a subset of nonnegative parameters ($v_n(\omega, l)$ and diagonal coefficients of $\mathbf{R}_n(\omega)$) and another subset of complex parameters (off-diagonal coefficients of $\mathbf{R}_n(\omega)$).

To estimate the first subset, it is required to factorize nonnegative observed components of mixture signals into nonnegative components related to both audio signals and corresponding propagation channels. This estimation step can be performed applying supervised NTF. However, the

second subset is estimated observing complex components of mixture signals. As it is known, nonnegative decomposition can not be applied to factorize complex components. In this sense the second subset can be directly estimated without factorization. However, a scaling ambiguity is generated as a result of estimating the first subset using factorization, and estimating the second subset without factorization.

We will show that it is possible to factorize a complex component into the multiplication of a nonnegative component and a complex component in a supervised scenario, if the nonnegative component is known and kept fixed during the factorization. This factorization step can be seen as estimation without factorization of the second subset observing complex components of the mixture signals. However, the difference is in the scaling of the estimation, which is controlled in this case by the factorization divergence factor $\beta$. Based on the above justification we apply supervised NMF to estimate the second subset of complex parameters. Figure 6.1 shows the flowchart of the proposed method.

In a separate step, observing the mixture signals, we propose that the basis vectors $\mathbf{u}_n(\omega), n = 1, .., N$, are either **extracted**, or **detected** from a redundant **library** containing trained vectors. Both the extraction and detection steps are performed by decomposing the nonnegative observed components of the mixture signals using respectively supervised or unsupervised NTF. Furthermore, exploiting the time-frequency sparsity of audio source signals and preserving their basic spectral structure continuity, $\beta$ is tuned for each task of training, detecting/extracting the spectral basis matrices, and for estimating the parameters of the model.

We highlight that NTF should in principle perform better than NMF when there is spatial redundancy among multiple signal observations, because in NTF multichannel observations are jointly processed in a parallel way. Since the dataset to train spectral bases consists of several spoken

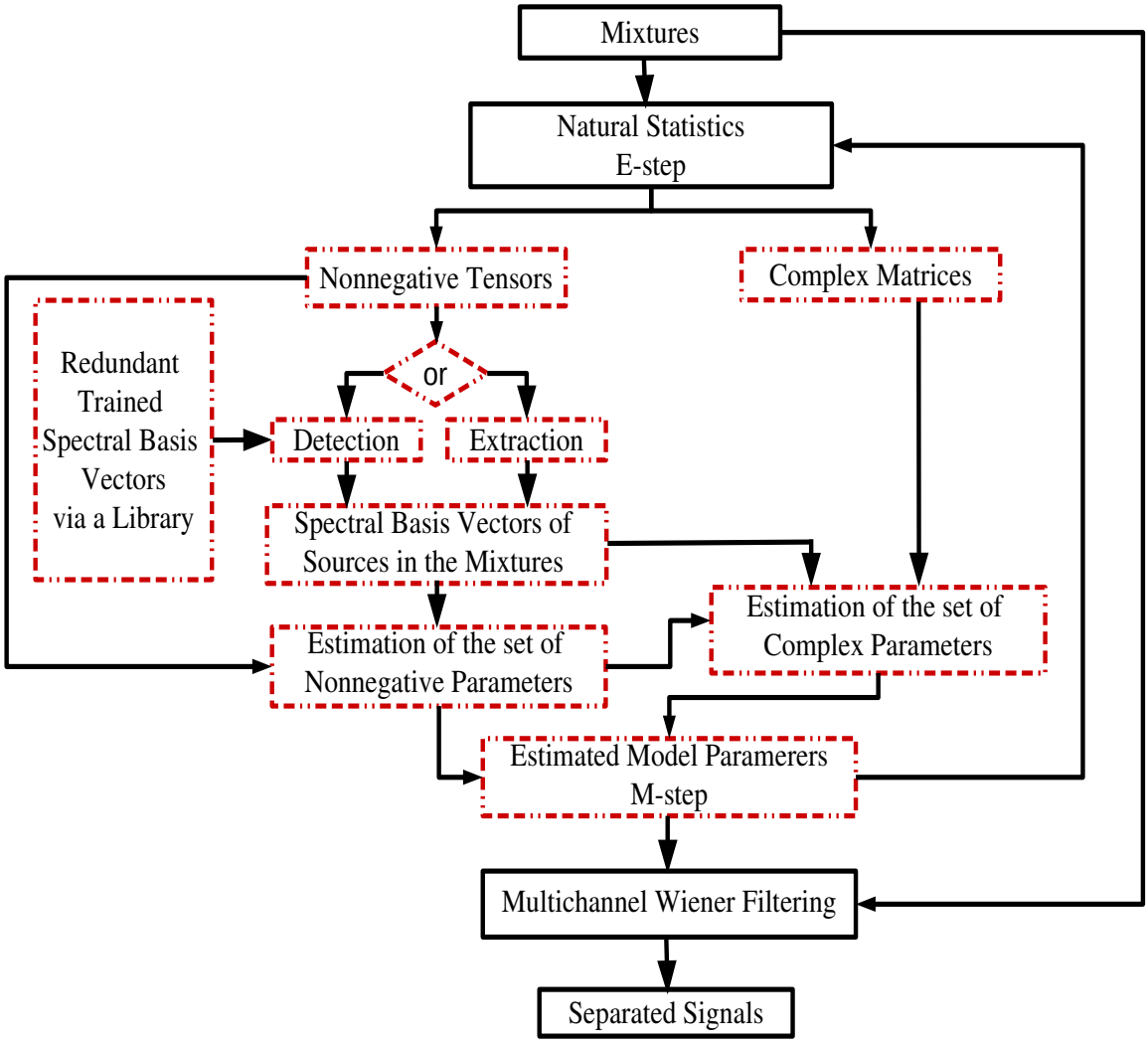


Figure 6.1: A flowchart of the proposed method. Highlighted blocks refer to novel contributions.

sentences, there is not enough spatial redundancy among the sentences, as a result of time-frequency sparsity. For this reason, training using NTF is not much effective, so the bases are trained applying NMF by concatenating power spectra of the training data in one matrix. This training leads to a good representation of the source signals by means of a few spectral basis vectors. On the contrary, for the test data, the same spoken sentences are propagating through several channels. Hence, the spatial redundancy among multiple signal observations is very high. To exploit this redundancy, the multichannel observations are arranged in a 3 D tensor. In this sense, NTF is applied to extract or detect the spectral basis matrices, and to estimate the subset of nonnegative parameters.

## 6.1 Method

As an alternative study to the one performed in Section 5.3.1, in this section we give an example of minimizing the residual artifacts in an observation using semi-supervised factorization by adjusting the value of $\beta$. Let us assume that we observe a corrupted copy $\hat{v}_n(\omega, l)$ of the true source variance $v_n(\omega, l)$, where the corruption derives from other source signals, multipath propagation or noisy environments. In terms of a given reference spectral vector $\mathbf{u}_n(\omega)$ of the source variance $v_n(\omega, l)$, applying semi-supervised NMF, the corrupted source variance can be approximately represented as

$$\hat{v}_n(\omega, l) = \mathbf{u}_n^T(\omega)\hat{\mathbf{w}}_n(l) \approx \mathbf{u}_n^T(\omega)\mathbf{w}_n(l) + \epsilon(\omega, l), \tag{6.1}$$

where $\hat{\mathbf{w}}_n(l)$ indicates the weight coefficient vector of $\hat{v}_n(\omega, l)$ in the factorization domain, and $\epsilon(\omega, l)$ encompasses the corruption error, as well as the factorization error. If the basis vector $\mathbf{u}_n(\omega)$ well describes the true source variance $v_n(\omega, l)$, a properly good estimation of the vector $\mathbf{w}_n(l)$ could be obtained. Following this model and observing $\hat{v}_n(\omega, l)$, given $\mathbf{u}_n(\omega)$, the

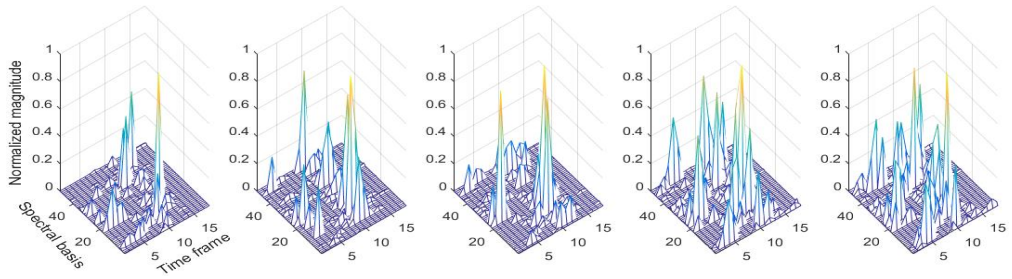


Figure 6.2: Examples of controlling the sparsity of $\mathbf{W}_n$ of the corrupted power spectrum by selecting the value of $\beta$, from the left to the right, respectively, original $\mathbf{W}_n$ (training) with $\beta = 0.9$, estimated with $\beta = 0.1$, estimated with $\beta = 0.3$, estimated with $\beta = 0.6$, and estimated with $\beta = 0.9$.

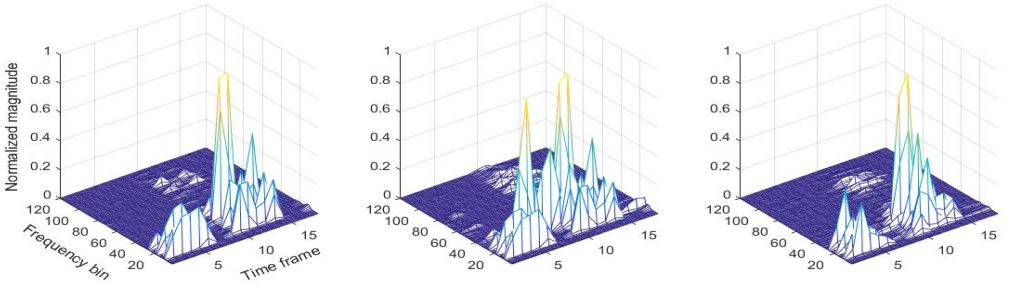


Figure 6.3: Normalized power spectra of true, corrupted, and reconstructed signals, from the left to the right respectively.

vector $\mathbf{w}_n(l)$ can be obtained by applying an efficient factorization algorithm. Figures 6.2 and 6.3 show an example of using the $\beta$-divergence in minimizing the influence of an additive interfering signal from a mixture of an original signal and the interfering one. We trained 50 spectral basis vectors $\mathbf{u}_n(\omega)$ using the power spectrum of a male speech signal, applying the $\beta$-divergence with $\beta = 0.9$ and using the multiplicative update (MU) rule. The male speech signal was linearly mixed with a second female speech signal using a mixing vector with coefficients $[1,\ 0.7]$.

Observing the power spectrum of the mixture of the two speech signals, given the trained vectors $\mathbf{u}_n(\omega)$ of male speech signal, we want to reconstruct the power spectrum of the signal by estimating the coefficient

vectors $\mathbf{w}_n(l)$. For the male speech signal reconstruction using NMF in a semi-supervised scenario applying the $\beta$-divergence, the vectors $\mathbf{w}_n(l)$ can be constrained to be sparse by tuning the value of $\beta$. As we observe in Figure 6.3, the sparsity of the estimated vectors $\mathbf{w}_n(l)$ can be governed. Accordingly, the value of $\beta$ can be tuned to minimize the impact of residual artifacts in a signal observation, which results in a better estimation. Figure 6.3 shows the reconstructed power spectrum using $\beta = 0.3$.

Building on this idea, to estimate $\theta_n = \{\{v_n(\omega, l)\}_l, \mathbf{R}_n(\omega)\}_\omega$, the minimization function in (3.17) can be replaced by the $\beta$-divergence in a scenario that is similar to the semi-supervised estimation explained above. The difference is that the number of parameters to estimate is larger than the number of pre-known parameters. The minimization function based on the $\beta$-divergence is represented element-by-element as follows

$$\tilde{\theta}_n = \arg\min_{\theta_n} \sum_{m_1,m_2} \sum_{\omega,l} d_\beta[\tilde{r}_{\mathbf{c}_n}^{m_1 m_2}(\omega, l)/v_n(\omega, l) r_n^{m_1 m_2}(\omega)], \tag{6.2}$$

where $\tilde{r}_{\mathbf{c}_n}^{m_1 m_2}(\omega, l)$ and $r_n^{m_1 m_2}(\omega)$ are the $(m_1, m_2)$ coefficients of the matrices $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ and $\mathbf{R}_n(\omega)$, respectively, and $m_1, m_2 = 1, ..., M$. Substituting the source variance $v_n(\omega, l)$ by its decomposition, the parameters are estimated by solving the following minimization problem

$$\tilde{\theta}_n = \arg\min_{\theta_n} \sum_{m_1,m_2} \sum_{\omega,l} d_\beta[\tilde{r}_{\mathbf{c}_n}^{m_1 m_2}(\omega, l)/\mathbf{u}_n^T(\omega)\mathbf{w}_n(l) r_n^{m_1 m_2}(\omega)]. \tag{6.3}$$

To minimize the degree of freedom of the decomposition, one or more of the coefficients/vectors might be assumed to be known (prior information). Since the focus of this work is on exploiting source-based information, we assume that the frequency-dependent spectral basis vector $\mathbf{u}_n(\omega)$ is pre-known, and the task is to estimate by decomposition the time-varying coefficient vector $\mathbf{w}_n(l)$ and the frequency-dependent entry $r_n^{m_1 m_2}(\omega)$. Hence the set of parameters is redefined again as $\theta_n = \{\{\mathbf{w}_n(l)\}_l, \{\mathbf{R}_n(\omega)\}_\omega\}$.

Since both $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ and $\mathbf{R}_n(\omega)$ are complex covariance matrices, the nonnegativity constraint of the factorization can be applied only on their diagonal coefficients. For this reason, we split $\theta_n$ into two subsets, a subset of nonnegative parameters $\theta_n^{diag} = \{\{\mathbf{w}_n(l)\}_l, \{r_n^{m_1 m_2}(\omega)\}_{\omega,m_1,m_2}\}$, $m_1 = m_2$, and a subset of complex parameters $\theta_n^{off} = \{r_n^{m_1 m_2}(\omega)\}_{\omega,m_1,m_2}$, $m_1 \neq m_2$, where $\theta_n = \theta_n^{diag} \cup \theta_n^{off}$.

The diagonal entries of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ are multiple observations, where each observation is represented as a corrupted copy of the source variance times the inter-channel propagation intensity. From one observation to another, the source variance is fixed, however, the inter-channel intensity changes. Since the multiple observations are represented by one common component accompanied by multiple components, we propose to use NTF, fixing $\mathbf{u}_n(\omega)$, to update the estimation of $\theta_n^{diag}$.

Each parameter of $\theta_n^{off}$ is individually estimated using NMF, observing each off-diagonal coefficient of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ separately, by fixing $\mathbf{u}_n(\omega)$ and $\mathbf{w}_n(l)$ (from the previous step). In this estimation step, NMF with nonnegative constraints is used to update one complex parameter $r_n^{m_1 m_2}(\omega)$, $m_1 \neq m_2$ out of three parameters, fixing the other two nonnegative parameters $\mathbf{u}_n(\omega)$ and $\mathbf{w}_n(l)$. In practice, the complex parameter is not used to update the nonnegative parameters, so the step is only a scaled update of $r_n^{m_1 m_2}(\omega)$ to keep the scale of computing $r_n^{m_1 m_2}(\omega)$, $m_1 \neq m_2$ the same as the scale of computing $r_n^{m_1 m_2}(\omega)$, $m_1 = m_2$, where the computation scaling parameter is $\beta$.

### 6.1.1 Tensor/matrix representation of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ and $\mathbf{R}_n(\omega)$

Over all the time-frequency points and the signal observations, a tensor $\tilde{\mathbf{V}}_{\mathbf{c}_n}^M$ of size $\Omega \times L \times M$ can be built from the diagonal elements of the matrix $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$. Over the diagonal of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, the $m$-th element is denoted as $\tilde{r}_{\mathbf{c}_n}^m(\omega, l)$, accordingly, the $m$-th slice of the tensor of observations is defined

as $\tilde{\mathbf{V}}^{m}_{\mathbf{c}_n} = [\{\tilde{r}^{m}_{\mathbf{c}_n}(\omega, l)\}_{\omega,l}]_{\Omega \times L}$. The same is done in order to define a tensor with diagonal slices of spatial information, $\mathbf{V}^{m}_{\mathbf{R}_n} = diag\ [\{r^{m}_{n}(\omega)\}_{\omega}]_{\Omega \times \Omega}$, where $r^{m}_{n}(\omega)$ indicates the $m$-th diagonal element of $\mathbf{R}_n(\omega)$. In terms of the tensor $\tilde{\mathbf{V}}^{M}_{\mathbf{c}_n}$ and the matrices $\mathbf{U}_n$ and $\mathbf{W}_n$, the minimization problem of the subset of nonnegative of parameters in (6.3) is described as

$$\tilde{\theta}^{diag}_{n} = \arg\min_{\theta^{diag}_{n}} \sum_{m} \sum_{\omega,l} d_{\beta}[\tilde{\mathbf{V}}^{m}_{\mathbf{c}_n} / \mathbf{V}^{m}_{\mathbf{R}_n} \mathbf{U}_n \mathbf{W}_n], \tag{6.4}$$

where the subset of parameters to estimate is defined as $\theta^{diag}_{n} = \{\mathbf{W}_n, \mathbf{V}^{m}_{\mathbf{R}_n}\}$. Observing the tensor $\tilde{\mathbf{V}}^{M}_{\mathbf{c}_n}$ and given the matrix of source spectral bases $\mathbf{U}_n$, the task is to estimate both the tensor of spatial information $\mathbf{V}^{M}_{\mathbf{R}_n}$ and the matrix of time-varying activation coefficients $\mathbf{W}_n$.

What remains to complete the representation is to rearrange the off-diagonal coefficients of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ and $\mathbf{R}_n(\omega)$. The off-diagonal coefficients of each matrix are the complex conjugation of each other centred around their diagonal coefficients, i.e. the $(m_1, m_2)$ coefficient is the complex conjugation of the $(m_2, m_1)$ coefficient. As a result, half of the coefficients in $\theta^{off}_{n}$ need to be estimated, then the second half is obtained by calculating the complex conjugation of the estimated one. Over all the frequencies, the $(m_1, m_2)$ complex coefficients of $\mathbf{R}_n(\omega)$ are arranged in a diagonal matrix as $\mathbf{V}^{m_1 m_2}_{\mathbf{R}_n} = diag\ [\{r^{m_1 m_2}_{n}(\omega)\}_{\omega}]_{\Omega \times \Omega}$. Over all the time-frequency points, a matrix from the $(m_1, m_2)$ complex coefficients of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ is defined as $\tilde{\mathbf{V}}^{m_1 m_2}_{\mathbf{c}_n} = [\{\tilde{r}^{m_1 m_2}_{\mathbf{c}_n}(\omega, l)\}_{\omega,l}]_{\Omega \times L}$. Given $\mathbf{U}_n$, and the estimation of $\mathbf{W}_n$ obtained from the solution of the previous minimization function (6.4), the estimation minimizer is then defined as

$$\tilde{\theta}^{off}_{n} = \arg\min_{\theta^{off}_{n}} \sum_{\omega,l,m_1 \neq m_2} d_{\beta}[\tilde{\mathbf{V}}^{m_1 m_2}_{\mathbf{c}_n} / \mathbf{V}^{m_1 m_2}_{\mathbf{R}_n} \mathbf{U}_n \mathbf{W}_n], \tag{6.5}$$

where the subset of the complex off-diagonal parameters to estimate is represented as $\theta^{off}_{n} = \{\mathbf{V}^{m_1 m_2}_{\mathbf{R}_n}\}$, $m_1 \neq m_2$.

### 6.1.2 Tensor/matrix update

Tensor/matrix factorization (NMF/NTF) is achieved by minimizing the $\beta$-divergence. To estimate the set $\theta_n$, the MU rule is applied to optimize the minimization functions in (6.4) and (6.5). The rule consists of updating each scalar parameter in $\theta_n$ by multiplying its value at a previous iteration by the ratio of the negative and positive parts of the derivative of the $\beta$-divergence with respect to the parameter (see appendix A).

1. To estimate the first subset of parameters $\theta_n^{diag}$, the tensor of multichannel observations $\tilde{\mathbf{V}}_{\mathbf{c}_n}^M$ is decomposed using NTF, *which accounts for the spatial redundancy among the observations*. Given $\mathbf{U}_n$, the MU rule to estimate the tensor slice $\mathbf{V}_{\mathbf{R}_n}^m$ and the matrix $\mathbf{W}_n$ is obtained by minimizing the $\beta$-divergence (appendix A)

$$\mathbf{W}_n \leftarrow \mathbf{W}_n \circ \frac{\sum_m (\mathbf{V}_{\mathbf{R}_n}^m \mathbf{U}_n)^T [\tilde{\mathbf{V}}_{\mathbf{c}_n}^m \circ (\mathbf{V}_{\mathbf{R}_n}^m \mathbf{U}_n \mathbf{W}_n)^{\beta^s - 2}]}{\sum_m (\mathbf{V}_{\mathbf{R}_n}^m \mathbf{U}_n)^T (\mathbf{V}_{\mathbf{R}_n}^m \mathbf{U}_n \mathbf{W}_n)^{\beta^s - 1}}, \tag{6.6}$$

$$\mathbf{V}_{\mathbf{R}_n}^m \leftarrow \mathbf{V}_{\mathbf{R}_n}^m \circ \frac{[\tilde{\mathbf{V}}_{\mathbf{c}_n}^m \circ (\mathbf{V}_{\mathbf{R}_n}^m \mathbf{U}_n \mathbf{W}_n)^{\beta^s - 2}](\mathbf{U}_n \mathbf{W}_n)^T}{(\mathbf{V}_{\mathbf{R}_n}^m \mathbf{U}_n \mathbf{W}_n)^{\beta^s - 1} (\mathbf{U}_n \mathbf{W}_n)^T}, \tag{6.7}$$

   where $\circ$ indicates element-wise multiplication, $\beta^s$ denotes the value of $\beta$ used for estimating the parameters. The division is element-wise. As it is noted, each slice $\mathbf{V}_{\mathbf{R}_n}^m$ is independently updated. However, the matrix $\mathbf{W}_n$ is jointly updated using slices of the tensors $\tilde{\mathbf{V}}_{\mathbf{c}_n}^M$ and $\mathbf{V}_{\mathbf{R}_n}^M$, which allows to exploit the multichannel spatial redundancy.

2. As the matrix $\mathbf{W}_n$ is estimated, it is accompanied with the matrix $\mathbf{U}_n$ in order to estimate the second subset $\theta_n^{off}$ in the same way we estimated the spatial information $\mathbf{V}_{\mathbf{R}_n}^m$ in the previous step, but using different observations

$$\mathbf{V}_{\mathbf{R}_n}^{m_1 m_2} \leftarrow \mathbf{V}_{\mathbf{R}_n}^{m_1 m_2} \circ \frac{[\tilde{\mathbf{V}}_{\mathbf{c}_n}^{m_1 m_2} \circ (\mathbf{V}_{\mathbf{R}_n}^{m_1 m_2} \mathbf{U}_n \mathbf{W}_n)^{\beta^s - 2}](\mathbf{U}_n \mathbf{W}_n)^T}{(\mathbf{V}_{\mathbf{R}_n}^{m_1 m_2} \mathbf{U}_n \mathbf{W}_n)^{\beta^s - 1} (\mathbf{U}_n \mathbf{W}_n)^T}. \tag{6.8}$$

In this step, the complex coefficients of $\mathbf{R}_n(\omega)$, i.e. $\mathbf{V}_{\mathbf{R}_n}^{m_1 m_2}$, are updated, which seems that we use NMF to update these complex coefficients, observing the complex coefficients of $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$, given the estimated coefficient matrix $\mathbf{W}_n$ and the basis matrix $\mathbf{U}_n$.

In practice, we do this to keep the scaling in equations (6.7) and (6.8) unchanged, otherwise a scaling ambiguity would be generated between the estimated diagonal and off-diagonal coefficients of $\mathbf{R}_n(\omega)$. Furthermore, the matrix update in (6.8) looks like factorization, but, it is only scaled estimation. On the other side, to maintain the diagonal representation of the tensor slice $\mathbf{V}_{\mathbf{R}_n}^{m}$ and the matrix $\mathbf{V}_{\mathbf{R}_n}^{m_1 m_2}$, they are initialized as matrices of zeros with diagonal entries of ones. Moreover, the matrix $\mathbf{W}_n$ is randomly initialized by values larger than zero.

## 6.2 Source-based prior information

As it was previously mentioned, the basis matrices, $\mathbf{U}_n, n = 1, .., N$, of the sources in the observed mixtures are assumed to be always known information. We propose that the matrices are either **extracted** in a separate step, or made available in advance in a pre-training step. In the second scenario, the matrices can be made either **directly** or **indirectly** available. In case that the matrices are indirectly available, we assume that a redundant **library** of trained spectral basis matrices is available. The library contains trained basis matrices of sources, both if they are in the observed mixtures and if they are not. Furthermore, we detect the basis matrices that best represent the sources in the mixtures.

As we will see, following the extraction scenario of the basis matrices, the proposed work is effective for BSS in mixing environments with low reverberation. However, the separation performance is improved using trained matrices in mixing environments with low and high reverberation.

### 6.2.1 Extraction of the prior information

To extract the matrix $\mathbf{U}_n$ observing the tensor $\tilde{\mathbf{V}}^M_{\mathbf{c}_n}$, we formulate a minimization function based on the $\beta$-divergence as follows

$$\{\tilde{\mathbf{U}}_n, \tilde{\mathbf{W}}^m_n\} = \arg \min_{\tilde{\mathbf{U}}_n, \tilde{\mathbf{W}}^m_n} \sum_m \sum_{\omega,l} d_\beta[\tilde{\mathbf{V}}^m_{\mathbf{c}_n} / \tilde{\mathbf{U}}_n \tilde{\mathbf{W}}^m_n], \tag{6.9}$$

where $\tilde{\mathbf{U}}_n$ is estimation of the spectral basis matrix of the $n$-th source. $\tilde{\mathbf{W}}^m_n$ denotes a time-varying activation coefficient slice corresponding to the observed slice $\tilde{\mathbf{V}}^m_{\mathbf{c}_n}$. In this formulation, the spatial diversity from one observed slice to another is represented by the matrices $\tilde{\mathbf{W}}^m_n, m = 1, .., M$, while the redundant information is kept in the matrix $\tilde{\mathbf{U}}_n$. As it was stated, the value of $\beta$ plays a role in controlling the sparsity of factorization. We build our idea to extract the matrix $\mathbf{U}_n$ on minimizing the above function by selecting a suitable value of $\beta$. In case that the value of $\beta$ is a positive large number, the low energy points of $\tilde{\mathbf{V}}^M_{\mathbf{c}_n}$ are weighted by small values, while a small value of $\beta$ increases the contribution of low energy points.

Accordingly, we can say that by assigning $\beta$ a suitable value, the high energy points belonging to the $n$-th source are only used to train the basis vectors of the matrix $\tilde{\mathbf{U}}_n$, while the other points associated with artifacts and reverberation are ignored. The matrix is extracted by applying the MU to minimize the $\beta$-divergence in (6.9) as (appendix A)

$$\tilde{\mathbf{U}}_n \leftarrow \tilde{\mathbf{U}}_n \circ \frac{\sum_m [\tilde{\mathbf{V}}^m_{\mathbf{c}_n} \circ (\tilde{\mathbf{U}}_n \tilde{\mathbf{W}}^m_n)^{\beta^e - 2}](\tilde{\mathbf{W}}^m_n)^T}{\sum_m (\tilde{\mathbf{U}}_n \tilde{\mathbf{W}}^m_n)^{\beta^e - 1}(\tilde{\mathbf{W}}^m_n)^T}, \tag{6.10}$$

$$\tilde{\mathbf{W}}^m_n \leftarrow \tilde{\mathbf{W}}^m_n \circ \frac{\tilde{\mathbf{U}}^T_n [\tilde{\mathbf{V}}^m_{\mathbf{c}_n} \circ (\tilde{\mathbf{U}}_n \tilde{\mathbf{W}}^m_n)^{\beta^e - 2}]}{\tilde{\mathbf{U}}^T_n (\tilde{\mathbf{U}}_n \tilde{\mathbf{W}}^m_n)^{\beta^e - 1}}. \tag{6.11}$$

$\beta^e$ denotes the value of $\beta$ used for extracting the matrix $\tilde{\mathbf{U}}_n$. Each column of the matrix $\tilde{\mathbf{U}}_n$ is normalized to sum up to 1, while iterating the above two steps of factorization. The slices $\tilde{\mathbf{W}}^m_n, m = 1, .., M$ are not needed anymore.

### 6.2.2 Training of the prior information

The matrix $\mathbf{U}_n$ can be pre-trained in advance on a separate set of training audio signals applying NMF. For clean training audio signals of the $n$-th source, the power spectra of the sources are concatenated in one matrix $\mathbf{V}_n^t$. As commonly applied, the minimization function based on the $\beta$-divergence is represented as follows

$$\{\mathbf{U}_n, \mathbf{W}_n^t\} = \arg \min_{\mathbf{U}_n, \mathbf{W}_n^t} \sum_{\omega,l} d_\beta[\mathbf{V}_n^t / \mathbf{U}_n \mathbf{W}_n^t], \tag{6.12}$$

The factorization of $\mathbf{V}_n^t$ is performed by minimizing the above function using the MU rule by alternating the following two steps (appendix A)

$$\mathbf{U}_n \leftarrow \mathbf{U}_n \circ \frac{[\mathbf{V}_n^t \circ (\mathbf{U}_n \mathbf{W}_n^t)^{\beta^t - 2}](\mathbf{W}_n^t)^T}{(\mathbf{U}_n \mathbf{W}_n^t)^{\beta^t - 1}(\mathbf{W}_n^t)^T}, \tag{6.13}$$

$$\mathbf{W}_n^t \leftarrow \mathbf{W}_n^t \circ \frac{\mathbf{U}_n^T[\mathbf{V}_n^t \circ (\mathbf{U}_n \mathbf{W}_n^t)^{\beta^t - 2}]}{\mathbf{U}_n^T(\mathbf{U}_n \mathbf{W}_n^t)^{\beta^t - 1}}. \tag{6.14}$$

$\beta^t$ denotes the value of $\beta$ used for training the matrix $\mathbf{U}_n$. The activation coefficient matrix $\mathbf{W}_n^t$ is not needed any more. In this scenario the identities of multiple speakers must be known in advance. Accordingly, the spectral basis matrices $\mathbf{U}_n, n = 1, .., N$, are predefined and fixed for all the speakers. Furthermore, the source order must be also known in advance.

To increase the flexibility of the proposed method, a redundant library of trained spectral basis matrices of all available sources can be built. Then the basis matrices matching the source signals in the observed mixtures, are detected. To constitute the library $\mathbf{U}_{lib}$ for a number $Z$ of source signals, where $Z > N$, the spectral basis matrices are trained and sequentially arranged side by side such as

$$\mathbf{U}_{lib} = [\mathbf{U}_1 \cdots |\mathbf{U}_z| \cdots |\mathbf{U}_Z]. \tag{6.15}$$

### 6.2.3 Detection of the matched prior information

As it was previously stated, the tensor $\tilde{\mathbf{V}}^M_{\mathbf{c}_n}$ approximately represents corrupted copies of the source power spectrum $\mathbf{V}_n$ weighted by propagation inter-channel intensities. Using NMF, the spectrum can be decomposed as $\mathbf{V}_n = \mathbf{U}_n\mathbf{W}_n$. The factorization can be expanded by using the library and involving a diagonal matrix $\mathbf{D}_{lib}$ as $\mathbf{V}_n = \mathbf{U}_{lib}\mathbf{D}_{lib}\mathbf{W}_{lib}$. The coefficients of the diagonal matrix $\mathbf{D}_{lib}$ can be seen to define the contribution of each spectral basis vector of the library $\mathbf{U}_{lib}$. Assuming that the matrix $\mathbf{U}_n$ is included in the library $\mathbf{U}_{lib}$, the coefficient of $\mathbf{D}_{lib}$ that are associated with the matrix $\mathbf{U}_n$, will have the largest values over all the other coefficients. Accordingly, the matrix $\mathbf{U}_n$ that best represents the source power spectrum $\mathbf{V}_n$ can be identified in the library $\mathbf{U}_{lib}$ by observing the diagonal coefficients of the matrix $\mathbf{D}_{lib}$.

This idea can be extended to the case of multiple observations. However, in this work, the tensor of multiple observations $\tilde{\mathbf{V}}^M_{\mathbf{c}_n}$ is composed by weighted and corrupted versions of the source power spectrum $\mathbf{V}_n$. The weights are inter-channel intensities, and the corruption is due to residual from other source signals, reverberant environments or additive noise. As a result, to perform successful detection, an efficient factorization algorithm is required that tries to compensate the impact of the weights and the corruption. To detect the matched matrices, we recall the conventional Nonnegative Tensor Factorization (NTF) formulation. Observing the tensor $\tilde{\mathbf{V}}^M_{\mathbf{c}_n}$, the minimization function based on the $\beta$-divergence is represented as follows

$$\{\mathbf{D}^m_{lib}, \mathbf{W}_{lib}\} = \arg\min_{\mathbf{D}^m_{lib}, \mathbf{W}_{lib}} \sum_m \sum_{\omega,l} d_\beta[\tilde{\mathbf{V}}^m_{\mathbf{c}_n} / \mathbf{U}_{lib}\mathbf{D}^m_{lib}\mathbf{W}_{lib}], \tag{6.16}$$

where the $m$-th slice $\mathbf{D}^m_{lib}$ is a diagonal matrix of size $ZK \times ZK$ of a $ZK \times ZK \times M$ tensor $\mathbf{D}^M_{lib}$. Given the library $\mathbf{U}_{lib}$, the matrix $\mathbf{D}^m_{lib}$ and the

activation coefficient matrix $\mathbf{W}_{lib}$ are obtained by applying the MU rule to minimize the $\beta$-divergence in (6.16) as follows (appendix A)

$$\mathbf{D}_{lib}^{m} \leftarrow \mathbf{D}_{lib}^{m} \circ \frac{\mathbf{U}_{lib}^{T}[\tilde{\mathbf{V}}_{\mathbf{c}_n}^{m} \circ (\mathbf{U}_{lib}\mathbf{D}_{lib}^{m}\mathbf{W}_{lib})^{\beta^d-2}]\mathbf{W}_{lib}^{T}}{\mathbf{U}_{n}^{T}(\mathbf{U}_{lib}\mathbf{D}_{lib}^{m}\mathbf{W}_{lib})^{\beta^d-1}\mathbf{W}_{lib}^{T}}, \tag{6.17}$$

$$\mathbf{W}_{lib} \leftarrow \mathbf{W}_{lib} \circ \frac{\sum_m(\mathbf{U}_{lib}\mathbf{D}_{lib}^{m})^{T}[\tilde{\mathbf{V}}_{\mathbf{c}_n}^{m} \circ (\mathbf{U}_{lib}\mathbf{D}_{lib}^{m}\mathbf{W}_{lib})^{\beta^d-2}]}{\sum_m(\mathbf{U}_{lib}\mathbf{D}_{lib}^{m})^{T}(\mathbf{U}_{lib}\mathbf{D}_{lib}^{m}\mathbf{W}_{lib})^{\beta^d-1}}. \tag{6.18}$$

$\beta^d$ is the value of $\beta$ used to detect the matched spectral basis matrices. We can detect the matched spectral basis matrix $\mathbf{U}_z$ that best represents the $n$-th source spectral basis matrix $\mathbf{U}_n$, observing the tensor $\mathbf{D}_{lib}^{M}$. We start by averaging along the elements of the diagonal slices of the tensor $\mathbf{D}_{lib}^{M}$, converting the tensor into a vector $\mathbf{d}$ of size $ZK$, whose entries define the average contribution of each spectral vector in multiple observations

$$\mathbf{d} = \frac{1}{M}\sum_{m=1}^{M} \mathrm{diag}(\mathbf{D}_{lib}^{m}). \tag{6.19}$$

The vector $\mathbf{d}$ is divided into $Z$ sub-vectors $\mathbf{d}_z(k)$, each one is associated with a spectral basis matrix $\mathbf{U}_z$, and defines the contribution of each basis vector in $\mathbf{U}_z$. To detect the optimal basis matrix $\mathbf{U}_{z^*}$ that best represents the $n$-th source signal, the index $z^*$ of the optimal matrix is selected as follows

$$z^* = \arg\max_{z} \sum_{k} \mathbf{d}_z(k), \quad z = 1, 2, ..., Z \tag{6.20}$$

Observing the tensors $\tilde{\mathbf{V}}_{\mathbf{c}_n}^{M}$, $n = 1, ..., N$, we detect $N$ optimal spectral basis matrices. This proposed detection algorithm alternates with the separation one, in order to correct wrong detections that may have occurred in a previous iteration. The wrong detections may occur because of the residual from the other source signals in the observed mixtures, and the coherence between the trained spectral basis matrices in the library.

## 6.3 Full description

The full algorithm is summarized as follows

---

**Training**: $\mathbf{U}_{lib}$ as in Section 6.2.2

**Input**: $\mathbf{x}(\omega, l)$

**Initialize**: $\tilde{\mathbf{c}}_n(\omega, l)$ as in Section 4.1.7, $\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l) = \mathbf{I}$

**Iterate**: *till convergence*

Compute $\tilde{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ as in (4.1)

Build the tensor $\tilde{\mathbf{V}}^M_{\mathbf{c}_n}$ and the matrix $\tilde{\mathbf{V}}^{m_1 m_2}_{\mathbf{c}_n}$ as in Section 6.1.1

**Extraction or Detection**:

Using $\tilde{\mathbf{V}}^M_{\mathbf{c}_n}$, Extract $\mathbf{U}_n$, $n = 1, ..., N$ as in Section 6.2.1

Using $\tilde{\mathbf{V}}^M_{\mathbf{c}_n}$, Detect $\mathbf{U}_n$, $n = 1, ..., N$ as in Section 6.2.3

**Estimation**:

Fixing $\mathbf{U}_n$, Factorize $\tilde{\mathbf{V}}^M_{\mathbf{c}_n}$:

to estimate $\theta_n^{diag} = \{\mathbf{W}_n, \mathbf{V}^m_{\mathbf{R}_n}\}$ as in (6.6) and (6.7)

Fixing $\mathbf{U}_n$ and $\mathbf{W}_n$, Factorize $\tilde{\mathbf{V}}^{m_1 m_2}_{\mathbf{c}_n}$:

to estimate $\theta_n^{off} = \{\mathbf{V}^{m_1 m_2}_{\mathbf{R}_n}\}$ as in (6.8)

**Separation**:

Rearrange $\mathbf{V}^m_{\mathbf{R}_n}$ and $\mathbf{V}^{m_1 m_2}_{\mathbf{R}_n}$ in $\mathbf{R}_n(\omega)$

Compute $v_n(\omega, l) = \mathbf{u}_n^T(\omega)\mathbf{w}_n(l)$

$\mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l) = v_n(\omega, l)\mathbf{R}_n(\omega)$

$\mathbf{G}_n(\omega, l) = \mathbf{\Sigma}_{\mathbf{c}_n}(\omega, l)\mathbf{\Sigma}^{-1}_{\mathbf{x}}(\omega, l)$

$\tilde{\mathbf{c}}_n(\omega, l) = \mathbf{G}_n(\omega, l)\mathbf{x}(\omega, l)$

**Return**

**Output**: $\tilde{\mathbf{c}}_n(\omega, l)$

---

## 6.4 Experiments

The experiments were carried out to investigate the effect of the number of spectral basis vectors $K$, the training divergence factor $\beta^t$, the detection divergence factor $\beta^d$, the extraction divergence factor $\beta^e$, and the estimation divergence factor $\beta^s$. Three different datasets were used for this experimental evaluation, including simulated and live-recorded data.

- The first one is a simulated dataset that was used to analyse the detection algorithm, as well as to identify the separation performance using trained spectral basis matrices.
- To evaluate the separation performance in real mixing environments, using the trained basis matrices, a second live-recorded dataset was recorded in an acoustically insulated room.
- The third dataset consists of simulated and live-recorded data from the SISEC evaluation campaign. This was used as a reference dataset to identify the performance of the algorithm in blind and informed scenarios, where the extraction algorithm proposed in section 6.2.1 is adopted to extract the spectral basis matrices in order to perform BSS. This was also used to assess the performance that can be obtained in the informed case, where the matrix $\mathbf{U}_n$ associated with each source signal is available.

For the multichannel underdetermined source separation problem, $M = 2$ observed mixtures and $N = 3$ speech signals were used. A smoothing factor $\mu = 0.1$ was adopted in (3.12). The discrete time-frequency representation of the observed mixtures $\mathbf{x}(\omega, l)$ was obtained through STFT using a Hanning analysis window with length of 128 ms (i.e. 2048 samples at 16 kHz sampling rate) and shift of 64 ms. The bi-dimensional window $\gamma$ for the computation of the empirical covariance matrix $\hat{\mathbf{R}}_{\mathbf{c}_n}(\omega, l)$ in (4.2)

is a Hanning window of size $3 \times 3$. The algorithm was iterated till convergence, which is achieved in less than 50 iterations, while the decomposition algorithms were iterated 100 times.

### 6.4.1 Simulated scenario

A room of size $4.45 \times 3.35 \times 2.5$ meters and an array of 2 omnidirectional microphones spaced of 0.2 m are considered. The array is located in the middle of the room and it has the same height (1.4 m) as the three sources. The distance between source positions and the central point between the microphones was randomly chosen between 0.8 and 1.2 m, with several source direction of arrivals (DOAs). The minimum angular distance between two neighboring sources is 25 degrees, and the maximum is 40 degrees. Synthetic room impulse responses (RIRs) are simulated through ISM [54] with a sampling frequency of 16 kHz for three reverberation times: $T_{60} = 130,\ 250$, or 380 ms. Six native Italian speakers are considered as our audio sources, 3 males and 3 females. For each speaker, 20 clean speech signals with average lengths of 8.75 s were produced. For each speaker, the signals are divided into 5 signals for testing data and 15 signals used to train the spectral basis matrices $\mathbf{U}_z$, $z = 1, ..., 6$. Six male-female combinations of mixtures were generated. This resulted in a total of 30 test observed mixtures for each reverberation time ($T_{60}$).

#### Analysis of the detection algorithm

To build the redundant library $\mathbf{U}_{lib}$ in (6.15) for $Z = 6$ trained spectral basis matrices, the power spectra of the training signals were computed and concatenated in the matrix $\mathbf{V}_z^t$, $z = 1, 2, ..., 6$. Applying NMF, each matrix was factorized with $K = 15$. The training divergence factor $\beta^t$ in (6.13) and (6.14) was assigned a value of 0.9, quite close to the KL divergence. The $z$-th spectral basis matrix $\mathbf{U}_z$ of size $1025 \times 15$ was obtained, and

integrated into the library $\mathbf{U}_{lib}$ of 6 trained spectral basis matrices of total size $1025 \times 90$. Let us now consider one specific case of mixing conditions, on source-to-microphone distance of 1 m, and in a mixing environment with reverberation time ($T_{60}$) of 250 ms, in which mixtures of two male and one female speech signals were generated. In the specific example here considered the indexes of the basis matrices, chosen from the library and involved in the mixtures, are $\{2, 4, 6\}$.

The detection divergence factor used in (6.17) and (6.18) was $\beta^d = 0.3$. Moreover, the separation divergence factor in (6.6), (6.7) and (6.8) was $\beta^s = 0.3$. The contribution of each spectral basis vector and spectral basis matrix was computed at each iteration of the separation phase as in (6.19) and (6.20). As it can be observed in Figures 6.4 and 6.5, the normalized likelihood of each spectral basis vector and matrix associated with a certain target source in the observed mixtures increases while iterating the separation algorithm; therefore, the optimal index of each spectral basis matrix becomes more identifiable with respect to the other indexes.

Applying the proposed detection algorithm on all mixtures under test, we observed that, as in the separation process, the accuracy of the algorithm depends on the value of $\beta^d$, the configuration of the mixing process and the construction of the spectral basis matrices. Low values of $\beta^d$ perform better than large ones, especially in mixing environments with high reverberation. If the mixing process involves low reverberation, the algorithm works with very high efficiency. However, if the mixing environment is highly reverberant, wrong detection may happen, especially if there is much residual from the other source signals, and if there is high redundant correlation between the pre-trained spectral basis matrices.

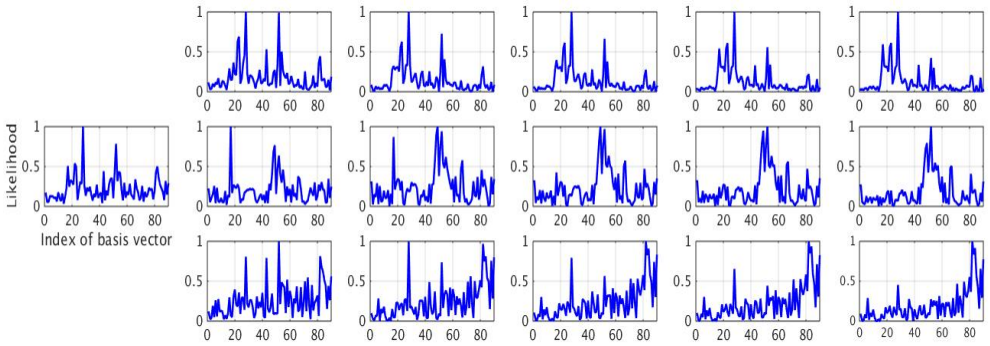


Figure 6.4: Normalized Likelihoods of each basis vector in $\mathbf{U}_{lib}$ as a function of separation iterations. The first graph on the left represents the mixtures, then columns from left to right correspond to iterations, while the 3 rows refer to each one of the 3 sources.

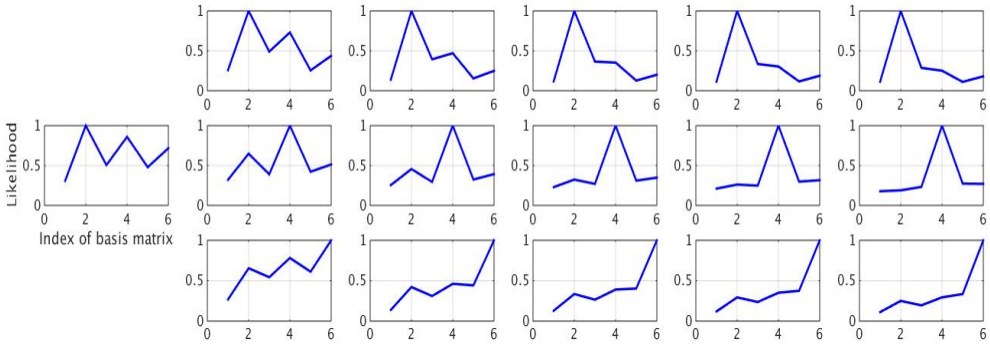


Figure 6.5: Normalized Likelihoods of each basis matrix $\mathbf{U}_z$ as a function of separation iterations. The first graph on the left represents the mixtures, then columns from left to right correspond to iterations, while the 3 rows refer to each one of the 3 sources.

**Source separation**

The average separation performance measurement (SDR) as a function of $T_{60}$, $K$, and $\beta^s$, with fixed $\beta^t = 0.9$, is reported in Table 6.1. At low reverberation when $T_{60} = 130$ ms, the best separation performance is obtained when the value of $K$ is large (between 25 and 40) and $\beta^s$ is assigned moderate values (between 0.3 and 0.6). On the other side, a good performance is achieved in high reverberation ($T_{60} = 380$ ms), when $K$ is small (between 15 and 25) and $\beta^s$ is also assigned small values (between 0.1 and

Table 6.1: Average SDR (dB) of the simulated scenario as a function of $K$, $T_{60}$ and $\beta^s$, $\beta^t = 0.9$.

| $T_{60}$ $(ms)$ | 130 | | | 250 | | | 380 | | | Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | 15 | 25 | 40 | 15 | 25 | 40 | 15 | 25 | 40 | 15 | 25 | 40 |
| $\beta^s = 0.9$ | 8.53 | 9.18 | 9.04 | 5.50 | 5.61 | 5.32 | 3.58 | 3.46 | 3.23 | 5.87 | 6.08 | 5.86 |
| $\beta^s = 0.6$ | 9.03 | 9.71 | 9.64 | 6.57 | 6.51 | 5.92 | 4.73 | 4.48 | 4.08 | 6.78 | 6.90 | 6.55 |
| $\beta^s = 0.3$ | 8.64 | 9.26 | 9.66 | 6.60 | 6.85 | 6.84 | 5.02 | 4.95 | 4.44 | 6.75 | 7.02 | 6.98 |
| $\beta^s = 0.1$ | 7.88 | 8.48 | 9.09 | 6.26 | 6.52 | 6.96 | 4.63 | 4.67 | 4.81 | 6.26 | 6.56 | 6.95 |

0.3). On the average, the best performance is obtained when $\beta^s$ is in the range between 0.3 and 0.6, but this range also depends on the value of the adopted $\beta^t$. In general, the experiments show the importance of choosing proper values of $\beta$ for training the basis matrices $\mathbf{U}_n$ and estimating the model parameters $\theta$ to guarantee an effective performance.

### 6.4.2 Live-recorded dataset

A room of size $6.5 \times 3 \times 2.2$ meters and 2 omnidirectional microphones spaced 16 cm are considered. The microphones are located close to the center of one of the room walls and has the same height (1.5 m) as 3 loudspeakers considered as speech sources. The distance between source positions and the central point between the microphones is 1.5 m, with DOAs at 55, 90 and 125 degrees. The measured reverberation time of the room is about 220 ms. The mixtures were recorded at sampling frequency of 16 kHz. The same speech source signals and trained spectral basis matrices, used in the simulated scenario in section 6.4.1, were used for this evaluation on the live-recorded data. Table 6.2 shows the corresponding results as a function of $K$ and $\beta^s$, with fixed $\beta^t = 0.9$. The performance is close to what obtained in the simulated experiments at $T_{60} = 250$ ms. The best separation performance is obtained when $K$ is moderately large (between 20 and 40), and $\beta^s$ is assigned small values (between 0.1 and 0.3)

Table 6.2: Average SDR (dB) of the live-recorded dataset as a function of $K$ and $\beta^s$, $\beta^t = 0.9$.

| $K$ | 10 | 20 | 30 | 40 |
|---|---|---|---|---|
| $\beta^s = 0.9$ | 4.52 | 4.82 | 5.10 | 5.43 |
| $\beta^s = 0.6$ | 5.60 | 5.70 | 5.79 | 5.88 |
| $\beta^s = 0.3$ | 5.72 | 6.62 | 6.78 | 6.62 |
| $\beta^s = 0.1$ | 5.67 | 6.46 | 6.42 | 6.43 |

### 6.4.3 Dataset of SISEC

The development dataset *dev*1 of SISEC (under-determined speech and music mixtures) was used to further assess the performance of the proposed algorithm. The dataset consists of 4 synthetic convolutive and 4 live-recorded stereo mixtures of 3 Japanese and English speech signals. All the mixtures are 10 s long sampled at 16 kHz. The synthetic convolutive filters are generated with the Roomsim toolbox [20]. They simulate 2 omnidirectional microphones placed 1 m apart in a room of dimension $4.45 \times 3.35 \times 2.5$ with reverberation times 130 and 250 ms, which corresponds to the setting employed for live-recorded mixtures. The distance between source positions and the central point between the microphones varies between 0.8 and 1.2 m. For all mixtures the DOAs vary between 60 and 300 angular degrees, with a minimal spacing of 15 degrees. This dataset was used to evaluate the proposed algorithm in a blind scenario, where the extraction algorithm of the spectral basis matrices $\mathbf{U}_n, n = 1, .., N$, proposed in section 6.2.1 is adopted, as well as in an informed scenario, where the true spectral basis matrices are available.

On the available 8 observed mixtures (male and female synthetic convolutive and live-recorded with two different reverberation times of 130 and 250 ms) and in an informed case, we computed the basis matrices

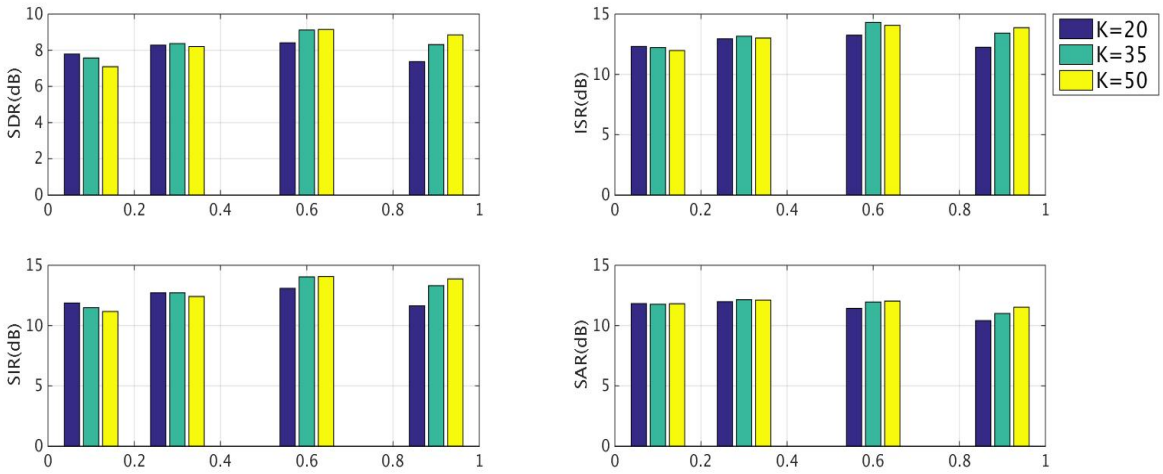


Figure 6.6: The average separation performance of the informed case of the SISEC dataset as a function of $K$ and $\beta^s$. $\beta^t = 0.9$. The horizontal axis indicates the value of the tested $\beta^s$=[0.1 0.3 0.6 0.9].

$\mathbf{U}_n$ of the available source signals. The matrices were with different sizes $K = [20, 35, 50]$, and were trained with $\beta^t = 0.9$. For the separation, the selected values of $\beta^s$ were again $0.1, 0.3, 0.6$, or $0.9$. Fig. 6.6 shows the average separation performance, for all values of $K$. The best performance is obtained when the value of $\beta^s$ equals 0.6. On the average, large values of $K$ between 35 and 50 benefit the separation performance when $\beta^s$ is assigned large values also (between 0.6 and 0.9). While when $\beta^s$ is low (between 0.1 and 0.3), it is preferable to adopt small values of $K$.

Table 6.3 shows an example of the detailed performance measurements of informed and blind source separation. In this experiment, only the live-recorded stereo mixture of 3-female speech signals, with reverberation time $T_{60} = 130$ ms, was used to evaluate the performance. The matrix $\mathbf{U}_n$ is either assumed to be pre-known with $K = 35$, or to be extracted with $K = 50$, the best performing sizes of $\mathbf{U}_n$. For the extraction, the matrix $\mathbf{U}_n$ is updated as proposed in Section 6.2.1. On the average, with the prior information, we gain around 1.5 dBs of SDR compared to the blind case.

Table 6.3: Detailed performance of informed and blind separation of live-recorded mixtures of three females from SISEC, $T_{60} = 130$ ms.

| $\mathbf{U}_n$ | Informed | | | Blind | | |
|---|---|---|---|---|---|---|
| Div. Factors | $\beta^t = 0.9, \beta^s = 0.3$ | | | $\beta^e = 0.6, \beta^s = 0.6$ | | |
| Source images | $\tilde{\mathbf{c}}_1$ | $\tilde{\mathbf{c}}_2$ | $\tilde{\mathbf{c}}_3$ | $\tilde{\mathbf{c}}_1$ | $\tilde{\mathbf{c}}_2$ | $\tilde{\mathbf{c}}_3$ |
| SDR | 11.27 | 10.42 | 10.22 | 9.63 | 8.67 | 9.17 |
| ISR | 15.13 | 14.56 | 17.76 | 12.90 | 12.82 | 16.50 |
| SIR | 17.26 | 18.02 | 13.12 | 15.28 | 13.97 | 11.95 |
| SAR | 14.99 | 12.99 | 13.68 | 14.32 | 12.33 | 12.76 |

**Blind source separation**

We evaluated the performance in a blind scenario using 4 observed mixtures (2 synthetic convolutive and 2 live-recorded, all with reverberation time of 130 ms). The matrix $\mathbf{U}_n$ is extracted as proposed in Section 6.2.1. Tables 6.4 and 6.5 report the average SDR in terms of $K$, $\beta^e$, and $\beta^s$. For the synthetic convolutive mixtures, when $K$ is small, moderate values of $\beta^s$ (between 0.3 and 0.6) perform good, if $\beta^e$ is assigned moderate values (between 0.6 and 0.9). By increasing $K$ and keeping $\beta^e$ in its moderate range, a good performance is obtained when $\beta^s$ is assigned small values (between 0.1 and 0.3). The best average SDR (7.07 dBs) is achieved when $K = 35$, $\beta^e = 0.6$ and $\beta^s = 0.1$.

On the other hand, for the live-recorded mixtures, large values for both $\beta^e$ and $\beta^s$ perform better than small ones, when $K = 20$. Better performance is achieved for large values of $\beta^s$ when $\beta^e = 0.6$ and $K = 35$. Furthermore, the best performance is obtained when $K = 50$, $\beta^e$ between 0.6 and 0.9, and $\beta^s$ in its moderate range (between 0.3 and 0.6). To conclude, in mixing environments with low reverberation ($T_{60} = 130$ ms), moderately large values of $K$ and moderate values of $\beta^s$ benefit the separation performance not only when the separation system is fed by the basis matrices as in the previous subsections, but also in the blind scenario.

Table 6.4: Average SDR of blind separation of mixtures from SISEC, $T_{60} = 130$ ms.

| Env. | Synthetic convolutive mixtures | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $K$ | 20 | | | 35 | | | 50 | | |
| $\beta^e$ | 1.2 | 0.9 | 0.6 | 1.2 | 0.9 | 0.6 | 1.2 | 0.9 | 0.6 |
| $\beta^s = 0.9$ | 5.53 | 5.97 | 6.00 | 5.70 | 5.79 | 5.63 | 5.78 | 5.80 | 5.68 |
| $\beta^s = 0.6$ | 5.92 | 6.32 | 6.49 | 5.71 | 6.22 | 5.97 | 6.18 | 6.17 | 5.76 |
| $\beta^s = 0.3$ | 5.51 | 6.79 | 6.75 | 5.66 | 6.01 | 6.81 | 5.71 | 6.90 | 6.08 |
| $\beta^s = 0.1$ | 5.07 | 5.55 | 5.75 | 5.44 | 6.50 | 7.07 | 5.34 | 6.36 | 6.20 |

Table 6.5: Average SDR of blind separation of mixtures from SISEC, $T_{60} = 130$ ms.

| Env. | Live-recorded mixtures | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $K$ | 20 | | | 35 | | | 50 | | |
| $\beta^e$ | 1.2 | 0.9 | 0.6 | 1.2 | 0.9 | 0.6 | 1.2 | 0.9 | 0.6 |
| $\beta^s = 0.9$ | 7.46 | 7.54 | 5.83 | 7.74 | 7.58 | 7.95 | 7.55 | 7.62 | 7.80 |
| $\beta^s = 0.6$ | 7.49 | 7.27 | 6.98 | 7.66 | 7.48 | 7.76 | 7.61 | 7.90 | 8.16 |
| $\beta^s = 0.3$ | 6.07 | 7.06 | 6.76 | 6.31 | 7.20 | 7.31 | 7.23 | 7.62 | 7.98 |
| $\beta^s = 0.1$ | 5.26 | 5.59 | 5.41 | 5.48 | 6.76 | 6.74 | 6.26 | 7.10 | 7.11 |

**Performance comparison**

Using the live-recorded dataset, we compared the performance of the proposed method in informed and blind scenarios with the method proposed in the previous chapter and two recently developed blind source separation methods [22, 67]. Tables 6.6, 6.7, 6.8 and 6.9 show the comparison results of the four methods, denoted as "Proposed III", "Proposed II","Nesta" [67], and "Cho" [22]. For the informed case, the spectral basis matrices $\mathbf{U}_n$ are trained as proposed in Section 6.2.2. The matrices are extracted as proposed in Section 6.2.1 for blind source separation.

Over all the mixtures under test, the prior information benefits the performance, the informed case of the proposed method in this chapter performs better than the informed case of the proposed method in the previous chapter. In the informed case, moderate values of $K$ (around 35)

and moderate small values of $\beta^s$ (around 0.3), when $\beta^t = 0.9$, perform the best for mixtures of female voices. However, large size pre-trained prior information ($K = 50$) is recommended for mixtures of male voices, and in this case $\beta^s$ should be assigned moderate large values (around 0.6).

In the blind case, in mixing environments with low reverberation ($T_{60} = 130\ ms$), the proposed method performs better than the methods "Nesta" and "Cho", in both cases of mixtures of female voices and mixtures of male voices. However, the method "Proposed II" outperforms the proposed method in case of mixtures of female voices. In mixing environments with moderate reverberation ($T_{60} = 250\ ms$) and in case of mixtures of female voices, the method "Proposed II" performs the best, moreover, the proposed method outperforms the other two methods. However, due to the weak sparsity and to the overlap at low frequency bands, the proposed methods "Proposed II" and "Proposed III" fail for mixtures of male voices.

For the extraction, in mixing environments with low reverberation, large values of $\beta^e$ (around 1.2) and moderate large values of $K$ (around 20) perform the best for mixtures of male voices. However, it is better to adopt large values of $K$ (around 50) and moderate values of $\beta^e$ (around 0.6) for mixtures of female voices. In mixing environments with moderate reverberation, the best performance is obtained when $K$ is moderately large (around 35) and $\beta^e$ is large (between 0.9 and 1.2) for both cases of male and female voices. On the other hand, values of $\beta^s$ in the interval between 0.3 and 0.6 still perform the best for all mixing conditions and separation strategies.

## 6.5 Conclusion

In this work, we tackled the problem of underdetermined audio source separation in reverberant environments. The proposed work adopts local

Table 6.6: Performance comparison of blind separation of live-recorded stereo mixtures of three female speech signals from SISEC, $T_{60} = 130$ ms.

| Method | Proposed III | | Proposed II | | Nesta [67] | Cho [22] |
|---|---|---|---|---|---|---|
| $\mathbf{U}_n$ | Inf. | Ext. | Inf. | Ext. | | |
| $K$ | 35 | 50 | 50 | 25 | | |
| Train/Extract | $\beta^t$=0.9 | $\beta^e$=0.6 | $\beta^t$=0.5 | $\beta^1$=0.5 | | |
| Estimation | $\beta^s$=0.3 | $\beta^s$=0.6 | $\beta^1$=0.5,$\beta^2$=0.5 | $\beta^1$=0.5,$\beta^2$=0.5 | | |
| SDR | 10.70 | 9.20 | 10.10 | 9.70 | 7.70 | 8.40 |
| ISR | 15.80 | 14.10 | 15.00 | 14.90 | 10.50 | 13.00 |
| SIR | 16.10 | 13.70 | 16.00 | 15.70 | 13.30 | 12.60 |
| SAR | 14.00 | 13.10 | 13.40 | 12.60 | 11.80 | 12.10 |

Table 6.7: Performance comparison of blind separation of live-recorded stereo mixtures of three male speech signals from SISEC, $T_{60} = 130$ ms.

| Method | Proposed III | | Proposed II | | Nesta [67] | Cho [22] |
|---|---|---|---|---|---|---|
| $\mathbf{U}_n$ | Inf. | Ext. | Inf. | Ext. | | |
| $K$ | 50 | 20 | 50 | 50 | | |
| Train/Extract | $\beta^t$=0.9 | $\beta^e$=1.2 | $\beta^t$=0.9 | $\beta^1$=0.9 | | |
| Estimation | $\beta^s$=0.6 | $\beta^s$=0.6 | $\beta^1$=0.5,$\beta^2$=0.9 | $\beta^1$=0.9,$\beta^2$=0.5 | | |
| SDR | 9.10 | 7.20 | 7.70 | 6.95 | 6.50 | 6.50 |
| ISR | 14.10 | 12.20 | 12.40 | 12.10 | 9.30 | 11.40 |
| SIR | 13.80 | 11.60 | 13.20 | 12.00 | 10.90 | 10.00 |
| SAR | 12.20 | 10.10 | 10.30 | 9.00 | 9.60 | 10.50 |

Table 6.8: Performance comparison of blind separation of live-recorded stereo mixtures of three female speech signals from SISEC, $T_{60} = 250$ ms.

| Method | Proposed III | | Proposed II | | | |
|---|---|---|---|---|---|---|
| $\mathbf{U}_n$ | Inf. | Ext. | Inf. | Ext. | | |
| $K$ | 35 | 15 | 50 | 15 | Nesta | Cho |
| Train/Extract | $\beta^t$=0.9 | $\beta^e$=0.9 | $\beta^t$=0.5 | $\beta^1$=0.5 | [67] | [22] |
| Estimation | $\beta^s$=0.3 | $\beta^s$=0.3 | $\beta^1$=0.9, $\beta^2$=0.9 | $\beta^1$=0.5,$\beta^2$=0.9 | | |
| SDR | 9.40 | 6.40 | 9.30 | 6.80 | 6.00 | 6.10 |
| ISR | 14.40 | 11.20 | 13.90 | 11.00 | 8.90 | 10.90 |
| SIR | 14.10 | 10.60 | 14.30 | 10.80 | 10.60 | 9.00 |
| SAR | 12.40 | 9.90 | 12.20 | 10.80 | 8.70 | 10.00 |

Table 6.9: Performance comparison of blind separation of live-recorded stereo mixtures of three male speech signals from SISEC, $T_{60} = 250$ ms.

| Method | Proposed III | | Proposed II | | | |
|---|---|---|---|---|---|---|
| $\mathbf{U}_n$ | Inf. | Ext. | Inf. | Ext. | | |
| $K$ | 50 | 35 | 50 | 50 | Nesta | Cho |
| Train/Extract | $\beta^t$=0.9 | $\beta^e$=1.2 | $\beta^t$=0.9 | $\beta^1$=0.5 | [67] | [22] |
| Estimation | $\beta^s$=0.6 | $\beta^s$=0.6 | $\beta^1$=0.5,$\beta^2$=0.9 | $\beta^1$=0.5,$\beta^2$=0.9 | | |
| SDR | 8.30 | 5.00 | 7.10 | 5.80 | 5.20 | 6.00 |
| ISR | 13.10 | 9.49 | 11.50 | 10.54 | 8.30 | 10.50 |
| SIR | 12.70 | 8.36 | 11.60 | 10.40 | 9.00 | 9.10 |
| SAR | 11.10 | 8.23 | 9.70 | 7.80 | 8.00 | 9.20 |

Gaussian modeling of the mixing process. The work describes a new estimation algorithm of the parameters of the model by applying nonnegative tensor/matrix factorization, given source-based prior information. Following this direction, the parameters are jointly estimated, and the related artifacts are consequently reduced. To perform the estimation, spectral basis matrices of power spectra of source signals in observed mixtures are assumed to be available as prior information. In a separate step, the basis matrices are either extracted or detected. Using nonnegative tensor factorization, we propose a new method for extracting the matrices, and in this case the algorithm fully works in a blind scenario. However, to obtain a better separation performance, in the other case the matrices are made indirectly available through a pre-trained redundant library of spectral basis matrices. Furthermore, using nonnegative tensor factorization, we propose a new method to detect the basis matrices that best represent the power spectra of the source signals in the observed mixtures.

For each of the training, detection, extraction and estimation phases, the factorization is performed using the $\beta$-divergence and applying the widely used multiplicative update rules. By tuning the value of $\beta$, we can govern the sparsity of factorization. Controlling the sparsity is an important issue because it is known that the speech signals are sparse in their nature, and so we can minimize any residual artifacts. Accordingly, we tested several values of $\beta$ for each task in order to identify the best performing ones. Experiments show that the choice of $\beta$ is a really critical aspect. We found that the best choice of $\beta$ to estimate the parameters, is in the interval between 0.3 and 0.6, in both cases of extracting the basis matrices and of detecting the trained ones. On the other hand, to extract the basis matrices, we found that for mixtures of female voices, the best choice of $\beta$ is between 0.6 and 0.9. However large values ($> 0.9$) perform better for mixtures of male voices.

The experimental results show that the proposed method can work with approximately the same efficiency in both cases of simulated and real environments. In the informed case, the proposed algorithm perform better than the proposed one in the previous chapter. In the blind case, the proposed method outperforms two of the recently proposed blind source separation algorithms, and provides comparable results to the one proposed in the previous chapter.

# SECOND PART

Exploiting spatial information about propagation channels used to generate observed mixtures, for source separation and speaker diarization, by applying the sparse modeling theory.

# Chapter 7

# Sparse source separation

Sparse modeling basically exploits a prior assumption that audio source signals are sparse in their nature in a known domain such as the time-frequency domain. In this work, a dictionary composed of the mixing parameters is built using Room Impulse Responses (RIRs) between multiple points of the space and an array of microphones. For example, the planar area of a room can be sampled into a finite set of points and the RIRs are measured and then arranged in the dictionary. Exploiting the sparse nature of audio signals, given the dictionary and observed mixtures of audio signals, applying a sparse modeling algorithm, we propose to detect trained mixing parameters that match the real parameters which characterize the observed mixtures [34]. Furthermore, the detected parameters are exploited to perform source separation in order to retrieve the original signals in the mixtures applying $l_0$-norm minimization [80] (see Figure 7.1).

To do this we propose and analyze an efficient greedy algorithm based on the Orthogonal Matching Pursuit (OMP) [61, 75] and focus on the mismatch between ideal and non-ideal mixing conditions. It is shown that the overall detection capability is considerably degraded if the time-frequency sparseness condition of speech signals is not ideally fulfilled or when there is a mismatch between the real mixing parameters and the trained ones.

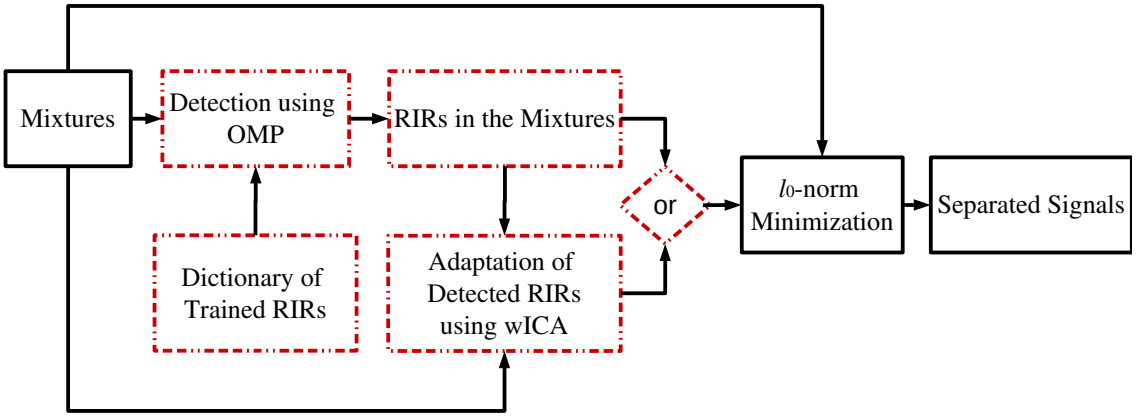


Figure 7.1: Flowchart of the proposed method. Highlighted blocks refer to novel contributions.

An efficient normalization strategy is applied which improves the work of the proposed MP and consequently the overall source detection and separation performance. Although the effect of this mismatch can be mitigated by the normalization, still its effect is crucial and it should be reduced as much as possible. In response to this need, dictionary adaptation with the observed data is a possible viable solution. In this study we fuse the concept of model-based spatial dictionary and blind mixing system estimation in a single framework, through the effective combination of sparse modeling and Independent Component Analysis (ICA) [66]. Unsupervised ICA based on the weighted Natural Gradient is exploited to adapt the original dictionary with the observed data. The ICA adaptation is based on the assumption of sparse spatio-temporal representation of the acoustic sources.

## 7.1 Observations and dictionary representation

Assuming that the source signals are sparse in the STFT domain and each time-frequency point is dominated by one source, the observed mixtures $\mathbf{x}(\omega, l)$ can be approximately represented as

$$\begin{aligned}\mathbf{x}(\omega, l) &= [x_1(\omega, l), \cdots, x_M(\omega, l)]^T \\ &\approx [h_{1,n(\omega,l)}(\omega), \cdots, h_{M,n(\omega,l)}(\omega)]^T s_{n(\omega,l)}(\omega, l),\end{aligned} \tag{7.1}$$

where $n(\omega, l)$ is the index of the most dominant source. Under this assumption, a convenient representation of the source signals can be obtained by computing the cross correlation between the $m$-th observed mixture at the $m$-th microphone and an observed mixture at a reference microphone, e.g. $m = 1$. Therefore, the $n$-th source dominance is approximated as

$$\begin{aligned}r_m(\omega, l) &= x_m(\omega, l) x_1(\omega, l)^H \\ &\approx h_{m,n(\omega,l)}(\omega) s_n(\omega, l) s_n(\omega, l)^H h_{1,n(\omega,l)}(\omega)^H.\end{aligned} \tag{7.2}$$

Applying **absolute normalization**, a normalized copy of $r_m(\omega, l)$ is defined as

$$\begin{aligned}r_m^{norm}(\omega, l) &= \frac{r_m(\omega, l)}{|r_m(\omega, l)|} = \frac{x_m(\omega, l) x_1(\omega, l)^H}{|x_m(\omega, l) x_1(\omega, l)^H|} \\ &\approx \begin{cases} \frac{h_{m,n(\omega,l)}(\omega) h_{1,n(\omega,l)}(\omega)^H}{|h_{m,n(\omega,l)}(\omega) h_{1,n(\omega,l)}(\omega)^H|} + \epsilon(\omega, l), & s_n(\omega, l) s_n(\omega, l)^H > 0 \\ 0, & s_n(\omega, l) s_n(\omega, l)^H = 0 \end{cases}\end{aligned} \tag{7.3}$$

where $\epsilon(\omega, l)$ is the error of a non strict sparse signal assumption or the residual from other source signals in the observed mixtures $\mathbf{x}(\omega, l)$. *Note that in the ideal sparseness, i.e. when $\epsilon(\omega, l) = 0$, $r_m^{norm}(\omega, l)$ does not depend on the source spectral variance $s_{n(\omega,l)}(\omega, l) s_{n(\omega,l)}(\omega, l)^H$ and then can be ideally represented by a finite small number of models, which are only related to the propagation channel characteristics.* Then, it is suitable to lead to an efficient sparse representation of the source signals.

According to equation (7.3), up to the error $\epsilon(\omega, l)$, at each frequency $\omega$, $r_m^{norm}(\omega, l)$ can only represent the cross-channel propagation characteristics. The propagation channel between the position of a source and the location of a microphone is in general difficult to model since it depends on the geometrical description of all the reflective surfaces and on their sound absorption characteristic. The propagation channel characteristics can be represented using measured RIRs. Therefore, the atoms of the dictionary, representing values assumed by $\frac{h_{m,n(\omega,l)}(\omega)h_{1,n(\omega,l)}(\omega)^H}{|h_{m,n(\omega,l)}(\omega)h_{1,n(\omega,l)}(\omega)^H|}$, can be designed using the measured RIRs.

Possible source spatial positions are approximated by selecting a finite set of points, e.g. on a two-dimensional grid. The RIR from the $o$-th location to the $m$-th microphone is obtained. Furthermore, the discrete Fourier transform is applied in order to compute the frequency representation of the impulse response $h_m^o(\omega)$. Finally, the normalized atom vector is defined as

$$\mathbf{d}_m^o = \left[\frac{h_m^o(1)h_1^o(1)^H}{|h_m^o(1)h_1^o(1)^H|} \cdots \frac{h_m^o(\Omega)h_1^o(\Omega)^H}{|h_m^o(\Omega)h_1^o(\Omega)^H|}\right]^T, \tag{7.4}$$

$$\mathbf{d}^o = [\mathbf{d}_2^o; \cdots ; \mathbf{d}_M^o]. \tag{7.5}$$

The over-complete dictionary including all the atom vectors is defined as $\mathbf{D} = [\mathbf{d}^1| \cdots |\mathbf{d}^O]$.

In order to match the atom definition, the time-frequency representations of the observations in equation (7.3) can be inserted into a single vector depending on the frame $l$ such as

$$\mathbf{r}_m(l) = [r_m^{norm}(1, l) \cdots r_m^{norm}(\Omega, l)]^T. \tag{7.6}$$

$$\mathbf{r}(l) = [\mathbf{r}_2(l); \cdots ; \mathbf{r}_M(l)]. \tag{7.7}$$

## 7.2 Detection of matched atoms

We focus on a greedy algorithm based on the Orthogonal Matching Pursuit (OMP) introduced in Section 3.6, and with a modification in the atom selection procedure. Since we observe multiple frames, in principle the projection could be averaged over all the observations. However, this strategy may lead to wrong results due to the intrinsic correlation between the atoms related to very close spatial locations. Here, in order to mitigate this problem and exploit also the source temporal sparsity, we adopt an extended selection procedure which still considers all the observed frames at the same time. In case of temporal sparsity assumption, large coherence between $\mathbf{r}(l)$ and one of the active atoms is detected inside each time frame $l$. Starting by initializing the iterative residual as $\mathbf{z}_0(l) = \mathbf{r}(l)$, we compute the inner product of the columns of the current residual and the atoms of the dictionary and select the one maximizing it inside each time frame $l$

$$\mathrm{o}_l^{\mathrm{match}} = \arg\max_o |(\mathbf{d}^o)^H \ \mathbf{z}_{i-1}(l)|. \tag{7.8}$$

We consider all the indexes $\mathrm{o}_l^{\mathrm{match}}$ obtained over all the frames and sort them in descending order in a vector $\mathbf{q}$, according to their frequencies of occurrences. Finally, for the first $J$-atoms in $\mathbf{q}$, the cumulative inner product is calculated and the atom leading to the highest integrated value is chosen

$$\mathrm{j}^{\mathrm{match}} = \arg\max_j \sum_l |(\mathbf{d}^{q_j})^H \ \mathbf{z}_{i-1}(l)|, \quad j = 1, 2, ..., J \tag{7.9}$$

This strategy avoids that wrong atoms matching noisy frames of the observed data, but with low projection value, would be erroneously detected.

The final algorithmic procedure is described below. In this work the OMP strategy was adopted in place of the standard MP, i.e. the residual is updated computing the projection orthogonal to the subspace spanned by all the atoms estimated until the current iteration.

---

**Initialize**: $\mathbf{z}_0(l) = \mathbf{r}(l),\ \Gamma_0 = \phi,\ \mathbf{D}_{\Gamma_0} = [\mathbf{0}]$.

**Iterate**: For $i = 1$; $i = i{+}1$; till stopping criterion,

Find the index $\mathrm{j}^{\mathrm{match}}$ of the best matching atom with $\mathbf{z}_{i-1}(l)$, $\forall l$ as in (7.9),

Update sub-dictionary by new atom $\mathbf{D}_{\Gamma_i} = [\mathbf{D}_{\Gamma_{i-1}} | \mathbf{d}^{\mathrm{j}^{\mathrm{match}}}]$,

Update the sub-space by new atom index $\mathbf{\Gamma}_i = \mathbf{\Gamma}_{i-1} \cup \mathrm{j}^{\mathrm{match}}$,

Orthogonal projection : $\hat{\mathbf{p}}_i(l) = \mathbf{D}^{\dagger}_{\Gamma_i} \mathbf{z}_{i-1}(l), \forall l$,

Update residual : $\mathbf{z}_i(l) = \mathbf{z}_{i-1}(l) - \mathbf{D}_{\Gamma_i} \hat{\mathbf{p}}_i(l), \forall l$,

Normalize each element of $\mathbf{z}_{i-1}(l), \forall l$ to unit magnitude.

**Return**

---

Here $\mathbf{D}_{\Gamma_i}$ is the sub-dictionary of the selected matched atoms in the $i$-th iteration spanned by the atoms indexed in th subspace $\mathbf{\Gamma}_i$ of the sparse dictionary $\mathbf{D}$, $\mathbf{D}^{\dagger}_{\Gamma_i} = (\mathbf{D}^H_{\Gamma_i} \mathbf{D}_{\Gamma_i})^{-1} \mathbf{D}^H_{\Gamma_i}$ is the pseudo-inverse of $\mathbf{D}_{\Gamma_i}$, and $\mathbf{D}^H_{\Gamma_i}$ is the conjugate transposition of $\mathbf{D}_{\Gamma_i}$. We can choose between two different stopping criteria for $i_{stop}$:

1) Repeat until a predefined level of sparsity $G$, i.e. $i == G$.

2) Repeat until the reduction of the total residual from the previous iteration is smaller than a certain threshold.

## 7.3 Dictionary adaptation

The mismatch between the true mixing parameters used to generate the observed mixtures $\mathbf{x}(\omega, l)$ and the trained atoms in the dictionary $\mathbf{D}$ is the main cause of poor performance of sparse modeling. On the other side, blind techniques are able to estimate the mixing system without specific geometrical knowledge and then better adapt to the observed mixtures

$\mathbf{x}(\omega, l)$. However, their robustness is limited by low convergence, high estimation variance and signal conditions not well fitting the general hypothesis of independence in short-time. We propose to combine both the approaches in order to compensate their individual weak points, leading to a semi-blind estimation method.

We start with the hypothesis that there is only one source dominating a specific STFT frame. Therefore, each instant is used to update only the atom related to the dominating source. For this purpose we use a modification of the weighted Natural Gradient (wNG) proposed in [67]. The main idea behind wNG is to re-weight the gradient according to the likelihood of dominance of a source in a given frame in order to selectively estimate the mixing parameters related to different spatial locations. Following this idea, we select the atom in the dictionary best matching with the observed frame $l$

$$\tilde{o} = \arg\max_{o} \Pr(o, l), \quad \Pr(o, l) = |(\mathbf{d}^{o})^{H}\ \mathbf{r}(l)|, \tag{7.10}$$

and normalize the respective projection as

$$\overline{\Pr}(\tilde{o}, l) = \frac{\Pr(\tilde{o}, l) - \Pr_{\tilde{o}}^{min}}{\Pr_{\tilde{o}}^{max} - \Pr_{\tilde{o}}^{min}}, \tag{7.11}$$

where $\Pr_{\tilde{o}}^{min}$ and $\Pr_{\tilde{o}}^{max}$ are the minimum and maximum projection of the atom $\tilde{o}$ with all the previously observed data frames. The normalized projection is then a weight with values ranging from 0 to 1, indicating the dominance of the source at the location $\tilde{o}$ at the frame $l$.

A weighting matrix $\mathbf{P}^{\tilde{o}}$ is defined as a diagonal matrix with the first element of value $\overline{\Pr}(\tilde{o}, l)$ and the remaining elements set to $1 - \overline{\Pr}(\tilde{o}, l)$. A squared $M \times M$ mixing matrix, describing the source propagating from the

location related to the atom $o$ at the frequency bin $\omega$, is initialized as

$$\widehat{\mathbf{h}}^{o}(\omega) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ d_2^o(\omega) & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ d_M^o(\omega) & 0 & \cdots & 1 \end{bmatrix}, \quad \forall o \tag{7.12}$$

where $d_m^o(\omega)$ indicates an element of the vector $\mathbf{d}_m^o$. According to the weighted NG, for each frame $l$, the atom selected in (7.10) and its corresponding mixing system are updated as follows

$$\mathbf{y}(\omega, l) = [\widehat{\mathbf{h}}^{\tilde{o}}(\omega)]^{-1}\mathbf{x}(\omega, l) \tag{7.13}$$

$$\Delta\mathbf{h}(\omega) = \widehat{\mathbf{h}}^{\tilde{o}}(\omega)(\mathbf{I} - \Phi(\mathbf{y}(\omega, l))\mathbf{y}(\omega, l)^H)\mathbf{P}^{\tilde{o}} \tag{7.14}$$

$$\widehat{\mathbf{h}}^{\tilde{o}}(\omega) = \widehat{\mathbf{h}}^{\tilde{o}}(\omega) - \eta\Delta\mathbf{h}(\omega) \tag{7.15}$$

$$\mathbf{d}_m^{\tilde{o}} = \left[\frac{\widehat{h}_{m1}^{\tilde{o}}(1)\widehat{h}_{11}^{\tilde{o}}(1)^H}{|\widehat{h}_{m1}^{\tilde{o}}(1)\widehat{h}_{11}^{\tilde{o}}(1)^H|}, \cdots, \frac{\widehat{h}_{m1}^{\tilde{o}}(\Omega)\widehat{h}_{11}^{\tilde{o}}(\Omega)^H}{|\widehat{h}_{m1}^{\tilde{o}}(\Omega)\widehat{h}_{11}^{\tilde{o}}(\Omega)^H|}\right]^T \tag{7.16}$$

$$\mathbf{d}^{\tilde{o}} = [\mathbf{d}_2^{\tilde{o}}; \cdots; \mathbf{d}_M^{\tilde{o}}] \tag{7.17}$$

where $\eta$ is the step-size and $\Phi(\cdot)$ is a non-linear function. In practice, the weighting matrix induces the gradient to update the first column of $\widehat{\mathbf{h}}^{\tilde{o}}(\Omega)$ when the source located in $\tilde{o}$ is dominant.

The above adaptation structure differs from that of traditional on-line determined BSS which updates a single mixing/demixing matrix, in order to split the observed mixtures into their individual components. In contrast the proposed algorithm realizes a semi-blind spatio-temporal learning, i.e. the learning proceeds not only in time but also in the spatial domain, according to the prior knowledge given by the geometry. Therefore, the

learning can continue even when the source of interest is silent but some localized noise sources are active, so that a *learning from noise* becomes possible. This is an attractive property which can considerably increases speed and robustness of separation when compared to any blind method.

## 7.4 Experiments

A room with size $8 \times 6 \times 3$ meters and an array of 2 omni directional microphones spaced of 0.2 m are considered. Microphones are located in the middle of the room and have the same height as the sources (1.5 m). Synthetic RIRs are simulated through ISM between multiple locations in the room and the microphones, over a grid of two-dimensional points with a spatial resolution of 0.2 m (i.e. a total of $N_{atoms} = 546$ atoms), with a sampling frequency of $f_s = 16$ kHz. The average reverberation time of the simulated RIRs is of about 250 ms (i.e. the length of the RIRs is about 4096 samples). Time-domain mixtures of $N = 4$ speech sources were generated by individually convolving the full length simulated RIRs with the original source signals and adding the source image contributions to each microphone. The discrete time-frequency representation of the mixture $\mathbf{x}(\omega, l)$ was obtained through STFT with Hanning analysis windows with length $N_{bins}$. The dictionary $\mathbf{D}$ is built by truncating the original impulse responses at the length $N_{bins}$ and applying the discrete Fourier transform (DFT) to obtain a discrete frequency representation. Due to the Hermitian symmetry, only half of the frequency bins were used in both mixtures and dictionary atoms.

### 7.4.1 Mismatch analysis

In this section we analyze through simulation the correlation between observations and atoms, in presence of two causes of mismatch between the

ideal model and the real conditions. Two different conditions generating mismatch are considered:

- Non-ideal source signal sparseness
- Atoms representing the true mixing system are not included in the dictionary

In a first analysis, we assume that the mixing system of each source is exactly modeled by an atom included in the dictionary, i.e. mixtures are obtained by convolution of the source signals with randomly selected RIRs, corresponding to four atoms in the dictionary. In this analysis we compare the effect of ideal and non-ideal time-frequency source sparseness. Having knowledge of the individual source signals recorded at the microphones, an ideal sparse representation of $\mathbf{r}(l)$ is obtained by first computing the index of the most dominant source for each time-frequency point

$$M(\omega, l) = \arg\max_n |s_n(\omega, l)|^2, \tag{7.18}$$

where $|s_n(\omega, l)|^2$ indicates the power of the $n$-th source. The ideal sparseness is simulated by modeling the observation as

$$r_2^{norm}(\omega, l) = \frac{h_2^{o(M(\omega,l))}(\omega) h_1^{o(M(\omega,l))}(\omega)^H}{|h_2^{o(M(\omega,l))}(\omega) h_1^{o(M(\omega,l))}(\omega)^H|}, \tag{7.19}$$

where $o(M(\omega, l))$ is the index of the atom in the dictionary ideally representing the mixing system of the source indexed in $M(\omega, l)$. Note that in the two-channel case, the observation is represented by a single component for each frequency bin $\omega$ and time frame $l$. Then, modeling the observation vector as in (7.6) and (7.7), we compute the magnitude of the inner product between the atoms related to each source and the frames of $\mathbf{r}(l)$

$$\zeta_n(l) = |(\mathbf{d}^{o(n)})^H \mathbf{r}(l)|, \tag{7.20}$$

where $o(n)$ indicates the index in the dictionary of the atom representing

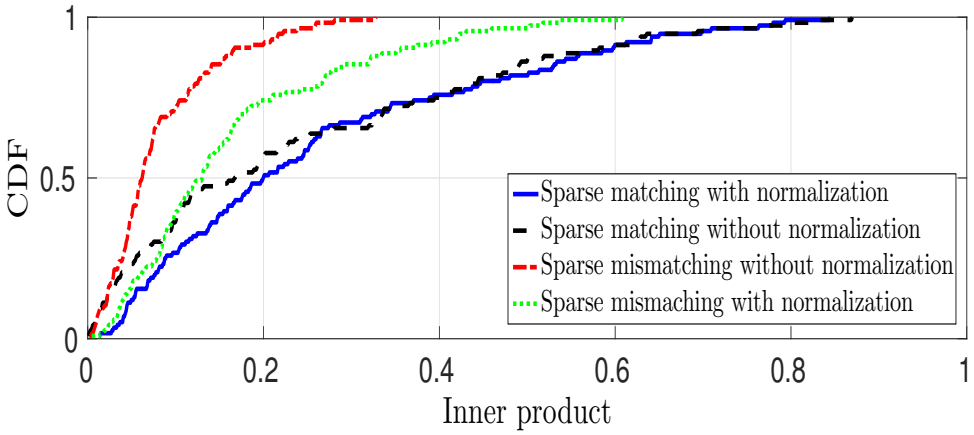


Figure 7.2: Cumulative distribution function of the correlation $\zeta_n(l)$ in case of ideal and non ideal TF source sparseness.

the $n$-th source. Figure 7.2 shows the cumulative distribution function (CDF) of $\zeta_n(l)$ for all $n$ and $l$, when the inner product is computed for both cases of ideal and non-ideal sparse sources with and without applying the absolute normalization in (7.3). It is straightforward to observe that for an ideal sparse representation, with and without absolute normalization, the CDF slowly approaches to 1 for large inner products, which means that there are frames ideally matching the atoms of the dictionary $\mathbf{D}$.

On the other hand, the source signal overlap in the non-ideal sparseness cases is responsible of a large mismatch between the magnitude of the observations and atoms, and consequently the CDF quickly saturates for a value close to 0.3 of correlation. However, when the normalization is applied, the correlation between atoms and observations is better preserved and the CDF gets closer to that obtained in the ideal sparse case.

In a second analysis, we analyze the behavior of the CDF when there is ideal source sparseness, i.e. the error in (7.3) is ideally 0, but the dictionary is represented by atoms non ideally matching the true mixing systems. Similarly to (7.20), we compute the inner product of each active atom and the ideal sparse ratio represented in (7.19). To simulate the mismatch,

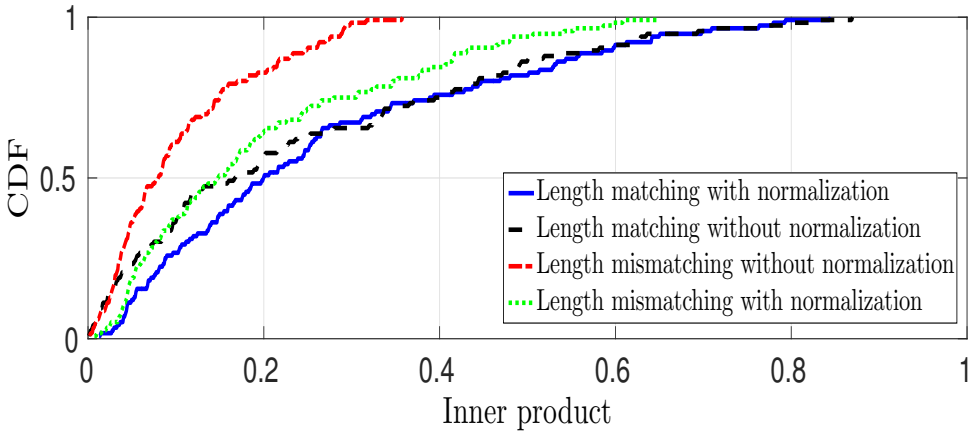


Figure 7.3: Cumulative distribution function of the correlation $\zeta_n(l)$ in case of match/mismatch between atoms and true mixing systems.

the atoms are generated from the simulated RIRs but reduced to a time-domain length of 256 samples. Figure 7.3 shows that the correlation can be seriously degraded in case of mismatch between atoms and true mixing conditions. Nevertheless, the absolute normalization seems considerably able to mitigate the effect of such errors.

### 7.4.2 Source detection and separation

In these simulation experiments, we test the ability of the proposed method to detect the active atoms in order to localize and separate the audio sources. The separation is carried out using the detected active mixing system and applying the $l_0$-norm minimization proposed in [80]. The performance of detection and separation was tested by counting the number of correct source location detection and the source-to-distortion ratio (SDR). To be fair, the measured results were averaged over 20 experiments simulated with different random source positions.

As before, in a first experiment we consider the ideal case of time-frequency source sparseness (the sparse ratios were obtained as in (7.19)) and ideal match between the true mixing system and the atoms in the

dictionary. In practice, the atoms in the dictionary were modeled from the true mixing systems setting their length to the length of the analysis frame $N_{bins} = 4096$, in order to describe the entire full length RIRs. Consistently with the mismatch analysis results, we obtained 100% of true detection and a very high average SDR of 9.24 dB. This result shows that in ideal conditions, the proposed OMP algorithm is able to correctly detect the sources and then is potentially an attractive method for source localization and separation.

In a second experiment we evaluate the performance in the real case, i.e. considering the observations obtained from the real mixtures and modifying the length of the RIRs used to model the dictionary. Specifically the length of the RIRs for the simulation of the atoms was set to the length of the analysis frame $N_{bins}$.

As expected, tables 7.1 and 7.2 show that the robustness can be considerably improved through the normalization which is able to correctly detect all the sources with a sufficient accurate atom definition (i.e. when the length is at least $\geq 1024$ samples). Using the detected mixing system, the original sources $s_n(t)$ could be separated with the $l_0$-norm minimization in [80], leading to a remarkable SDR (see table (7.2)), considered the difficult conditions (i.e. underdetermined scenario, large microphone-source distance). The average SDR is considerably higher with normalization since it allows a higher true atom detection, and grows up monotonically as a function of the length of the detected mixing system. In fact, the sparse demixing in frequency-domain applied by the $l_0$-norm minimization method, becomes more effective as the mixing system modeled by the atoms approaches the true one.

Finally, it can be noted that even when the detection rate is very low, a certain level of separation is still obtained (see the positive SDR values obtained without normalization). This is possible because of the intrinsic

Table 7.1: Percentage of successful detection.

| $N_{bins}$ | 256 | 512 | 1024 | 2048 | 4096 |
|---|---|---|---|---|---|
| *Without Norm.* | 5% | 0% | 5% | 10% | 10% |
| *With Norm.* | 35% | 35% | 90% | 95% | 100% |

Table 7.2: Average of separation performance SDR in dBs.

| $N_{bins}$ | 256 | 512 | 1024 | 2048 | 4096 |
|---|---|---|---|---|---|
| *Without Norm.* | 4.35 | 5.25 | 5.40 | 6.24 | 5.91 |
| *With Norm.* | 4.46 | 5.14 | 6.97 | 8.77 | 9.24 |

correlation between atoms related to multiple close locations. That is, even though it is not possible to correctly detect the true source locations, to a certain extent, wrong detected atoms are still able to describe the sources with a sparse signal representation.

### 7.4.3 Dictionary adaptation

For the evaluation, we simulated mismatched sets of RIRs for the generation of the dictionary and for the generation of the mixtures. The first set was obtained by simulating RIRs assuming uniform absorption coefficients over all the room surfaces (walls, ceiling and floor) and with a reverberation time of about $T_{60} = 50$ ms. The second set was simulated in a similar way but sampling the spatial locations with a random offset (between 0 and 5 cm) with respect to the atom locations and using a larger reverberation time $T_{60} = 250$ ms. In this way we generated a double mismatch between the RIRs in the dictionary and those underlying the mixtures. Indeed, this is a realistic condition that one would observe in real-world because the source locations cannot be exactly restricted to the points sampled in the dictionary and the accuracy of the modeled RIRs is always limited by the used geometrical model.

For the generation of the mixtures three different datasets were consid-

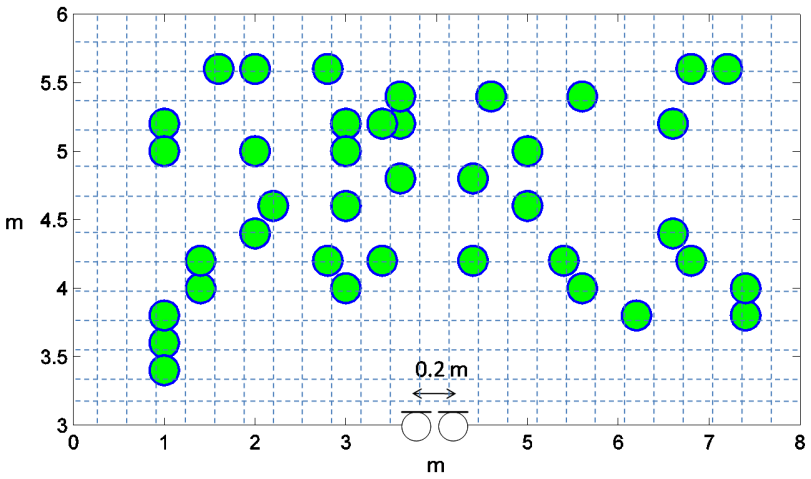


Figure 7.4: Simulation setup: green circles indicate the true locations of the sources in the mixtures while cross points in the grid indicate the spatial locations modeled by the original dictionary.

ered: a set for adaptation used for updating the atoms in the dictionary; two sets for evaluating the separation performance. The adaptation set was generated by creating mixtures of acoustic sources using domestic noise signal samples in the Freesound[1] and the Logic Pro libraries, and added in order to generate a time-varying degree of overlap (for a maximum of three sources overlapping in time). It consists of two hundred mixtures of 12 seconds each for a total of about 40 minutes. Time-domain mixtures were generated by individually convolving simulated RIRs for a given set of locations (see Figure 7.4), with the original source signals and adding the source image contributions at each microphone. The mixtures for the first evaluation dataset were generated by using a speech signal selected from the TIMIT database and three random domestic noise signals.

The second evaluation dataset was generated by using only speech signals, for a total of 4 overlapping speakers. Both test sets consist in 20 mixtures of about 15 s. In all the datasets source locations were randomly modified for each mixture. The discrete time-frequency representation of

[1]http://www.freesound.org/

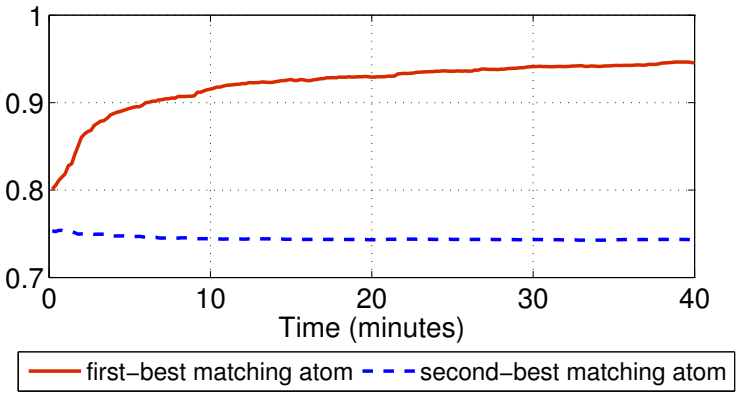


Figure 7.5: Average inner product between the true mixing systems with the first best matching atom (solid line) and the second best matching atom (dotted line)

the mixture $\mathbf{x}(\omega, l)$ was obtained through STFT with Hanning windows of length $L$, shifted of 512 samples. In the weighted Natural Gradient the adaptation step-size was set to $\eta = 0.02$ and the non-linear function to $\Phi(x) = \tanh(10 \cdot |x|)\frac{x}{|x|}$.

#### System identification

Figure 7.5 shows the average projection obtained after having adapted the dictionary with a certain amount of data and showing the projection when considering the first and second best matching atom. At the time instant 0 the average projection corresponds to the performance evaluated with the original unadapted one. It can be noted that as the learning process proceeds over time the average projection of the first atom approaches the unity, which means that each true mixing system will eventually have a close match with one of the adapted atom. On the other hand, the second best matching atom remains unaltered during the learning which means that the discrimination between the atoms increases with the learning, which is a desirable feature for MP-based detection algorithms.

Table 7.3: Mean (standard deviation) performance in dBs for separated signals with and without dictionary adaptation for test dataset with 1 speech + 3 noise random signals. Performance only refers to the target speech signal.

| Metric | Adapted dictionary | Original Dictionary |
|---|---|---|
| SDR | 8.2(2) | 2.9(4.5) |
| $\Delta$ SIR | 2.9(3.5) | -4(7) |

**Signal performance evaluation**

To complete the analysis we report the signal separation performance in terms of Signal-to-Distortion ratio (SDR) and Delta Signal-to-Interference-Ratio ($\Delta$SIR), as defined in [77]. In the evaluation both the original dictionary and the updated dictionary were considered. The signals were separated using the $l_0$-norm minimization [80], applied to each time-frequency point independently by defining the full estimated mixing system as

$$\widetilde{\mathbf{H}}(\omega) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ d_2^{o_1}(\omega) & d_2^{o_2}(\omega) & \cdots & d_2^{o_N}(\omega) \\ \cdots & \cdots & \cdots & \cdots \\ d_M^{o_1}(\Omega) & d_M^{o_2}(\Omega) & \cdots & d_M^{o_N}(\Omega) \end{bmatrix}, \tag{7.21}$$

where $d_M^{o_n}(\omega)$ indicates the $\omega$-th element of the $n$-th atom detected by the OMP algorithm. Tables 7.3 and 7.4 show the performance with and without adaptation when the separation algorithm is applied to both the test datasets, reporting mean and standard deviation. In the dataset with a single speech plus multiple noise sources, performance refers to the speech signal only, while for the other dataset the performance for each speaker is reported. In the first dataset a sensible average improvement in SDR can be observed compared with the original dictionary and the low deviation also indicates a very stable separation result. It is also worth noting that the SIR improvement is not so large even with the adaptation because the average SIR is already high at the input, although it may become very low

Table 7.4: Mean (standard deviation) performance in dBs for separated signals with and without dictionary adaptation for test dataset with 4 speech signals. S1, S2, S3 and S4 indicate the performance averaged over multiple locations but for the same source.

| Metric | S1 | S2 | S3 | S4 | Avg |
|---|---|---|---|---|---|
| SDR | 6.4(1.6) | 6.4(2.7) | 4(2.9) | 3.8(3.4) | **5.1**(**2.9**) |
| $\Delta$ SIR | 14.6(2.4) | 3.3(3) | 7.2(3.2) | 13(3.3) | **9.5**(**5.4**) |
| Adapted dictionary | | | | | |
| Metric | S1 | S2 | S3 | S4 | Avg |
| SDR | 3.2(3.2) | 2.4(4.4) | 4.7(5.4) | 2.3(4.4) | **3.1**(**4.4**) |
| $\Delta$ SIR | 10.9(4.6) | -1.7(5.3) | 6.9(6.6) | 11.2(4.4) | **6.8**(**7.3**) |
| Original dictionary | | | | | |

in some instants where an impulse noise source becomes active. However if the adaptation is not applied a degradation of SIR is also observed (see the negative value), because separation with wrong demixing systems tends to cancel the signal of the target speech.

With the second dataset the improvement becomes even more clear. In fact, performance is averaged over all the speech sources with signals of comparable average power. It is important to mention that sources are at considerable distance from the microphones, averagely around 2.5 meters for a maximum of about 4 meters, and therefore the direct-to-reverberant ratio is low making the estimation of the full mixing system very difficult. Furthermore, since for each mixture the source locations were randomly chosen, sources may be close to each other.

## 7.5 Conclusions

We discussed on a modified greedy algorithm, based on the orthogonal matching pursuit, which is able to detect the mixing systems of multiple sources recorded by a microphone array. After detection, separation of the original source signals was carried out through $l_0$-norm minimization. The

proposed method is based on a sparse multichannel representation of the source mixing parameters. A redundant dictionary, of atoms representing each spatial location, was built with prior knowledge on the room and array geometry. The effect of mismatch between ideal and real conditions was analyzed and the proposed dictionary on-line adaptation with the incoming data through a weighted Natural Gradient, led to higher detection performance. It was shown that the spatial-temporal adaptation mitigates the mismatch between the true mixing systems and the simulated geometrical models, which is otherwise cause of high distortion in the separated signals.

# Chapter 8

# Speaker Diarization

In this work, speaker diarization is referred to the multichannel audio processing operation to determine "who spoke when" in real meetings, when multiple speakers alternate in a discussion. The difficulty of achieving the diarization task in real meetings comes from not only the recording environments but also the speaker activities. In real environments, the recordings are corrupted by background noise plus the influence of reverberation. Moreover, the activities of speakers contain large overlapping speech, and speaker turns occur frequently.

## 8.1 Review

Most speaker diarization systems fit into one of two categories: the bottom-up and top-down approaches [65]. The bottom-up approach trains a number of clusters or models representing the different speakers and aims at merging and reducing the number of clusters until only one remains for each speaker. In contrast, the top-down approach first models the entire audio stream with a single model and adds new models to it until the full number of speakers are deemed to be accounted for. Both bottom-up and top-down approaches are mainly based on Hidden Markov Models (HMMs) where each state is a Gaussian Mixture Model (GMM) and corresponds to

a speaker. Transitions between these states correspond to speaker turns. The procedure of speaker diarization consists of:

- Acoustic beamforming that jointly processes the multiple microphones used to record the meeting.
- Speech activity detection that involves the labeling of speech and non-speech segments.
- Segmentation and clustering that aims at splitting the audio stream into speaker homogeneous segments.

In the multichannel scenario, acoustic beamforming is applied as a front-end pre-processing step in order to mix multiple observations into a unique enhanced signal [5]. Acoustic beamforming can be performed applying, for instance, maximum likelihood (ML) [64], generalized sidelobe canceller (GSC) [43], or minimum variance distortionless response (MVDR) [82]. An alternative approach for acoustic beamforming is to utilize estimated source Time-Difference-Of-Arrivals (TDOAs) and fuse spatial and cepstral information as in [73].

A common approach for speech activity detection is to assume that speech and non-speech segments follow certain models, which can be pre-trained with external speech and non-speech data [88]. The models may optionally be adapted to specific meeting conditions [18]. Labeling of speech and non-speech segments is then performed applying an Expectation Maximization (EM) algorithm with two Gaussian components. Furthermore, temporal smoothing techniques are applied on the binary labels to discard short duration non-speech regions of the audio stream.

After these two pre-processing steps, speaker diarization algorithms diverge into two main directions, i.e. those that apply segmentation to the audio stream, and those that do not apply such segmentation. Both algorithmic approaches exploit a certain characteristic that the speaker labels

exhibit, which is the temporal continuity. Step-by-step algorithms exploit this continuity to turn the problem into a typical unsupervised clustering task. They represent each segments using a statistical model (a single Gaussian or a GMM) and they apply clustering techniques to group them into speakers. On the contrary, the integrated algorithms exploit the temporal continuity by assuming that the transitions between speakers follow a stochastic process which can be modeled by a (first-order and time-independent) Markov chain. Since the labels are not directly observed, an observation model should be added, to link each distinct label (or state) with the observations. The overall model is therefore a HMM, where the observation model (i.e. the part of model that accounts for the state-emission probabilities) is usually a GMM for each state.

## 8.2 Introduction to the proposed method

From the above short review, we can conclude that the speaker diarization problem has been tackled by extracting either spectral or spatial information, or combinations of them [5, 7, 65, 89]. In the conventional spatial feature-based methods, clustering of time-difference of arrivals (TDOAs) of multiple sources between multiple microphone observations has been widely applied [7, 45, 72]. The estimation of the TDOAs is known to be sensitive to the presence of noise and reverberation, and therefore, the diarization performance is limited. It has been proven that directional clustering of normalized observation vectors with the Watson mixture model (WMM) [14] performs well in the reverberant and noisy conditions [47]. Its performance has been confirmed for Blind Source Separation (BSS), while its applicability to a meeting situation is still under investigation.

In this work, we employ the WMM-based clustering for speaker diarization in real meetings. The WMM is represented as the weighted sum of

Watson distributions. For the WMM-based clustering, good estimation of the model parameters is essential. For the reliable parameter estimation, motivated by [30, 34, 66], we also utilize a dictionary of spatial feature models [33]. This dictionary consists of parameters of Watson distributions modeling spatial features for multiple possible speaker locations. The parameters are pre-trained by using training data composed of reverberant speech signals from these locations. Using the spatial dictionary and observing mixture signals, mixture weights modeling activity of each possible speaker location are estimated. To perform the speaker diarization, the estimated mixture weights are post-processed applying cluster merging and temporal smoothing. We will show that the proposed pre-trained WMM via a spatial dictionary realizes robust clustering in real meetings.

## 8.3 Watson mixture model

Directional statistics are primarily concerned with normalized vectors or equivalently vectors residing on the surface of a hypersphere of a unit radius. The $n$-th vector of source spatial images $\mathbf{c}_n(\omega, l)$ of size $M \times 1$ can be normalized, resulting in normalized spatial features related to the $n$-th source position, as [83]

$$\mathbf{c}_n^s(\omega, l) = \frac{\mathbf{c}_n(\omega, l)}{||\mathbf{c}_n(\omega, l)||}, \tag{8.1}$$

where $||.||$ denotes the Euclidean norm. The normalized vector of spatial features $\mathbf{c}_n^s(\omega, l)$ is said to follow the multivariate Watson distribution if its probability density function is given by [63]

$$\begin{aligned} &p(\mathbf{c}_n^s(\omega, l)|\mathbf{a}_n(\omega), \kappa_n(\omega)) = \\ &\frac{(M-1)!}{2\pi^M \mu(1, M, \kappa_n(\omega))} \exp\left(\kappa_n(\omega)|\mathbf{a}_n(\omega)^H \mathbf{c}_n^s(\omega, l)|^2\right), \end{aligned} \tag{8.2}$$

where $\mu$ is the Kummer function. The distribution is rotationally symmetric around the mean orientation $\mathbf{a}_n(\omega)$, which is also a unit norm vector. As the concentration parameter $\kappa_n(\omega)$ increases, the distribution tends to get more spread out around $\mathbf{a}_n(\omega)$. The parameters $\mathbf{a}_n(\omega)$ and $\kappa_n(\omega)$ are estimated in sense of Maximum-likelihood (ML) as in [47, 63].

Assuming that each time-frequency point of the vectors $\mathbf{x}(\omega, l)$ in (3.3) is dominated by one source, the normalized spatial features of the vectors can be approximately represented as

$$\mathbf{x}^s(\omega, l) \approx \mathbf{c}_n^s(\omega, l), \quad n = 1, .., N. \tag{8.3}$$

The directional clustering of the normalized spatial features $\mathbf{x}^s(\omega, l)$ is accomplished by building either binary or soft clustering masks. To cluster the spatial features $\mathbf{x}^s(\omega, l)$ using the WMM, they are probabilistically represented by a mixture of Watson distributions as follows [14]

$$p(\mathbf{x}^s(\omega, l)|\theta) = \sum_{k=1}^{K} \alpha_k(l) p(\mathbf{x}^s(\omega, l)|k, \mathbf{a}_k(\omega), \kappa_k(\omega)), \tag{8.4}$$

where $p(\mathbf{x}^s(\omega, l)|k, \mathbf{a}_k(\omega), \kappa_k(\omega))$ is the Watson distribution of $\mathbf{x}^s(\omega, l)$ as defined in (8.2), replacing $\mathbf{c}_n^s(\omega, l)$ by $\mathbf{x}^s(\omega, l)$. $k$ denotes the cluster number out of total number $K$ of clusters, and $\alpha_k(l)$ indicates the mixture weight encoding the source activity of the $k$-th cluster at the time-frame $l$. If the number of source signals $N$ in the observed mixtures $\mathbf{x}(\omega, l)$ is unknown, which is the case in this work, $K$ is assigned a value larger than $N$. During the clustering, the features $\mathbf{x}^s(\omega, l)$ will be adaptively described by only $N$ active clusters, while the remaining ones are considered as clusters of noise or inactive clusters. The set of parameters of the WMM is defined as

$$\theta = \{\{\alpha_k(l)\}_l, \{\mathbf{a}_k(\omega), \kappa_k(\omega)\}_\omega\}_k. \tag{8.5}$$

Clustering of the vectors $\mathbf{x}^s(\omega, l)$ is performed by estimating the parameters of the WMM applying an EM algorithm in order to optimize the likelihood function in (8.4) (see [14, 47, 63] for details).

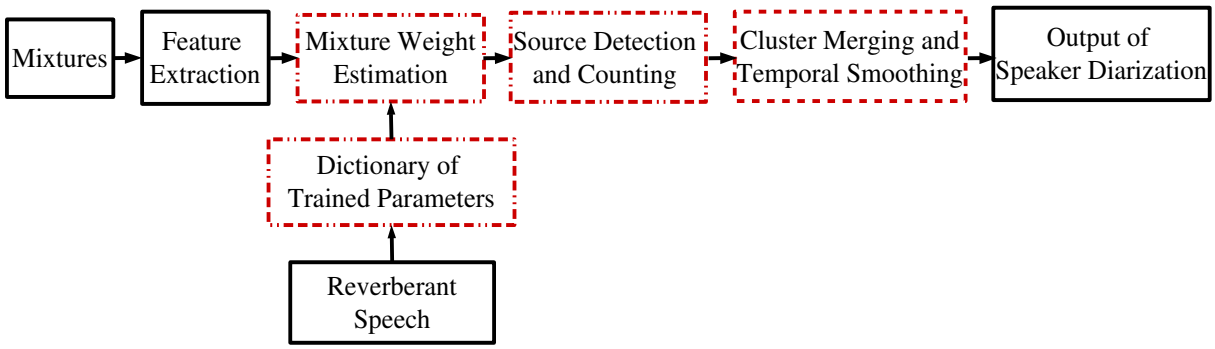


Figure 8.1: Proposed speaker diarization method.

## 8.4 Method

Figure 8.1 shows the processing flow of the proposed method. The method consists of two phases: training and testing. In the training phase, The parameters $\mathbf{a}_k(\omega)$ and $\kappa_k(\omega)$ of the Watson distribution in (8.2) are trained for each $k$-th possible speaker spatial location using training source spatial images $\mathbf{c}_k^t(\omega, l)$. As a result, a spatial dictionary composed of the trained parameters of the Watson distributions of $K$ possible speaker locations is available. In the testing phase, the mixture weight $\alpha_k(l)$ of each $k$-th possible speaker location is estimated by using the dictionary and the observed mixtures $\mathbf{x}(\omega, l)$. To do this, the set of the parameters of the WMM is divided into two disjoint subsets, i.e. $\theta = \theta^t \cup \theta^e$. The subset of parameters to train is defined as $\theta^t = \{\mathbf{a}_k(\omega), \kappa_k(\omega)\}_{k,\omega}$, and the subset of parameters to estimate is represented as $\theta^e = \{\alpha_k(l)\}_{k,l}$.

### 8.4.1 Training of model parameters in spatial dictionary

Let us assume that a speaker is at one of $K$ possible locations, where $K \gg N$. Training data consists of source spatial images $\mathbf{c}_k^t(\omega, l)$ for each $k$-th location. The spatial dictionary is trained using these data by means

of Maximum-Likelihood (ML) estimation of the subset $\theta^t$ applying the following procedures (see [47, 63] for details):

1. Obtain normalized spatial features $\mathbf{c}_k^{ts}(\omega, l)$ of the $k$-th possible speaker location by normalizing the training source images $\mathbf{c}_k^t(\omega, l)$ as in (8.1).

2. Compute a time-invariant empirical spatial covariance matrix of the normalized spatial features as

$$\mathbf{R}_k(\omega) = \frac{1}{L}\sum_{l=1}^{L} \mathbf{c}_k^{ts}(\omega, l)[\mathbf{c}_k^{ts}(\omega, l)]^H. \tag{8.6}$$

3. Derive the largest eigenvalue $\lambda_k(\omega)$ and a corresponding normalized eigenvector as estimation of $\mathbf{a}_k(\omega)$ by the eigenvalue decomposition of the matrix $\mathbf{R}_k(\omega)$.

4. Compute $\kappa_k(\omega)$ by

$$\kappa_k(\omega) \approx \frac{M\lambda_k(\omega) - 1}{2\lambda_k(\omega)(1 - \lambda_k(\omega))}\left[1 + \sqrt{1 + \frac{4(M+1)\lambda_k(\omega)(1-\lambda_k(\omega))}{M-1}}\right]. \tag{8.7}$$

### 8.4.2 Estimation of mixture weights in testing phase

Given the trained set $\theta^t = \{\mathbf{a}_k(\omega), \kappa_k(\omega)\}_{k,\omega}$, the Watson distributions of all $K$ locations or clusters are computed by observing the spatial features $\mathbf{x}^s(\omega, l)$ in (8.3), i.e. $p(\mathbf{x}^s(\omega.l)|k, \mathbf{a}_k(\omega), \kappa_k(\omega))$, $k = 1, ..., K$. By considering Bayes' rule, soft clustering masks are obtained as [47]

$$\gamma_k(\omega, l) = \frac{\tilde{\alpha}_k(l)p(\mathbf{x}^s(\omega, l)|k, \mathbf{a}_k(\omega), \kappa_k(\omega))}{\sum_{\tau=1}^{K}\tilde{\alpha}_\tau(l)p(\mathbf{x}^s(\omega, l)|\tau, \mathbf{a}_\tau(\omega), \kappa_\tau(\omega))}, \tag{8.8}$$

where $\tilde{\alpha}_k(l)$ is the current estimation of the $k$-th mixture weight, which is updated by averaging over the whole frequency bins as follows

$$\tilde{\alpha}_k(l) = \frac{1}{\Omega}\sum_{\omega=1}^{\Omega}\gamma_k(\omega, l). \tag{8.9}$$

The final estimation of $\theta^e = \{\alpha_k(l)\}_{k,l}$ is obtained by iterating (8.8) and (8.9) for two or three times. The masks can also be used to obtain estimation of the source spatial images as follows

$$\tilde{\mathbf{c}}_k(l,\omega) = \gamma_k(\omega,l)\mathbf{x}(\omega,l). \tag{8.10}$$

In practice, if one source dominates all the frequency bins at a time-frame $l_d$, the value of $\tilde{\alpha}_k(l_d)$ associated with the cluster of the dominant source is equal to one. Furthermore, if two sources are active inside the frame, two optimal clusters indexed by $k_1^*$ and $k_2^*$ are detected as active and $\tilde{\alpha}_{k_1^*}(l_d) + \tilde{\alpha}_{k_2^*}(l_d) = 1$. This means that inside each time-frame $l$, the sum of $\tilde{\alpha}_k(l)$ over $k$ is equal to one. However, what will happen if a frame is with noise activity, or with activities received from spatial locations that are not trained? To answer the question, in a pre-processing step, observing $\mathbf{x}(\omega,l)$, we propose to define one more cluster containing time-frequency points of noise activity. The cluster is modeled by training parameters of a Laplace distribution [4], which is accompanied with the trained Watson distributions to estimate the masks in (8.8). Accordingly, one more mixture weight is estimated in (8.9), and one more cluster is computed in (8.10).

### 8.4.3 Modeling of noise

To describe time-frames with noise activity, let us denote the additional estimated weight by $\tilde{\alpha}_{K+1}(l)$ and the additional computed cluster as $\tilde{\mathbf{c}}_{K+1}(l,\omega)$. Observing $\mathbf{x}(\omega,l)$, we recall the computation of an empirical covariance matrix of observed mixtures in (3.9)

$$\tilde{\mathbf{R}}_{\mathbf{x}}(l,\omega) = \frac{\sum_{\tilde{l},\tilde{\omega}} \gamma(\tilde{l}-l,\tilde{\omega}-\omega)\mathbf{x}(\tilde{l},\tilde{\omega})\mathbf{x}^H(\tilde{l},\tilde{\omega})}{\sum_{\tilde{l},\tilde{\omega}} \gamma(\tilde{l}-l,\tilde{\omega}-\omega)}, \tag{8.11}$$

Using this computation of $\tilde{\mathbf{R}}_{\mathbf{x}}(l,\omega)$ accounts for source activities in a block of time-frequency points. The diagonal coefficients of $\tilde{\mathbf{R}}_{\mathbf{x}}(l,\omega)$ convey information about the amount of source activities received at the microphones.

On the other hand, for the sparsity of speech signals, a few time-frequency points are active, and the majority of the points are with low activities. This means using a peaky distribution centered around zero can well describe the noise activity. Accordingly, $p(\tilde{\mathbf{c}}_{K+1}(l,\omega)|a,b(\omega))$ is described by a Laplace distribution trained on the trace of $\tilde{\mathbf{R}}_{\mathbf{x}}(l,\omega)$ as

$$p(\tilde{\mathbf{c}}_{K+1}(l,\omega)|a,b(\omega)) = \frac{1}{2b(\omega)}\exp(-\frac{|\mathrm{tr}(\tilde{\mathbf{R}}_{\mathbf{x}}(l,\omega)) - a|}{b(\omega)}), \tag{8.12}$$

where $a$ and $b(\omega)$ are location and diversity parameters, and tr(.) denotes the trace of a matrix. Since $\mathrm{tr}(\tilde{\mathbf{R}}_{\mathbf{x}}(l,\omega))$ rates from zero to infinity, $p(\tilde{\mathbf{c}}_{K+1}(l,\omega)|a,b(\omega))$ is a positive-sided distribution. To center the distribution around zero, $a$ is set equal to zero. The parameter $b(\omega)$ is empirically obtained as

$$b(\omega) = \frac{1}{L}\sum_{l=1}^{L}\mathrm{tr}(\tilde{\mathbf{R}}_{\mathbf{x}}(l,\omega)). \tag{8.13}$$

## 8.5 Experiments

### 8.5.1 Experimental conditions

To evaluate the proposed work, a real live recorded conversation dataset in a meeting room was considered ( see Figure 8.2). There are a table with chairs and other furniture in the room. An array of 8 omni-directional microphones spaced 4 cm is present on the center of the table, and from 4 to 6 participants are seating on the chairs. The measured reverberation time ($T_{60}$) of the room is about 500 ms. For the testing and estimation of $\theta^e$, real natural conversations were recorded. While a presenter is talking, interruption is coming from the participants by opening discussions about the talk topic in the presence of background noise and interferences. Exhibition audience noise is simulated to be present outside of the room

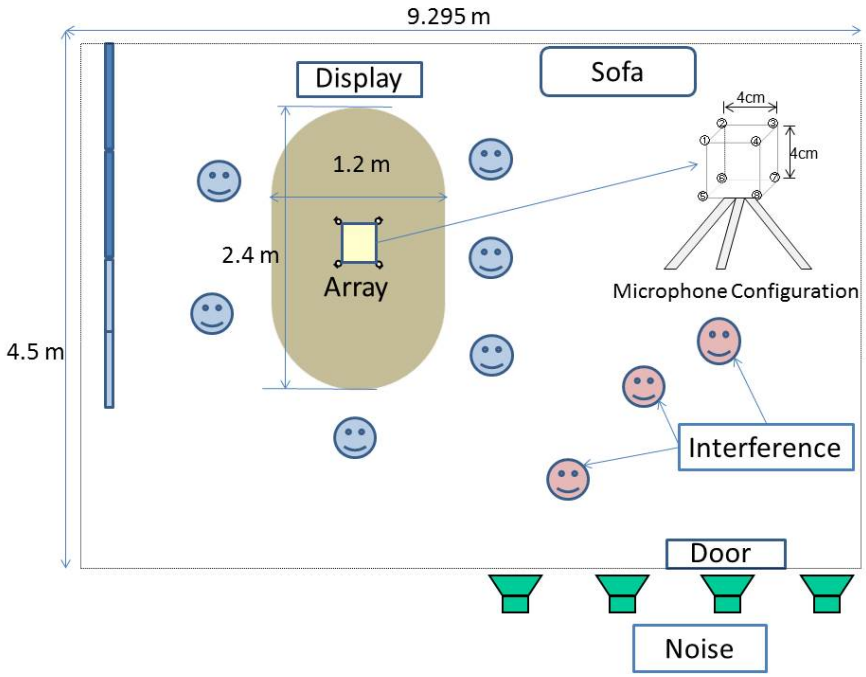


Figure 8.2: Experimental setup for speaker diarization.

and controlled using the door. The number of testing recordings is 8, with lengths varying between 15 and 20 minutes.

For the training and estimation of $\theta^t$, the space around the table is sampled into 72 points with angular distance of 5° and impulse responses are measured at the points. Then $\mathbf{c}_k^t(\omega, l)$ are simulated using a speech signal of length 10 s. As the spatial locations of the speakers in the meetings are randomly chosen, exact spatial matching between the trained and tested locations is not urgently required.

Table 8.1 reports some statistics about the recordings. From the left column to the right, for every one of the recordings, the table states the recording number, the number of speakers, the total overlap in seconds, the percentage of overlap, the number of speaker turn-takings, the average of speaker turn-takings per minute, the total number of utterances, and the average number of utterances per minute.

Table 8.1: Statistics of the recordings.

| Rec. | No. Speakers | Overlap (Second) | Overlap (%) | No. Spk-turns | Spk-turns /min | No. Utters. | Utters. /min |
|---|---|---|---|---|---|---|---|
| 1 | 6 | 165.82 | 18.25 | 326 | 21.50 | 501 | 33.10 |
| 2 | 6 | 227.74 | 21.22 | 471 | 26.30 | 645 | 36.10 |
| 3 | 6 | 229.74 | 23.62 | 352 | 21.70 | 513 | 31.60 |
| 4 | 5 | 235.09 | 25.90 | 433 | 28.57 | 565 | 37.30 |
| 5 | 5 | 296.73 | 32.82 | 633 | 42.00 | 745 | 49.40 |
| 6 | 6 | 152.27 | 16.53 | 270 | 17.60 | 413 | 26.90 |
| 7 | 4 | 286.46 | 31.60 | 496 | 32.80 | 631 | 41.80 |
| 8 | 4 | 326.00 | 35.93 | 515 | 34.10 | 633 | 41.90 |

### 8.5.2 Implementation issues

Given one of the recordings, the signal is divided into time-blocks each of length one second, which allows for on-line processing. For the computation of the time-frequency representation of the signal through the Short Time Fourier Transform (STFT) with a sampling rate of 16 kHz, the frame length is 1024 samples (64 ms) with a shift of 256 samples (16 ms). In the implementation, it is required to apply temporal smoothing and merge neighboring clusters. Considering the continuity of speech signals, a sliding averaging window of size 60 time-frames is applied in order to smooth the estimation of $\theta^e$ over the frames. Observing the smoothed mixture weights $\tilde{\alpha}_k(l)$, the active clusters are detected and the sources are counted and localized by considering the peaks to obtain a time-varying set of active cluster indexes $\mathbf{k}^*(l) = \{k_1^*(l), k_2^*(l), .., k_N^*(l)\}$.

Several kinds of artifacts such as the reverberation, the coherence between the trained models of neighboring spatial locations, the mismatch between the trained and tested locations, etc., could disturb the stability of the method. As a result of such disturbance, exact match between one particular trained cluster and one tested location at each time frame over

all the signal length cannot be captured. Accordingly, multiple neighboring clusters related to one location are detected to be active from one time-frame to another, fluctuating around the main detected active cluster. For this reason, the source activity is estimated in a wide angular range by merging the contribution of neighboring clusters together. For example, in this experiment, we collect the contribution of 3 estimated weights in an angular range of size 10°; i.e. $\mathbf{k}^*_{\pm 1}(l) = \{k^*_1(l) \pm 1, k^*_2(l) \pm 1, .., k^*_N(l) \pm 1\}$. One more processing step has been applied to remove short fragments and pauses using hangover smoothing. Given $\mathbf{k}^*_{\pm 1}(l)$, the smoothed source activities $\tilde{\alpha}_{\mathbf{k}^*_{\pm 1}}(l)$ are used for speaker diarization.

### 8.5.3 Speaker diarization results

The performance of the proposed method was evaluated by computing the diarization error rate (DER),

$$\text{DER} = \frac{\text{Wrongly estimated speaker time length}}{\text{Entire speaker time length}} \times 100[\%],$$

which was established by NIST [2]. To calculate the DER, the estimated activities were compared to a manually annotated dataset. Table 8.2 reports information about the background noise with the DER of the proposed method (Prop.) compared to a baseline method (Base), which is based on the combination of voice activity detection with TDOA estimation (see Section III-C [45]). As it is observed the proposed method outperforms the baseline method over all the testing recordings. As it is expected the performance of the proposed method varies as a function of the overlap and the background noise level. Increasing either the overlap or the noise level, or both of them, reduces the diarization performance.

Table 8.2: Diarization results. Door is: 0 for closed and 1 for opened; Level is: 0 for noiseless, 1 for low, and 2 for high.

| Recording | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Noise | Door | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| | Level | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 2 |
| DER(%) | Baseline | 46.81 | 64.58 | 62.63 | 23.84 | 47.46 | 67.17 | 73.56 | 70.93 |
| | Proposed | 9.34 | 12.21 | 15.61 | 17.17 | 18.86 | 18.88 | 24.80 | 27.69 |

## 8.6 Conclusion

This chapter presented a review of the speaker diarization as well as a proposed method for speaker diarization in real meetings. The core idea of the method is to train parameters of probabilistic models using directional spatial features, where each model statistically describes a spatial cluster of a possible speaker location. The Watson distribution is adopted to model the spatial features of each cluster. Mixtures of audio signals are probabilistically modeled by the Watson mixture model (WMM), which is expressed in terms of the trained Watson distributions and mixture weights that define the source activities of the spatial locations. Furthermore, we proposed to on-line train a Laplace distribution to model the noise activity. Given the trained models, the observed mixtures are clustered and the mixture weights are estimated. The speaker diarization is performed by observing post-processed mixture weights. The post-processing of the estimated weights involves cluster merging and temporal smoothing. The elaborated results show the consistency and superiority of the proposed method over a recently developed baseline method.

# Chapter 9

# Conclusion and Future Research

Source separation has been tackled during the last two decades aiming at producing high quality separated signals. Conventionally, the separation is performed by exploiting source statistical properties such as independence, non-gaussianity, etc. In this work we considered the case where multiple microphones distributed inside an echoing room are used to pick up mixtures of audio signals produced by multiple speakers at several spatial positions. In this scenario, audio source separation is achieved by exploiting the spatial diversity of multiple observations at the multiple microphones together with source statistical properties. Advanced spectral and sparse modeling theories are applied for further exploitation of source spectral-temporal redundancy and sparseness. A separation system can be informed by prior information about a specific mixing problem, in order to increase the quality and to accelerate the separation.

On the other side, to arrive at multiple microphones, an audio signal propagates through multiple channels. In this work, the audio signals are represented by their spectral descriptions and the propagation channels are represented by their spatial descriptions. A separation system can be fed by either spectral descriptions of audio signals in the mixtures, or spatial descriptions of propagation channels used to generate the mixtures. Three

modalities of making these descriptions available are considered, i.e. a) the descriptions are on-line trained during the separation, b) the descriptions are pre-trained and made directly available, c) or the descriptions are pre-trained and made indirectly available through a redundant dictionary or library. In the latter, either the spectral descriptions, best representing audio signals in the mixtures, or the spatial descriptions, best representing propagation channels used to generate the mixtures, are detected during the separation. The spectral descriptions of audio signals are exploited by applying the spectral modeling theory, and the spatial descriptions of propagation channels are exploited by applying the sparse modeling theory.

## 9.1 Spectral information for source separation

Using spectral modeling based on Nonnegative Matrix Factorization (NMF), the power spectrum of an audio signal is represented as the product of two matrices, i.e. a spectral basis matrix containing constitutive parts of the spectrum, and an activation coefficient matrix containing time-varying weights. The factorization can be performed either in an unsupervised scenario where the two matrices are unknown, or in a supervised scenario where one of the two matrices is known in an approximate representation of the true one. As recently proposed, spectral basis matrices of multiple audio signals can be assumed to be known and identified as source spectral descriptions.

It has been proven that model-based audio source separation methods achieve good performance. Gaussian model-based audio source separation represents observed mixtures of audio signals by a probabilistic model parametrized by spectral and spatial parameters, i.e. source variances encoding power spectra of audio signals, and spatial covariance matrices encoding propagation channels. Source separation is performed by first

estimating the parameters of the Gaussian model, later applying multi-channel Wiener filtering. The model parameters are estimated in sense of Maximum-Likelihood (ML) by optimizing the model with respect to each one of the parameters applying a Generalized Expectation-Maximization (GEM) algorithm.

Conventionally, the spectral and spatial parameters of the model are dependently estimated, where each parameter is updated using the other estimated one. As a result, the estimation error is accumulated from one parameter to the other one, and from one estimation iteration to another. To reduce the dependency, we proposed either to estimate the spectral parameter regardless of the spatial one, or to jointly estimate them:

- In the first estimation method, the source variance is estimated by computing the singular value decomposition of a matrix of multiple observations as in Chapter 4 and 5. The estimated source variance can be factorized applying supervised NMF in case that the separation system is informed by pre-trained source spectral basis matrices, as proposed in Chapters 4 and 5. However, unsupervised NMF can be applied to factorize the estimated source variance in a blind scenario and we consider this step as on-line training of the spectral basis matrices, as proposed in Chapter 5. To estimate the spatial covariance matrix, we started by factorizing absolute values of the matrix of multiple observations applying supervised NMF or NTF by using the trained spectral basis matrices as in Chapter 4, or by generating weighted basis matrices from vectors of either the pre-trained or on-line trained spectral basis matrices as in Chapter 5. The main purpose of this factorization step is to find compact representations of the multiple observations using the spectral basis matrices. The spatial covariance matrix is then estimated using the factorization of the source variance, the factorization of the absolute values of the matrix

of multiple observations, and the phase information of the matrix of multiple observations as in Chapter 4 and 5.

- A stable estimation could be recognised if the parameters are jointly updated applying supervised NMF/NTF as proposed in Chapter 6, given either extracted or pre-trained source spectral basis matrices. Following this direction, we proposed to rearrange elements of the matrices of multiple observations in tensors of nonnegative parameters, and matrices of complex parameters. Due to the implicit requirement of nonnegativity, we split the model parameters into two subsets, i.e. a subset of nonnegative parameters to be estimated by factorizing the tensors, and a subset of complex parameters to be estimated by factorizing the matrices. In fact, the purpose is not to factorize complex matrices using NMF, but is only an updating step to keep the scaling matched between the two estimated subsets. Besides using the tensors of nonnegative parameters to estimate the subset of nonnegative parameters, they are also used either to extract the spectral basis matrices, or to detect them from a redundant library of pre-trained spectral basis matrices.

All the above factorization steps are performed by minimizing the $\beta$-divergence applying the widely used Multiplicative Update (MU) rules [52]. A deep study on selecting the size of the spectral basis matrices and the values of the divergence factor $\beta$ was conducted in order to identify the best performing combinations. We found that these values are data-dependent and should be optimized for each mixing condition in order to obtain the best separation performance. Large size spectral basis matrices mostly benefit the performance, however, better performance could be obtained if $\beta$ is assigned a suitable value when the size of the matrices is small. Experiments were carried out to assess the performance of the proposed

methods, and it was found that they outperform other recently developed state-of-the-art methods.

Although the separation system can perform better, in some mixing conditions, with the first proposed estimation method than with the second one, the separation convergence with the second method is more guaranteed than with the first one. Both the estimation methods suffer when the source signals are highly overlapped, that was confirmed by the low separation performance obtained when they were used to separate mixtures of male speech signals. In the detection step of the pre-trained spectral basis matrices from the library, best representing source signals in observed mixtures, wrong detection may happen. The main cause of that wrong detection is the high redundant coherence between the pre-trained spectral basis matrices.

## 9.2 Spatial information for source separation and speaker diarization

Sparse modeling assumes an ability to describe a signal by a small number of values using a pre-defined dictionary. In this work, the dictionary is composed of trained spatial descriptions of propagation channels representing a finite set of source spatial positions. Given the dictionary, source separation is performed by first detecting the active spatial source positions using a modified sparse modeling algorithm (orthogonal matching pursuit) [21] as proposed in Chapter 7. Then later the detected spatial descriptions associated with the active positions are used to obtain the source signals applying $l_0$-norm minimization [80]. In case that there is mismatch between the spatial descriptions in the dictionary and the ones used to generate the observed mixtures, an unsupervised dictionary adaptation step using weighted Independent Component Analysis (wICA) [67]

is proposed in order to reduce such mismatch.

**Spatial dictionary for source separation:** By working at the signal level, a column of the dictionary is represented as the cross correlation between each pair of parameters of multiple propagation channels obtained between a spatial position and multiple microphones. To match the observed mixtures with each column of the dictionary, they are also represented as the cross correlation between each pair of mixture signals. To increase the matching probability, each element of both the dictionary and the observations is normalized using its absolute value. This normalization step is seen as important for improving the detection accuracy as well as for increasing the separation performance.

On the other side, for speaker diarization, observed mixtures of audio signals are probabilistically modeled by a Watson mixture model (WMM) [14] as proposed in Chapter 8. Given the dictionary, speaker diarization is achieved by estimating the source activity of each spatial position by optimizing the model using an Expectation-Maximization (EM) algorithm [27] and applying soft spectral masking [47]. In case of the presence of background noise and interferences, a Laplace distribution [4] is proposed to model the accompanying corruption generated by the noise and the interferences .

**Probabilistic spatial dictionary for speaker diarization:** By working at the probabilistic level, a column of the dictionary is represented as a Watson distribution [63] with trained controlling parameters. The Watson distribution is well known to probabilistically describe directional statistics. The distribution models a unit norm vector by an exponential function controlled by a mean orientation vector and a concentration parameter. The parameters of the distribution related to a particular spatial position are trained using unit norm vectors of reverberant speech signal received at multiple microphones from that position. To match the signal

representation used to train the distributions, the observed mixtures are also normalized using the Euclidean norm of a vector.

Based on the trained spatial dictionaries, according to the obtained experimental results, consistent and robust source separation and speaker diarization can be achieved. However, the unsupervised dictionary adaptation step for source separation still needs to be reconsidered again in order to obtain improved performance. For speaker diarization, the training of normalized mean orientation vectors of Watson distributions is not enough to well describe the propagation environments with high reverberation.

## 9.3 Perspective on future research directions

In the future, the effectiveness and robustness of the proposed methods can be improved by:

- Looking for more efficient spectral modeling based on Nonnegative Matrix and Tensor Factorization (NMF/NTF) algorithms. These algorithms should be stable to work in difficult mixing conditions, and weak source spectral-temporal sparseness. Constraining the sparsity of factorization as in [19] and using other factorization cost functions such as the $\alpha$-divergence as in [24] could be good research directions leading to improved solutions. Training spectral basis matrices with high source reconstruction and discrimination properties or updating the trained matrices during the separation for better properties as in [85] could also be a possible trend to minimize the coherence between the trained matrices, and to increase the performance. Furthermore, self adaptation of the factorization divergence factor $\beta$ as a function of the signal level is a special need to increase the flexibility of the proposed methods.

- Seeking for a robust on-line update method of the spatial dictionary

for source separation, in which the mismatch between the trained parameters in the dictionary and the parameters used to generate observed mixtures is reduced. To meet this objective, a dictionary update scheme by an iterative projection method complemented by a rotation of the dictionary as proposed in [12] could be a good idea aiming at reducing such mismatch and improving the overall detection and separation performance.

- Learning spatial covariance matrices that is a more consistent option to describe propagation environments with high reverberation. The Bingham distribution [49] provides a model to describe directional statistics using spatial covariance matrices. In this case, the Watson distribution could be replaced by the Bingham distribution, and the Watson mixture model (WMM) is replaced by the Bingham mixture model (BMM), for more robust speaker diarization.

The proposed methods find their viable applications in several fields including:

- Audio surveillance and tracking of multiple speakers.
- Hearing aids by increasing the accuracy of source localization and by reducing the environmental nuisance.
- Improving the quality of audio communications during a phone call or teleconference by reducing the surrounding interferences.
- Increasing the performance of automatic speech recognition and transcription systems by using the proposed methods as front-end aiming to produce enhanced signals.

# Bibliography

[1] http://sisec2010.wiki.irisa.fr/tiki-index8f77.html?page=Underdetermined-+speech+and+music+mixtures.

[2] http://www.nist.gov/speech/test_beds/mr_proj/.

[3] H. S. Stern A. Gelman, J. B. Carlin and D. B. Rubin. *Baysian Data Analysis*. Chapman and Hall /CRC, 2003.

[4] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55. U.S. Government Printing Office, Washington, D.C., 1964.

[5] X. Anguera, C. Wooters, and J. Hernando. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7):2011–2022, 2007.

[6] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda. Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. *Acoustic Society of Technology*, 22(2):149157, 2001.

[7] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino. A DOA based speaker diarization system for real meetings. In *Proceedings of HSCMA*, 2008.

[8] S. Araki, T. Nakatani, H. Sawada, and S. Makino. Stereo source separation and source counting with MAP estimation with dirichlet prior considering spatial aliasing problem. In *Proceeding of ICA*, 2009.

[9] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and Vandergheynst. Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation. In *Proceedings of ISSPA*, 2010.

[10] A. Asaei, M. E. Davies, H. Bourlard, and V. Cevher. Computational methods for sparse component analysis of convolutive speech mixtures. In *Proceeding of ICASSP*, pages 2425–2428, 2012.

[11] A. Asaie, H. Bourlard, and V. Cevher. Model-based compressive sensing for multi-party distant speech recognition. In *Proceeding of ICASSP*, 2011.

[12] D. Barchiesi and M. D. Plumbley. Learning incoherent dictionaries for sparse approximation using iterative projections and rotations. *IEEE Transactions on Signal Processing*, 61(8):2055–2065, 2013.

[13] A. Basu, I. R. Harris, N. L. Hjort, and M. C. Jones. Robust and efficient estimation by minimizing a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[14] A. Bijral, M. Breitenbach, and G. Grudic. Mixture of Watson distributions: A generative model for hyperspherical embeddings. In *AISTATS*, volume 2 of *JMLR*, pages 35–42, 2007.

[15] J. Bobin, J. L. Starck, Y. Moudden, and M. J. Fadili. Blind source separation: the sparsity revolution, 2008.

[16] P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. *IEEE Transactions on Signal Processing*, 81(11):2353–2362, 2001.

[17] P. Boll. Underdetermined blind separation of delayed sound sources in the frequency domain. *Neurocomputing*, 55:627–641, 2003.

[18] N. W. D. Evans C. Fredouille. The LIA RT'07 speaker diarization system. In *Lecture Notes on Computer Science, CLEAR 2007 and RT 2007, Multimodal Technologies for Perception of Humans*, volume 4625/2008, pages 520–532, 2008.

[19] C. Caiafa and A. Cichocki. Block sparse representations of tensors using kronecker bases. In *Proceedings of 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[20] D. Campbell. Roomsim toolbox [online]. available:. `http://www.mathworks.com/matlabcentral/fileexchange/5184`.

[21] H. Cheng. *Sparse Representation, Modeling and Learning in Visual Recognition 2015: Theory, Algorithms and Applications.* Springer, 2015.

[22] J. Cho and C. D. Yoo. Underdetermined convolutive bss: Bayes risk minimization based on a mixture of super-gaussian posterior approximation. *IEEE Transactions on Audio, Speech and Language Processing*, 23(5):828–839, 2015.

[23] A. Cichocki and S. I. Amari. *Adaptive Blind Signal and Image Processing:learning algorithms and applications.* John Wiley and Sons, New York, NY, USA, 2002.

[24] A. Cichocki, S. Cruces, and S. Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13:134–170, 2011.

[25] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative Matrix and Tensor factorizations: Applications to Exploratory Multiway Data Analysis*. John Wiley and Sons, New York, NY, USA, 2009.

[26] A. Deleforge, F. Forbes, and R. Horaud. Variational EM for binaural sound-source separation and localization. In *Proceedings of ICASSP*, 2013.

[27] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

[28] S. Doclo and M. Moonen. On the output SNR of the speech-distortion weighted multichannel wiener filter. In *Proceedings of IEEE SPL*, 2005.

[29] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47:2845–2862, 2001.

[30] N. Q. K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech and Language Processing*, 18(7):1830–1840, 2010.

[31] N. Q. K. Duong, E. Vincent, and R. Gribonval. Spatial location priors for gaussian model based reverberant audio source separation. *Journal of Advanced Signal Processing (EURASIP)*, pages 149–162, 2013.

[32] E. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE*

*Transactions on Audio, Speech, and Language Processing*, 32(6):1109–1121, 1984.

[33] M. Fakhry, N. Ito, S. Araki, and T. Nakatani. Modeling audio directional statistics using a probabilistic dictionary for speaker diarization in real meetings. In *Proceedings of IWAENC*, 2016.

[34] M. Fakhry and F. Nesta. Underdetermined source detection and separation using a normalized multichannel spatial dictionary. In *Proceedings of IWAENC*, 2012.

[35] M. Fakhry, P. Svaizer, and M. Omologo. Reverberant audio source separation using partially pre-trained nonnegative matrix factorization. In *Proceedings of IWAENC*, 2014.

[36] M. Fakhry, P. Svaizer, and M. Omologo. Audio source separation using a redundant library of source spectral bases for nonnegative tensor factorization. In *Proceedings of ICASSP*, 2015.

[37] M. Fakhry, P. Svaizer, and M. Omologo. Estimation of the spatial information in gaussian model based audio source separation using weighted spectral bases. In *Proceedings of EUSIPCO*, 2016.

[38] C. Fevotte, N. Bertin, and J. L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence with applications to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[39] C. Fevotte and J. F. Cardoso. Maximum likelihood approach for blind audio source separation using time-frequency gaussian models. In *Proceedings of WASPAA*, 2005.

[40] C. Fevotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 2011.

[41] D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *Proceedings of the Irish Signals and Systems Conference (ISSC)*, 2008.

[42] S. Gorlow and S. Marchand. Informed source separation: Underdetermined source signal recovery from instantaneous stereo mixture. *Proceedings of WASPAA*, pages 309–312, 2011.

[43] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30:27–34, 1982.

[44] S. Araki H. Sawada and S. Makino. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. *IEEE transactions on Audio, Speech, and Language Processing*, 19(3):516–527, 2011.

[45] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, and A. Nakamura. Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2):499–513, 2012.

[46] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, New York, NY, USA, 2001.

[47] N. Ito, S. Araki, and T. Nakatani. Permutation-free convolutive blind source separation via full-band clustering based on frequency-independent source presence priors. In *Proceeding of ICASSP*, 2013.

[48] A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. *Proceedings of ICASSP*, 2000.

[49] J. T. Kent, P. D. L. Constable, and F. Er. Simulation for the complex bingham distribution. *Statistics and Computing*, 14(1):53–57, 2004.

[50] M. Kim and P. Smaragdis. Mixtures of local dictionaries for unsupervised speech enhancement. *IEEE Signal Processing Letters*, 22(3):293–297, 2015.

[51] K. H. Knuth. Informed source separation: A bayesian tutorial. In *Proceedings of EUSIPCO*, 2005.

[52] R. Kompass. A generalized divergence measure for nonnegative matrix factorization. *Neural Computation*, 19(3):780–791, 2007.

[53] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[54] E. A. Lehmann and A. M. Johansson. Prediction of energy decay in room impulse responses simulated with an image-source model. *The Journal of the Acoustic Society of America*, 124(1):269–277, 2008.

[55] Y. Q. Li, S. Amari, A. Cichocki, and D. W. C. Ho. Underdetermined blind source separation based on sparse representations. *IEEE Transactions on Signal Processing*, 54(2):2423–437, 2006.

[56] N. Linh-Trung, A. Belouchrani, K. Abed-Meraim, and B. Boashash. Separating more sources than sensors using time-frequency distributions. *Journal of Applied Signal Processing*, 2005(17):2828–2844, 2005.

[57] A. Liutkus, J. Pinel, R. Badeau, L. Girni, and G Richard. Informed source separation through spectrogram coding and data embedding. *Signal Processing*, 92(8):1937–1949, 2012.

[58] B. Loesch and B. Yang. Blind source separation based on time-frequency sparseness in the presence of spatial aliasing. In *Proceedings of LVA/ICA*, 2010.

[59] S. Makino, T. W. Lee, and H. Sawada. *Blind Speech Separation*. Springer, Berlin, 2007.

[60] J. Màlek, Z. Koldovskỳ, and P. Tichavskỳ. Semi-blind source separation based on ica and overlapped speech detection. In *Proceedings of LVA/ICA*. 2012.

[61] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[62] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis. Model-based expectation maximization source separation and localization. *IEEE Transactions on Audio, Speech, Language Processing*, 18(2):382394, 2010.

[63] K. Marida and I. Dryden. The complex Watson distribution and shape analysis. *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)*, 61(4):913–926, 1999.

[64] L. S. Michael, R. Bhiksha, and M. S. Richard. Likelihood-maximizing beamforming for robust hands-free speech recognition. *IEEE Transactions on Speech and Audio Processing*, 12(5):489–498, 2004.

[65] X. A. Mir, S. Bozonnet, N. W. D. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2):356–370, 2012.

[66] F. Nesta and M. Fakhry. Unsupervised spatial dictionary learning for sparse underdetermined multichannel source separation. In *Proceedings of ICASSP*, 2013.

[67] F. Nesta and M. Omologo. Convolutive underdetermined sources separation through weighted interleaved ica and spatio-temporal correlation. In *Proceedings of LVA/ICA*, 2012.

[68] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2006.

[69] M. Omologo and P. Svaizer. Use of the cross-power-spectrum phase in acoustic event location. *IEEE Transactions on Speech and Audio Processing*, 5:288–292, 1997.

[70] A. Ozerov and C. Fevotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):550–563, 2010.

[71] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(4):1118–1133, 2012.

[72] J. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multi-microphone meetings using only between-channel difference. *MLMI 2006, LNCS 4299*, pages 257–264, 2006.

[73] J. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Transactions on Computing*, 56(9):1189–1224, 2007.

[74] M. Parvaix and L. Girni. Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *IEEE TASLP*, 19(6):1721–1733, 2011.

[75] Y. Pati, R. Rezaifar, and P. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet

decomposition. In *Proceedings of the 27th Asilomar Conference Signals, Systems and Computer*, 1993.

[76] N. Roman, D. Wang, and G. Brown. Speech segregation based on source localization. *Acoustic society of America*, 114(4):22362252, 2003.

[77] H. Sawada, S. Araki, R. Mukai, and S. Makino. Blind extraction of dominant target sources using ICA and time-frequency masking. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2165–2173, 2006.

[78] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada. A multichannel mmse-based framework for speech separation and noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1913–1928, 2013.

[79] D. Sun and G. Mysore. Universal speech models for speaker independent single channel source separation. In *Proceedings of ICASSP*, 2013.

[80] E. Vincent. Complex nonconvex $l_p$ norm minimization for underdetermined source separation. In *Proceedings of ICA*, pages 430–437, 2007.

[81] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong. The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges. *Signal Processing*, 92(8):1928–1936, 2012.

[82] A. S. Vorobyov. Principles of minimum variance robust adaptive beamforming design. *Signal Processing*, 93(12):3264–3277, 2013.

[83] D. H. Tran Vu and R. Haeb-Umbach. Blind speech separation employing directional statistics in an expectation maximization framework. In *Proceeding of ICASSP*, 2010.

[84] D. L. Wang and G. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. John Wiley and Sons, 2006.

[85] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe. Discriminative nmf and its application to single-channel source separation. In *Proceedings of INTERSPEECH*, 2014.

[86] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran. Speech denoising using nonnegative matrix factorization with priors. In *Proceedings of ICASSP*, 2008.

[87] S. Winter, W. Kellerman, H. Sawada, and S. Makino. Map based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization. *Journal of Advanced Signal Processing (EURASIP)*, pp. Article ID 24 717, 12pp, 2007.

[88] C. Wooters, J. Fung, B. Peskin, and X. Anguera. Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system. In *In Proceedings of the RT-04F Workshop*, 2004.

[89] C. Wooters and M. Huijbregts. The ICSI RT07s speaker diarization system. In *Proceedings of the Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, pages 509–519, 2008.

[90] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.

# Appendix A

# The MU rule for the $\beta$-divergence

The divergence in (3.18) between the element $a$ and its decomposition $bcd$ can de defined as

$$d_\beta(a/bcd) = \frac{a^\beta + (\beta-1)(bcd)^\beta - \beta a(bcd)^{\beta-1}}{\beta(\beta-1)}. \tag{A.1}$$

To minimize the divergence with respect to, for example, the element $b$, the partial derivative of $d_\beta(a/bcd)$ is computed

$$g = \frac{\beta(\beta-1)cd(bcd)^{(\beta-1)} - \beta(\beta-1)acd(bcd)^{\beta-2}}{\beta(\beta-1)} \tag{A.2}$$

So the positive part of the derivative is defined as

$$g_+ = cd(bcd)^{(\beta-1)}, \tag{A.3}$$

and the negative part is represented as

$$g_- = acd(bcd)^{\beta-2}. \tag{A.4}$$

The MU rule to update the element $b$ is described in terms of $g_+$ and $g_-$ as

$$b = b\frac{g_-}{g_+} = b\frac{acd(bcd)^{\beta-2}}{cd(bcd)^{(\beta-1)}}. \tag{A.5}$$

And so, we can define the update rule using MU for each element. Moreover, this element-wise update rule can be easily extended for matrix factorization, respecting the dimensions.

For tensor factorization, let us assume that $\mathbf{A}^M$ is a tensor of size $1 \times 1 \times M$. By decomposing the tensor into two elements $b$ and $d$, and a tensor $\mathbf{C}^M$, the divergence can be described as

$$\sum_m d_\beta(a^m/bc^m d) = \sum_m \frac{(a^m)^\beta + (\beta-1)(bc^m d)^\beta - \beta a^m (bc^m d)^{\beta-1}}{\beta(\beta-1)}, \quad \text{(A.6)}$$

where $a^m$ and $c^m$ are the *m-th* elements of $\mathbf{A}^M$ and $\mathbf{C}^M$, respectively. To minimize the divergence with respect to, for example, the element $b$, the partial derivative $g$ is computed

$$g = \sum_m c^m d(bc^m d)^{(\beta-1)} - \sum_m a^m c^m d(bc^m d)^{\beta-2}. \quad \text{(A.7)}$$

Accordingly, the MU rule to update the element $b$ is described as

$$b = b\frac{\sum_m a^m c^m d(bc^m d)^{\beta-2}}{\sum_m c^m d(bc^m d)^{(\beta-1)}}. \quad \text{(A.8)}$$

Respecting the dimensions, this update rule can be easily extended to decompose a tensor into tensors and matrices.