POLITEHNICA UNIVERSITY OF BUCHAREST

Doctoral School of Electronics, Telecommunications and Information Technology

Decision No 971 from 08-12-2022

Ph.D. THESIS SUMMARY

Ing. Victor POPA

SEPARAREA OARBĂ A SURSELOR SONORE DIN MIXTURI BINAURALE

BLIND SOURCE SEPARATION FROM BINAURAL MIXTURES

THESIS COMMITTEE

Prof. Dr. Ing. Gheorghe BREZEANU Politehnica University of Bucharest	President
Prof. Dr. Ing. Ion MARGHESCU Politehnica University of Bucharest	PhD Supervisor
Prof. Dr. Ing. Daniela TĂRNICERIU „Gheorghe Asachi" Technical University of Iaşi	Reviewer
Prof. Dr. Ing. Corneliu RUSU Technical University of Cluj-Napoca	Reviewer
Prof. Dr. Ing. Cristian NEGRESCU Politehnica University of Bucharest	Reviewer

BUCHAREST 2022

Content

Chapter 1

Introduction

Blind source separation from mixtures is a topic addressed in a wide range of fields such as multimedia signal processing, communications, medicine, industrial engineering, and others. The concept of "blind separation" refers to the fact that both the mixing system and the original sources are unknown, although in most applications a number of general assumptions are made about the type of sources involved in the mixing system. For separation purposes, only the final mixing result is accessible, which may have one or more dimensions.

1.1 Presentation of the field of the doctoral thesis

Blind Source Separation (BSS) has been studied for decades and research is still ongoing. There are many applications that can be developed based on this technology, but the main focus is on processing audio signals, such as solving the "cocktail party" problem using artificial intelligence, extracting target speech in a noisy environment for better speech recognition results, separating each musical instrument from an audio recording for music analysis. Table 1.1 shows the main methods for blind source separation according to a series of criteria.

Table 1.1 *Classification of the main methods for blind source separation*

	Single channel $M = 1$	Multichannel N - number of sources M - number of channels/mixtures		
		Overdetermined $N < M$	Determined $N = M$	Underdetermined $N > M$
Without training	NMF	PCA All after dimension reduction	DUET ICA, IVA MNMF ILRMA	GMM and other statistical clustering, MNMF
With training	Deep Neural Networks (DNN)			

1.2 Scope of the doctoral thesis

We will start from both female and male voice signals and combine them into a series of mixtures obtained by using reverberant binaural recordings that simulate the spatial perception of the human auditory system in real rooms.

A main objective is to measure and analyze a series of customized binaural recordings and to extract the head related transfer functions and corresponding impulse responses. Using these and other responses from the literature, mixtures will be generated by simulating the simultaneous placement at various angles of a predetermined number of sound sources in a virtual audio space.

A second objective is to determine the angle of incidence of a source in two-microphone recordings applied to reverberant binaural recordings. Determining these angles will aid the separation process for sound source separation methods developed later.

The third objective is to implement and develop blind source separation methods for determined and underdetermined mixtures based on the classification of points on the short-time Fourier transforms of the mixtures. The source modeling will be as Gaussian mixtures and methods will be developed to optimize the parameters of the probability distributions so as to achieve a high-performance classification of the points on the short-time Fourier transforms.

The last objective is to implement blind source separation methods based on data nonnegativity, thus testing the separation performance for reverberant binaural mixtures.

Finally, all methods will be compared in order to determine the optimal methods for separating mixtures according to the amount of reverberation in the respective rooms.

Methods for blind separation of speech from binaural mixtures may have applications for:

- Deepening the understanding of how the human auditory system separates sound sources;
- Pre-processing of speech signals to increase intelligibility before they are used by automatic speech transcription systems or for voice command systems;
- Reducing reverberation for speech signals recorded in real rooms;
- Accentuating certain speakers (such as the moderator) in audio conferencing applications.

1.3 Content of the doctoral thesis

The second chapter will review the methods for processing mixtures in order to prepare them for the separation algorithms used in the thesis. The first phase will present the audio signals used as well as how to process them using the short-time Fourier transform with perfect reconstruction. In order to have control over how the mixtures are created, a method for measuring the head related transfer functions and the corresponding impulse responses will be developed. These will prove useful for placing sound sources in a virtual acoustic environment. At the same time, the method for signal whitening as a pre-processing step of mixtures will be presented, a method useful for algorithms based on statistical source modelling. In the last part of the

chapter, the main quantities used to evaluate the separation performance of the algorithms will be discussed.

Chapter three will investigate the method of determining the direction of arrival of a sound source in a reverberant environment using adaptive eigenvalue decomposition. The method will prove useful in determining sound source location information that will improve the performance of the methods subsequently presented.

Chapter four will investigate the expectation maximization method for estimating sound sources from mixtures, which is based on the premise of modelling the spectrograms of observed mixtures as mixtures of Gaussian distributions. The observations analyzed will be represented by simulated binaural recordings either using binaural impulse responses measured in various rooms or head related impulse responses and binaural recordings measured with the method described in Chapter 2.

Chapter five will explore and develop a new method for estimating sound sources from mixtures, similar to the one presented in Chapter 4, with the difference that the modelling is performed with Gaussian mixtures, but the model parameters are conditioned with a series of prior distributions. The resulting algorithm is based on the variational Bayes model and has a number of advantages over the one based on expectation maximization.

In chapter six, methods based on the non-negativity of the observed mixtures will be analyzed and implemented. The first part of the chapter will present the non-negative matrix factorization method for the single-channel case, discussing its applications and implementation, and the second part will investigate its extension to the multi-channel case, discussing its performance for both the determined and underdetermined mixtures case, by testing the same reverberant binaural mixtures.

Chapter seven will explore the separation method based on independent low-rank matrix analysis as a combination of independent vector analysis and non-negative matrix factorization, discussing its performance in the case of reverberant determinate mixtures.

Finally, the concluding chapter will present the comparative results of the methods as well as personal contributions and future developments.

Chapter 2

Generation and analysis of binaural mixtures

2.1 Short-time Fourier analysis with perfect reconstruction

Part of the research in sound source separation is based on the decomposition of signals corresponding to the observations into a time and frequency representation. To achieve this, the short-time Fourier transform of discrete signals using a time-sliding window is used. The described perfect reconstruction method is based on the overlap-add method.

The short-term Fourier transform is then obtained by applying the discrete Fourier transform (TFD) to each block of N elements:

$$S(k,m) = TFD\{x(n) \cdot w_a(n - mp)\} \tag{1.1}$$

In the case of the Hann window, perfect reconstruction can be achieved if the step is $N/4$, and N is a multiple of 4.

2.2 Measurement of head related impulse responses

Head Related Transfer Function (*HRTF*) and the corresponding *Head Related Impulse Response (HRIR*) play a decisive role in sound localization for humans, and an individualized set of measurements can improve localization accuracy. Measurement proves difficult in the absence of an anechoic chamber.

2.2.1 The measurement system

The measurement system consists of a rotating platform on which the listener is placed, a set of calibrated binaural microphones worn by the listener and a loudspeaker placed at a fixed distance. The acoustic impulse response is measured with the ESS method [1] using the enhanced implementation in [2]. The recorded signal, denoted by $y_{L,R}(n)$ is then processed according to the block diagram shown in figure 2.6.

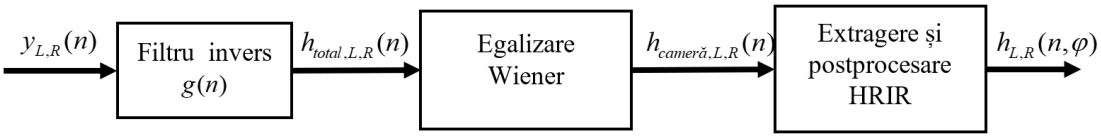

Figure 2.1 *HRIR measurement and processing chain*

2.2.2 Speaker equalization

From the measured impulse response, only the direct sound portion was selected and an equalization procedure similar to that described in [3] and improved in [4] was used. The aim is to obtain a flat frequency response for the linear part of the acoustic system composed of audio amplifier, loudspeaker and microphone. The general structure of the Wiener filter is shown in Figure 2.7, where $h(n)$ is the initial loudspeaker response, $h_B(n)$ is the impulse response of a predetermined band stop filter, $h_f(n)$ is the equalized loudspeaker response, $d(n)$ is the desired response and $e(n)$ is the error signal.

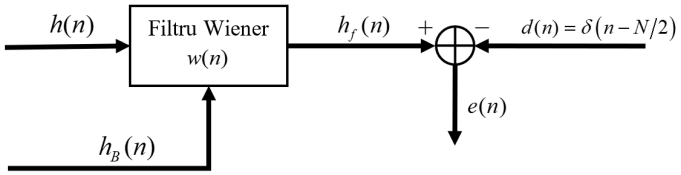

Figure 2.2 *Diagram of the loudspeaker equalization block using Wiener filtering*

2.2.3 Head related impulse response extraction and post-processing

In figure 2.9 we show the main parts of an impulse response: the direct sound, the early reflections and the reverberation part. Therefore, we could separate the direct sound corresponding to the impulse response of the head if the early reflections are delayed with respect to the direct sound by at least $t_{\min} = 4$ ms. If the distance between the sound source and the microphone is d_{dir}, we can calculate the minimum propagation path length for the first reflection, d_{refl}, by:

$$d_{refl} = d_{dir} + c \cdot t_{\min} \tag{1.2}$$

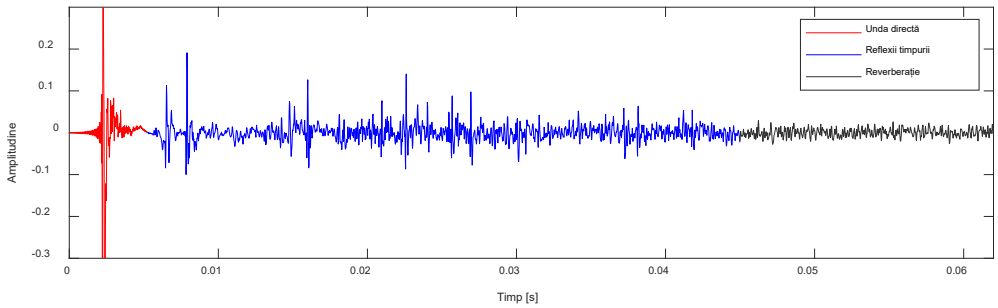

Figure 2.3 *Parts of an acoustic impulse response measured in a room*

2.2.4 Measured head related transfer functions

The test signal used to measure the HRTFs was generated with a frequency band between 20 Hz and 20 kHz and a duration of 3 seconds, taking advantage of the benefits discussed in [5]. The test signal is repeated 36 times, one repetition for each 10° angle increment, with a 5 second pause in between. A set of HRTFs measured with the described method can be seen in figure 2.11.

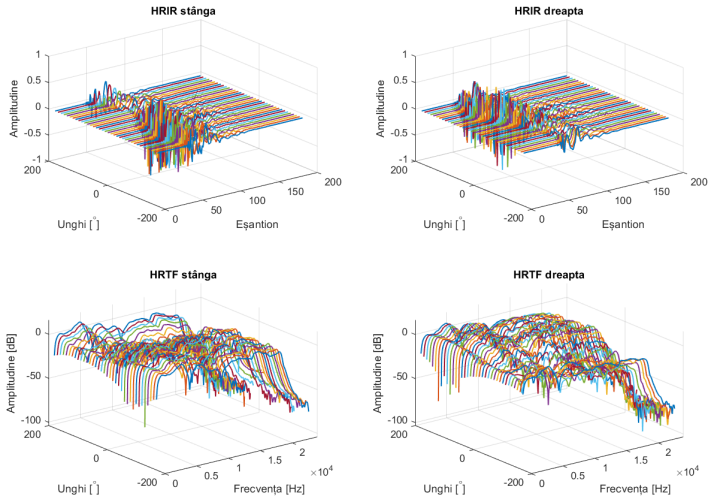

Figure 2.4 *Examples of measured HRIR and HRTF*

2.3 Analysis of head transfer functions

A first analysis will be by determining the short-time Fourier transform of the initial speech signals. The signals are recorded at a sampling frequency of 16 kHz, and a Hann window of 1024 samples and 75% overlap was used for the short-time Fourier analysis, thus guaranteeing perfect reconstruction. To obtain the interaural indices we express the ratio of the short-time Fourier transforms of the two channels:

$$\frac{X_L(\omega,t)}{X_R(\omega,t)} = 10^{\frac{\alpha(\omega,t)}{20}} \cdot e^{j\phi(\omega,t)} \tag{1.3}$$

Where $\alpha(\omega,t)$ will be the interaural level difference, ILD, expressed in dB, and $\phi(\omega,t)$ is the interaural phase difference, IPD, expressed in radians.

2.4 Evaluation of mix separation performance

Evaluating the performance of sound source separation algorithms is a self-contained problem that is addressed in the literature in an attempt to establish a common and unanimously accepted norm to describe the performance of different sound source separation algorithms and to provide a way to compare them [6].

2.4.1 Performance evaluation for signals with time-invariant gain

Source to Distortion Ratio (SDR):

$$SDR := 10\log_{10}\frac{\left\|s_{\text{țintă}}\right\|^2}{\left\|e_{interf}+e_{zg}+e_{artef}\right\|^2} \tag{1.4}$$

Source to Interference Ratio (SIR):

$$SIR := 10\log_{10}\frac{\left\|s_{\text{țintă}}\right\|^2}{\left\|e_{interf}\right\|^2} \tag{1.5}$$

Source to Noise Ratio (SNR):

$$SNR := 10\log_{10}\frac{\left\|s_{\text{țintă}}+e_{interf}\right\|^2}{\left\|e_{zg}\right\|^2} \tag{1.6}$$

Source to Artifacts Ratio (SAR):

$$SAR := 10\log_{10}\frac{\left\|s_{\text{țintă}}+e_{interf}+e_{zg}\right\|^2}{\left\|e_{artef}\right\|^2} \tag{1.7}$$

2.4.2 Performance evaluation for signals with distortions introduced by time-invariant filters

If time-invariant filtering is allowed as a distortion, then the target signal $s_{\text{țintă}}$ is no longer an amplified (or attenuated) version of the original signal s_i, but the result of filtering. In other words, $s_{\text{țintă}}$ is obtained by summing a set of delayed variants of the original signal with certain amplitudes expressed by:

$$s_{\text{țintă}}(n) = \sum_{k=0}^{L-1} h(k)\cdot s_i(n-k) \tag{1.8}$$

2.5 Conclusions

Following discussions on the implementation of the short-time Fourier transform with perfect reconstruction we chose as parameters for the overlap-add method the Hann window with 75% overlap between consecutive frames.

A fast measurement method for extracting HRIRs and their associated HRTFs from impulse responses measured in reverberant rooms was presented. From preliminary tests, spatial cues were preserved and the resulting impulse responses are found to be very similar to those measured under anechoic conditions.

Chapter 3

Determining the direction of arrival of a sound source using adaptive eigenvalue decomposition

3.1 The eigenvalue decomposition method

This method belongs to the category of blind channel identification algorithms for SIMO acoustic systems, and is a special case for one source and two microphones.

3.1.1 System of equations

We will consider the ideal case where no noise is present and exploit the commutativity property of convolution [7]:

$$(y_1 * h_2)(n) = (s * h_1 * h_2)(n) = (y_2 * h_1)(n) \tag{1.9}$$

$$(y_1 * h_2)(n) - (y_2 * h_1)(n) = 0 \tag{1.10}$$

3.1.2 Adaptive algorithm

The final block diagram of the adaptive algorithm can be seen in figure 3.3.

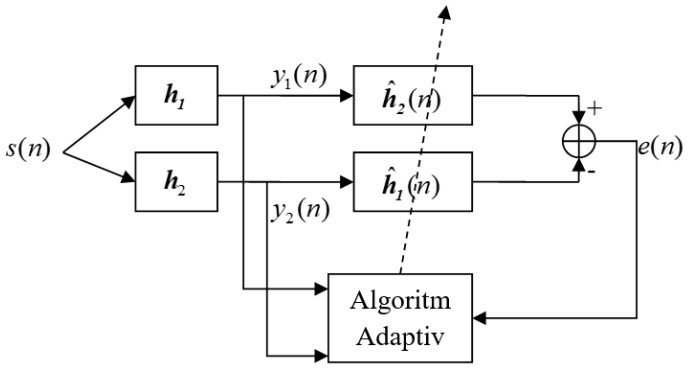

Figure 3.1 *Block diagram of the adaptive algorithm*

The algorithm can be described as follows:

Initialization:

N

$\hat{\mathbf{h}}_2(0) = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}^T$, with 1 on position $N/2$

$\hat{\mathbf{h}}_1(0) = \begin{bmatrix} 0 & 0 & \dots & 0 \end{bmatrix}^T$

$\hat{\mathbf{h}}(0) = \begin{bmatrix} \hat{\mathbf{h}}_2(0)^T & -\hat{\mathbf{h}}_1(0)^T \end{bmatrix}^T$

For $n = 0, \dots, N-1$

$e(n) = \hat{\mathbf{h}}(n)^T \cdot \mathbf{y}(n)$

$$\hat{\mathbf{h}}(n+1) = \frac{\hat{\mathbf{h}}(n) - \mu e(n)\mathbf{y}(n)}{\left\| \hat{\mathbf{h}}(n) - \mu e(n)\mathbf{y}(n) \right\|}$$

After convergence:

$$\tau_{12} = \tau_2 - \tau_1 = \left(N/2 - \arg\left\{ \min\left\{ \hat{\mathbf{h}}_1(n) \right\} \right\} \right) / F_s$$

3.2 Experimental results

To test the algorithm the following configuration was used:

- 2 cardioid microphones placed at a distance $d = 0{,}1$ m;
- 2 sound sources placed at angles $\theta = 90°$ and $\theta = 135°$, and distance from sources to microphones $r = 2{,}3$ m;
- Reverberant room with $T_{60} \cong 0.53s$;
- A voice signal was reproduced and recorded with $F_s = 48$ kHz.

To test the detection of multiple sources we will further apply the algorithm for mixtures of 2, 3, 4 and 5 sources in binaural mixtures made in the room described in the previous chapter. The angles chosen for each mixture are:

- Mix of 2 sources: $-60°$ and $60°$;
- Mix of 3 sources: $-60°$, $0°$ and $60°$;
- Mix of 4 sources: $-90°$, $-30°$, $30°$ and $90°$;
- Mix of 5 sources: $-90°$, $-45°$, $0°$, $45°$ and $90°$;

After applying the algorithm, the histogram of angles was determined to better visualize the detected angles. The results can be seen in figure 3.15.

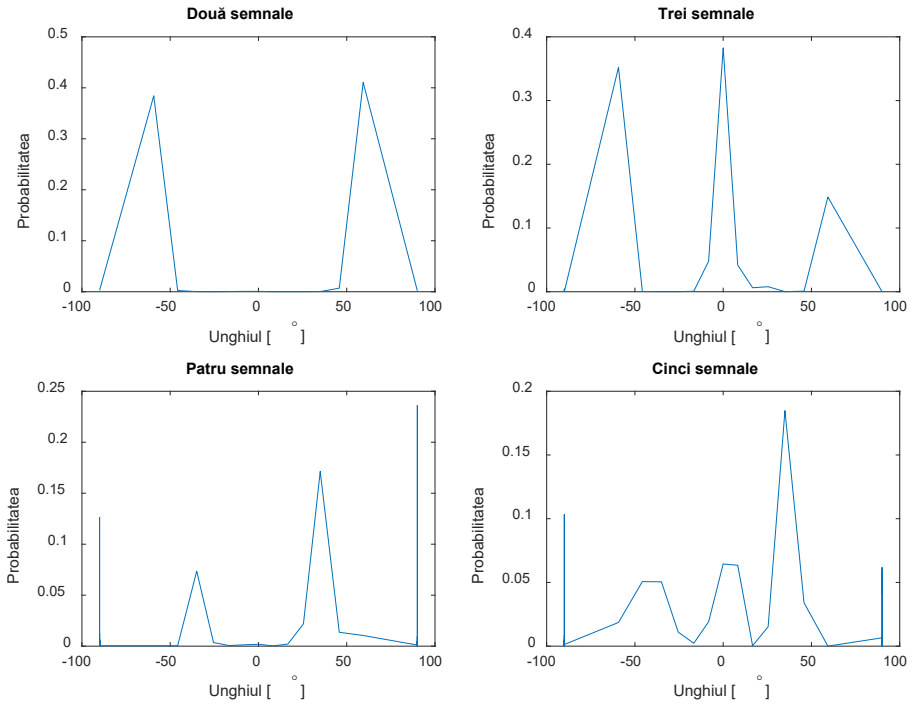

Figure 3.2 *Histogram of detected angles using the adaptive eigenvalue decomposition algorithm for binaural mixtures of 2,3,4 and 5 sources*

3.3 Conclusions

The adaptive algorithm based on eigenvalue decomposition gives satisfactory results when only one source is present at a time and is robust even in the presence of strong reverberation. As a comparison, the GCC algorithm was applied for the same signal, but the correct detection was obtained starting with a window size of 5000 samples.

The algorithm tends to prefer directional noise signals because they better satisfy the variety condition, leading to the conclusion that a solution for detecting speech signals would be to pre-whiten them. In the presence of two sources it converges to the one that best meets the algorithm's conditions, i.e. it focuses to the channel with the smaller eigenvalue.

When applying the algorithm to binaural mixtures of two or more sources, we observed that the algorithm correctly detects all the angles involved and proves to be a good measure for determining the number of sources and their angles. The method will be useful in the preprocessing and initialization step of the algorithms for separating sound sources from binaural mixtures that will be presented in the following chapters.

Chapter 4

Expectation maximization algorithm for sound source separation from binaural mixtures

The main idea is to model the *Interaural Phase Difference (IPD*), *Interaural Level Difference (ILD*) as in [8] and the mixing vectors as in [9] at each time-frequency unit with Gaussian mixture models for each source and each observation type, similar to the algorithm in [10].

4.1 Modelling spectrogram values using binaural cues

The interaural spectrogram, i.e. the ratio of the spectrograms corresponding to the left and right channels, is determined as follows:

$$X_L(\omega,t) = TFTS\{x_L(n)\} \tag{1.11}$$

$$X_R(\omega,t) = TFTS\{x_R(n)\} \tag{1.12}$$

$$\frac{X_L(\omega,t)}{X_R(\omega,t)} = 10^{\frac{\alpha(\omega,t)}{20}} \cdot e^{j\phi(\omega,t)}, \tag{1.13}$$

where $X_L(\omega,t)$ and $X_R(\omega,t)$ are the short-time Fourier transforms for the two signals recorded at each discrete frequency ω and discrete time t.

We model the observations for ILD and IPD as Gaussian distributions with mean $\mu_i(\omega)$ and variance $\eta_i^2(\omega)$ for $\alpha(\omega,t)$, respectively, mean $\xi_{i,\tau}(\omega)$ and variance $\sigma_{i,\tau}^2(\omega)$ for $\hat{\phi}(\omega,t;\tau)$ [8] in the following way:

$$\begin{aligned} p\left(\alpha(\omega,t) \mid \mu_i(\omega),\eta_i(\omega)\right) &= \mathcal{N}\left(\alpha(\omega,t) \mid \mu_i(\omega),\eta_i^2(\omega)\right) \\ \mathcal{N}\left(\alpha(\omega,t) \mid \mu_i(\omega),\eta_i^2(\omega)\right) &= \frac{1}{\eta_i(\omega)\sqrt{2\pi}} \exp\left(-\frac{\left(\alpha(\omega,t)-\mu_i(\omega)\right)^2}{2\cdot\eta_i^2(\omega)}\right) \end{aligned} \tag{1.14}$$

$$p\left(\hat{\phi}(\omega,t;\tau)\,|\,\xi_{i,\tau}(\omega),\sigma^2_{i,\tau}(\omega)\right)=\mathcal{N}\left(\hat{\phi}(\omega,t;\tau)\,|\,\xi_{i,\tau}(\omega),\sigma^2_{i,\tau}(\omega)\right)$$

$$\mathcal{N}\left(\hat{\phi}(\omega,t;\tau)\,|\,\xi_{i,\tau}(\omega),\sigma^2_{i,\tau}(\omega)\right)=\frac{1}{\xi_{i,\tau}(\omega)\sqrt{2\pi}}\exp\left(-\frac{\left(\hat{\phi}(\omega,t;\tau)-\xi_{i,\tau}(\omega)\right)^2}{2\cdot\sigma^2_{i,\tau}(\omega)}\right)$$

(1.15)

4.2 Modelling spectrogram values using mixing vectors

To model such vectors for each source, we follow the idea described in [11] and we use the following multidimensional complex Gaussian probability density:

$$\begin{aligned}p\left(\mathbf{X}(\omega,t)\,|\,\mathbf{a}_i(\omega),\gamma_i(\omega)\right)&=\mathcal{N}\left(\mathbf{X}(\omega,t)\,|\,\mathbf{a}_i(\omega),\gamma_i^2(\omega)\right)\\&=\frac{1}{\left(\pi\cdot\gamma_i^2(\omega)\right)^2}\exp\left(-\frac{\left\|\mathbf{X}(\omega,t)-\left(\mathbf{a}_i^H(\omega)\cdot\mathbf{X}(\omega,t)\right)\cdot\mathbf{a}_i(\omega)\right\|^2}{\gamma_i^2(\omega)}\right)\end{aligned}\quad(1.16)$$

where $\mathbf{a}_i$ is a two-dimensional-complex mean and is a centroid with unit norm $\|\mathbf{a}_i\|=1$, and γ_i^2 is the variance.

4.3 Combining the binaural model with the vector model in an implementation of the expectation maximization algorithm

We can thus express the log-likelihood function of the model:

$$\mathcal{L}(\Theta)=\sum_{\omega,t}\ln\left(p\left(\alpha(\omega,t),\phi(\omega,t),\mathbf{X}(\omega,t)\middle|\Theta\right)\right)\quad(1.17)$$

4.3.1 The expectation maximization algorithm

The expectation maximization algorithm involves the iterative determination of the optimal parameters that increase at each iteration the log-likelihood that a time-frequency point belongs to a source and is divided into two steps:

1. **The expectation step (E)** in which the expectation of the hidden random variable $z_{i,\tau}(\omega,t)$ is calculated based on the observations and the estimated parameters $\hat{\Theta}$, equivalent to model evaluation;

$$
\begin{aligned}
q_{i,\tau}(\omega,t) &= p\left(z_{i,\tau}(\omega,t)\middle|\alpha(\omega,t),\phi(\omega,t),\mathbf{X}(\omega,t),\hat{\Theta}\right) \\
&\propto p\left(z_{i,\tau}(\omega,t),\alpha(\omega,t),\phi(\omega,t),\mathbf{X}(\omega,t)\middle|\hat{\Theta}\right) \\
&= \psi_{i,\tau}(\omega)\cdot\mathcal{N}\left(\alpha(\omega,t)\,|\,\mu_i(\omega),\eta_i^2(\omega)\right) \\
&\cdot\mathcal{N}\left(\hat{\phi}(\omega,t;\tau)\,|\,\xi_{i,\tau}(\omega),\sigma_{i,\tau}^2(\omega)\right) \\
&\cdot\mathcal{N}\left(\mathbf{X}(\omega,t)\,|\,\mathbf{a}_i(\omega),\gamma_i^2(\omega)\right).
\end{aligned}
\tag{1.18}
$$

2. **Maximization step (M)** in which Q is maximized with respect to Θ based on the expectation value of $z_{i,\tau}(\omega,t)$ determined in the previous step. To simplify the notations we define the weighted averaging operator:

$$
\mathcal{M}_{a,b}\{x\} \overset{Def}{=} \frac{\sum_{a,b} x\cdot q_{i,\tau}(\omega,t)}{\sum_{a,b} q_{i,\tau}(\omega,t)}\ , \tag{1.19}
$$

Using this operator we can derive the update relations for the model parameters at the **maximization step**:

$$
\mu_i(\omega) = \mathcal{M}_{t,\tau}\{\alpha(\omega,t)\} \tag{1.20}
$$

$$
\eta_i^2(\omega) = \mathcal{M}_{t,\tau}\left\{\left(\alpha(\omega,t)-\mu_i(\omega)\right)^2\right\} \tag{1.21}
$$

$$
\xi_{i,\tau}(\omega) = \mathcal{M}_t\left\{\hat{\phi}(\omega,t;\tau)\right\} \tag{1.22}
$$

$$
\sigma_{i,\tau}^2(\omega) = \mathcal{M}_t\left\{\left(\hat{\phi}(\omega,t;\tau)-\xi_{i,\tau}(\omega)\right)^2\right\} \tag{1.23}
$$

$$
\mathbf{R}_i(\omega) = \sum_{t,\tau} q_{i,\tau}(\omega,t)\cdot\mathbf{X}(\omega,t)\cdot\mathbf{X}^H(\omega,t) \tag{1.24}
$$

$$
\gamma_i^2(\omega) = \mathcal{M}_{t,\tau}\left\{\left\|\mathbf{X}(\omega,t)-\left(\mathbf{a}_i^H(\omega)\cdot\mathbf{X}(\omega,t)\right)\cdot\mathbf{a}_i(\omega)\right\|^2\right\} \tag{1.25}
$$

$$
\psi_{i,\tau}(\omega) = \frac{1}{T}\sum_t q_{i,\tau}(\omega,t) \tag{1.26}
$$

4.3.2 Frequency permutation alignment using posterior probability based on mixing vectors

Taking into account the modelling given by binaural indices and mixing vectors, a posterior alignment of the source order can be performed, as described in [9]. This involves determining the posterior probability:

$$
\begin{aligned}
q_{k,\tau}(\omega,t) &= p\left(z_{k,\tau}(\omega,t)\middle|\alpha(\omega,t),\phi(\omega,t),\mathbf{X}(\omega,t),\hat{\Theta}\right) \\
&= p\left(C_k\middle|\alpha(\omega,t),\phi(\omega,t),\mathbf{X}(\omega,t),\hat{\Theta}\right)
\end{aligned}
\tag{1.27}
$$

which suggests that the source with the number k, described by the class C_k, is dominant within the observation set $\Psi(\omega,t) \overset{Not}{=} \{\alpha(\omega,t), \phi(\omega,t), \mathbf{X}(\omega,t)\}$.

4.3.3 Alignment of frequency permutation by appropriate initialization of parameters

We consider an alternative approach using binaural index information, described and implemented in [10] and [12]. Since the expectation maximization algorithm can be initialized either from the expectation step or from the maximization step, and also there is usually no prior information about the mixing system, we first perform the mask initialization for the spectrogram and then estimate the initial values of $\mathbf{a}_i(\omega)$ and $\gamma_i(\omega)$ based on the masked spectrogram [10]. The block diagram of the expectation maximization algorithm is shown in Figure 4.6.

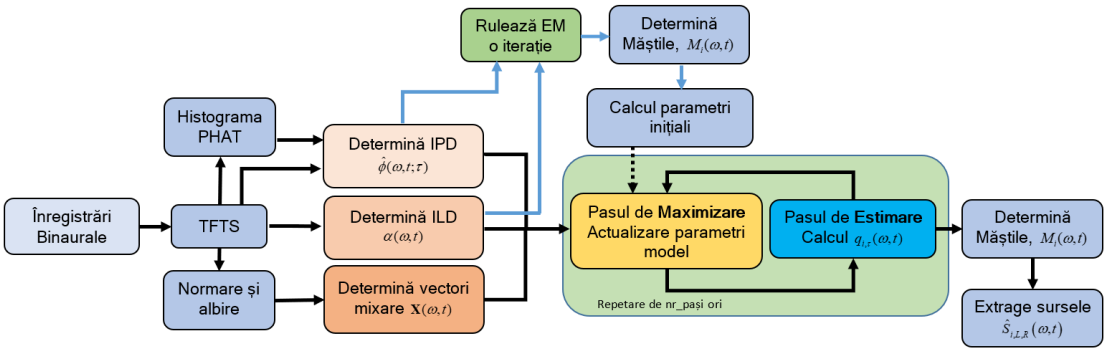

Figure 4.1 *Block diagram of the expectation maximization algorithm with permutation alignment.*

4.4 Experimental results

The first set of experiments was conducted using the set of speech recordings from the TIMIT database [13]. To perform the set of experiments, 15 uttered sentences were randomly chosen from the database and their duration was cut to 2.5 s. In order to investigate the language independence of the algorithm, 7 uttered sentences with female and male speakers were also used. To obtain the binaural mixtures, convolution between binaural impulse responses and recordings was then performed. The first set of binaural impulse responses measured in the rooms are those determined by Hummersone [14]. In addition to these we also added the head related impulse responses, denoted later with HRIR, and a room in the Faculty of Electronics, Telecommunications and Information Technology, UPB, denoted with E. For the experiment with 3 randomly chosen voices in English or Romanian, placed at the angles $-60°$, $0°$ and $60°$, using the 7 types of rooms and then averaging the SDR over all the results we obtain the performance shown in Figure 4.15. The Romanian

signals were filtered with a bank of parametric filters so that on average the spectrum is similar to the English one.

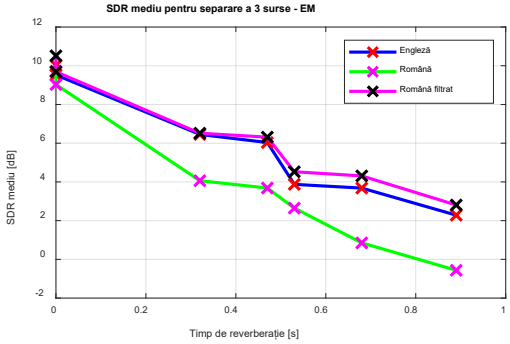

Figure 4.2 *Average separation performance of three sources as a function of reverberation time - expectation maximization (EM) algorithm.*

4.5 Conclusions

In this chapter, an implementation of the expectation maximization algorithm based on the modelling of short-time Fourier transforms of binaural records has been presented. In order to avoid phase ambiguities leading to spatial frequency aliasing, a method of associating the delay time difference with a discrete set of values was presented. The observations were then modeled using Gaussian distributions and an optimization algorithm based on expectation maximization was developed. In order to avoid the problem of permutation of source order between different frequencies, a method of initializing responsibilities based on a combination of the PHAT histogram and the adaptive eigenvalue decomposition algorithm presented in the previous chapter was presented.

Chapter 5

Variational Bayes algorithm for separating sound sources from binaural mixtures

Unlike the expectation maximization algorithm described in the previous section, Bayesian inference aims to fully model the probability distribution of the posterior parameters described in the expectation maximization algorithm framework by means, variances and mixing weights.

5.1 General formulation of variational inference

Given the data set $\mathbf{x}$, the hidden data set $\mathbf{z}$ and the parameters of the statistical model characterizing the observations and the hidden data, Θ, we want to estimate the probability $p(\mathbf{x},\mathbf{z},\Theta)$. For the available data, we can define the marginal log-likelihood also called model proof as:

$$\mathcal{L}_{\Theta}(\mathbf{x}) = \ln \iint p(\mathbf{x},\mathbf{z},\Theta)\, d\mathbf{z} d\Theta . \tag{1.28}$$

To introduce parameter dependence we define a posterior variational probability distribution $q(\mathbf{z},\Theta|\mathbf{x},\theta)$ that approximates the true posterior $p(\mathbf{z},\Theta|\mathbf{x})$, where θ refers to a set of ***hyperparameters*** that model the parameter distributions`. We can thus *define the log-likelihood variational* dependence of the set of hyperparameters:

$$\begin{aligned}\mathcal{L}_{\theta}(\mathbf{x}) &= \iint q(\mathbf{z},\Theta|\mathbf{x},\theta) \ln\left(\frac{p(\mathbf{x},\mathbf{z},\Theta)}{q(\mathbf{z},\Theta|\mathbf{x},\theta)}\right) d\mathbf{z} d\Theta + \\ &+ \iint q(\mathbf{z},\Theta|\mathbf{x},\theta) \ln\left(\frac{q(\mathbf{z},\Theta|\mathbf{x},\theta)}{p(\mathbf{z},\Theta|\mathbf{x})}\right) d\mathbf{z} d\Theta \\ &= L_{\inf}(\mathbf{x}) + D_{KL}(q\|p)\end{aligned} \tag{1.29}$$

where $L_{\inf}(\mathbf{x})$ is the same lower bound as in (5.4), and $D_{KL}(q\|p)$ is the Kullback-Leibler divergence between the posterior variational probability density $q(\mathbf{z},\Theta|\mathbf{x},\theta)$ and the true posterior probability $p(\mathbf{z},\Theta|\mathbf{x})$.

5.2 Statistical modelling of the observation set

5.2.1 Modelling the observations extracted from the spectrograms

The three sets of observations, $\mathbf{X}(\omega,t)$, $\alpha(\omega,t)$ and $\phi(\omega,t)$, with $\phi(\omega,t)$ transformed to $\hat{\phi}(\omega,t;\tau)$, will be combined into a joint probability, where T is the total number of time frames, Ω is the number of frequency channels, I is the number of sources, and Υ is the number of interaural time differences. The probability distribution of the latent variables conditioned by the mixing coefficients will be expressed as [15]:

$$p(\mathbf{Z}\,|\,\boldsymbol{\gamma}_x)=\prod_{t=1}^{T}\prod_{i=1}^{I}\prod_{\tau=1}^{\Upsilon}\gamma_{x,i}^{z_{t,i,\tau}} \tag{1.30}$$

The distribution of the set of observations in the spectrograms, conditioned by the latent variables will be expressed as a product of complex Gaussian distributions:

$$p(\mathbf{X}\,|\,\mathbf{Z},\boldsymbol{\mu}_x,\boldsymbol{\lambda}_x)=\prod_{t=1}^{T}\prod_{i=1}^{I}\prod_{\tau=1}^{\Upsilon}\mathcal{N}_c\left(\mathbf{x}_t\,|\,\boldsymbol{\mu}_{x,i},\lambda_{x,i}^{-1}\right)^{z_{t,i,\tau}}, \tag{1.31}$$

The modeling of the $\boldsymbol{\alpha}=\{\alpha_1,\ldots,\alpha_T\}$ interaural level difference is similarly done with a mixture of Gaussians, with mean $\boldsymbol{\mu}_\alpha=\{\mu_{\alpha,i}\}$ and precision $\boldsymbol{\lambda}_\alpha=\{\lambda_{\alpha,i}\}$, having mixing coefficients $\boldsymbol{\gamma}_\alpha=\{\gamma_{\alpha,i}\}$:

$$p(\mathbf{Z}\,|\,\boldsymbol{\gamma}_\alpha)=\prod_{t=1}^{T}\prod_{i=1}^{I}\prod_{\tau=1}^{\Upsilon}\gamma_{\alpha,i}^{z_{t,i,\tau}} \tag{1.32}$$

$$p(\boldsymbol{\alpha}\,|\,\mathbf{Z},\boldsymbol{\mu}_\alpha,\boldsymbol{\lambda}_\alpha)=\prod_{t=1}^{T}\prod_{i=1}^{I}\prod_{\tau=1}^{\Upsilon}\mathcal{N}\left(\alpha_t\,|\,\mu_{\alpha,i},\lambda_{\alpha,i}^{-1}\right)^{z_{t,i,\tau}} \tag{1.33}$$

Observations corresponding to the interaural phase difference will be modelled similarly to the interaural level difference using a Gaussian mixture:

$$p\left(\mathbf{Z}\,|\,\boldsymbol{\gamma}_{\hat{\phi}}\right)=\prod_{t=1}^{T}\prod_{i=1}^{I}\prod_{\tau=1}^{\Upsilon}\gamma_{\hat{\phi},i,\tau}^{z_{t,i,\tau}} \tag{1.34}$$

$$p(\hat{\boldsymbol{\Phi}}\,|\,\mathbf{Z},\boldsymbol{\mu}_{\hat{\phi}},\boldsymbol{\lambda}_{\hat{\phi}})=\prod_{t=1}^{T}\prod_{i=1}^{I}\prod_{\tau=1}^{\Upsilon}\mathcal{N}\left(\hat{\phi}_{t,\tau}\,|\,\mu_{\hat{\phi},i,\tau},\lambda_{\hat{\phi},i,\tau}^{-1}\right)^{z_{t,i,\tau}} \tag{1.35}$$

5.2.2 Prior distributions

The mixing coefficients, having the structure of a multinomial distribution, will have a Dirichlet distribution as a prior of the form:

$$p(\boldsymbol{\gamma})=Dir(\boldsymbol{\gamma}\,|\,\mathbf{a}_0)=B(\mathbf{a}_0)\prod_{i=1}^{I}\prod_{\tau=1}^{\Upsilon}\gamma_{i,\tau}^{a_0-1}, \tag{1.36}$$

For the mean and precision of Gaussian distributions we use Gauss-Gamma prior distributions of the form:

$$p(\boldsymbol{\mu}_x,\boldsymbol{\lambda}_x)=p(\boldsymbol{\mu}_x\,|\,\boldsymbol{\lambda}_x)p(\boldsymbol{\lambda}_x)=\prod_{i=1}^{I}\mathcal{N}_c\left(\boldsymbol{\mu}_{x,i}\,|\,\mathbf{m}_{x,0},(\lambda_{x,i}\beta_{x,0}\mathbf{I})^{-1}\right)\mathcal{G}(\lambda_{x,i}\,|\,b_{x,0},c_{x,0}) \tag{1.37}$$

$$p(\boldsymbol{\mu}_\alpha, \boldsymbol{\lambda}_\alpha) = p(\boldsymbol{\mu}_\alpha \mid \boldsymbol{\lambda}_\alpha) p(\boldsymbol{\lambda}_\alpha) = \prod_{i=1}^{I} \mathcal{N}\left(\mu_{\alpha,i} \mid m_{\alpha,0}, (\lambda_{\alpha,i}\beta_{\alpha,0})^{-1}\right) \mathcal{G}(\lambda_{\alpha,i} \mid b_{\alpha,0}, c_{\alpha,0}) \quad (1.38)$$

$$\begin{aligned} p(\boldsymbol{\mu}_{\hat{\phi}}, \boldsymbol{\lambda}_{\hat{\phi}}) &= p(\boldsymbol{\mu}_{\hat{\phi}} \mid \boldsymbol{\lambda}_{\hat{\phi}}) p(\boldsymbol{\lambda}_{\hat{\phi}}) \\ &= \prod_{i=1}^{I} \prod_{\tau=1}^{\Upsilon} \mathcal{N}\left(\mu_{\hat{\phi},i,\tau} \mid m_{\hat{\phi},0}, (\lambda_{\hat{\phi},i,\tau}\beta_{\hat{\phi},0})^{-1}\right) \mathcal{G}(\lambda_{\hat{\phi},i,\tau} \mid b_{\hat{\phi},0}, c_{\hat{\phi},0}) \end{aligned} \quad (1.39)$$

5.3 Optimization process

Optimization of model parameter values involves maximizing the log-likelihood similar to the discussion in the first part of the chapter, using (5.1).

Updating the responsibilities

By responsibilities we mean the probability that a time-frequency point belongs to a certain source and comes from a certain direction. From the equation (5.42) we observe that the a priori distribution is averaged over $z_{t,i,\tau}$ and can be expressed by:

$$\ln q^{opt}(\mathbf{Z}) = \sum_{t=1}^{T} \sum_{i=1}^{I} \sum_{\tau=1}^{\Upsilon} z_{t,i,\tau} \ln \rho_{t,i,\tau} + const\,, \quad (1.40)$$

where

$$\begin{aligned} \ln \rho_{t,i,\tau} = {} & \mathrm{E}_{\gamma_{i,\tau}}\left[\ln \gamma_{i,\tau}\right] - \ln \pi - \ln 2\pi + \mathrm{E}_{\lambda_{x,i}}\left[\ln \lambda_{x,i}\right] + \frac{1}{2}\mathrm{E}_{\lambda_{\alpha,i}}\left[\ln \lambda_{\alpha,i}\right] + \\ & + \frac{1}{2}\mathrm{E}_{\lambda_{\hat{\phi},i,\tau}}\left[\ln \lambda_{\hat{\phi},i,\tau}\right] - \mathrm{E}_{\boldsymbol{\mu}_{x,i},\lambda_{x,i}}\left[\lambda_{x,i} \ \|\mathbf{x}_t - \left(\boldsymbol{\mu}_{x,i}^H \mathbf{x}_t\right)\boldsymbol{\mu}_{x,i} \ \|^2\right] - \\ & - \frac{1}{2}\mathrm{E}_{\boldsymbol{\mu}_{\alpha,i},\lambda_{\alpha,i}}\left[\lambda_{\alpha,i}\left(\alpha_t - \mu_{\alpha,i}\right)^2\right] - \frac{1}{2}\mathrm{E}_{\boldsymbol{\mu}_{\hat{\phi},i,\tau},\lambda_{\hat{\phi},i,\tau}}\left[\lambda_{\hat{\phi},i,\tau}\left(\hat{\phi}_{t,\tau} - \mu_{\hat{\phi},i,\tau}\right)^2\right] \end{aligned} \quad (1.41)$$

Updating the hyperparameters

Tthe hyperparameter update for the mixing coefficients is done by:

$$a_{i,\tau} = \sum_{t=1}^{T} r_{t,i,\tau} + a_0 \quad (1.42)$$

The hyperparameters for the mixing vector observations are obtained:

$$\begin{aligned} \boldsymbol{\beta}_{x,i} &= \beta_{x_0}\mathbf{I} - \sum_{t=1}^{T}\sum_{\tau=1}^{\Upsilon} r_{t,i,\tau}\left(\mathbf{x}_t \mathbf{x}_t^H\right) \\ \mathbf{m}_{x,i} &= \boldsymbol{\beta}_{x,i}^{-1}\beta_{x_0}\mathbf{I}\mathbf{m}_{x,0} \\ b_{x,i} &= b_{x,0} + \sum_{t=1}^{T}\sum_{\tau=1}^{\Upsilon} r_{t,i,\tau} \\ c_{x,i} &= c_{x,0} + \sum_{t=1}^{T}\sum_{\tau=1}^{\Upsilon} r_{t,i,\tau}\left(\mathbf{x}_t^H \mathbf{x}_t\right) + \mathbf{m}_{x,0}^H \beta_{x,0}\mathbf{I}\mathbf{m}_{x,0} - \mathbf{m}_{x,i}^H \boldsymbol{\beta}_{x,i}\mathbf{m}_{x,i} \end{aligned} \quad (1.43)$$

Similarly, we determine the hyperparameters for ILD and IPD:

$$
\begin{aligned}
\beta_{\alpha,i} &= \beta_{\alpha_0} + \sum_{t=1}^{T}\sum_{\tau=1}^{\Upsilon} r_{t,i,\tau} \\
m_{\alpha,i} &= \beta_{\alpha,i}^{-1}\left(\sum_{t=1}^{T}\sum_{\tau=1}^{\Upsilon} r_{t,i,\tau}\alpha_t + m_{\alpha,0}\beta_{\alpha,0}\right) \\
b_{\alpha,i} &= b_{\alpha,0} + \frac{1}{2}\sum_{t=1}^{T}\sum_{\tau=1}^{\Upsilon} r_{t,i,\tau} \\
c_{\alpha,i} &= c_{\alpha,0} + \frac{1}{2}\sum_{t=1}^{T}\sum_{\tau=1}^{\Upsilon} r_{t,i,\tau}\alpha_t^2 + \frac{1}{2}m_{\alpha,0}^2\beta_{\alpha,0} - \frac{1}{2}m_{\alpha,i}^2\beta_{\alpha,i}
\end{aligned}
\tag{1.44}
$$

Calculation of the lower bound $L_{\inf}\left(\mathbf{X},\boldsymbol{\alpha},\hat{\boldsymbol{\Phi}}\right)$:

$$
\begin{aligned}
L_{\inf}\left(\mathbf{X},\boldsymbol{\alpha},\hat{\boldsymbol{\Phi}}\right) &= \sum_{\mathbf{Z}}\int q(\mathbf{Z},\boldsymbol{\Theta})\ln\left(\frac{p\left(\mathbf{X},\boldsymbol{\alpha},\hat{\boldsymbol{\Phi}},\mathbf{Z},\boldsymbol{\Theta}\right)}{q\left(\mathbf{Z},\boldsymbol{\Theta}\right)}\right)d\Theta \\
&= \mathrm{E}\left[\ln p\left(\mathbf{X},\boldsymbol{\alpha},\hat{\boldsymbol{\Phi}},\mathbf{Z},\boldsymbol{\Theta}\right)\right] - \mathrm{E}\left[\ln q\left(\mathbf{Z},\boldsymbol{\Theta}\right)\right]
\end{aligned}
\tag{1.45}
$$

5.4 Variational Bayes algorithm

The algorithm is shown in figure 5.1 as a block diagram.

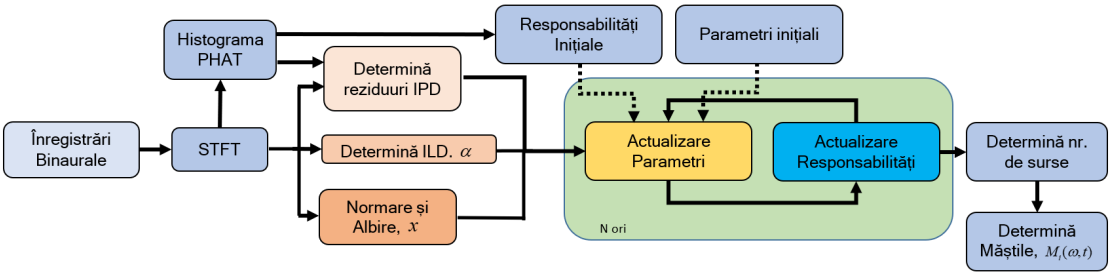

Figure 5.1. *Variational Bayes algorithm*

5.5 Experimental results

We run the variational Bayes algorithm for the 7 room types by randomly choosing sources from the given set, for a number of initial sources from 2 to 5 to determine the accuracy of correctly identifying the number of sources in the mixtures for each room type. The results are presented in Table 5.1.

Table 5.1 *Accuracy [%] of the determination of the number of sources in the mixture as a function of chamber and number of initial sources*

Accuracy [%]	Camera						
No. sources	N	HRIR	A	B	C	D	E
2	100	100	98	97	92	85	95
3	100	100	95	93	86	76	91
4	93	96	81	76	69	74	73
5	62	68	53	45	38	35	42

Re-running the experiment on two-source mixtures for the variational Bayes algorithm we obtain the performance as a function of reverberation time from figure 5.6. The results for three-source mixtures are shown in figure 5.7.

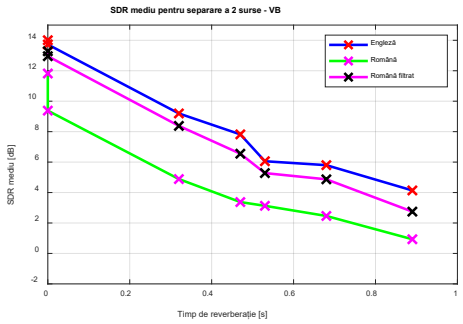

Figure 5.2 *Average separation performance of two sources as a function of reverberation time - variational Bayes (VB) algorithm.*

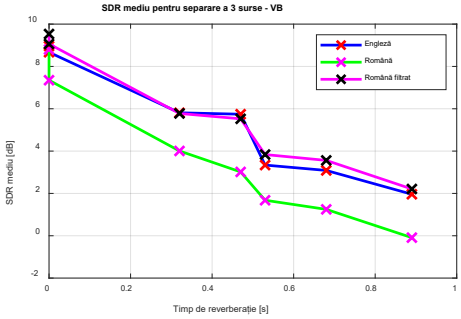

Figure 5.3 *Average separation performance of three sources as a function of reverberation time - variational Bayes (VB) algorithm.*

5.6 Conclusions

A method for separating reverberant speech from binaural mixtures based on variational Bayes theory was presented. The proposed algorithm benefits from the fact that Bayesian inference is less sensitive to improper initializations and also automatically determines the number of sources, which can be an advantage in real recordings. The convergence of the algorithm was studied by determining the variational lower bound of the log-likelihood function and the influence of the angle between source signals on the degree of separation was investigated.

Chapter 6

Blind source separation using non-negative matrix factorization

6.1 Single-channel non-negative matrix factorization

Nonnegative matrix factorization (NMF) is a commonly used method of dimensional reduction of a quantity of data and is used to represent non-negative data [16]. Starting from a matrix $\mathbf{V}$ of size $\Omega \times T$ which will represent a form of the signal spectrogram with Ω the number of points in frequency and T the number of points on the discrete time axis, NMF is used to determine the factorization:

$$\mathbf{V} \approx \mathbf{WH}, \tag{1.46}$$

Solving the equation (6.1) involves determining the matrices $\mathbf{W}$ and $\mathbf{H}$ after minimization:

$$\min_{\mathbf{W},\mathbf{H}\geq 0} D\left(\mathbf{V}\middle|\mathbf{WH}\right), \tag{1.47}$$

where $D\left(\mathbf{V}\middle|\mathbf{WH}\right)$ is a cost function.

In the following we will focus specifically on the Itakura-Saito divergence (IS) as a cost function for NMF given by:

$$d_{IS}\left(x\middle|y\right) = \frac{x}{y} - \ln\frac{x}{y} - 1 \tag{1.48}$$

6.1.1 Properties of the Itakura-Saito divergence

Link to the β divergence

IS divergence is a case of the β divergence which is defined as:

$$d_{\beta}\left(x\middle|y\right) \overset{def}{=} \begin{cases} \dfrac{1}{\beta\left(\beta-1\right)}\left[x^{\beta} + \left(\beta-1\right)y^{\beta} - \beta x y^{\beta-1}\right] & , \beta \in \mathbb{R}\setminus\{0,1\} \\ x\ln\dfrac{x}{y} + \left(y-x\right) & , \beta = 1 \\ \dfrac{x}{y} - \ln\dfrac{x}{y} - 1 & , \beta = 0 \end{cases} \tag{1.49}$$

Scale invariance property

For any value of β the following property occurs:

$$d_\beta\left(\gamma x|\gamma y\right)=\gamma^\beta d_\beta\left(x|y\right) \tag{1.50}$$

This implies that the IS divergence, for which $\beta=0$, is scale invariant because $d_\beta\left(\gamma x|\gamma y\right)=d_\beta\left(x|y\right)$. Scale invariance means that the small and large coefficients of $\mathbf{V}$ are given the same relative importance in the cost function in the sense that a poor factorization fit for a low power coefficient $\mathbf{V}_{\omega,t}$ will cost as much as a poor fit for a higher power coefficient $\mathbf{V}_{\omega',t'}$.

6.1.2 Non-negative matrix factorization algorithm

The resulting update rules lead to the algorithm below.

IS-NMF algorithm with multiplicative updates

Input: non-negative matrix $\mathbf{V}$

Output: optimal $\mathbf{W}$ and $\mathbf{H}$ so that $\mathbf{V}\approx\mathbf{WH}$

1. **Initialization** $\mathbf{W}$ and $\mathbf{H}$ with random non-negative values
2. **For** $i=1:nr\_iter$

$$\mathbf{H}\leftarrow\mathbf{H}\circ\frac{\mathbf{W}^T\left(\left(\mathbf{WH}\right)^{\circ(-2)}\circ\mathbf{V}\right)}{\mathbf{W}^T\left(\mathbf{WH}\right)^{\circ(-1)}}$$

$$\mathbf{W}\leftarrow\mathbf{W}\circ\frac{\left(\left(\mathbf{WH}\right)^{\circ(-2)}\circ\mathbf{V}\right)\mathbf{H}^T}{\left(\mathbf{WH}\right)^{\circ(-1)}\mathbf{H}^T}$$

Normalization for $\mathbf{W}$ and $\mathbf{H}$

End for

6.1.3 Using auxiliary cost functions for optimization

One method to obtain the optimal values for $\mathbf{W}$ and $\mathbf{H}$ by minimizing the cost function is the majorization-minimization algorithm [17], [18].

Let θ be a set of objective variables. For a cost function, hereafter called an objective function $\mathcal{C}\left(\theta\right)$, we create an auxiliary function $\mathcal{C}^+\left(\theta,\tilde{\theta}\right)$ with a set of auxiliary variables $\tilde{\theta}$ satisfying the following two conditions:

1. The auxiliary function is greater than or equal to the objective function:

$$\mathcal{C}^+\left(\theta,\tilde{\theta}\right)\geq\mathcal{C}\left(\theta\right) \tag{1.51}$$

2. Minimizing the auxiliary function according to the auxiliary variables leads to the objective function:

$$\min_{\tilde{\theta}}\mathcal{C}^+\left(\theta,\tilde{\theta}\right)=\mathcal{C}\left(\theta\right) \tag{1.52}$$

Using these two conditions one can develop an algorithm for minimizing the objective function $\mathcal{C}(\theta)$ in an indirect way by minimizing the auxiliary function $\mathcal{C}^{+}\left(\theta,\tilde{\theta}\right)$ iteratively. We obtain the multiplicative updates similar to those in the general algorithm in the previous section:

$$w_{\omega,n} \leftarrow w_{\omega,n}\sqrt{\frac{\sum_{t=1}^{T}\frac{h_{n,t}\left|x_{\omega,t}\right|^{2}}{\hat{x}_{\omega,t}^{2}}}{\sum_{t=1}^{T}\frac{h_{n,t}}{\hat{x}_{\omega,t}}}} \tag{1.53}$$

$$h_{n,t} \leftarrow h_{n,t}\sqrt{\frac{\sum_{\omega=1}^{\Omega}\frac{w_{\omega,n}\left|x_{\omega,t}\right|^{2}}{\hat{x}_{\omega,t}^{2}}}{\sum_{\omega=1}^{\Omega}\frac{w_{\omega,n}}{\hat{x}_{\omega,t}}}} \tag{1.54}$$

6.1.4 Experimental results for single-channel non-negative matrix factorization

The separation experiment was performed on a test audio signal with a sampling frequency of $F_s = 16000$ Hz and duration of 4.7 s. Since the theory behind the non-negative matrix factorization is based on the assumption that the observations (in our case the time-frequency points of the spectrogram) are distributed according to a super-Gaussian, or Gaussian, distribution, we chose in the first phase a piano sound signal, more precisely the sequence of musical notes E, D, C, D, E, E, E.

Using the NMF algorithm and multiplicative update rules (6.40) and (6.41) the bases **W** and **H** are determined iteratively in 100 steps (figure 6.3). The matrices composing the bases were randomly initialized with positive values between 0 and 1.

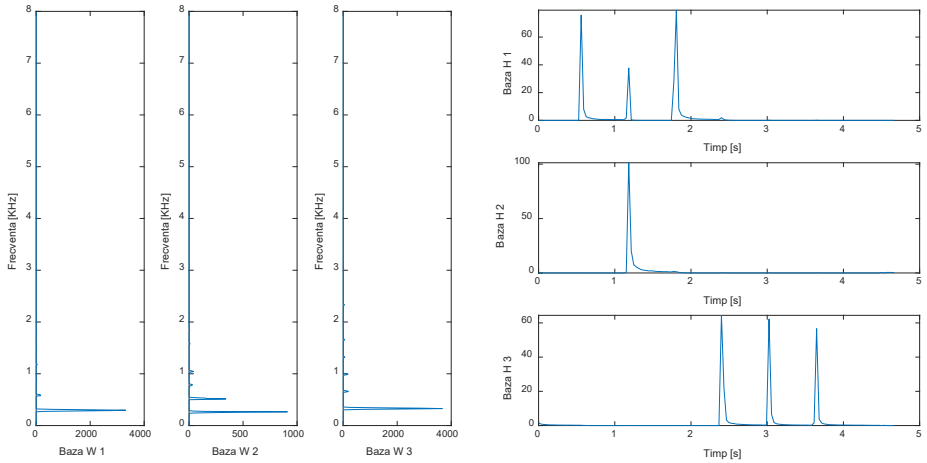

Figure 6.1 *NMF basis for the piano signal*

After testing the non-negative matrix factorization for a mixture of speech signals, the bases were determined and the signals were extracted. The extraction is

not performed correctly because the points on the voice spectrogram do not classify as a super-Gaussian distribution.

6.2 Multichannel non-negative matrix factorization

Given the success of NMF for single-channel source separation, there have been several attempts to extend it to multichannel mixtures [19].

6.2.1 Gaussian Modelling

Given the multichannel mixing equation and independence assumptions, the STFT mixing coefficients can be modeled with a Gaussian distribution:

$$\mathbf{x}_{\omega,t} \sim \mathcal{N}_c\left(0, \sum_{n=1}^{N} \mathbf{R}_{n,\omega,t} v_{n,\omega,t}\right) \tag{1.55}$$

6.2.2 Spectral modelling

Spectral modelling can be achieved by NMF modelling of each individual source and consists of structuring the variances of the sources $v_{n,\omega,t}$:

$$v_{n,\omega,t} = \sum_{k=1}^{K_n} w_{n,\omega,k} h_{n,k,t} \,, \tag{1.56}$$

where K_n depends on the source n and is smaller than Ω and T, and $w_{n,\omega,k}$ and $h_{n,k,t}$ are non-negative. In addition, in order to achieve an association between the K NMF components and the N sources a new non-negative matrix of size $N \times K$, denoted by $\mathbf{G} = \left[g_{n,k} \right] \in \mathbb{R}_{+}^{N \times K}$, will be created [20], and the variances of the sources are represented as:

$$v_{n,\omega,t} = \sum_{k=1}^{K} w_{\omega,k} h_{k,t} g_{n,k} \tag{1.57}$$

6.2.3 Combining the Gaussian and spectral models for multichannel NMF

In order to implement multichannel NMF we will consider the short-time Fourier transforms of the channels, denoted by $\left[x_{m,\omega,t} \right]$ and structure each time-frequency point into a M-dimensional vector of the form $\mathbf{x}_{\omega,t} = \left[x_{1,\omega,t}, \ldots, x_{M,\omega,t} \right]^T \in \mathbb{C}^M$ based on (6.42). We will determine:

$$\mathbf{X}_{\omega,t} = \mathbf{x}_{\omega,t} \mathbf{x}_{\omega,t}^H = \begin{bmatrix} \left| x_{1,\omega,t} \right|^2 & \cdots & x_{1,\omega,t} x_{M,\omega,t}^* \\ \vdots & \ddots & \vdots \\ x_{M,\omega,t} x_{1,\omega,t}^* & \cdots & \left| x_{M,\omega,t} \right|^2 \end{bmatrix} \tag{1.58}$$

This will allow us to approximate the outer product with a K rank structure:

$$\mathbf{X}_{\omega,t} \approx \sum_{k=1}^{K} \mathbf{G}_{\omega,k} w_{\omega,k} h_{k,t} = \hat{\mathbf{X}}_{\omega,t} \tag{1.59}$$

Similar to the presentation in section 6.1.3 we define the objective function by log of the total probability:

$$\mathcal{C}\left(\mathbf{W},\mathbf{H},\mathbf{G}\right) = -\log p\left(\mathbf{X}\middle|\mathbf{W},\mathbf{H},\mathbf{G}\right) = \sum_{\omega=1}^{\Omega}\sum_{t=1}^{T}\left[\mathbf{x}_{\omega,t}^{H}\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{x}_{\omega,t} + \log\det\hat{\mathbf{X}}_{\omega,t}\right] \tag{1.60}$$

As a result, the model (6.49) and the objective function (6.52) become [21]:

$$\hat{\mathbf{X}}_{\omega,t} = \sum_{k=1}^{K}\left(\sum_{n=1}^{N} z_{k,n}\mathbf{G}_{\omega,n}\right) w_{\omega,k} h_{k,t} \tag{1.61}$$

$$\mathcal{C}\left(\mathbf{W},\mathbf{H},\mathbf{G},\mathbf{Z}\right) = \sum_{\omega=1}^{\Omega}\sum_{t=1}^{T}\left[\mathbf{x}_{\omega,t}^{H}\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{x}_{\omega,t} + \log\det\hat{\mathbf{X}}_{\omega,t}\right], \tag{1.62}$$

With $\mathbf{Z} = \left[z_{k,n}\right]$ and the multidimensional matrix $\mathbf{G} = \left\{\mathbf{G}_{\omega,n}\right\}_{\substack{\omega=1..\Omega \\ n=1..N}}$.

6.2.4 Multi-channel NMF algorithm based on minimizing an auxiliary cost function

We define the auxiliary cost function for multichannel NMF relative to the objective cost function (6.52) as follows:

$$\begin{aligned}\mathcal{C}^{+}\left(\mathbf{W},\mathbf{H},\mathbf{G},\mathbf{R},\mathbf{Q}\right) = &\sum_{\omega=1}^{\Omega}\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{\mathbf{x}_{\omega,t}^{H}\mathbf{R}_{\omega,t,k}\mathbf{G}_{\omega,k}^{-1}\mathbf{R}_{\omega,t,k}\mathbf{x}_{\omega,t}}{w_{\omega,k}h_{k,t}} + \\ &+\sum_{\omega=1}^{\Omega}\sum_{t=1}^{T}\left[\operatorname{tr}\left(\hat{\mathbf{X}}_{\omega,t}\mathbf{Q}_{\omega,t}^{-1}\right) + \log\det\mathbf{Q}_{\omega,t} - M\right]\end{aligned}, \tag{1.63}$$

The updates for the objective function variables are determined by equating the partial derivatives of the auxiliary function $\mathcal{C}^{+}\left(\mathbf{W},\mathbf{H},\mathbf{G},\mathbf{R},\mathbf{Q}\right)$ to zero in terms of $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{G}$, giving:

$$\begin{aligned} w_{\omega,k} &\leftarrow w_{\omega,k}\sqrt{\frac{\sum_{t=1}^{T} h_{k,t}\mathbf{x}_{\omega,t}^{H}\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{G}_{\omega,k}\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{x}_{\omega,t}}{\sum_{t=1}^{T} h_{k,t}\operatorname{tr}\left(\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{G}_{\omega,k}\right)}} \\ h_{k,t} &\leftarrow h_{k,t}\sqrt{\frac{\sum_{\omega=1}^{\Omega} w_{\omega,k}\mathbf{x}_{\omega,t}^{H}\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{G}_{\omega,k}\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{x}_{\omega,t}}{\sum_{\omega=1}^{\Omega} w_{\omega,k}\operatorname{tr}\left(\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{G}_{\omega,k}\right)}} \\ \mathbf{G}_{\omega,k} &\leftarrow \mathbf{A}^{-1}\#\left(\mathbf{G}_{\omega,k}\mathbf{B}\mathbf{G}_{\omega,k}\right)\end{aligned} \tag{1.64}$$

In (6.60) $\mathbf{A} = \sum_{t=1}^{T} h_{k,t}\hat{\mathbf{X}}_{\omega,t}^{-1}$, $\mathbf{B} = \sum_{t=1}^{T} h_{k,t}\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{X}_{\omega,t}\hat{\mathbf{X}}_{\omega,t}^{-1}$ and the operator $\mathbf{X}\#\mathbf{Y}$ is defined as the geometric mean of two positive semidefinite matrices:

$$\mathbf{A}\#\mathbf{B} = \mathbf{A}\left(\mathbf{A}^{-1}\mathbf{B}\right)^{\frac{1}{2}} \tag{1.65}$$

Up to this point, the cost function (6.52) was minimized. In order to also benefit from the grouping of components determined in correlated sources based on the spatial structure we will continue with the optimization of the objective cost function (6.54) where we will obtain the following updates for the parameters:

$$\begin{aligned}
w_{\omega,k} &\leftarrow w_{\omega,k}\sqrt{\frac{\sum_{l=1}^{L} z_{l,k}\sum_{t=1}^{T} h_{k,t}\,\mathrm{tr}\left(\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{X}_{\omega,t}\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{G}_{\omega,l}\right)}{\sum_{l=1}^{L} z_{l,k}\sum_{t=1}^{T} h_{k,t}\,\mathrm{tr}\left(\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{G}_{\omega,l}\right)}} \\
h_{k,t} &\leftarrow h_{k,t}\sqrt{\frac{\sum_{l=1}^{L} z_{l,k}\sum_{\omega=1}^{\Omega} w_{\omega,k}\,\mathrm{tr}\left(\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{X}_{\omega,t}\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{G}_{\omega,l}\right)}{\sum_{l=1}^{L} z_{l,k}\sum_{\omega=1}^{\Omega} w_{\omega,k}\,\mathrm{tr}\left(\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{G}_{\omega,l}\right)}} \\
z_{l,k} &\leftarrow z_{l,k}\sqrt{\frac{\sum_{\omega=1}^{\Omega}\sum_{t=1}^{T} w_{\omega,k}h_{k,t}\,\mathrm{tr}\left(\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{X}_{\omega,t}\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{G}_{\omega,l}\right)}{\sum_{\omega=1}^{\Omega}\sum_{t=1}^{T} w_{\omega,k}h_{k,t}\,\mathrm{tr}\left(\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{G}_{\omega,l}\right)}} \\
\mathbf{G}_{\omega,l} &\leftarrow \mathbf{A}^{-1}\left(\mathbf{A}\mathbf{G}_{\omega,l}\mathbf{B}\mathbf{G}_{\omega,l}\right)^{\frac{1}{2}}
\end{aligned} \tag{1.66}$$

6.2.5 Extracting source estimates from multichannel NMF results

This can be done using a single channel Wiener filter for the m-th channel in the mix in the following way:

$$\left\{\mathbf{y}_{\omega,t}^{(l)}\right\}_m = \frac{\left\{\mathbf{G}_{\omega,l}\right\}_{mm}\sum_{k=1}^{K} z_{l,k}w_{\omega,k}h_{k,t}}{\sum_{l=1}^{L}\left\{\mathbf{G}_{\omega,l}\right\}_{mm}\sum_{k=1}^{K} z_{l,k}w_{\omega,k}h_{k,t}}\left\{\mathbf{x}_{\omega,t}\right\}_m \tag{1.67}$$

Alternatively, a multi-channel Wiener filter can be used:

$$\mathbf{y}_{\omega,t}^{(l)} = \left(\sum_{k=1}^{K} z_{l,k}w_{\omega,k}h_{k,t}\right)\mathbf{G}_{\omega,l}\hat{\mathbf{X}}_{\omega,t}^{-1}\mathbf{x}_{\omega,t}, \tag{1.68}$$

6.2.6 Multichannel NMF experimental results

Using the same mixtures as in the previous chapters, the separation of two sources from binaural recordings was tested. The number of NMF bases chosen was 20 for each source and was chosen after testing several values.

After testing the multichannel NMF performance on the 20 mixtures of 2 randomly chosen voice sources we get the results from figure 6.13. In order to check also the ability to separate 3 sources the test was repeated, and it is observed in figure

6.14 that the performance is much poorer than the results of the previous methods, especially for reverberant mixtures.

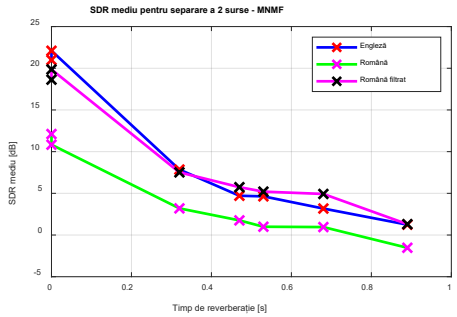

Figure 6.2 *Average separation performance of two sources as a function of reverberation time - multichannel NMF algorithm (MNMF).*

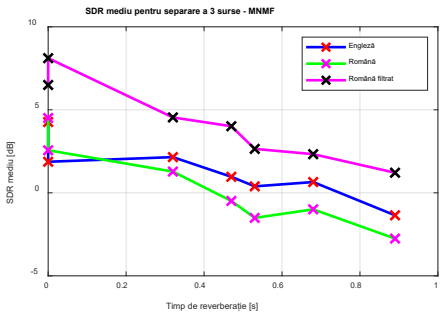

Figure 6.3 *Average separation performance of three sources as a function of reverberation time - multichannel NMF algorithm (MNMF).*

6.3 Conclusions

In the single-channel case, good separation is observed for harmonic signals (such as those generated by musical instruments), with a spectrogram characterized by mixtures of super-Gaussian distributions, but for speech signals, whose spectral content is more complex, the performance is very low.

For the multichannel case, very good separation is observed under anechoic conditions, and for reverberant rooms the performance is very similar to those in the previous chapters.

Chapter 7

Independent low-rank matrix analysis

Independent Low-Rank Matrix Analysis (*ILRMA*) is proving to be a solution to the separation of multichannel mixtures and arises from the combination of *Independent Vector Analysis* (*IVA*) and non-negative matrix factorization.

7.1 Independent vector analysis - IVA

Independent vector analysis is an extension of the *Independent Component Analysis (ICA*) method and is applied to determined mixtures where the number of microphones equals the number of sources in the mixture, in our case $N = M$ [22].

The separation method using ICA involves determining the square matrix $\mathbf{W}_\omega$ of dimension M which linearly transforms the mixtures $\mathbf{x}_{\omega,t}$ into estimates of the original sources $\mathbf{y}_{\omega,t} = \left[y_{1,\omega,t}, \ldots, y_{N,\omega,t} \right]^T \in \mathbb{C}^N$ in the following way:

$$\mathbf{y}_{\omega,t} = \mathbf{W}_\omega \mathbf{x}_{\omega,t} \tag{1.69}$$

In the case of IVA the notion of independence is extended to vector variables. These vectors are defined over all frequencies corresponding to a time moment as $\mathbf{y}_{n,t} = \left[y_{n,1,t}, \ldots, y_{n,\Omega,t} \right]^T$.

7.2 Extending IVA with NMF to obtain ILRMA

A first way to implement ILRMA is to extend the independent vector analysis by introducing non-negative matrix factorization [23].

Using the notations for factorization matrices $\mathbf{W}_n = \left[w_{n,\omega,k} \right]$ and $\mathbf{H}_n = \left[h_{n,k,t} \right]$ we can define the objective function:

$$\mathcal{C}\left(\mathcal{T}, \{\mathbf{W}_n\}, \{\mathbf{H}_n\}\right) = \sum_{n=1}^{N} \sum_{\omega=1}^{\Omega} \sum_{t=1}^{T} \left[\frac{\left| y_{n,\omega,t} \right|^2}{\hat{y}_{n,\omega,t}} + \log\left(\hat{y}_{n,\omega,t} \right) \right] - 2T \sum_{\omega=1}^{\Omega} \log\left| \det\left(\mathbf{T}_\omega \right) \right| \tag{1.70}$$

Objective function minimization (7.23) can be achieved by alternating NMF updates similar to (6.40) and (6.41) with solving the HEAD problem [24]. First we focus on the first term in (7.23) and determine the updates for $\mathbf{W}_n = \left[w_{n,\omega,k} \right]$ and $\mathbf{H}_n = \left[h_{n,k,t} \right]$. We will note that for each n we can obtain the updates by replacing $\left| x_{\omega,t} \right|^2$ and $\hat{x}_{\omega,t}$ with $\left| y_{n,\omega,t} \right|^2$, respectively $\hat{y}_{n,\omega,t}$ in (6.40) and (6.41), obtaining the updates:

$$w_{n,\omega,k} \leftarrow w_{n,\omega,k} \sqrt{\frac{\sum_{t=1}^{T} \frac{h_{n,k,t} \left| y_{n,\omega,t} \right|^2}{\hat{y}_{n,\omega,t}^2}}{\sum_{t=1}^{T} \frac{h_{n,k,t}}{\hat{y}_{n,\omega,t}}}} \tag{1.71}$$

$$h_{n,k,t} \leftarrow h_{n,k,t} \sqrt{\frac{\sum_{\omega=1}^{\Omega} \frac{w_{n,\omega,k} \left| y_{n,\omega,t} \right|^2}{\hat{y}_{n,\omega,t}^2}}{\sum_{\omega=1}^{\Omega} \frac{w_{n,\omega,k}}{\hat{y}_{n,\omega,t}}}} \tag{1.72}$$

7.3 Multichannel NMF restriction to achieve ILRMA

A second method to obtain the ILRMA method is by restricting the multichannel NMF spatial property defined by the $\mathbf{G}_{n,\omega} \in \mathbb{C}^{M \times M}$ matrix to be of rank 1. Starting then from the MNMF (6.53) model, we can simplify it such that:

$$\hat{\mathbf{X}}_{\omega,t} = \mathbf{G}_{\omega} \mathbf{D}_{\omega,t} \mathbf{G}_{\omega}^{H}, \tag{1.73}$$

Where $\mathbf{G}_{\omega} = \left[\mathbf{g}_{1,\omega}, \ldots, \mathbf{g}_{N,\omega} \right]$ and $\mathbf{D}_{\omega,t}$ is a diagonal square matrix of dimension $N \times N$ whose n-th element on the diagonal has the value:

$$\hat{y}_{n,\omega,t} = \sum_{k=1}^{K} z_{k,n} w_{\omega,k} h_{k,t} \tag{1.74}$$

Further restricting the mixing system to a determined one by $N = M$ to obtain the separation matrix $\mathbf{T}_{\omega}$ from the mixing matrix $\mathbf{G}_{\omega}$:

$$\mathcal{C}\left(\mathcal{T}, \mathbf{W}, \mathbf{H}, \mathbf{Z}\right) = \sum_{\omega=1}^{\Omega} \sum_{t=1}^{T} \sum_{n=1}^{N} \left[\frac{\left| y_{n,\omega,t} \right|^2}{\hat{y}_{n,\omega,t}} + \log\left(\hat{y}_{n,\omega,t} \right) \right] - 2T \sum_{\omega=1}^{\Omega} \log \left| \det\left(\mathbf{T}_{\omega} \right) \right| \tag{1.75}$$

After minimizing the objective function without using auxiliary functions we obtain the updates for the source partitioning variable $z_{k,n}$:

$$z_{k,n} \leftarrow z_{k,n} \sqrt{\frac{\sum_{\omega=1}^{\Omega} \sum_{t=1}^{T} \frac{w_{\omega,k} h_{k,t} \left| y_{n,\omega,t} \right|^2}{\hat{y}_{n,\omega,t}^2}}{\sum_{\omega=1}^{\Omega} \sum_{t=1}^{T} \frac{w_{\omega,k} h_{k,t}}{\hat{y}_{n,\omega,t}}}} \tag{1.76}$$

$$w_{\omega,k} \leftarrow w_{\omega,k} \sqrt{\frac{\sum_{n=1}^{N} \sum_{t=1}^{T} \frac{z_{k,n} h_{k,t} \left| y_{n,\omega,t} \right|^2}{\hat{y}_{n,\omega,t}^2}}{\sum_{n=1}^{N} \sum_{t=1}^{T} \frac{z_{k,n} h_{k,t}}{\hat{y}_{n,\omega,t}}}} \tag{1.77}$$

$$h_{k,t} \leftarrow h_{k,t} \sqrt{\frac{\sum_{\omega=1}^{\Omega} \sum_{N=1}^{N} \frac{w_{\omega,k} z_{k,n} \left| y_{n,\omega,t} \right|^2}{\hat{y}_{n,\omega,t}^2}}{\sum_{\omega=1}^{\Omega} \sum_{n=1}^{N} \frac{w_{\omega,k} z_{k,n}}{\hat{y}_{n,\omega,t}}}} \tag{1.78}$$

7.4 Experimental results

Similar to the tests in the previous chapters we test the separation performance using the 20 English and 10 Romanian mixtures of two spoken sentences each. The results for the ILRMA algorithm without the partitioning function can be seen in figure 7.4, for the one with partitioning function we can see them in figure 7.5.

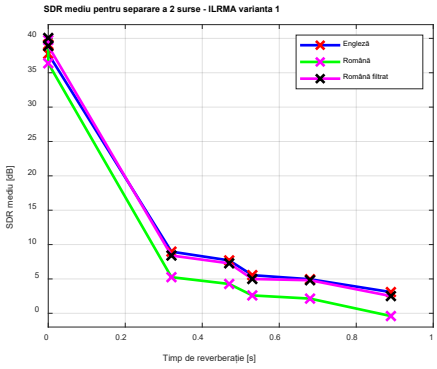

Figure 7.1 *Average separation performance of two sources as a function of reverberation time - ILRMA algorithm variant 1*

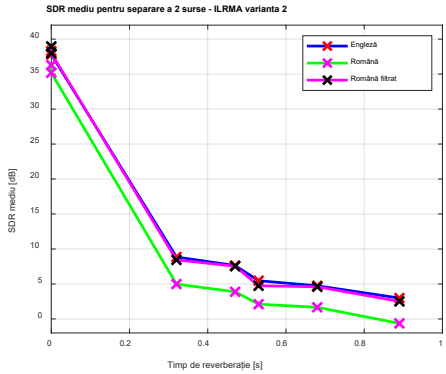

Figure 7.2 *Average separation performance of two sources as a function of reverberation time - ILRMA algorithm variant 2*

7.5 Conclusions

The separation performance of the method is very good for mixtures under anechoic conditions and moderate for reverberant mixtures. However a run time analysis for the method will allow a decision to be made on the overall performance of the method.

Chapter 8

Conclusions

8.1 Obtained results

In this thesis, four methods of separating speech sound sources from reverberant binaural mixtures were presented.

The male and female speech signals used were extracted from the TIMIT dataset [13], which contains English spoken sentences. The Romanian female and male speech set were represented by 7 sentences recorded under studio conditions. All audio signals were resampled at the sampling frequency $F_s = 16$ kHz. The composition of the mixtures by convolution was performed with a set of binaural recordings made by Hummersone [14] with a binaural head placed in an anechoic chamber and in a set of typical rooms with different reverberation times. We have noted the binaural head related impulse response with the term "Anechoic" to distinguish it from other head related impulse responses. The rooms were denoted A, B, C and D in ascending order of reverberation time. To this set of impulse responses were also added custom head related impulse responses (of a real person) measured and extracted with the method described in Chapter 2 and in [25] and the set of binaural responses associated with the room in which they were measured. The impulse response of the head was noted in the results with HRIR, and the respective room was noted with E. The acoustic properties of the rooms, together with the notations used for each are given in Table 8.1.

Table 8.1 *Acoustic properties of rooms expressed by Initial Time Delay Gap (ITDG), Direct-to-Reverberant Ratio (DRR) and reverberation time* T_{60}

	ITDG [ms]	DRR [dB]	T_{60} [s]
Anechoic	2	-	0
HRIR	2	-	0
Room A	8,72	6,09	0,32
Room B	9,66	5,31	0,47
Room C	11,9	8,82	0,68
Room D	21,6	6,12	0,89
Room E	9,32	5,78	0,53

Because all the methods presented are based on a spatial separation of the sound sources in the final mix we first analyzed the separation performance as a function of the angle between two sources. The performance was evaluated by determining the source distortion ratio (SDR) since it encompasses all artefacts, noises and interferences that may occur in the separation process. After applying the methods some of the results were represented in figure 8.1.

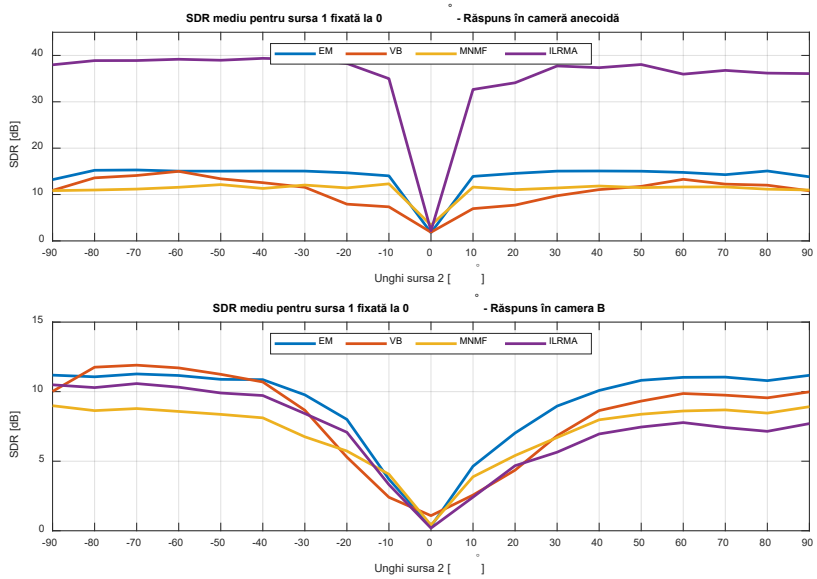

Figure 8.1 *Separation performance as a function of source angle in SDR for fixed source; Top: in anechoic chamber; Bottom: in room B.*

Note that under anechoic conditions the minimum angle between sources can be $10° - 20°$, and under reverberant conditions the minimum angle should be at least $30°$. The following analysis was performed on a mixture of two sound sources placed at angles of $-30°$ and $30°$ in the 7 room types. The results for the separation of two sources in English can be seen in figure 8.2 and for Romanian in figure 8.3. It can be seen that the MNMF and ILRMA methods have a very good separation performance in the case of non-reverberant determined mixtures. Under reverberant conditions the EM method is the best performing, and the VB method follows on par with ILRMA.

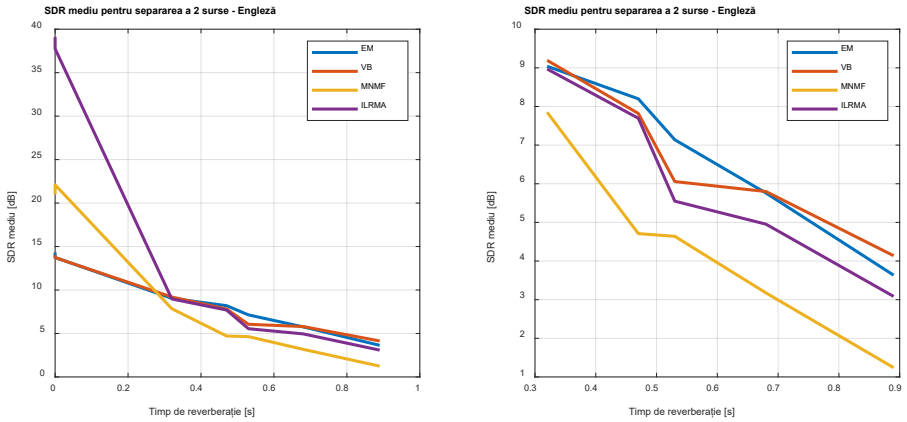

Figure 8.2 *Separation performance for two-source mixtures in English. Average SDR [dB] over recovered sources and over mixtures; Left: for all rooms; Right: For reverberant rooms*

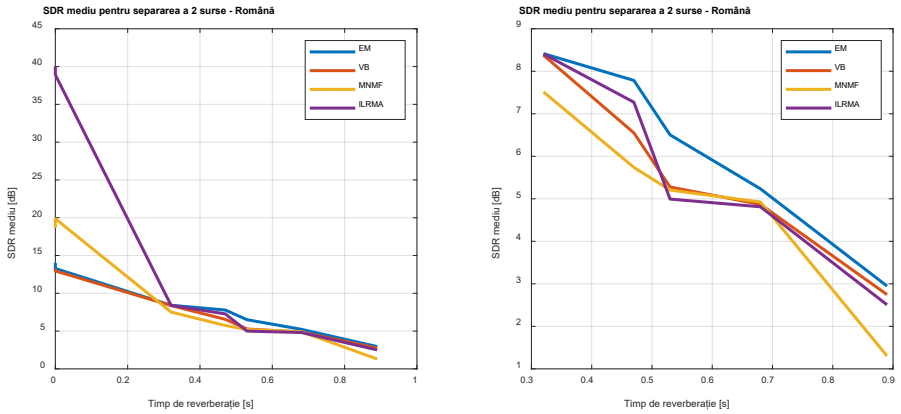

Figure 8.3 *Separation performance for two-source mixtures in Romanian. Average SDR [dB] over recovered sources and over mixtures; Left: for all rooms; Right: For reverberant rooms*

The results for three-source mixtures can be seen in figure 8.4 for English and in figure 8.5 for Romanian.

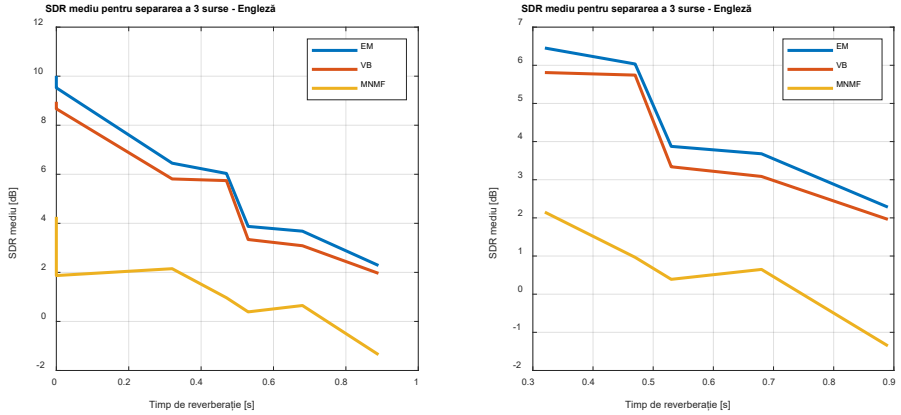

Figure 8.4 *Separation performance for three-source mixtures in English. Average SDR [dB] over recovered sources and over mixtures; Left: for all rooms; Right: For reverberant rooms*

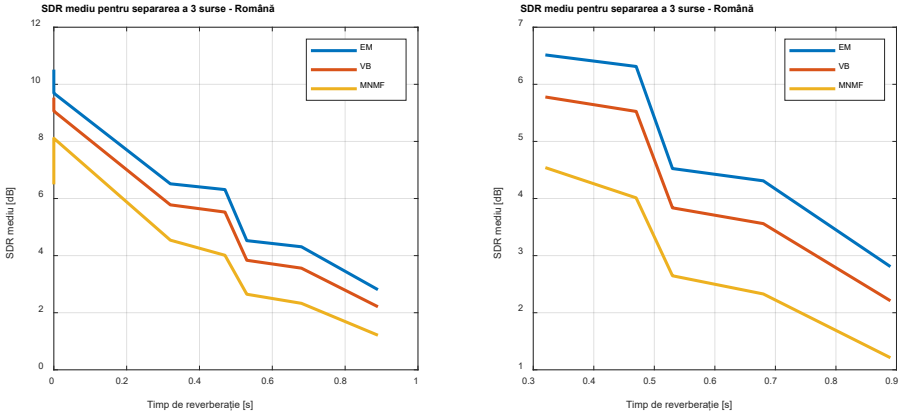

Figure 8.5 *Separation performance for three-source mixtures in Romanian. Average SDR [dB] over recovered sources and over mixtures; Left: for all rooms; Right: For reverberant rooms*

Following the analysis of the separation of sound sources from underdetermined mixtures we observe that the MNMF method has lower performance than the EM and VB methods, with SDR differences between 2 and 5 dB. A conclusion would be that the MNMF method does not perform well for underdetermined mixtures, being much better for determined ones. As a comparison between the EM and VB methods, the statistical modelling between the two is very similar, as is the initialization, the difference being that the VB method uses prior distributions for the parameters of the Gaussian distributions. We observe a systematic performance difference between the two of about 0.5 dB, with the best performing being the EM method. However, the advantage of the VB method is that it does not require knowledge of the number of sound sources, this is determined by the algorithm.

In the following we make a comparison with similar results from literature. A number of results can be found in [10] and improved in [26], performed by Atiyeh Alinaghi using a method similar to the expectation maximization method presented in this thesis, with the difference that in this thesis an initialization of the responsibilities including the results of the adaptive eigenvalue decomposition is used, while the method implemented by Atiyeh uses an additional source to retrieve garbage artefacts. These results were noted on the graph of the English mixtures with EM-At. The testing method was changed because in the mentioned articles the SDR value determined refers to a target source, which we assumed to be the source with the highest SDR, the others being considered as interference type signals. Similarly the results of the method developed by Mandel in [8] also based on expectation maximization using a model similar to Atiyeh's, but only with binaural cues, and Sawada's results based on the method described in [27] also based on expectation maximization, but for mixing vectors. Madel's method was plotted with EM-Man and Sawada's with EM-Saw. The separation performance for two-source mixtures can be seen in Figure 8.6 and the separation performance for 3-source mixtures in figure 8.7.

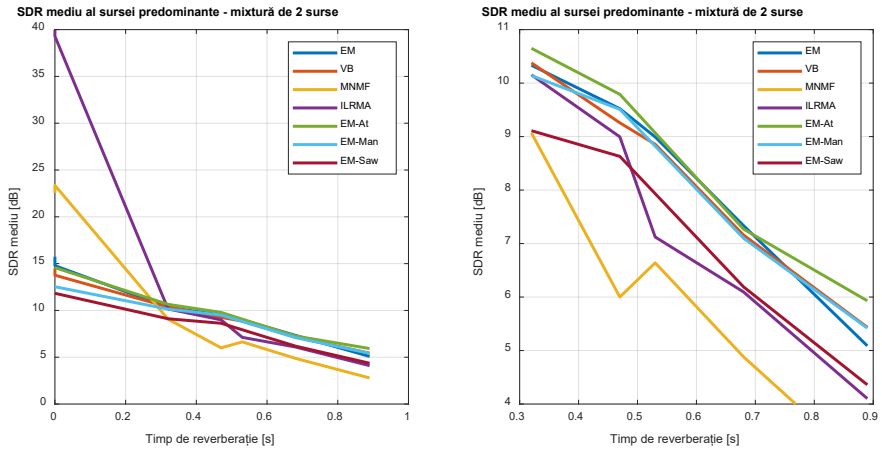

Figure 8.6 *Separation performance for two-source mixtures in English. Average SDR [dB] for the predominant source; Left: for all rooms; Right: For reverberant rooms*

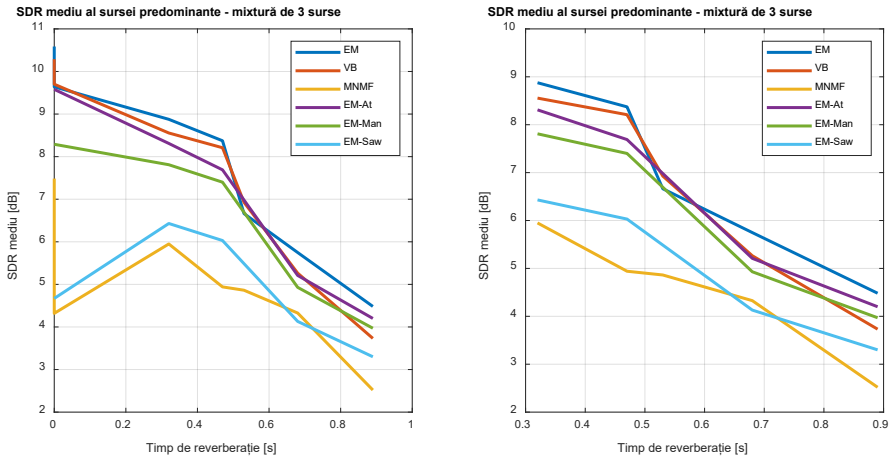

Figure 8.7 *Separation performance for two-source mixtures in English. Average SDR [dB] for the predominant source; Left: for all rooms; Right: For reverberant rooms*

From the comparative results presented we observe that the EM method implemented in this thesis, the VB method, the EM-At method and the EM-Man method have similar results in terms of separation performance for both determined and underdetermined mixtures. Among them the EM method seems to have better results in the underdetermined case than the others. A distinction must be made between the anechoic and reverberant cases because the ILRMA and MNMF methods perform very well, well above the others, in the anechoic determined case, but the performance in reverberant environments is mediocre, below the other methods. The EM-Saw method seems to perform worse than the other methods, being above MNMF in the determined and underdetermined reverberant case. Looking at the average results for reverberant rooms we see that the EM and VB methods have very close average performance, the difference being $\pm 0,2$ dB, with VB performing marginally better in the underdetermined case. The differences from the best performing results in the literature are about 0.3 dB.

To provide an additional degree of comparison between the methods implemented in this paper we have determined in table 8.3 the average execution times for each method following the separation of a mixture of two or three sources.

Table 8.2 *Average run times of each method for separating a mixture*

Method	**No. Sources**	**Method Iterations**	**Running time [s]**
EM	2	20	9,95
EM	3	20	13,39
VB	2	40	22,44
VB	3	40	34,06
MNMF	2	200	46,18
MNMF	3	200	64,49
ILRMA v1	2	50	3,82
ILRMA v2	2	50	4,36

After analyzing the results obtained, we can establish a number of criteria for choosing the right method depending on the type of binaural mixture:

- For anechoic mixtures, or with low reverberation times in the determined case the best method is ILRMA, having both very good results and low runtime;
- For reverberant determined or underdetermined mixtures where the number of sources is known, both EM and VB can be applied, but the EM method is faster and, in some cases, more efficient;
- For overdetermined, determined or underdetermined mixtures in which the number of sources is not known, the VB method can be applied which has good separation results and does not require knowledge of the number of sources in the mixture;
- The methods work similarly regardless of the language in which the sentences are spoken, with low-frequency spectral content having a greater influence on the separation process.

8.2 Original contributions

In order to perform the measurements in Chapter 2 we have created a software library, published in the conference paper [C1], which implements the method of determining the acoustic impulse response using the exponential sine-sweep test signal which has a number of advantages in terms of separating the linear response from higher order harmonics. Also, we have presented the method for obtaining synchronized acoustic impulse responses. This software was used to determine the acoustic parameters of the room in which part of the tests were performed.

In the conference paper [C2] we developed a method to reduce the measurement time for an acoustic impulse response using the ESS method by sectioning an impulse response and overlapping the two parts in time which streamlined the measurements for the head related impulse response in Chapter 2.

In the review article [R1] we have developed an impulse response equalization method for a loudspeaker and analyzed its performance. The equalization was used for the loudspeaker used to measure head related impulse responses in Chapter 2.

In the conference paper [C7] we developed a method for extracting head related impulse response from binaural acoustic impulse responses measured in reverberant rooms, the major advantage being the elimination of the need for an anechoic chamber with the preservation of good quality of the measured responses. This method was presented in Chapter 2.

In Chapter 3 we investigated a method for determining the direction of arrival of sound sources using adaptive eigenvalue decomposition and experimentally analyzed its detection performance. In addition, by making a histogram for the detected angles in a binaural mixture we proved that the method can detect multiple angles as long as the mixtures of speech-like signals respect the sparsity condition.

In Chapter 4 we presented and implemented a method for separating sound sources from literature-determined and underdetermined reverberant binaural mixtures based on expectation maximization of a Gaussian mixture model applied for both mixing vectors and binaural cues. A performance enhancement of this method was provided by combining the initialization based on the PHAT histogram with the direction of arrival method based on adaptive eigenvalue decomposition presented in Chapter 3.

In Chapter 5 we have designed, developed and implemented a new method for separating sound sources from underdetermined binaural mixtures based on a variational Bayes algorithm in which the observations of mixing vectors and binaural indices are, in the first phase, modeled as a mixture of Gaussian distributions. In the second phase prior distributions are assigned to the means and variances of these Gaussian distributions, thus creating the variational model. We then determined and implemented the hyperparameter update rules to achieve convergence. The model as well as the separation results obtained with this method have been published in the conference paper [C3].

In Chapter 6 we developed and implemented the well-known method of separating sound sources from binaural mixtures by multichannel nonnegative matrix factorization that is based on the covariance matrix decomposition of the short-term Fourier transforms of observations into smaller dimensional matrices describing the time and frequency activities of base components.

In Chapter 7 we presented and implemented the independent low-rank matrix analysis which is a method of blind source separation for determined mixtures, based either on combining independent vector analysis with non-negative matrix factorization, or on reducing to rank 1 the matrix describing the spatial property within the multichannel non-negative matrix factorization method.

All methods have been implemented and a comparison has been made regarding the separation performance of binaural mixtures determined or underdetermined under reverberant or anechoic conditions.

8.3 List of original publications

Books:

[B1] **V. Popa**, C. Negrescu, ”Măsurarea și caracterizarea sistemelor acustice : îndrumar de laborator”, Editura Politehnica Press, București, 2013, ISBN 978-606-515-505-3

[B2] C. Negrescu, A. Ciobanu, **V. Popa**, „Inginerie audio – Îndrumar de laborator”, Editura Politehnica Press, București, 2013, ISBN 978-606-515-527-5

[B3] R. M. Udrea, C. A. Cojocariu, **V. Popa**, ”Servere de conținut, procesoare de flux şi terminale multimedia : îndrumar de laborator”, Editura Politehnica Press, București, 2013, ISBN 978-606-515-504-6.

Articles published in scientific journals:

[R1] T. M. Culda, **V. Popa**, C. Cojocariu, D. Stanomir, C. Negrescu, "The Influence of Loudspeaker Performance in Loudspeaker Equalization Using Wiener Approach", U.P.B. Sci. Bull., Series C, Vol. 74, Iss. 4, P. 219-228, 2012, ISSN 1454-234x

Conference papers:

[C1] **V. Popa**, T. M. Culda, C. Negrescu, "Software Library for Managing a Complete Audio Chain Intended for Determining Acoustic Room Parameters and Spatial Impression", The Annual Symposium Of The Institute Of Solid Mechanics (SISOM), May 2012 (**BDI**: IndexCopernicus, MathSciNet, Reaserchgate)

[C2] T. M. Culda, **V. Popa**, D. Stanomir, C. Negrescu, "Reducing Time in Acoustic Impulse Response Measurements Using Exponential Sine Sweeps", International Symposium On Signals, Circuits And Systems (ISSCS), P. 1-4, Iași, July 2013 (**ISI, IEEE**)

[C3] **V. Popa**, W. Wang, A. Alinaghi, "Underdetermined Model-Based Blind Source Separation of Reverberant Speech Mixtures Using Spatial Cues in a Variational Bayesian Framework", Intelligent Signal Processing (ISP), London, December 2013, DOI: 10.1049/Cp.2013.2074 (**IEEE, Scopus**)

[C4] A. Ciobanu, V. A. Niță, **V. Popa**, "Forgery Detection Based on Reverberation Time Estimation in Multiple Bands", 13th International Symposium on Electronics and Telecommunications (ISETC), Timisoara, November 2018, WOS:000463031500068 (**ISI, IEEE**)

[C5] V. A. Niță, **V. Popa**, "A Framework for Privacy Assurance in a Public Video-Surveillance System", 2019, International Symposium on Signals, Circuits and Systems (ISSCS), Iasi, Romania, 2019, Pp. 1-4, WOS:000503459500066 (**ISI, IEEE**)

[C6] V. A. Niță, **V. Popa**, "Bringing Technology into Fencing Training. the Art of Counterattacking", 2019 11th International Symposium on Advanced Topics in Electrical Engineering, ATEE 2019, WOS:000475904500146 (**ISI, IEEE**)

[C7] **V. Popa**, B. Moroșanu, C. Negrescu, "Head related transfer function measurement in reverberant rooms", 11th edition of the International Conference "Advanced Topics in Optoelectronics, Microelectronics and Nanotechnologies" (ATOM-N), Constanta, Romania, 2022, WOS indexing pending.

[C8] B. Morosanu, **V. Popa**, C. Negrescu, Ionuț Fîciu, "Control Room Design for Subjective Audio Critical Listening", 11th edition of the International Conference "Advanced Topics in Optoelectronics, Microelectronics and Nanotechnologies" (ATOM-N), Constanta, Romania, 2022, WOS indexing pending.

PhD reports:

[RC1] **V. Popa**, "Determinarea direcției de sosire a unei surse sonore într-un mediu reverberant folosind descompunerea adaptivă a valorii proprii" - Primul raport de cercetare științifică, coordonator științific Prof.dr. ing. Ion Marghescu.

[RC2] **V. Popa**, "Pregătirea și evaluarea mixturilor de semnale audio folosind transformata Fourier pe termen scurt. Evaluarea performanțelor de separare" – Al doilea raport de cercetare științifică, coordonator științific Prof.dr. ing. Ion Marghescu.

[RC3] **V. Popa**, "Algoritmul maximizării mediei statistice pentru separarea surselor sonore din mixturi subdeterminate" – Al treilea raport de cercetare științifică, coordonator științific Prof.dr. ing. Ion Marghescu.

[RC4] **V. Popa**, "Algoritmul Bayes variațional pentru separarea mixturilor binaurale" – Al patrulea raport de cercetare științifică, coordonator științific Prof.dr. ing. Ion Marghescu.

[RC5] **V. Popa**, "Tehnici de separare oarbă a surselor sonore folosind factorizarea matriceală nenegativă" – Al cincilea raport de cercetare științifică, coordonator științific Prof.dr. ing. Ion Marghescu.

8.4 Perspectives for further developments

A first further development could be to extend the variational Bayes model to reverberant underdetermined mixtures with multiple microphones placed in rooms. The modelling could be achieved by replacing the binaural indices with a matrix set expressing the spatial characteristics as level as well as phase difference for the respective observations. Given the separation mode of the method, the possibility of using it as a method to reduce the reverberation of the recorded signals to improve their intelligibility could also be investigated.

Also based on the variational method, its use in room acoustics could be investigated by extracting basic acoustic parameters using either recordings of instruments or voices in the room or using acoustic impulse responses measured with multiple microphones. Determining these acoustic parameters would be particularly useful for live performances as it would allow the response of the audio reproduction system to be equalized in real time according to the conditions encountered in the venue.

The use of other acoustic methods could also be investigated to develop robust algorithms for measuring and interpreting room acoustic characteristics.

Another direction of development could focus on combining the presented methods with deep neural networks, either as a pre-processing part for them or as reinforcement or optimization measures for the learning process.

Bibliography

[1] A. Farina, "Advancements in Impulse Response Measurements by Sine Sweeps," *Journal of The Audio Engineering Society*, 2007.

[2] V. Popa, T. M. Culda, and C. Negrescu, "Software Library for Managing a Complete Audio Chain Intended for Determining Acoustic Room Parameters and Spatial Impression," presented at The Annual Symposium Of The Institute Of Solid Mechanics (SISOM), Bucuresti, 2012.

[3] O. Kirkeby and P. Nelson, "Digital Filter Design for Inversion Problems in Sound Reproduction," *Journal of the Audio Engineering Society*, vol. 47, Jan. 1999.

[4] T. M. Culda, V. Popa, C. Cojocariu, D. Stanomir, and C. Negrescu, "The influence of loudspeaker performance in loudspeaker equalization using wiener approach," *UPB Scientific Bulletin, Series C: Electrical Engineering*, vol. 74, pp. 219-228, Jan. 2012.

[5] T. M. Culda, V. Popa, D. Stanomir, and C. Negrescu, "Reducing time in acoustic impulse response measurements using exponential sine sweeps," in *International Symposium on Signals, Circuits and Systems ISSCS2013*, 2013, pp. 1-4.

[6] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462-1469, Jul. 2006.

[7] Y. Huang, J. Benesty, and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for real time acoustic source localization system," 1999, pp. 937-940 vol.2.

[8] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-Based Expectation-Maximization Source Separation and Localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382-394, 2010.

[9] H. Sawada, S. Araki, and S. Makino, "A Two-Stage Frequency-Domain Blind Source Separation Method for Underdetermined Convolutive Mixtures," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 139-142.

[10] A. Alinaghi, W. Wang, and P. J. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 209-212.

[11] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833-1847, 2007.

[12] V. Popa, W. Wang, and A. Alinaghi, "Underdetermined model-based blind source separation of reverberant speech mixtures using spatial cues in a variational Bayesian framework," in *IET Intelligent Signal Processing Conference 2013 (ISP 2013)*, 2013, pp. 1-6.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM." NIST, 1993.

[14] C. Hummersone, "A psychoacoustic engineering approach to machine sound source separation in reverberant environments," Ph.D. thesis, Music and Sound Recording, University of Surrey, UK, 2011.
[15] J. Taghia, N. Mohammadiha, and A. Leijon, "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 253-256.
[16] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural computation*, vol. 21, pp. 793-830, Oct. 2008.
[17] K. Lange, D. R. Hunter, and I. Yang, "Optimization Transfer Using Surrogate Objective Functions," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 1-20, 2000.
[18] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ILRMA originating from ICA and NMF," *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e12, 2019.
[19] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel Extensions of Non-Negative Matrix Factorization With Complex-Valued Data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971-982, 2013.
[20] J. Nikunen and T. Virtanen, "Direction of Arrival Based Spatial Covariance Model for Blind Sound Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 727-739, 2014.
[21] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 257-260.
[22] T. Kim, "Real-Time Independent Vector Analysis for Convolutive Blind Source Separation," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 57, pp. 1431-1438, Aug. 2010.
[23] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626-1641, 2016.
[24] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 189-192.
[25] V. Popa, B. Moroșanu, and C. Negrescu, "Head related transfer function measurement in reverberant rooms," presented at the ATOM-N, Constanța, 2022.
[26] A. Alinaghi, P. J. Jackson, Q. Liu, and W. Wang, "Joint Mixing Vector and Binaural Model Based Stereo Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1434-1448, 2014.
[27] H. Sawada, S. Araki, and S. Makino, "Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516-527, Mar. 2011.