POLITECNICO DI MILANO
Polo Territoriale di Como
Corso di Laurea Magistrale in Ingegneria Informatica
Dipartimento di Elettronica, Informazione e Bioingegneria

# STUDY OF MULTIPLE DICTIONARIES IN EXEMPLAR-BASED NMF FOR SPEECH ENHANCEMENT

Relatore: Prof. Fabio Antonacci
Correlatore: Dr. Jordi Janer (Universitat Pompeu Fabra, Barcelona)

Tesi di Laurea di:
Marco Lunardon, 783465

Anno Accademico 2013-2014

# Abstract

Growing in importance, especially over the last years, speech enhancement has been an important research topic due to the fact that it is required in many applications in the daily life. Speech enhancement and noise reduction aim to improve the speech quality, intelligibility and overall perceptual clarity of a noisy signal by removing the unwanted noise using several techniques.

The traditional noise reduction techniques, such as the Wiener filtering or the Spectral Subtraction do not work satisfactorily in the presence of real non-stationary background noise. In order to overcome this problem, we decided to use a different technique, the Non-Negative Matrix Factorization (NMF) jointly with a sparse representation method. The NMF is a class of algorithm where a matrix $V$ is factorized into two matrices, $W$ and $H$, with the property that all of them have no negative elements. The NMF has applications in many fields like computer vision, document clustering and recommender systems but also in spectral analysis, becoming widely used as a source separation technique. Recently, NMF has also been applied to estimate the clean speech from a noisy observation. We use this technique in order to obtain the weight matrix $H$ of a matrix $W$, called dictionary, that multiplied approximate the noisy observation $V$. Once obtained, the activation matrix $H$ can be used, together with a dictionary $W$, for the reconstruction of the enhanced input noisy signal.

The purpose of this thesis is to give a proof-of-concept for a later development of a more general real-time sparse NMF algorithm for speech enhancement of any speech signal. The algorithm devised is able to reconstruct an enhanced version of every speech signal corrupted by non-stationary real background noise using different training signals. In particular, we investigate the importance of the training dictionary, obtained from the training signal, in the factorization part. We use three approaches. The first consists in using the same noisy and clean utterances respectively for the NMF decomposition and for the reconstruction. The second approach uses directly

the clean speech for both NMF factorization and reconstruction. The third approach also uses the clean speech for both tasks, but with an additional refinement using vocal features on the speech dictionary.

To do so, we investigate the fundamental aspects to consider for the NMF factorization and the enhanced reconstruction of a noisy observation, such as the dictionary size, the bases dimension and the sparsity constraint. Comparing different settings of these features, turns out that the second approach, that uses clean dictionaries, obtains the best results. However, with more specific study over the vocal feature extraction, the third approach can be faster and as good as the actual best.

# Sommario

Negli ultimi anni lo speech enhancement ha gradualmente visto aumentare la propria importanza, fino a divenire un importante argomento di ricerca in virtù delle sue numerose applicazioni nella vita quotidiana. Speech enhancement e noise reduction mirano al miglioramento della qualità della voce, dell'intelligibilità e della chiarezza percepita di un segnale rumoroso mediante molteplici tecniche di rimozione del rumore indesiderato.
I metodi tradizionali di riduzione del rumore, come Wiener filtering o Spectral Subtraction non ottengono risultati soddisfacenti in presenza di rumori di fondo reali e non stazionari. Per superare questa limitazione, si è deciso di utilizzare una tecnica diversa basata sulla Non-Negative Matrix Factorization (NMF), congiuntamente ad un metodo di sparse representation. La NMF consiste in un gruppo di algoritmi nei quali la matrice $V$ viene fattorizzata in due matrici $W$ e $H$, dove tutte e tre le matrici hanno la proprietà di essere composte da elementi non negativi.
Lo scopo di questo elaborato è quello di fornire un'adeguata base teorica per il successivo sviluppo di un algoritmo più generale di *sparse* NMF in tempo reale, che sia in grado di migliorare il parlato di un segnale vocale. Un algoritmo così concepito è in grado di ricostruire una versione migliorata di qualsiasi segnale vocale, deteriorato da un rumore di fondo reale e non stazionario, utilizzando registrazioni diverse nella fase di training. In particolare, verrà analizzata l'importanza del dizionario ottenuto in questa fase di training in vista della successiva fattorizzazione non negativa (NMF).
Verranno adottati tre approcci distinti. Il primo di questi consiste nell'usare la stessa frase registrata in due momenti differenti, prima senza disturbi e poi in ambiente rumoroso, rispettivamente per la ricostruzione e per la decomposizione. Il secondo approccio utilizza direttamente la registrazione pulita tanto per la decomposizione quanto per la ricomposizione. Infine, il terzo approccio fa uso anch'esso della registrazione priva di disturbi per entrambi i compiti, ma allo stesso tempo ricorre all'estrazione di alcune caratteristiche vocali per un ulteriore perfezionamento del dizionario ottenuto in fase

di training.

Per fare ciò sono stati indagati gli aspetti fondamentali da tenere in considerazione per garantire la miglior fattorizzazione e ricostruzione, come ad esempio il metodo di selezione delle basi e la loro dimensione, la dimensione dei dizionari usati e gli sparsity constraint. Utilizzando in congiunto diverse configurazioni di questi fattori all'interno dei tre diversi approcci, si giungerà alla conclusione che la soluzione migliore consiste nel secondo dei tre metodi applicati. Tuttavia, attraverso uno studio più specifico dell'estrazione e della classificazione delle caratteristiche vocali, si dimostrerà che anche il terzo approccio può portare agli stessi risultati, riuscendo ad essere più veloce del metodo corrente.

# Contents

# Chapter 1

# Introduction

What does “improving speech quality” mean? It is very hard and complex to explain, however, it can be summarized as the improvement in intelligibility, and, overall, perceptual clarity and pleasantness of the degraded speech signal.

Speech enhancement and noise reduction aim to do this: improve the speech quality of a noisy signal by removing the background noises with a wide variety of techniques. Over the last years, this subject has been an important research topic due to the fact that it is required in many situations in daily life. The most common example is telephone communication. Just think about a phone call in which one subject is in a noisy environment such as a street or inside a car. The noise reduction can attenuate or totally eliminate the background disturbance and make the communication clearer. It also applies to the communication over the internet, (for instance Skype calls), and in other areas such as speech/speaker recognition and transcription systems, hearing aids, cochlear implants and restoration of degraded registrations. At the same time noise reduction is a very challenging and complex problem due to several reasons. First of all, the nature of the noise changes significantly from application to application, and also changes over time. It is therefore very difficult, if not impossible, to develop a versatile algorithm capable of working satisfactorily in all the scenarios. Secondly, the scope of a noise reduction system depends on a specific context. For example, in some applications we want to increase the intelligibility or the overall speech perception quality, while in other cases, we try to get better accuracy for an ASR system or simply to reduce the listeners’ fatigue. Therefore, considering the complex characteristics of speech and the large amount of restrictions, it gets even more complicated to satisfy all objectives at once. Traditional noise reduction methods, such as Spectral Subtraction and Wiener

filtering, are based on strong stationary assumption and do not work satisfactorily in the presence of real non-stationary background noise [1].
In this work, in order to overcome this problem and all the other issues already described, we decided to exploit the properties of the already known Non-Negative Matrix Factorization (NMF) [2] technique for source separation and speech enhancement. We applied this technique to an exemplar-based sparse representation approach that has recently gained greater interest in a broad range of signal processing.
In this approach, the observed signal is decomposed into a combination of a small number of elementary parts, called atoms or bases, and their weights called activations. Following the idea depicted by Gemmeke in [3], the observed speech is decomposed into speech atoms, noise atoms and their weights. The collection of all the bases is usually called dictionary and these bases are grouped together for each source, such as speech and noise. By only using the atoms and the weight, related to a specific source dictionary, an estimation of the desired signal can be reconstructed [3–7]. As proposed by Takashima [8] the reconstruction can be performed with the weights of the speech bases extracted from the test signal together with the desired dictionary bases, the target dictionary, to reconstruct an enhanced version of the test signal. In this thesis we investigate the effectiveness of combining two approaches, the one proposed in [3] and in [8], in order to obtain en enhanced speech signal of any input signal. This will be performed without using the same sentence in the training and in the enhancement part, differently from how it was done till now or how it was done in [8, 9]. We want to develop a more general supervised single-channel speech enhancement algorithm based on a sparse exemplar-based NMF. To do so, we will focus our attention over the NMF factorization, specifically over the dictionary used and its reconstruction. We propose three new approaches in order to improve the results obtained with the previous methods. This way, we achieve an improvement of the two above described enhancement methods.

## 1.1 Objectives

The goal of this thesis is to develop an algorithm able to reconstruct an enhanced version of any speech signal corrupted by non-stationary real background noise with a supervised single-channel sparse non-negative matrix factorization technique. For this purpose we use three different approaches. The first one uses a noisy speech dictionary for the NMF factorization and a clean speech dictionary for the full-band denoised reconstruction task. The clean and the noisy dictionaries are obtained from the same utterance, recorded by the same user, but in a noisy environment. Ideally this noisy speech signal can be captured just before the NMF speech enhancement process. It means that the user can record a short training signal just before using this system. The fundamental point is that this sentence has to be the same as the clean one already provided, since the system needs to align the two of them. This is the same method proposed by Takashima [8]. On the other hand, the second approach only uses a clean speech dictionary for both tasks (factorization and reconstruction). So there is no need to capture a short training audio before starting the process. The last approach also uses a clean speech dictionary for the factorization and the reconstruction, but the dictionary is refined in order to contain also speech features. These features will be used to control simultaneous activations of different kinds of exemplars, as explained in section 3.1.3.

In order to do so, first we investigate the fundamental aspects to take into account for the best result, such as: the dictionary dimension needed for a good decomposition; the bases selection; the sparsity constraint and the reconstruction techniques.

Using jointly all these methods, we carried out many experiments, mixing various settings, and verifying which of these configurations provide the best results.

## 1.2 Structure of the thesis

The work is divided in five chapters that are organized in the following way:

- Chapter 2: State of the Art. We start by introducing the basic theoretical background and the fundamental assumptions on which this thesis and the past works rely. First we provide a general classification of the speech enhancement techniques organized by the assumption on which they are based. Secondly we depict the most relevant techniques for speech enhancement developed nowadays with their properties and drawbacks. Moreover we describe in detail some particular aspects that we will use in our system.

- Chapter 3: Methods. In this chapter we discuss the three approaches used in this work, showing the various options and settings used for each one, based on the theory and others author's solutions previously explained.

- Chapter 4: Evaluation. We describe the validation methods used in order to verify the quality of the enhancement obtained with the chosen approach. Afterwards we show the results obtained for each approach and for each employed characteristics of the enhanced system.

- Chapter 5: Conclusions and future works. Finally we expose the conclusions obtained and we outline the possible future developments and directions of research.

# Chapter 2

# State of the Art

## 2.1 Theoretical Background

Since we live in a natural environment where noise is inevitable and ubiquitous, speech signals can rarely be recorded in pure form and are generally contaminated by the acoustic background noise. As a result, it is essential for speech processing and communication systems to apply effective noise reduction in order to extract the desired speech signal from its corrupted observation without affecting the speech signal quality.
There are quite a lot of solutions for this problem, but they can be classified according to the assumptions on which they are based as:

- **Number of channels available**
  Speech enhancement techniques can be divided into multichannel and single channel systems. The first one uses multiples microphones and it is able to exploit, in addition to spectrum, the spatial information to estimate the desired clean speech from a noisy signal. This technique uses the fact that speech source is quite stationary and therefore, by using beamforming techniques, the system can suppress non-stationary interferences more effectively than any single sensor system as proved by [10, 11]. In general, the more microphones, the easier the task of noise reduction gets.
  The second system uses just a single acquisition channel; temporal and spectral information of speech and noise are extracted from a single noisy signal.
  Multiple microphone systems are very powerful, but they require a larger number of microphones to be effective and this has an important cost, other than the fact that they are not common in conventional devices.

For a practical reason, in this work we decided to use the more flexible single channel system approach of less hardware requirements and lower costs.

- **Supervised or Unsupervised**
  In the unsupervised method, a statistical model is assumed for the speech and noise and the clean speech is estimated from the noisy signal without any a priori information on the noise. No supervision or labeling is required for the speech and the noise. The main difficulty is the estimation of the noise Power Spectral Density (PSD) that can be very complex if the background noise is non-stationary. This approach is used by spectral subtraction, Wiener filtering, short-time spectral estimators and others.
  Differently, the supervised methods use a model for both the noise and the speech signal estimated using training samples of each category. The noise reduction task is carried out with the use of these speech and noise models combined.
  The main advantage is that there is no need to previously estimate the noise Power Spectral Density (PSD). These methods give better results due to the wider a priori information provided to the system by the user or a built-in classification algorithm and due to the ad hoc training of the system for each specific signal. This approach is used by HMM and NMF methods.

- **Statistical relationship between noise and speech**
  In this case we consider how the noise is assumed respect to the speech. If the noise is assumed as uncorrelated or even independent, or if the noise is taken as correlated, such as echo and reverberation. In this work we consider the noise and the speech as uncorrelated.

- **Speech and noise mixing model**
  This difference identifies how the noise and the speech are mixed: if the noise is considered additive, multiplicative, or convolutional with respect to the speech. Noise reduction approaches are usually based on the assumption that the speech and the noise are additive. This assumption is not valid in a real noise case. However, it is usually assumed to be true since it makes the problem simpler and leads to satisfactory results in practice, as proved by all the papers and the already developed algorithms [2–5]. This means that we can assume the speech signal to be modeled as the sum

$$y(t) = s(t) + n(t), \tag{2.1}$$

where $s(t)$ represents the pure speech signal, $n(t)$ is the additive noise and $y(t)$ represents the degraded speech signal.
Multiplicative noise also refers to a similar model but in this case the unwanted random signal gets multiplied into the relevant signal: $y[t] = s[t]n[t]$. One common way to remove multiplicative noise is to transform it into the better known additive noise model.
Differently to the previous cases, the convolutional noise is a type of noise originated by some differences in transmission channels caused by the changes of the distance between the mouth and the microphone, of the microphone characteristics or of the recording environment. In this model, each source contributes to the total sum with multiple delays corresponding to the multiple paths by which an acoustic signal propagates till the microphone. These differences cause a model mismatch between the training and testing conditions.

## 2.2 Spectral Subtraction

This is one of the earliest, simplest and less expensive, in terms of computation requirements, speech enhancement class of algorithms. This method is based on the principle that the enhanced speech can be obtained by subtracting the estimated spectral components of the noise from the spectrum of the input noisy signal. This technique can be implemented in power or magnitude spectral subtraction and it makes the assumption that the noise is additive, uncorrelated and stationary or slowly varying in a short-term. So we assume the speech signal model, as explained before in (2.1), is

$$y(t) = s(t) + n(t). \tag{2.2}$$

Where $s(t)$ represents the pure speech signal, $n(t)$ is the uncorrelated additive noise and $y(t)$ represents the degraded speech signal. In the frequency domain, the noisy signal model becomes

$$Y(f) = S(f) + N(f), \tag{2.3}$$

where $Y(f), S(f)$ and $N(f)$ are the Fourier transforms of the noisy signal $y(t)$, the clean signal $s(t)$ and the noise $n(t)$ respectively, and $f$ is the frequency variable. The incoming signal $s(t)$ is buffered and divided into segments of $K$ samples. Each segment is windowed and then transformed via discrete Fourier transform (DFT) to $K$ spectral samples. The windowing alleviate the effects of the discontinuities at the endpoints of each segment and can be expressed in the frequency domain as:

$$Y_k(f) = S_k(f) + N_k(f). \tag{2.4}$$

Where $k$ is the segment index, $Y_k(f), S_k(f)$ and $N_k(f)$ denote the short-time DFT magnitudes taken of $y(t), s(t)$ and $n(t)$, respectively. If an estimate of the noise spectrum $\hat{N}_k$ can be obtained, then an approximation of speech $\hat{S}_k$ can be achieved from $Y_k$, expressed as:

$$\hat{S}_k(f) = Y_k(f) - \hat{N}_k(f). \tag{2.5}$$

Due to the fact that quite often the pure noise spectrum is not available, it can be estimated during period when no speech is present in the input signal. Most single channel spectral subtraction methods use a voice activity detector (VAD) to determine when one frame contains silence or not in order to get an accurate noise estimate. The noise is assumed to be short-term stationary, so that noise of the silence frames can be used to remove noise from speech frames.

The phase used for the reconstruction is the same of the observed signal and is kept untouched. This assumption rely on the fact that audible noise is mainly due to distortion of the spectrum, and that the phase distortion is largely inaudible in human perception, as said by Vaseghi in [12]. At the same time, due to random variations of noise, spectral subtraction can lead to negatives values. But magnitude or power spectrum are non-negative values and so they have to be mapped into non-negative values. This rectification process distorts the restored signal introducing the so called musical noise, due to their narrow-band spectrum and the tone-like characteristics. This is a perceptual phenomena that occurs when isolated peaks are left in a spectrum after processing it. Particularly in silence sections these isolated components sound like musical tones and in speech present sections it produces a “warble” of the speech. This processing distortion becomes more noticeable as the SNR ratio decreases. This phenomenon can be explained by noise estimation errors leading to false peaks in the processed spectrum. When the enhanced signal is reconstructed in the time-domain, these peaks result in short sinusoids whose frequencies vary from frame to frame.

Most of the research, at the present time, is focused in ways to combat this problem [13]. It is almost impossible to minimize the musical noise without affecting the speech quality, and hence there is a trade-off between the amount of noise reduction and speech distortion. The success of spectral subtraction depends on the ability of the algorithm to reduce the noise variations and to remove the processing distortions.

## 2.3 Wiener Filter

Wiener first formulated the continuous-time, least mean square error, estimation problem in his classic work of interpolation, extrapolation and smoothing of time series in 1949 [14]. This approach uses a filter to produce an estimate of the clean signal from an observed signal corrupted by additive noise. It is a technique based on a statistical approach, where the Mean-Square Error (MSE) is minimized between the estimated signal magnitude spectrum $\hat{S}(f)$ and the original magnitude spectrum $S(f)$. Therefore, we obtain a filter to apply to the noisy signal to filtering out the noise and so reducing the background noise. The Wiener filtering can be considered one of the oldest noise reduction technique.

As the spectral subtraction, this method also relies on the assumption that the noise is stationary, uncorrelated and additive with known spectral characteristics with respect to the clean signal. Typically, a signal is not stationary, but, as we said before, we can assume it stationary in a short-time period without making excessive approximations. So, for an additive noise, the noisy signal model is the same as in (2.1).

The signal is then segmented into overlapped frames, each frame multiplied by a window, and the DFT is applied to the windowed data in order to convert the noisy signal into his frequency-domain version. Letting $r$ represent the frequency bin and $m$ the short-time frame indices, we simplify the notation of the DFT coefficients corresponding to frame $m$ of the noisy signal by $y_m = y_{rm}$, where $y_{rm}$ is the $r$ -th element of $y_m$. So the vector of the DFT coefficients of the clean speech and noise signals are represented by $\hat{y}_m$ and $n_m$, respectively. The clean speech DFT coefficients are estimated by an element-wise product of the noisy signal $\hat{y}_m$ and a weight vector $\hat{h}_m$

$$\hat{y}_m = h_m \circledast y_m, \tag{2.6}$$

where $\circledast$ denotes an element-wise product. The weight vector $h_m$ is the Minimized Mean-Square Error (MMSE) between the clean and estimated speech signal. Assuming that different frequency bins are independent, we can minimize the MSE for each individual frequency bin $r$ separately

$$h_{rm} = \underset{h_{rm}}{\operatorname{argmin}} E(|y_{rm} - \hat{y}_{rm}|^2). \tag{2.7}$$

Setting the partial derivative of the real and imaginary parts of $h_{rm}$ to zero, and assuming that the speech and noise signals are zero-mean and uncorrelated, the optimal weights are obtained as:

$$h_{rm} = \frac{E(|s_{rm}|^2)}{E(|s_{rm}|^2) + E(|n_{rm}|^2)}. \tag{2.8}$$

A fixed filter requires a priori knowledge of both the signal and the noise. If we know the signal and noise beforehand, we can design a filter that leaves undistorted the frequencies containing signal and rejects the frequency band occupied by noise.

On the other hand, if the filter coefficients are periodically recalculated for every block of $m$ signal samples, then the filter adapts itself to the characteristics of the signals within the blocks and becomes block-adaptive. This kind of filter has the ability to adjust its impulse response to filter out the signal in the input with little or no priori knowledge of the signal and the noise characteristic. Furthermore, considering the speech stationary over a relative small block of samples, it helps to take into account the non-stationarity of the signal.

Noise reduction is a variation of optimal filtering that involves producing an estimate of the noise and filtering the reference input subtracting this noise estimation from the input containing both signal and noise.

The problem here, as in most of unsupervised speech enhancement methods, is the need to estimate the noise power spectral density (PSD), $E(|n_{rm}|^2)$. The simplest approach for this purpose is to use a voice activity detector (VAD). In this approach, the noise PSD is updated during the speech pauses. These methods can be very sensitive to the performance of the VAD and cannot perform very well at the presence of a non-stationary noise. There are some alternative methods that use the statistical properties of the speech and noise signals to continuously track the noise PSD. Comparing several of these methods, it results that the MMSE approach was found to be the most robust noise estimator among the considered algorithms [15].

## 2.4 Hidden Markov Model

Enhancement methods that are based on stochastic models, as Hidden Markov models (HMM), have become very important in digital signal processing by modeling both clean speech and noise, and by accommodating the non stationarity of speech and noise with multiple states connected with the transition probabilities of a Markov chain. Using multiple states and mixtures for the noise, HMM enables the speech enhancement system to relax the assumption of noise stationarity [16–18].

The Markov Model consists on generalizations of a mixture model in which the system being modeled is assumed to be a Markov process, and a Markov process is a stochastic process that satisfies the conditional probability distribution of future states of the process to be dependent only upon the present state (Markov Property). In particular, a HMM is a Markov Model with unobserved (hidden) states that control the mixture component and have a probability distribution over the possible output. These hidden states are not directly visible, but we can observe the output depending on the state.

The HMM is designed to capture time varying signal's statistics and it can be considered as a generalization of the Gaussian Mixture Model (GMM). As it is difficult to formulate a continuously time varying model, the HMM shapes it through a state to state transition, therefore, this approach can be considered as the discretization of the continuous varying case.

A general HMM consists of several inter-connected states and each state is a GMM. The model jumps from one state to another according to the signal transition relationship between the states. Thus an HMM consists of a discrete Markov chain and a set of state-conditional probability distributions shown by $P(s_t|z_t = j)$ , $j \in \{1 \dots J\}$ where $J$ is the number of states in the HMM. The Markov chain itself is characterized by an initial probability vector over the hidden states, denoted by $q$, with $q_j = P(z1 = j)$ and a transition matrix between the states, denoted by $A$ with elements $a_{ij} = P(z_t = j|z_t1 = i)$ and a mixture weight, or the parameters of the output, represented with $\theta$. The HMM parameters can be expressed as:

$$\lambda = \{q, A, \theta\}. \tag{2.9}$$

The model parameters in HMM are usually estimated by the Maximum Likelihood (ML) estimation criterion. The ML estimate of $\hat{\Theta}$ is obtained by maximizing the likelihood of the observation with respect to $\Theta$ as

$$\hat{\Theta}_{ML} = \underset{\Theta}{\operatorname{argmax}}\{P(y|\Theta)\}. \tag{2.10}$$

Other two lately used estimation criteria are the Minimum Mean-Square Error (MMSE) criterion and the Maximum-A-Posteriori (MAP) criterion, both belonging to the bayesian estimator theory and expressed by

$$\hat{x}_{MMSE} = \underset{x}{\operatorname{argmin}}\{E[\|\hat{x} - x\|^2]\}. \tag{2.11}$$

The most significant difference between ML and the bayesian estimators is that in the classical estimation theory, the parameter to be estimated is treated as an unknown constant; in the Bayesian estimation theory, it is considered as a unknown random variable. If the prior information about the parameter is available, the Bayesian estimation theory enables the use of the prior information to produce more accurate estimate than the classical estimation theory.

The HMM structure of the model in speech enhancement is different from the model used for speech recognition. The objective in speech enhancement is to average out the noise signal and extract the general spectral characteristics of speech regardless of the phoneme or sentence pronounced. The goal is to distinguish speech from noise and not to distinguish different units of speech. We want to accommodate all the speech characteristics in a single, compact model. This model is not supposed to capture distinctive properties of speech within different utterances but to capture the global characteristics of speech. Furthermore, the temporal order of the states in the model does not need be constrainted since there is a single, global model for speech and different state sequences for the same state ensemble can represent distinct utterances. As a result, the speech model for enhancement is structured to be ergodic, so there are no constraints on the transition probabilities of the HMM. This makes the model less redundant since each distinct spectral shape of speech or noise needs to be represented only once in the model.

In general, there is a huge number of diversified types of noise with very time-varying spectral characteristics, but the HMM based enhancement systems are inherently relying on the type of training data for noise. Expectedly, such a system can handle only the type of noise that has been used for training the noise in HMM. Therefore, data from various noise types should be used for training the noise HMM. This creates the problem of a large model size for the noise HMM, making the search space expand linearly with the number of noise types with computation cost growing drastically. Furthermore, the unwanted large search space deteriorates the system performance by introducing more sources of error in the MMSE forward algorithm. There are some new researches about noise adaptation algorithm that aim to enable the system to handle arbitrary types of corrupting noise and to avoid over-growth in computational complexity as proposed by Sameti [19, 20], and Mohammadiha [21].

## 2.5 Non-Negative Matrix Factorization

Non-Negative Matrix Factorization (NMF) is a method in linear algebra in which a matrix $V$ is factorized into two matrices , $W$ and $H$, subject to the non-negative constraints: $V_{ij} >= 0, W_{ij} >= 0, H_{ij} >= 0$. In our case, W is the matrix of the bases vectors (feature space) and $H$ is the weight matrix, or the activations of these bases along the time

$$V \approx WH. \tag{2.12}$$

NMF was introduced as a concept by a Finnish group of researchers in the middle of the 90s with the name of Positive Matrix Factorization [22] and it became more widely known as Non-Negative Matrix Factorization after the publication of Lee and Seung (1999) [2], who investigated the properties of the algorithm and proposed some useful algorithms for its computation.

NMF can be applied to the statistical analysis of multivariate data in the following manner: given a set of $R$-dimensional data vectors, the vectors are grouped into the columns of a matrix $V \in \mathbb{R}^{RxM}$ where $M$ is the number of bases in the data set and $R$ is the number of features. This matrix is then approximately factorized into a matrix $W \in \mathbb{R}^{RxK}$ and a matrix $H \in \mathbb{R}^{KxM}$ where $K$ is the dimension of the decomposition.

The meaning of this decomposition is that the input matrix $V$ can be rewritten column by column as $v \approx Wh$ , where $v$ and $h$ are the corresponding columns of $V$ and $H$. In other words, each data vector $v$ is approximated by a linear combination of the columns of $W$ , weighted by the components of $h$. Therefore, $W$ contains the bases that are optimized for the linear approximation of the data in $V$, depending on the used algorithm. Since with any arbitrary invertible matrix $Q \in \mathbb{R}^{KxK}$ we get $V = WH = (WQ^{-1})(QV)$, the problem is not exactly solvable and $W$ and $H$ are not unique. For this reason it is commonly approximated numerically with some approximation constraints that characterize the different types of NMF algorithms. This differentiation emerges from the use of various cost functions for measuring the difference between $Y$ and $WH$ and possibly the regularization of the $W$ and/or the $H$ matrices. Since there could be many possible solutions, it is important to enforce additional constraints to ensure the uniqueness of the factorization. This control is achieved by enforcing sparsity constraints over the activation of each base of the matrix $W$, obtaining a more sparse and unique representation. Sparse representation idea states that, since relatively few bases vectors are needed to represent many data vectors, a good representation can only be achieved if these bases vectors are the ones that best fit the structure of the data analyzed. Because of this aspect, the NMF is particularly interesting, as it is able to perform $K$-means data clustering depending on the sparsity constraint adopted [23]. We will analyze more accurately this aspect in the dictionary creation section 2.7 and in the sparseness section 2.8.

Another important aspect is that the bases vectors $W_i$ can also be not orthogonal; they can overlap each other because they are treated as different basic elements. This allows us to use overlapping bases of more than one frame size (this aspect will also be analyzed later in section 2.6). But what makes this model particularly interesting and capable of giving good interpretability is the constraint that the matrices $V$, $W$, and $H$ are all non-negative, $V, W, H \in \mathbb{R}^{\geqslant 0}$. This point ensures that the frame vector $v_j$ of the $j$ column of matrix $V$, made by the factor matrices $W$ and the $j$ column of $H$, can be interpreted as constructive building blocks of the input as:

$$v_j \approx \sum_{i=1}^{K} (w_i h_{i,j}) = W h_j. \tag{2.13}$$

This interpretation is not applicable to decompositions that employ negative-valued entries, because in such decompositions, the elements of $W$ and $W$ can cancel each other [24]. Moreover negatives values are meaningless and hard to explain in real applications. When NMF is applied to data that was generated by mixing a number of non-negative sources, the NMF decomposition is able to discover and separate remarkably well the contributions of each source in the mixture data.

To employ the NMF factorization we need to transform our inputs into an additive non-negative representation. Since most natural signals tend to be sparse in the magnitude or power, by using these transforms we can often guarantee, with high probability, that the transform of the sum of two sources will be equal, or approximately equal, to the sum of the transform of the two sources separately, which can satisfy the additivity constraint.

### 2.5.1 Cost Function

To find an approximate factorization of $V$ as $WH$, we first need to define a cost function that quantifies the quality of the approximation and the success of the reconstruction. Such cost function can be constructed using some measure of distance between two non-negative matrices $A$ and $B$. One useful and very used measure is simply the square of the Euclidean distance between $A$ and $B$ [25]:

$$||A - B||^2 = E_{ij}(A_{ij} - B_{ij})^2. \tag{2.14}$$

Another well know and largely adopted measure is the generalized Kullback-Leibler divergence (KL):

$$d(A||B) = \sum_{ij} \left( A_{ij} \log \left( \frac{A_{ij}}{B_{ij}} \right) - A_{ij} + B_{ij} \right). \tag{2.15}$$

This is a non-symmetric measure of the difference between two probability distributions $A$ and $B$, expressed as $d(A||B)$, when $B$ is used to approximate $A$. Originally introduced by Solomon Kullback and Richard Leibler in 1951 [26] as the direct divergence between two distributions, it can not be called a “distance”, because it is not symmetric in $A$ and $B$. Therefore we will refer to it as the “divergence” of $A$ from $B$.

The distance measure should be chosen according to the properties of the data: Euclidean distance assumes additive Gaussian noise meanwhile KL assumes Poisson observation model where the variance scales linearly with the model.
In source separation methods, the KL divergence has been found to produce better results than, for example, the Euclidean distance [5].
There are other very interesting cost functions for the NMF, such as the Bregman divergences [27], the parametric generalized divergence introduced by Kompass [28] and the family of $\beta$-divergence [29]. However, the KL divergence coincides up to a factor with the $\beta$-divergence

introduced and analyzed by Eguchi and Kano [30], Fevotte [31] and the family of Csiszar divergences, to which Amari's $\alpha$-divergence belongs [32].
Another very interesting cost function for future works, which is a limit case for the $\beta$-divergence is the Itakura-Saito (IS) divergence [33, 34] expressed by

$$d_{is}(A||B) = \frac{A}{B} - \log\left(\frac{A}{B}\right) - 1. \tag{2.16}$$

The convergence of this algorithm is observed only in practice, and the proof is still an open problem. The IS-NMF allows to derive a new type of minimization method, derived from Space Alternating Expectation-Maximization (SAGE), a variant of the standard Expectation Maximization (EM) algorithm. This method leads to new update rules, which do not possess a multiplicative structure, but the EM guarantees that is still converging to a stationary point of the cost function.
Concluding, in this work we decided to adopt the well know and robust KL divergence, since it is the cost function with the best performance and has been found to produce better results than the other distances [5].

### 2.5.2 Multiplicative Update Rules

Although the functions $||V-WH||^2$ and $d(V||WH)$ are convex in $W$ or $H$, they are not convex in both variables together. Therefore, it is not possible to expect an algorithm to solve the problem of minimizing the Euclidean and the KL divergence in the sense of finding global minima. However, Lee and Seung [2] proposed a multiplicative update rule for each method in order to optimize the task of finding the local minima. Their multiplicative methods were found to provide a good compromise between speed and ease of implementation for solving this problem. They prove that the Euclidean distance $||V - WH||^2$ is not increasing under the update rules

$$H_{a\mu} \leftarrow H_{a\mu}\frac{(W^TV)_{a\mu}}{(W^TWH)_{a\mu}} \qquad W_{ia} \leftarrow W_{ia}\frac{(VH^T)_{ia}}{(WHH^T)_{ia}}, \tag{2.17}$$

and also the KL divergence $d(V||WH)$ is non increasing under the update rules

$$H_{a\mu} \leftarrow H_{a\mu}\frac{\sum_i W_{ia}V_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}} \qquad W_{ia} \leftarrow W_{ia}\frac{\sum_\mu H_{a\mu}V_{i\mu}/(WH)_{i\mu}}{\sum_\nu Ha\nu}. \tag{2.18}$$

At every iteration of our algorithms, the new value of $W$ or $H$ is found by multiplying the current value by some factor that depends on the quality of the approximation. It is proved that the quality of the approximation improves monotonically with the application of these multiplicative updated rules [2, 35]. In practice, this means that repeated iterations of the update rule guarantee the convergence to a locally optimal matrix factorization.

There are also other update methods proposed by other authors, such as the Alternative Non-Negative Least Square (ANLS) suggested by Paatero [25], a projected gradient method by Lin [36], active set method by Kim and Park [37] and a coordinate descent method by Hsieh [38].

### 2.5.3 Convolutive Non-Negative Matrix Factorization

The basic NMF described before works well with many audio tasks. However, it does not take into account the relative positions of each spectrum, thereby discarding the temporal relationship between multiple observations over close time intervals.

In the conventional NMF, each object is described by its spectrum and his corresponding activation in time, while for convolutive NMF (CNMF or NMFD), each object has a sequence of successive spectra and a corresponding activation pattern across the time. So, in the original NMF, an exemplar must accurately match the analyzed spectral characteristics in order to be used. When the window's length, the exemplar dimension, is increased, it becomes less likely that a matching exemplar will be found in a limited dictionary. To overcome this problem Smaragdis [39, 40], Gemmeke [3], Saedi [41], Hurmalainen [42], Weninger [43], Carlin [44] and O'Grady [45] introduced an extended version of NMF which deals with this issue. On the other hand, for convolutive NMF it is sufficient to find a single temporal position, where an exemplar matches the observed speech segment in order to use it, as demonstrated by Hurmalainen [42].

In the previous section, the model was expressed as in (2.12) while in the convolutive NMF, as in [39] the model is extended to:

$$V \approx \sum_{t=1}^{T} W_t \cdot \overset{t\rightarrow}{H}, \tag{2.19}$$

where $V \in \mathbb{R}^{\geqslant 0, NxM}$ is the input we want to decompose, and $W_t \in \mathbb{R}^{\geqslant 0, NxK}$ and $H \in \mathbb{R}^{\geqslant 0, KxM}$ are the bases and weights, or activation, matrices. $T$ is the length of each spectrum sequence. The $i$-th column of $W_t$ describes the spectrum of the $i$-th object at $t$ time steps after the object has begun. The operator $\overset{i\rightarrow}{(\cdot)}$ shifts the columns of its argument by $i$ slots to the right. So that:

$$M = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}, \overset{0\rightarrow}{M} = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}, \overset{1\rightarrow}{M} = \begin{pmatrix} 0 & a & b \\ 0 & d & e \end{pmatrix}, \overset{2\rightarrow}{M} = \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & e \end{pmatrix}, \ldots, \quad (2.20)$$

while the opposite operator $\overset{\leftarrow i}{(\cdot)}$ shift the columns in the opposite direction by $i$ spots as:

$$M = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}, \overset{\leftarrow 0}{M} = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}, \overset{\leftarrow 1}{M} = \begin{pmatrix} b & c & 0 \\ e & f & 0 \end{pmatrix}, \overset{\leftarrow 2}{M} = \begin{pmatrix} c & 0 & 0 \\ f & 0 & 0 \end{pmatrix}, \ldots. \quad (2.21)$$

The left columns of the matrix are appropriately set to zero in order to maintain the original size of the input.

Just as before, our goal is to find a set of $W_t$ and a $H$ to approximate $V$ as best as possible. So, introducing $\Lambda = \sum_{t=1}^{T} W_t \cdot \overset{t\rightarrow}{H}$ as the estimated approximation of $V$ we can defined the KL divergence adapted cost function as:

$$d(V||\Lambda) = \left\| V \otimes log(\frac{V}{\Lambda}) - V + \Lambda \right\|_F , \quad (2.22)$$

where $\otimes$ is an element wise multiplication and $|| \cdot ||_F$ is the Frobenius norm.

To optimize this model we can use the same strategy used in the conventional NMF presented before. This new cost function can be considered as a set of $T$ conventional NMF operations that are summed to produce the final result. Consequently, as opposed to updating two matrices ($W$ and $H$) as in normal NMF, $T$ matrices require an update, including all $W_t$ and $H$ plus some shifting to appropriately line up the arguments. Likewise, we define the inverse operation which shifts columns to the left. The resultant convolutive NMF update equations are:

$$H \leftarrow H \otimes \frac{W_t^T \cdot \overset{\leftarrow t}{\left[\frac{V}{\Lambda}\right]}}{W_t^T \cdot 1} \quad , \quad W_t = W_t \otimes \frac{\frac{V}{\Lambda} \cdot \overset{t\rightarrow}{H}^T}{1 \cdot \overset{t\rightarrow}{H}^T} \quad , \quad \forall t \in [0 \ldots T-1]. \quad (2.23)$$

Can be easily seen that for $T = 1$ this equation will reduce to conventional NMF. At every iteration we update both $H$ and $W_t$ for each $t$. This way, is possible to optimize the factors in parallel and account for their interplay. Due to the fast convergence properties of the multiplicative update there is the risk that $H$ can be more influenced by the last $W_t$ used for its update, rather than the entire ensemble of $W_t$ [36].

## 2.6 Multiple Frame Exemplars

In order to decode utterances of arbitrary length, and considering time continuity of speech, a multiple frame bases approach is commonly adopted. The methods proposed till now are mostly based on sliding windows, as utilized by Gemmeke [3]. This technique consists in dividing an utterance magnitude spectrogram into a number of overlapping, fixed-length windows, with the window length, in frames, equal to the $T$ size that we want to give to our exemplars. Sliding this window along the magnitude spectrum, using a shift of $\Delta$ frames, provides a sequence of windowed segments $W_b = [W_1, \ldots, W_B]$, where $B = M - T$ is the number of possibles windows in the utterance (with $\Delta = 1$), as $M$ magnitude length and $T$ window length both in frames. Each multiple frame window matrix is then reassembled into an observation matrix of dimension $E = NxT$ and placed in a vector of matrices $W \in \mathbb{R}^{ExB}$ where the observation vectors are the columns, going by the name of multiple frame exemplars.

In this work, we propose the same idea but slightly different in order to adopt the NMFD proposed by Smaragdis [39]. The windowing technique proposed reduces the total amount of windows that we need to calculate, by selectively computing them only over the chosen index in the magnitude spectrogram. This is done by selecting one frame and then creating the window around it, with the window dimensions equal to the exemplar size $T$. The central frame can be selected in two ways: the first one is randomly, as in [3, 5], and the second one is made by using a maximum energy factor. This second method selects the frames with highest quantity of energy in the training signal, in order to make sure that we are not considering low energy or silence frames. In fact, these kinds of frames do not have much speech information. The maximum energy (ME) calculation is done by computing the energy of each frame as:

$$ME(b) = \sum_{f=1}^{N} |X(f, b)|^2. \tag{2.24}$$

Once this value is obtained, we can sort all the frames and only take the first $K$ highest energy frames, discarding the others.
The index frame position $l$, selected with both methods, ranges from l $\in$ [floor(T/2) ... M-round(T/2)]. To avoid that high SNR background noise corrupts this picking, we decided to adopt this approach on the clean speech already aligned. Only then we can select the corresponding noisy speech frames (in the case we are using the first approach). Once the frame is selected, randomly or by maximization of energy factor, the window of $T$ frames is taken around it to preserve the temporal continuity and to allow the NMFD to better perform the recognition. The process is repeated for all the $K$ frames composing the dictionary (see 3.1.2). $K$ is also the rank of the NMF decomposition.
The multiple frames window $W_l$, also called multiple-frames exemplar, of dimension $\in \mathbb{R}^{NxT}$, is then positioned inside an observation matrix as a column. At the end of the process the observation matrix, also called dictionary, will be a three dimensional matrix of $\in \mathbb{R}^{NxKxT}$ composed of matrices as columns, representing each one an exemplar (as described in the figure 2.1). This operation is used for both the speech and the noise dictionary. This is made by following the order of the frames' index. Either way, the order is not important as we do this after the alignment, see alignment chapter (3.1.1), and due to the fact that the NMFD is able to recognize the contribution of each exemplar even if they are disordered.

## 2.7 Activation Decision

A flexible approach for modeling temporal correlations during the reconstruction task consist in imposing constraints on the model activations as proposed by Smaragdis and Févotte in [24]. Following this assumption, at the beginning of this work, we introduce the idea of deciding which activation to use and which to discard for the reconstruction, without making any assumptions on the sparseness of the decomposition. This was done through different approaches. The first one was to only select the biggest value in the activation matrix for each frame and to put to zero all the others, but, in high SNR background noise, this results in the selection of noise activations of silence frames. Therefore, we decided to use this method only on the speech part of the activation matrix. However, this introduces the problem

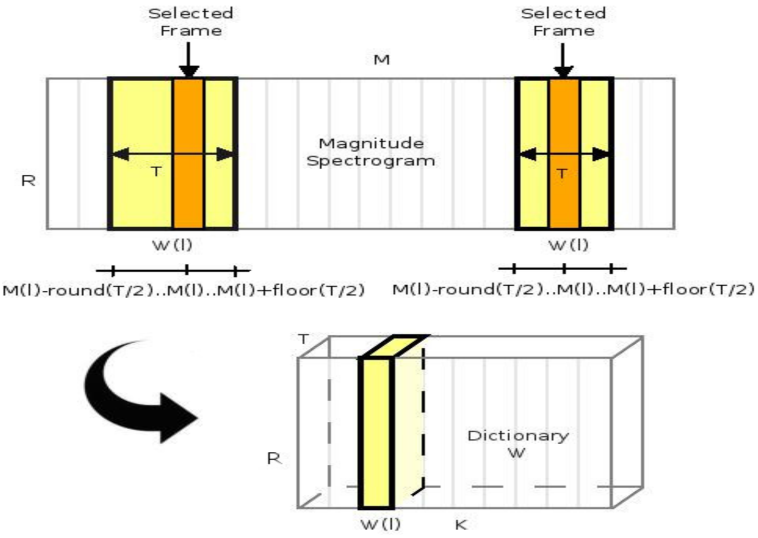


*Figure 2.1: Multiple frames exemplar l and dictionary W creation.*

of a forced selection of some frames inside the silence part that are not contributing on the correct reconstruction. This introduces some utterances that are not present in the originally sentence. Faced with this, we mixed the selection of the maximum activation value in the whole activation matrix, but discarding it if it is in the noise part of this matrix. Although it worked better, there is still the problem of errors in selection and unpleasant reconstruction.

Instead of focusing our attention on the activations matrix, we decide to get a more accurate and refined dictionaries that only contain voiced bases. By doing so we avoid silence exemplars that can be recognized as noise (see 3.1.2). This technique is used in the first and in the second approaches. In the third procedure, we introduce an additional improvement for handling simultaneous voiced/unvoiced exemplars activations. During the dictionary refinement stage, see 3.1.3, we obtain the descriptors of each base and classify it as voiced or unvoiced. After the NMF factorization, when we are performing the reconstruction, if there are simultaneous active bases in the same frame we check the class of the bases (voiced/unvoiced). For the considered frame, we compare the total activation energy of the voiced and unvoiced bases and we keep only the bases of the highest energy class and vice versa. So the unvoiced exemplars are not considered in the reconstruction when they are active in parallel with voiced exemplars. This is done in order to avoid wrong additional noisy bases activation (correspond-

ing to unvoiced phonemes) during the reconstruction.
Moreover, some sparsity constraint can be applied in the approximation of the activations matrix in order to obtain a better decomposed and unique solution, as proposed by numerous authors such as Gemmeke [3] and Virtanen [5].

## 2.8 Sparsity Constraint

In order to overcome the problem of non-uniqueness of the NMF solution and to make the NP-hard problem of NMF factorization a less ill-posed obstacle, several approaches can be used. Two common techniques are based on the incorporation of an additional constraint into the NMF. They are the sparsity technique (Kim [23] and Hoyer [46]) and the minimum volume technique (Miao [47], Zhou [48]). Another approach is the orthogonality of the bases matrix W (Ding, [49]), but this condition is rather restrictive and difficult to apply to speech. A practical solution commonly used in speech enhancement is to enforce a proper penalty term in the objective function [5, 41, 45, 46, 50]. The motivation for using this approach is based on the geometric interpretation of NMF. This version shows that sparse matrices correspond to more well-posed NMF problems whose solutions are sparser, as in [51].
A source is assumed to be sparse when it is non-active most of the time, which means that the the activation matrix is zero or nearly zero most of the time. In other words, it means that the observed signal can be represented by a linear combination of a small number of atoms that have non-zero weights.
The advantage of a sparse representation is that the probability of two or more activation patterns being active simultaneously is low. Thus, sparse representations provide themselves a good separability [52]. Previous researches have shown that $H$ can be extremely sparse [50]. This means that only a few non-zero entries are sufficient to represent the observed signal with enough accuracy. Statistical interpretation behind this can be found in [53].
However, sometimes the sparseness achieved normally by the NMF is not enough. In such cases it might be useful to control the degree of sparseness explicitly imposing some constraint on the NMF.
The first problem is to choose what should be sparse: the bases vectors w of the matrix $W$ or the $h$ coefficients of the weight matrix $H$? The answer depends on the specific application and does not have a

general solution. So the choice of which to constrain, or if both or none of they, must be made through experiments.
There are many solutions to this problem, like the one proposed by Hoyer [46], a measure based on the relationship between the $l1$-norm and $l2$-norm, and the one proposed by Gillis, based on a preprocessing of the input matrix $V$ to make it sparser [51]. But one increasingly popular and powerful constraint is the one that force the rows of $H$ to have a parsimonious activation pattern for each bases in the columns of $W$. This is induced by additional penalty term, such as a $l1$-norm penalty. This is the sparsity constraint introduced by Filed [54], which has been found to be effective in obtaining sparse solutions in [34] and [5]. This sparsity constraint is used by Virtanen [5], Gemmeke [3], Schmidt [6], O'Grady [45], Kim and Park [23]. Formally, it consists in the basic NMF cost function described before combined with a penalty term on the coefficients of the weighting matrix as:

$$g(V||\Delta) = d(V||\Delta) + ||\lambda \otimes H||_p \qquad s.t. H \in \mathbb{R}^{\geqslant 0}. \qquad (2.25)$$

The left part $d(V||\lambda)$ corresponds to KL divergence, while the right term is the additional constraint on $H$. This constraint enforces sparsity by penalizing the non-zero entries of $H$ using the $l_p$-norm of the activation matrix, weighted by element-wise multiplication with $\lambda$ [45, 50, 55]. In our work we use a $l_1$-norm as constraint. The parameter $\lambda$ gives the degree of sparsity and controls the trade off between sparseness and accurate reconstruction. As proposed by Gemmeke [3], and differently from the other authors that use a single scalar value to penalize all the entries equally, the sparseness parameter $\lambda$ can be defined for each exemplar. This allows different weights for speech and noise bases by setting $\lambda_k = [\lambda_1 \dots \lambda_{K_s} \dots \lambda_{K_s+K_n}]$.
Gemmeke's investigations over the influence of the sparsity at various SNR [56] revealed that, in the presence of strong background noise, the use of sparser solutions is beneficial. This sparser solution makes the separation of speech and noise easier, but at the same time, those same values can damage the performance at high SNR.

## 2.9 Exemplar-Based Sparse Representation

The exemplar-based approach proposed in this work is based on the square root of the signal spectro-temporal distribution of energy, called the magnitude spectrogram. When describing a simple clean speech signal $S$, the magnitude spectrogram is a matrix $\in \mathbb{R}^{NxM}$ with $N$ the number of frequency bins and $M$ the duration length in frames. The magnitude spectrogram is used to ensure the non-negativity of the data, as required by the NMF. This way it is easier to satisfy the commonly adopted additivity property of speech and noise, as explained in the equation (2.1). If the data is non-negative there is no problem with negative values summation.

As mentioned by Gemmeke in [3], any arbitrary speech spectrogram $S$ can be expressed as a linear non-negative combination of speech atoms with $k = [1, \ldots, K]$ denoting the atom, or frame index. These atoms, also called exemplars, are magnitude spectrograms describing segments of the reference speech signal $S$. In a typical speech enhancement situation, the exemplars are extracted from a previous trained database, called dictionary, as described in (3.1.2). So the dictionary is usually formed by a huge quantity of different exemplars. In order to properly reconstruct the reference signal we have to weigh the contribution of each exemplar in the dictionary. Considering the non-negative weight, or activation, value of each exemplar, we can write that for a general source frame $l$ as:

$$v_l \approx \sum_{k=1}^{K} W_k h_{k,l} = W h_l. \tag{2.26}$$

The $K$ exemplars are grouped into an exemplar matrix $W$ as $W_k = [W_1, W_2 \ldots W_K]$ and the activations stacked into $h_l$, a $K$-dimensional activation vector.

If the source is considered without noise, we can adapt this model to a pure speech signal (clean speech) with $K_s$ dimension of $W$ as $W_s = [W_{s_1}, W_{s_2} \ldots W_{s_{K_s}}]$ and changing $x_l$ to $s_l$ we obtain that: $s_l = W_s h_{s,l}$. Like in the clean speech, we can also assume that the noise magnitude spectrogram can be modeled in the same way, so it can be represented by the magnitude spectrogram $N$, as a linear combination of $K_n$ noise dictionary exemplars $W_n = [W_{n_1}, W_{n_2} \ldots W_{n_{Kn}}]$ being the noise exemplar index, $h_n$ being the activation of the noise exemplars and $W_n$ the dictionary containing the noise exemplars. So, based on the additivity of speech and noise, the reference noisy signal $y$ can now be modeled, for each frame $l$, as a linear combination of both speech and noise exemplar as:

$$
\begin{aligned}
x_l &= s_l + n_l \\
&\approx \sum_{k=1}^{Ks} W_{s_k} h_{s_{k,l}} + \sum_{j=1}^{Kn} W_{n_j} h_{n_{j,l}} \\
&= W_s h_{s,l} + W_n h_{n,l} \\
&= [W_s W_n] \begin{bmatrix} h_{s,l} \\ h_{n,l} \end{bmatrix} \qquad s.t. h_{s,l}, h_{n,l} \in \mathbb{R}^{\geqslant 0} \\
&= W h_l \qquad s.t. h_l \in \mathbb{R}^{\geqslant 0}.
\end{aligned}
\tag{2.27}
$$

The concatenated clean and noise dictionary matrix $W$ has dimension $\in \mathbb{R}^{NxK}$, where $K = K_s + K_n$ and the matrix $h_l$ contains the activation energy of the clean and the noise exemplars active during frame $l$. The vector $h_l$ holds the information of how the energy of the test signal x is decomposed over the exemplars of the dictionary matrix W. Considering the entire magnitude spectrogram of the reference signal X, we can rewrite the previous equation in a more compact matrix form

$$
\begin{aligned}
X &= \sum_{l=1}^{K} x_l \approx \sum_{l=1}^{K} W h_l \\
&\approx [W_s W_n] \begin{bmatrix} H_s \\ H_n \end{bmatrix} \qquad s.t. H_s, H_n \in \mathbb{R}^{\geqslant 0} \\
&= WH \qquad s.t. H \in \mathbb{R}^{\geqslant 0}.
\end{aligned}
\tag{2.28}
$$

In order to consider temporal continuity, the solution proposed by many authors [3, 5] is to use exemplars of T multiple frames, with $T > 1$, created by windowing the reference signal as showed in section 2.6. Extending this method to these multiple frames exemplar-based approach we obtain the same representation as before but with one more dimension in both clean and noise dictionaries. In this case, each exemplar of each dictionary is composed by multiple frames concatenated. If we want to obtain only a single exemplar l at a time we obtain that

$$
V_l \approx \sum_{t=1}^{T} W_t^{(l)} (\overset{t\rightarrow}{H}).
\tag{2.29}
$$

In a more specific speech plus noise case, they are arranged in a vector w of length $K = K_s + K_n$ composed by a set of $W \in \mathbb{R}^{NxT}$ matrix exemplars. In this sense we obtaining a vector of matrices $w(k) = [W_s(1), \ldots W_s(K_s), W_n(K_s + 1), \ldots W_n(K_s + K_n)]$. All together they compose a dictionary matrix $W \in \mathbb{R}^{NxK_sxT}$. Being $H$ the energy activation matrix of each exemplar we can rewrite the model 2.27 in a convolutive manner by

$$
\begin{aligned}
X &\approx \sum_{t=1}^{T}\sum_{l=1}^{Ks} W_s^{(l)} \overset{t\rightarrow}{H_s} + \sum_{t=1}^{T}\sum_{j=1}^{Kn} W_n^{(j)} \overset{t\rightarrow}{H_n} \\
&= \sum_{t=1}^{T}\sum_{l=1}^{K} [W_s^{(l)} W_n^{(l)}] \begin{bmatrix} \overset{t\rightarrow}{H_s} \\ \overset{t\rightarrow}{H_n} \end{bmatrix} \qquad s.t.H_s, H_n \in \mathbb{R}^{\geqslant 0} \\
&= \sum_{t=1}^{T} [W_s W_n] \begin{bmatrix} \overset{t\rightarrow}{H_s} \\ \overset{t\rightarrow}{H_n} \end{bmatrix} \\
&= \sum_{t=1}^{T} W \overset{t\rightarrow}{H} \qquad st.H \in \mathbb{R}^{\geqslant 0}.
\end{aligned}
\tag{2.30}
$$

In this kind of representation, the reference signal magnitude is provided and the dictionary is fixed and extracted from an already performed training section. The goal is to estimate which activity matrix $H$ of the over-complete set of exemplars $W$ best approximates the reference noisy signal $X$. To accomplish this scope, the matrix $H$ is first initialize as unit matrix and then is repetitively updated by minimizing the difference between the test and the estimated signal magnitudes. This iterative updating of the matrix $H$ is performed until reaching a convergence point after a pre-defined quantity of iterations.

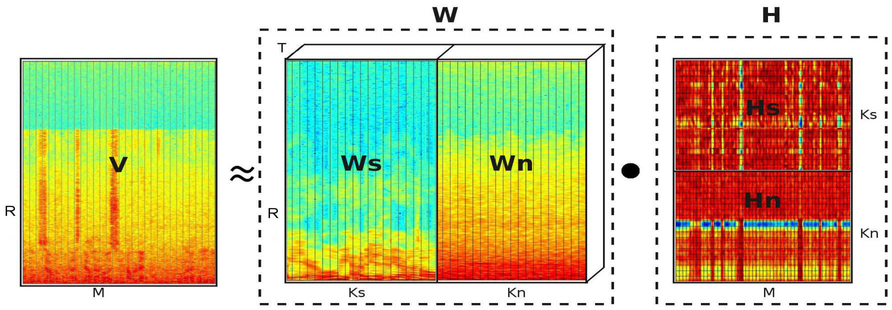


*Figure 2.2: Exemplar-based Non-Negative Matrix Factorization (NMF).*

In addition to this model, if we also add one of the before considered sparsity constraints, it is possible to say that the exemplar-based representation is sparse. Sparse representations are models that account for most of the information of a signal with a linear combination of a small number of elementary signals atoms, or exemplars. Since relatively few bases are needed to represent large spectrograms, a good approximation can be achieved if the exemplars used discover the composition that is latent inside the data. The aim of sparse representation is to reveal certain structures inside a signal and to represent it in a compact way. As stated by Gemmeke [3], Virtanen [5] and Schmidt [6], sparsity for speech is very important because it avoids the over-fitting of the representation of $H$ and forces the exemplars that are selected to be closer to the underlying, lower-dimensional speech unit in the observed speech signal.

This kind of representations have been increasingly recognized as providing extremely high performance for applications such as: noise reduction, compression, clustering, feature extraction, pattern classification and blind source separation [23, 52, 57, 58]. The generation of the sparse representation with an over-complete dictionary is non-trivial and falls inside the category of the ill-posed problems. So, the general problem of finding a representation with the smallest number of atoms from an arbitrary dictionary has been shown to be NP-hard.

In our work we give the possibility to choose over a simpler NMFD as proposed by Smaragdis [39] and implemented by Grindlay or the sparse NMFD as proposed and implemented by O'Grady [45]. This chance is given because, depending on the situation, additional sparsity constraint may worsen the result at lower SNR as proved by Gemmeke in [3].

# Chapter 3

# Methods

In this work our purpose is to enhance a reference corrupted speech signal with a real background noise using a supervised exemplar-based NMF technique. Our intention is to outperform this task by using three different approaches, that we will also call procedures.

– **Procedure 1**: The first procedure consists in using two parallel dictionaries already aligned in the training section with a DTW, as described in section 3.1.1. The first one is a noisy speech dictionary that we use, jointly with an estimated noise dictionary, for the NMF factorization and the speech activity estimation. The second one is is a clean speech dictionary that we use for the reconstruction of the enhanced reference signal. Once we have obtained the weight of each base from the NMF factorization of the test signal, we can discard the noise activities and only keep the speech activity matrix $H_s$. If we apply this activity matrix to the parallel clean speech dictionary we will obtain an enhanced version of the reference signal, as proposed by [8]. See figure 3.1.

– **Procedure 2**: The second procedure consists in using directly a clean speech dictionary, jointly with an estimated noise dictionary, for both the NMF factorization and the reconstruction. In this approach we will obtain the activations of the clean dictionary right from the NMF factorization. Once we have the total weight matrix we can discard the noise activations and use the speech activations together with the clean speech dictionary. This will reconstruct the enhanced test signal as shown in figure 3.2.

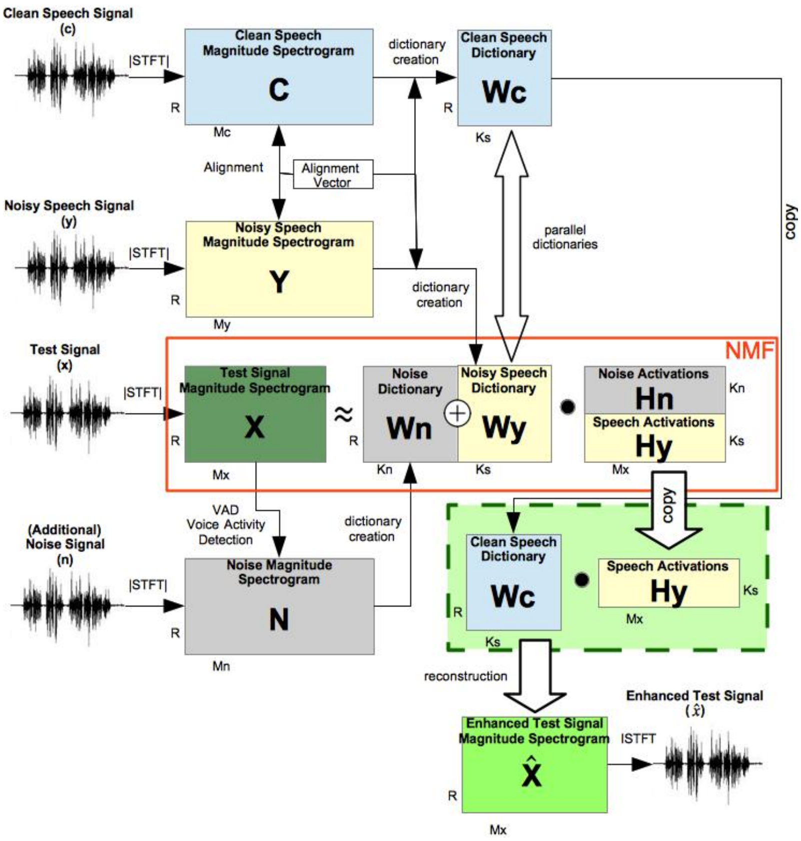


*Figure 3.1: Procedure 1.*

- **Procedure 3**: The third procedure uses the dictionaries in the same way as the second one but with an additional refinement over the clean speech dictionary. This si done in order to extract a voiced/unvoiced (V/U) identification vector for each exemplar of the dictionary. This (V/U) vector will be used in the reconstruction to avoid simultaneous activation of the voiced and unvoiced speech exemplars, that can lead to errors and noisy reconstruction, as described in section 2.7 and illustrated in figure 3.3.

Finally, we also propose a modification of the three procedures by using a white noise for the noise dictionary. The aim is to verify if the NMF can be noise independent avoiding the estimation of a noise dictionary from the signals provided, by using a white noise dictionary as parameter for the NMF factorization. By doing so, we will try to observe if the NMF is able to match the real noise exemplars of the noisy

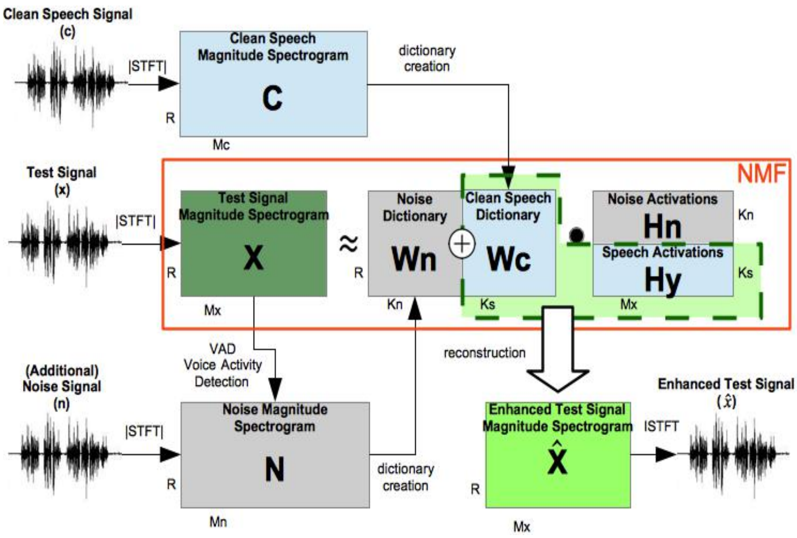


*Figure 3.2: Procedure 2.*

test signal with the white noise exemplars of the noise dictionary.

In all these cases, the training signals and the reference test signal that we want to enhance are difference utterances. However, in the first procedure, the noisy and the clean training signals have to be the same sentence since the system needs to align them.

In order to investigate the best way to solve the speech enhancement problem, we provide a Matlab script with several different combinations of all methods proposed up to this day and explained in this work. Nevertheless, the core of the system is fixed and based on the Non-Negative Matrix Deconvolution as described by Smaragdis [39] and implemented by Grindlay [59], with the possibility of inserting a sparsity constraint as proposed and implemented by O'Grady [45].

In addition, we integrate our work with the Audio Degradation Toolbox [60], in order to apply to any signal in use a controlled degradation for a better estimation of the quality of the reconstruction. This way we can add white noise or an external recorded noise at the desired SNR level. The external additional noise can be the same in the training and in the test signals; however, we will use different recorded noises in the two parts for better simulate a real background noise.

As we said for Spectral Subtraction (2.2) and Wiener Filtering (2.3), also the exemplar-based NMF represents speech as a linear combination of exemplars that are only achievable for short signal segments. By default, we use a Hamming window of length approximately 100

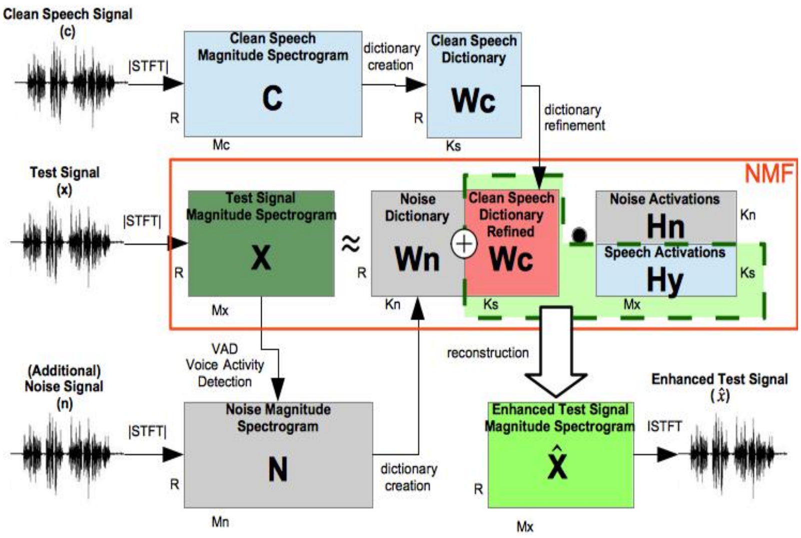


Figure 3.3: Procedure 3.

ms (4096) and an hop size of 10 ms (512) but it is possible to set these parameters accordingly with the type of signal and application that we are dealing with. At the beginning we also need to set the SNR level at which we want to work, in order to adjust the power of the various signals.

In this chapter, due to the large amount of options, whenever we mention the speech dictionary without any other specification, we are referring to the noisy or the clean speech dictionary that we are using in the particular procedure described in the relative section. For instance, when we use the reference letter $S$, we are talking about the speech dictionary taken into account in that particular procedure. Moreover we use the letter $y$, $c$ and $n$ to identify the noisy, the clean and the noise signals or with the same capital letters $Y$, $C$ and $N$ when we refer to their respective magnitude spectrograms. Other used synonyms are the words exemplar and base, as described in section 2.6, as well as the words reference and test. In this work we use these words to identify the same meanings, respectively in the first case the same object and in the second case the same input signal.

The common general process is divided in two fundamental parts: a training stage and the test execution.

## 3.1 Training

In the training part, we start by taking two equal utterance signals as input, one recorded without surrounding noise and called clean speech, and the other one recorded in a noisy environment, called noisy speech. Then, if required, we apply an additional filtering and more deterioration for simulate the desired harder condition. After performing this extra degradation we need to obtain a non-negative transformation of these matrices. This is done by computing the STFT and taking their absolute values, called magnitude spectrogram. We obtain the matrices $Y$ and $C$, respectively for the noisy and the clean speech signals, with dimension $\mathbb{R}^{RxM}$ with $R$ the frequency bins of the FFT and $M$ the time length in frames. Once we obtain this non-negative representation, and if we are following the first procedure, we need to align these two magnitude spectrograms in order to have parallel dictionaries.

### 3.1.1 Alignment

In the first procedure we need to map the noisy signal to the clean one because our purpose is to interchange the two dictionary for the reconstruction as described in [8, 9]. This step emerges because the two signals, even if they are records of the same sentence (remember that in the training part we use the same utterance, differently from the test part), they will never be perfectly aligned due to different pauses, timing in pronunciations or duration of each sub-word. For this reason we want to create two parallel dictionaries consisting of noisy and clean dictionaries with the same size and aligned. As this two dictionaries will be interchanged, we need to get the most accurate matching as possible between each one. Otherwise we will use activations of some bases that do not match with the corresponding bases in the other parallel dictionary.

This problem is solved by using a dynamic-programming technique called Dynamic Time Warping (DTW) for measuring similarity and accommodate difference in timing between the noisy and the clean magnitude spectrograms. This method calculates the optimal match between two temporal sequences which may vary in time or speed with certain restrictions [61]. In our case, the DTW used is the one implemented by Dan Ellis of Columbia University [62].

We propose that the alignment has to be done to the entire magni-

tude spectrograms of the noisy and the clean speech in the training part before the dictionary creation. Indeed, if the DTW tries to align the two dictionaries, the alignment will result forced and less precise. Sometimes the algorithm will stuck over a state for many following frames creating continuous repetition of the same sub-word. Moreover, if we are interested in using the maximal energy base selection, see chapter 2.6, we need to have the clean and noisy magnitudes spectrograms already aligned before creating the dictionary, or there will be no correctly matching cases.
Another issue is the alignment cost time. If the DTW deals with very large signals, then it is better to convert the magnitude spectrograms into mel-frequencies scale. The mel-frequency cepstrum coefficients (MFCC) are based on a non linear mel scale which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This gives a faster alignment time and the same alignment vector as the DTW applied directly to the STFT as showed in [3].
To convert the magnitude spectrogram to the mel scale we use a slightly modified version of the *melcepstr* function of the Voicebox [63]. This function provides an alignment array for the clean and the noisy magnitude spectrogram compared together applying a default 23 band mel-scale filterbank with a 13 cepstral coefficients. These values can be changed in order to best adapt the algorithm to any particular scope. Once the alignment array is obtained, the system is able to select the corresponding bases of both the noisy and the clean magnitude spectrograms that best match each others.

### 3.1.2 Speech Dictionary Creation

A crucial point for the accuracy of the NMF factorization is the speech dictionaries construction. In the first procedure, when we mention the speech dictionary, we are referring to the parallel noisy speech dictionary, while in the second and in the third approach we identify the clean speech dictionary. In this task, several problems emerge, such as the optimal dimension of the speech dictionary ($K_s$), the method to adopt for the bases selection and also the way to obtain the speech descriptors to sort the dictionary.
The $K_s$ dimension of the speech dictionary, as the $K_n$ dimension of the noise dictionary, can be chosen in two ways: as a fixed number of bases or as a percentage of the corresponding training magnitude

spectrograms. In both cases, if we are using a small $K_s$ dimension, smaller than the test signal magnitude, we will obtain a clustered representation, as described in sparsity constraint section in 2.8. This representation is more well-posed than the next one but in return, it has a lower quality reconstruction. This occurs because, with only a few bases, we expose the NMF factorization to a badly active components identification, resulting in a poor approximation of the test signal.

In the opposite case, if we are dealing with a very big dictionary we can approximate remarkably well each exemplar in the test signal. The problem here is the growing computational cost and the partition of energy over the detected active bases. The first issue is easy to understand and the second one arises due to the presence of multiple similar bases, that make the NMF split total energy on them. This occurs because, when the NMF is computing the divergence, the iteratively updating of the matrix $H$ makes active all these bases that contribute to the approximation. This makes the NMF split over them the proper quantity of the reference base energy in order to obtain the best possible estimation. It implies that a great number of bases are active, or own although a small energy, at the same time. If these bases inside the speech dictionary are part of unvoiced words, the noise can be approximated as unvoiced bases and so factorized as speech. The result is still a noisy audio where the noise is caused by the simultaneous activation of a great number of exemplars, even if with low energy, that slightly differ from each other. Moreover the bases can mix together with correct voiced and wrong noise bases recognized as unvoiced parts of words. This creates a new unwanted sort of background noise. In order to solve this issue we provide two possible solutions: one is a dictionary refinement and the second is the introduction of a sparsity constraint to the activation matrix. The first is analyzed in the third procedure explicitly while the second is applicable to every procedure. The dictionary refinement process is a way to control the wrong voiced/unvoiced bases recognition by selecting which are the fundamental bases to take into account in the reconstruction, see section 3.1.3. The sparsity constraint is an additional penalty cost applied on the KL divergence during the NMF factorization. It limits the number of active bases and enforces the subdivision of the energy just over a few bases, see section 2.8. This approach avoids the over fitting of active exemplars that alter the correct audio reconstruction.

Another complication lies in the fact that if the noise dictionary is bigger than the speech dictionary, so $W_n \geqslant W_s$, the NMF factorization will assign more weight to the noise exemplars than to the speech exemplars. This happens because it is more likely to approximate a noisy exemplar with more noise bases than with a single speech base. So the weight assigned to the noise dictionary will be much higher than the energy assigned to the speech dictionary. This fact makes the reconstruction poor in details and almost inaudible.

We concluded that the decision of the size $K_s$ of the speech dictionary is strongly dependent on the procedure and on the desired performance. However, they still have to be at least bigger or equal to the size of the noise dictionary and bigger than the test signal spectrogram. For this reason, we provide the user the possibility to choose the more adequate size for the speech dictionary expressed as an integer number of bases or a percentage of the noisy speech magnitude.

The problem of obtaining the speech descriptors does not appear here because we are dealing with a supervised technique. In fact, in this training part, for what concerns the speech, we already own the noisy and the clean speech descriptors. These descriptors are identified with the magnitude spectrograms of the two signals.

Once calculated, the two magnitude spectrograms and their alignment possess a large amount of information for each source that depend on the dimension of the provided signals.

Another issue is how to choose which bases to use and how to limit the size of the parallel dictionaries to be of the desired $K_s$ dimension. For this task we can choose between two methods. The first one is the random selection of the dictionary bases, as proposed by many authors [3, 5, 8]. This provides a sufficient approximation in order to achieve a good reconstruction for the noise reduction. But, since we are dealing with different sentences in the training and in the enhancing part, we propose a different method. This second method aims to choose the bases for the noisy and the clean speech dictionary that contain as more speech information as possible. This is done in order to better perform the activation recognition avoiding the useless excess of silences or redundancies of the same speech features. These bases are the ones with more energy in the clean speech magnitude spectrogram. Otherwise, a NMF with dictionary $W$ full of silences poorly approximates the speech parts inside the reference signal. At the same time we have to be careful about redundancies in the speech descriptors. The presence of similar bases causes the NMF to split the activation energy between the two same exemplars, which reduces the

final energy reconstruction and degrades the quality of the perception due to the introduction of echoes. This effect is produced by the simultaneous activation of the similar exemplars but with slightly different energy and timing. The goal is to describe the test signal magnitude the most accurate as possible with the less number of speech bases in the dictionaries, in order to reduce computational cost.
Once chosen the bases for the dictionaries, if we are in the first procedure, we can construct the parallel dictionaries using the alignment vector already calculated, as described in section 3.1.1. If we are in the second and in the third procedure, we will directly use the extracted clean dictionary without alignments, see figures 3.1, 3.2 and 3.3.
An improvement in the NMF identification of the compounding bases of a segment is that it allows us to use multiple-frames exemplars. As proved by $[3, 5, 41]$ using exemplars of $T \geqslant 1$ frames gives better results in terms of time continuity. In order to adopt this improvement, we apply a windowing technique, as described in section 2.6, around the selected single-frame base $l$ taking the $T$ frames around it as $W_l = [W_{l-floor(T/2)} \dots W_l \dots W_{l+round(T/2)}]$. This way we assemble multiple-frames exemplars and the resulting dictionary obtained will have a three dimensional shape $\mathbb{R}^{RxKxT}$, with $R$ the number of rows equivalent to the FFT length, $K$ the rank of the NMF factorization, and $T$ the size of each exemplar in frames. See section 2.6 and figure 2.1.

### 3.1.3 Dictionary Refinement

Only in the third approach we use a particular dictionary refinement in order to prevent wrong reconstruction. This refinement aims to avoid the mistaken noise detection as unvoiced speech frames. This error is caused by the unvoiced frames that, in a noisy environment, can be recognized as noise parts. If this occurs, an unwanted noise will be factorized as part of the speech bases in the final result, and the activation energy of these speech frames will also reconstruct the undesired noise part. To prevent this situation we introduce a method that takes the speech activations matrix and, in the case there are multiple active exemplars at the same time, deletes the unvoiced activation and only keeps the voiced frames for the dictionary construction. In order to do so we need to extract some speech parameters that allow us able to recognize the voiced/unvoiced nature of each frames. The features we focus on are the zero crossing rate (ZCR) and the energy of each frame of the speech signal. The energy of a signal is calculated as described before, see 2.24, while the zero crossing rate is the rate of sign changes from positive to negative, or the opposite, along a signal. This feature is very used in speech recognition and music information retrieval because it is fundamental for Voice Activity Detection (VAD) algorithms [64,65]. The zero crossing rate formal definition is given by

$$zcr = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]|w(n-m), \tag{3.1}$$

where:

$$sgn[x(n)] = \begin{cases} 1 & for : x(n) \geqslant 0 \\ 0 & for : x(n) < 0 \end{cases} , w(n) = \begin{cases} \frac{1}{2N} & for\,:\,0 \leqslant n \leqslant N-1 \\ 0 & otherwise. \end{cases} \tag{3.2}$$

More accurate parameters extracted from the speech signal can provide finer results [66]. Anyway, as proved by [67, 68] these two parameters are sufficient for the voiced/unvoiced decision. Obviously if we use this technique in a noisy signal it can fail. But, as we are in the third procedure, we are working with a clean speech dictionary. Therefore, there are no errors in the decision introduced by a noisy environment. So, using these two features, we can identify the silence and unvoiced frames and delete them from the activations matrix. This is done only when the unvoiced exemplars occur at the same instant of a voiced frame allowing a better and clearer reconstruction. In particular, firstly the silence frames are recognized by maximum energy detection and considered in the dictionary construction only if there is an excess of bases. This is done at first because they have a small number of zero crossing, just like the voiced parts. If the frame analyzed has small energy it is a silence frame, while the voiced/unvoiced characteristics are detected by the zero crossing rate. Once pushed aside the silence parts, if the frame considered has low ZCR, it means that it is a voiced frame and we have to keep it. On the other hand, if the frame has high ZCR we still keep it for the factorization but, when performing the reconstruction, we have to check if this base is active alone or in parallel with other voiced exemplars. In the first case we keep it for the reconstruction, while in the second case we will push down to zero its energy. This is done in order to eliminate wrong activations of the noise bases, detected as unvoiced bases, jointly with the speech during the reconstruction.

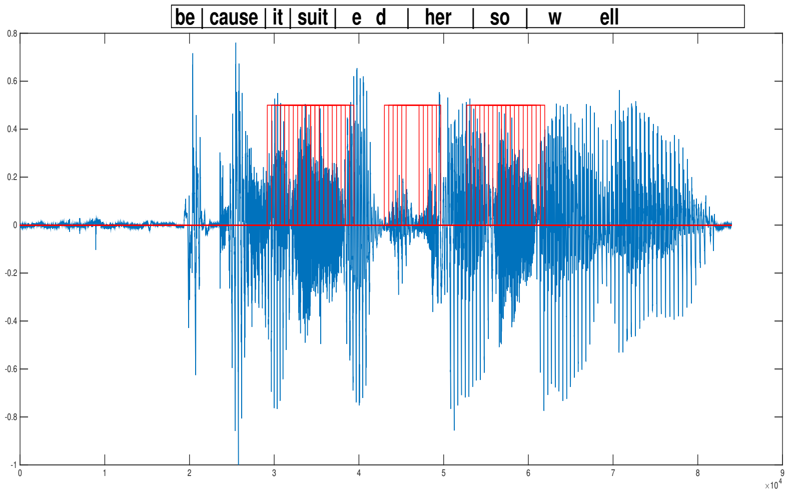


*Figure 3.4: Voiced/Unvoiced recognition.*

To implement this refinement we adopt the maximum energy base decision as described before (2.24) jointly with the MIRToolbox [69] for the zero crossing rate feature extraction. With this toolbox we can split the input time-domain signal into segments of the same duration as in the STFT. For each segment we compute the number of zero crossing. Once we obtain these two features we are able to discriminate if the frame we are dealing with is silence,unvoiced or a voiced part.

## 3.2 Enhancement Process

As in the training part, we start with the construction of the analysis data by taking the reference signal, also called the test signal, to enhance and by computing its magnitude spectrogram. In the enhancement part we do not need to align the test magnitude or to create any dictionary because we are directly using its magnitude spectrogram for the NMF factorization. The NMF is able to recognize the arrangement of exemplars that compose one reference signal segment also if the exemplars are just a few and disordered. In any case, for the reconstruction, we need the exemplars activations at each time step of the test magnitude spectrogram. In fact, if we do not perform the NMF on the entire spectrogram of the analyzed signal, we will loose the temporal information and the correct succession of the frames for the reconstruction. This activation weight along the time is given by the energy matrix $H$ of dimension $\mathbb{R}^{KxM}$, with $M$ time length in frames and $K$ the size of the NMF factorization. The dimension $K$ is also considered as the number of exemplars in which the reference signal have to be decomposed. Each of these $K$ exemplars can be considered as a different source in which we are interested to factorize the test signal. But at the same time we can also take into account sources composed by several exemplars. So, grouping multiple-exemplars per each source desired, we are able to recognize and separate multiple-exemplars sources inside the reference signal.

In our case we are using two sources, identified by each dictionary, for the NMF factorization: one is the speech dictionary proceeded in the training part and the other is the noise dictionary. The speech dictionary is obtained in different ways depending on the used approach, as described in section 3.1.2.

### 3.2.1 Noise Dictionary Creation

The noise dictionary has a very important impact on the final results and it can strongly influence the NMF factorization. This happens because if we are trying to decompose a noise segment not present in the noise dictionary, it will be approximated by the NMF with the simultaneous activation of others exemplars. These exemplars can also be speech exemplars, and so the NMF, without a good matching case, will split the energy into inappropriate speech exemplars. This leads us to obtain a residual noise energy in the reconstruction, the same problem found in the Wiener filtering, with a bad noise power density estimation.

The problem arises because in real-world applications we cannot know in advance the descriptors of the noise sources in order to create an accurate noise dictionary. This occurs due to the huge non-stationarity and unpredictability of real noise and also to the difficulty of describing it properly. For this reason we will give to our algorithm two possible approaches. The first one consists on providing an additional short record of the surrounding noise or to use a few seconds segment at the beginning of the test signal as noise reference. The idea is to supply the system with a sort of calibration for the surrounding noise. This pre-recorded noise signal will be the reference for the noise dictionary creation. This is a more reliable technique to obtain the noise descriptors for the dictionary creation and it is more robust against the problems of wrong descriptors identification in comparison with the next one. The disadvantage is that it is unable to follow the unpredictable mutation of noise along the time.

The second option is the one proposed by Schmidt [6] where it is assumed that the entire noisy signal is observed and the noise bases vectors are learned during the speech pauses. Due to the fact that we do not have any a priori knowledge of the noise dictionary, this method is called semi-unsupervised. The noise components are identified in the noisy speech from the Voice Activity Detector algorithm (VAD) which is able to distinguish between speech and noise parts inside a reference signal. This is optimum for high non-stationary noise since it is able to adapt the noise dictionary through time following the noise changes. At the same time, the biggest disadvantage is that the entire noise reduction system depends on this voice recognition detection and, as we can easily understand, with strong noisy signal it can fail. Fail in recognition means providing incorrect speech bases to the noise dictionary, setting speech segments as noise exemplars. In that case, NMF will wrongly identify the activations of this speech segment as noise part and in the reconstruction will miss this portion of vowel.

In this work we use the Voicebox implementation of VAD [63], and we try to strongly control the identification of noise by imposing some strict parameters. For our scope and for the large variance of real noise it is better to lose some noise information than to identify some speech portion as a noise base. In any case, for a correct identification, the VAD parameters have to be set depending on the situation and this is an annoying drawback.

Another possible approach consists on mixing these two techniques obtaining the noise descriptors together with a small pre-recorded surrounding noise signal combined with a noise detection with the VAD along the test signal.

This technique provides a more accurate noise dictionary following the dynamic changes of noise along time. Nevertheless, it still has the problem of wrong identification between speech and noise of the VAD algorithm, which can severely influence the final reconstruction result. Differently with the speech dictionary, here there is no need to use particular methods to select the bases or to have a noise dictionary aligned. When we have to create the dictionary, we can randomly choose the exemplars from all the noise descriptors. This is possi-

ble due to the unpredictable nature of real noises that make all the exemplars equally probable. At the same time, a real noise is not instantaneous and it can continue for several time intervals. In order to consider time continuity in the noise dictionary, we use $T$ multiple-frames exemplars, as described for the speech dictionary before 3.1.2. This produces a noise dictionary of three dimensions $\mathbb{R}^{RxKxT}$ where $T$ is the number of frames per each bases, or the windowing length in frames.
Another problem lies on the size of the noise dictionary $K_n$. The size of the speech dictionary is a critical trade off between accuracy in the approximation and speed of execution. Bigger noise and speech dictionaries are better at identifying exemplars activations and better at approximating the test signal, which provide a superior enhancement result. But at the same time, bigger dictionaries increase the execution time. For this reason we provide a settable $K_n$ dimension for the noise dictionary expressed, as in the speech case, as an integer number of bases or percentage of speech dictionary dimension. These dimensions are forced to be smaller or equal to the dimension of the speech dictionary, $K_n \leqslant K_s$, in order to avoid incorrect strong energy assignation to the noise bases, as delineated before.
Differently, if we are evaluating the system for the white noise recognition, the noise dictionary is created in the same way as described for the multiple-frames exemplar-based, but using as magnitude spectrogram the absolute value of a white noise signal's STFT. In this case we obtain a white noise dictionary.

### 3.2.2 Normalization

In order to take into account only the shape of the magnitude spectrum $X$ and to match arbitrary speech levels and SNR, the dictionaries $W_s$ and $W_n$ have to be normalized. This is another significant point in the process of speech enhancement and it can lead to different results. Some authors as Gemmeke [3] and Virtanen [5] proposed that the norm of each frame, or exemplar, equal unity and the norm of the rows are approximately equals. This is done by iteratively scaling each row and column so that its Euclidean norm equals unity. Other authors as Takashima and Aihara proposed that the sum of the magnitudes over the frequency bins equal unity [8,9]. We decide to provide both these normalization options in our script.
Differently, the reference signal magnitude does not have to be nor-

malized since this would mean that every frame in the test magnitude would have the same energy. But silence frames have less energy than the frames with speech information. Otherwise, if we normalize the test signal magnitude, the error calculation of the NMF will consider all the frames as equals and the result will say that the NMF can sufficiently approximate the test signal magnitude with the noise exemplars. This still lead to a very noisy reconstruction, and in some cases even worst than the test signal. So, we also provide the possibility to normalize (or not) the dictionary bases and to perform the NMF with unnormalized speech and noise dictionaries.

### 3.2.3 Exemplar-Based Sparse Representation

Although it may not be obvious at first, an arbitrary magnitude spectrogram can be represented as a sparse linear combination of smaller spectrograms. The experimental data obtained by Gemmeke [3] and Takashima [8] indicate that this is a reasonable assumption. The explanation lies on the fact that spectrograms of different pronunciations of the same word have approximately the same patterns of energy concentration, also called activity. This also applies for multiple exemplar spectrogram as proved by Takashima [8]. So, in order to obtain the activations $h$, we search for the linear combinations of the fixed exemplars $W$, which are able to represent the source $v$ with the signal model $v = Wh$ while using only a small number of non-zero entries in the activity matrix $H$.
In our case, the sources are two (speech and noise) and they are made of exemplars composed by multiple frames. In a matrix form, the model will be expressed by the non-negative linear combination

$$\begin{aligned} X &\approx W_s H_s + W_n H_n \\ &= [W_s W_n] \begin{bmatrix} H_s \\ H_n \end{bmatrix} \\ &= WH. \end{aligned} \tag{3.3}$$

Where $X$ is the test signal magnitude, $W$ is the concatenating representation of $W_s$ and $W_n$, respectively the speech and the noise dictionary, and $H$ is the jointly representation of $H_s$ and $H_n$, the speech and noise activation energy.
Using the test signal magnitude $X$ as reference and the speech and noise dictionaries together $(W)$ as NMF parameters, the NMF finds

the weight of each base along the time ($H$) to provide the best approximation of the test signal magnitude. This is done by iteratively updating the matrix $H$ in order to minimize the KL distance between the original reference magnitude $X$ and the estimated one $\hat{X}$. For this scope we adopt the update rule described in section 2.5.2 with 100 updating repetitions since, as proved by Lin in [35], it is a sufficient good trade-off between the quality of the approximation and speed of convergence. To perform this task we use the convolutive NMF algorithm as implemented by Grindlay in the *nmflib* library [59].

In addition, the user can choose to add a sparsity constraint to the NMF cost function in order to solve the problem of non-uniqueness of the solution and to obtain a sparser solution as described in chapter 2.8. This can be done for the matrix $W$ and for the matrix $H$ as mentioned in the sparsity constraint paragraph.

In our work we analyze both possibilities to obtain an unique and well-posed solution: either by reducing the dimension of the matrix $W$ or enforcing sparsity on the matrix *H*. The first case is obtained by creating a speech and noise dictionary $W$ smaller than the test magnitude, $K \leqslant M$. This way we perform data clustering forcing the decomposition to fit inside a small quantity of bases. The other solution is to use a larger speech and noise dictionary and to make the matrix $H$ sparse. This is done by applying an additional penalty term to the cost function limiting the decomposition of test energy into a small number of exemplars chosen from an over-complete dictionary. Formally, an over-complete dictionary is a collection of exemplars which number exceeds the dimension of the signal magnitude, so that any signal can be represented by more than one combination of different atoms. The over-complete dictionary is the cause of the ill-posed nature of the problem that we want to solve with a sparsity constraint. So, at each time step, only few bases of the whole dictionary are forced to be active, or non-zero, and contribute in the reconstruction task.

This method, creates an aggregate function by multiplying each constituent cost function by a weighting factor and summing these weighted costs at each iteration [3]. The sparsity constraint is a *l1*-norm regularization term and the weights of the sparsity constraints can be defined for each exemplar by defining $\lambda_K = [\lambda_1 \ldots \lambda_{K_s} \ldots \lambda_{K_s+K_n}]$ . In this work, the penalty for speech exemplars $\lambda_s = [\lambda_1 \ldots \lambda_{K_s}]$ was set by default all equally to 0.65, and those for noise exemplars $\lambda_n = [\lambda_{K_s} \ldots \lambda_{K_s+K_n}]$ were set to 0, as used in [3]. Both values can be adjusted by the user depending on the specific situation.

In this case, the system uses the NMFD with sparseness constraint realized by O'Grady [45], that manages the column shift at each repetition, see formula (2.30). The sparsity constraint solution over the matrix $H$ should be preferred. In fact, limiting the dictionary size will result in a poor reconstruction in terms of accuracy. Instead, a sufficiently large dictionary can better approximate a reference segment. Moreover, by enforcing sparse energy activation we obtain a sparse and unique solution without loosing accuracy.

The three procedures implemented in this work share the same core NMFD as described till now and the reconstruction method described in the next paragraph 3.2.4, but they differ in the dictionary that they use for the activity matrix estimation.

The first procedure adopts as NMF parameter the noisy speech dictionary extracted in the training part. This way, the activity matrix is obtained weighting the contribution of each noisy base in the dictionary minimizing the difference with the test signal magnitude. This is theoretically the most appropriate approach since it may be easier to approximate a reference corrupted exemplar with another exemplar that has a similar degradation. Anyway the results obtained will prove that it is not the best approach.

The second and the third procedure try to do the same, but they use directly the clean speech dictionary as NMF parameter. We want to obtain the activity matrix straight by minimizing the difference between the weighted clean speech dictionary and the reference signal magnitude.

The third procedure also incorporates a refinement into the speech dictionary creation in order to even better approximate the reference signal, as showed in figure 3.3.

In the white noise dictionary testing approach, the user has the possibility to choose if he wants to directly use the clean or the noisy dictionary as NMF speech parameter. Yet, for the estimation of the noise exemplars activity, it is used a white noise dictionary obtained in the same way as described in the noise dictionary extraction. This particular approach is used to analyze if the NMF is also able to match a real noise exemplar with a white noise exemplar and to assign the estimated weight of the first to the second.

### 3.2.4 Reconstruction

In our case, for the reconstruction, we are not interested in a time-varying filter for denoise the test signal as proposed by Gemmeke [3] and Raj [7]. This because we are dealing with fixed dictionaries obtained from different utterances compared to the one we want to enhance. Therefore, we are unable to obtain a filter that can perfectly match the reference speech and eliminate the background degradation without keeping noise residuals into account. This noise residual is the noisy speech energy not modeled by the linear combination of speech and noise exemplars.

Moreover we are not looking for a noise reduction task in a speech recognition system as in [3]. In fact, we are working for a speech enhancing method that can reconstruct an enhanced version of the test signal. This version must recreate energy content at frequencies that were not present in the test signal, and not just to reduce the background noise. The resulting enhanced signals will be a full-band signal also if the input is band-passed, as in telephone communications.

For our purpose we need a different approach, that is the one proposed in [8]. As shown by Takashima, the estimated activity matrix $H$ of two equal noisy and clean utterance magnitudes have the same high energy at similar positions in time. For this reason, when we are using parallel dictionaries of the same utterance, the estimated activations of the noisy speech dictionary can be used for weight the clean speech dictionary, also called target dictionary. In other words, in the first procedure, we can interchange the clean and noisy parallel speech dictionaries in order to use the activations of the noisy speech dictionary with the clean speech bases.

The idea is that applying the noisy speech activations to the parallel clean speech dictionary we are able to reconstruct an enhanced version of the test signal completely noiseless, as showed in figure 3.1.

This feature is not necessary for the second and the third procedure. In fact, we obtain the estimation of the activity applying directly the clean speech dictionary to the NMF decomposition. So, for the enhanced reconstruction, it is possible to use the clean speech dictionary jointly with the estimated weight matrix without any other assumptions. This because we multiply the estimated activations by the same dictionary that has produced them.

However, in the third procedure, we need a last refinement stage, as explained in section 2.7. In fact, we want to limit the error introduced by superposition of voiced and unvoiced exemplars by limiting their activation in the reconstruction. To do so, we compare the total energy of the voiced and unvoiced exemplars extracted in the refinement stage. For each frame we keep only the bases pertaining to the group with the highest energy, as showed in figure 3.5. This way we limit the amount of noise introduced by the unvoiced exemplars during voice segments but, at the same time, we maintain the correct and characteristic pronunciation of voiceless segments.

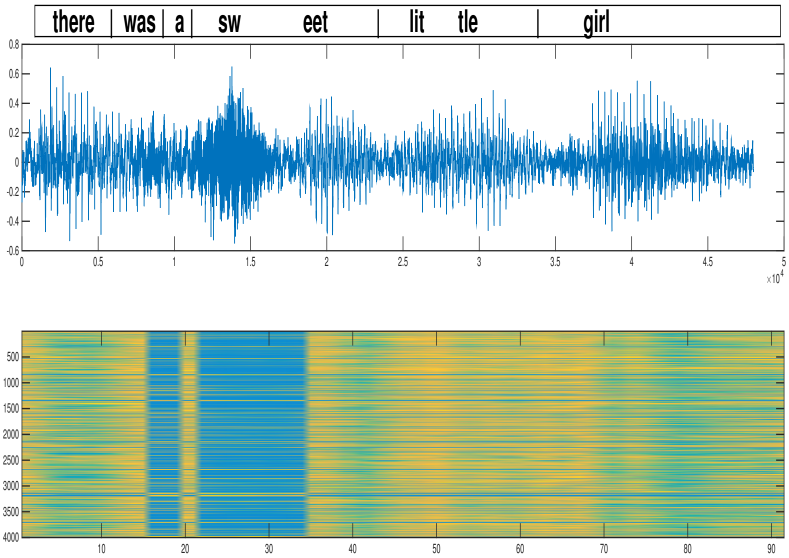


*Figure 3.5: Activation of voiced and unvoiced exemplars.*

In all our procedures we take a fixed dictionary, composed by speech and noise exemplars (each one obtained following the methods described in the previous sections 3.1.2 and 3.2.1), and we find its activation matrix $H$. Once we get the matrix $H$ we are able to isolate speech from noise by just selecting the $K_s$-dimension speech activation matrix and discarding the $K_n$-dimension noise matrix contribution. Using the activation of the wanted source on the corresponding desired dictionary we are able to reconstruct only the desired part of the reference signal.

In our case, the wanted source is the speech dictionary (noisy or clean, depending on the procedure) and the desired dictionary is the clean one. This way, we can reconstruct only the speech part by eliminating the background noise contribution. Therefore, since the desired dictio-

nary is the clean one, we are able to reconstruct an enhanced version of the reference speech part.
In particular, for reassemble one specific source mapped into the $l$-th base, we can multiply the clean magnitude dictionary ($Wc$) by the corresponding parallel estimated noisy weight ($Hs$), plus a shifting operator for the deconvolution, and we will obtain the enhanced base $l$ as: $\hat{X(l)} \approx W_c(l)h_{s_l}$. This gives us the enhanced magnitude of the single exemplar $l$.
If the source we want to extract is mapped by more exemplars, as our dictionaries are, we can apply the same formula to all the exemplars that build up the desired source. Therefore, the enhanced speech can be constructed by only using the $K_s$-dimension of the clean speech dictionary, as bases, weighted by the corresponding $K_s$-dimension of the activity matrix of the noisy speech exemplars. By doing so, we will obtain a noise reduced multiple-frames source magnitude expressed as:

$$\begin{aligned} \hat{X} &\approx \sum_{l=1}^{Ks} W_c^{(l)} h_{s_l} \\ &= W_c H_s. \end{aligned} \tag{3.4}$$

If we are using a dictionary with exemplars' size $T > 1$, the system uses a deconvolution method, like explained before, to reconstruct the enhanced magnitude from the three dimension clean speech dictionary (Wc) and the two dimension activation matrix (Hs), where $W_c\mathbb{R}^{RxKxT}$ and $H_s\mathbb{R}^{KxM}$. Employing the same column shift operator, with zero filling on the left, to the bases of the clean dictionary weighted by the activation matrix, we are able to reconstruct the enhanced magnitude spectrogram. This results in the convolutive reconstruction formula is obtained by

$$\hat{X} = \sum_{t=1}^{T} W_c \overset{t\rightarrow}{H_s}. \tag{3.5}$$

### 3.2.5 Phase Estimation

After obtaining the denoised magnitude representation we end up in a clean reconstruction and we have also recreate the energy content in the frequencies that were not present in the test signal. So, we are able to take as input a band-passed filtered signal, as in telephones lines or VoIP communications, and obtain a reconstruction also with the high frequency content. This involves reestimating the phase for those frequencies before computing the ISTFT.

The new issue is the phase estimation for the time domain reconstruction of a modified signal. To solve this last problem, the easiest way is to use directly the phase of the test signal, as used by [3, 5, 8, 9, 41]. In this case we can simply use the ISTFT to convert back to time domain the enhanced magnitude, by applying the same phase as the reference test signal. But since we want to reassemble an enhanced signal made with exemplars of a difference utterance, we may be interested in a better reconstruction. What we can do is estimate the phase from the magnitude spectrogram. Nowadays there are lots of investigations on faster and optimal techniques for recovering the phase from the magnitude [70], and it is still an open problem.

However, we decided to use a simple method that is the one proposed by Griffin and Lim (GL) in [71], which is the least square error reconstruction method (LSEE). Our decision is motivated by the fact that we do not need real-time phase reconstruction, since we are working in batch processing.

Anyway, in a future case in which some implementation will work in real time, we will be forced to change this method for a faster one, like the real time iterative spectrum inversion with look ahead (RTISI-LA) proposed by Zhu [72].

# Chapter 4

# Evaluation

## 4.1 Validation Techniques

In order to investigate the effectiveness and the accuracy of the enhancement process, we analyze the reconstructed output quality of the approaches described above. Since we need to use the same user speech in a clean and in a real noisy environment for the first procedure, we recorded the signals by our own. The real noisy signal was recorded with an iPhone on the street while the clean speech signal was recorded in a recording room with a Shure SM58. For the additional degradation noise we adopted real disturbs that can daily affect normal conversations, like real street noises. We used online material from the *Freesound* [73] database for the five different street noises for a stronger degradation. For each of the three procedures, plus the white noise test, we investigated the factorization and reconstruction improvement obtained as a function of Signal to Noise Ratio (SNR) and Signal to Error Ration (SER). The SNR is the measure used to compare the level of a desired signal to the level of the background noise, defined as the ratio of signal power to the noise power and often expressed in decibels. In this work we use the SNR to compare the obtained enhanced signal with the original noisy test signal. This way we obtain the improvement in dB that we reached with the enhancement process. For this particular purpose we use the *snrseg* function, already implemented in the VOICEBOX matlab toolbox [63].
The SER is a measure of the signal phase reconstruction used to evaluate the STFT magnitude spectral inversion quality. First introduced by Gillis [74], it was adopted by Zhu [75], Beauregard [76]

and Chami [77]. Chami defined it as:

$$SER = 10\log_{10}\frac{\sum\limits_{m=-\infty}^{\infty}\int\limits_{\bar{w}=-\pi}^{\pi}|X(mL,\bar{w})|^2\,d\bar{w}}{\sum\limits_{m=-\infty}^{\infty}\int\limits_{\bar{w}=-\pi}^{\pi}[|X(mL,\bar{w})|-|\hat{X}(mL,\bar{w})|]^2\,d\bar{w}}. \tag{4.1}$$

Where $X$ is the original test signal magnitude spectrogram and $\hat{X}$ is the enhanced signal obtained. We use this measure for evaluate the quality of the phase reconstruction obtained using the G&L method. Moreover, since we cannot measure the perceived quality of the reconstruction with these values, we also use a Mean Opinion Score (MOS) evaluation to obtain the user's perception of the quality achieved. The MOS is expressed as a single number in the range of 1 to 5, where 5 is the highest and 1 is the lowest perceived audio quality measurement. The MOS scale is graded using the following rating scheme:

| **MOS** | **Quality** | **Impairment** |
|---|---|---|
| **5** | Excellent | Imperceptible |
| **4** | Good | Perceptible but not annoying |
| **3** | Fair | Slightly annoying |
| **2** | Poor | Annoying |
| **1** | Bad | Very annoying |

This evaluation was performed over a sample of 10 people with different levels of english (from very low to middle-high) and with no particular knowledge in music. These conditions may have affected the outcomes of these tests and how people evaluated the quality of the sounds. This may explain the variety of the results obtained. Further research should include a bigger sample of people. The final MOS value is obtained by averaging the results of a set of subjective tests where a number of listeners rate the heard audio quality of the enhancement system output. The SNR and the SER provide us informations more related to the power and noise reduction, while, this MOS evaluation, allow us to verify if the enhancement process effectively improves the perceived audio quality.

## 4.2 Results

The results are obtained by comparing the three different procedures with all the proposed settings. In order to do this we performed the enhancement test with the five street signals and we organized together the outputs grouped by each setting used. For our tests we used a fixed length FFT (NFFT = 4096) with a hamming window of 4096 length, fixed hop size (HOPSIZE = 512), fixed number of filter for filter bank alignment (L = 23) and cepstral coefficients (CC = 13). In addition, we always used the same normalization technique: the 1-norm normalization of the exemplars proposed by Gemmeke [3]. We applied this normalization to the dictionaries bases because we found in others results, not showed here, that it produces slightly better outcomes.

For each procedure and each background degradation noise we performed the enhancement test and the results are assembled together in a statistical box plot like the one in figure 4.1.

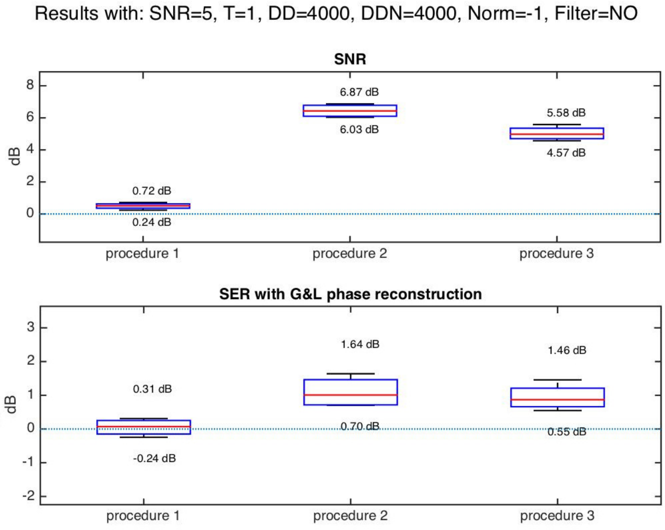


*Figure 4.1: General box plot obtained for each setting at different SNR and for each procedure using five different background degradation noise.*

The box represents the central 50% of the data and is lower and upper bounded by lines that represent the best and the worst enhancement obtained. The central red line indicates the median of the outcomes. Due to the huge variety of settings and quantity of results, we aimed for a clearer view by only considered the median result for each procedure of each setting and we grouped them in a comparative histogram. We divided the review of the results for each parameter taken into account and we analyzed its influence over the enhancement system. For all the tests described, we used a 64-bit machine Intel Core i7 with 2,2 GHz processor.

### 4.2.1 Exemplar Size

The exemplars size is the dimension $T$, expressed in number of frames, of each base used to create the dictionaries. As proved by Gemmeke [3], multiple frames exemplars help to keep time continuity in the reconstruction. This feature affects the resulting SNR and SER but mostly influence the perceived quality of the reconstruction. As shown in figure (4.2), the results obtained with a bigger $T$ prove that this feature give better results in terms of SNR and SER.
We also performed several tests with other background noise degradation to verify if these results are noise dependent and to verify the contribution of size $T$ in different degradation scenarios. The additional degradations used are a restaurant noise and a white noise. We tested the best approach obtained in the previous test, the second procedure, with these two noises. We compare the new results obtained with the previous outputs at different SNR levels and with different exemplar size $T$. The outcomes are showed in figure 4.3. All the MOS results with the three different degradation noises are resumed in the figure 4.4.
The results obtained in figures 4.2, 4.3 highlight how the resulting SNR and SER are influenced by the use of different size $T$ at various SNR levels. Moreover, in figure 4.4, it is possible to notice that the size of the exemplars also influence the user's perception. In fact, at low SNR, we obtain better perceived results for larger exemplar size $T$, while, at high SNR, a smaller size $T$ outperform the larger size $T$ approach. This happens also with different degradation noises as showed in figure 4.2 and 4.3. During the experiments, we focused over this particular fact, depicted in figure 4.2, where larger exemplars ( $T$=10 ) provide better results with lower SNR's while smaller exemplars ($T$ = 1) provide better results with higher SNR's. The reason is probably that, at low SNR, longer exemplars including more time content prevent confusion with noise exemplars, by imposing more constraints on the search of linear combination of bases. This does not happen with smaller exemplars that can be easily approximated with noise bases at low SNR. At the same time, using larger exemplars at higher SNRs decrease their performance because the factorization becomes less pre-

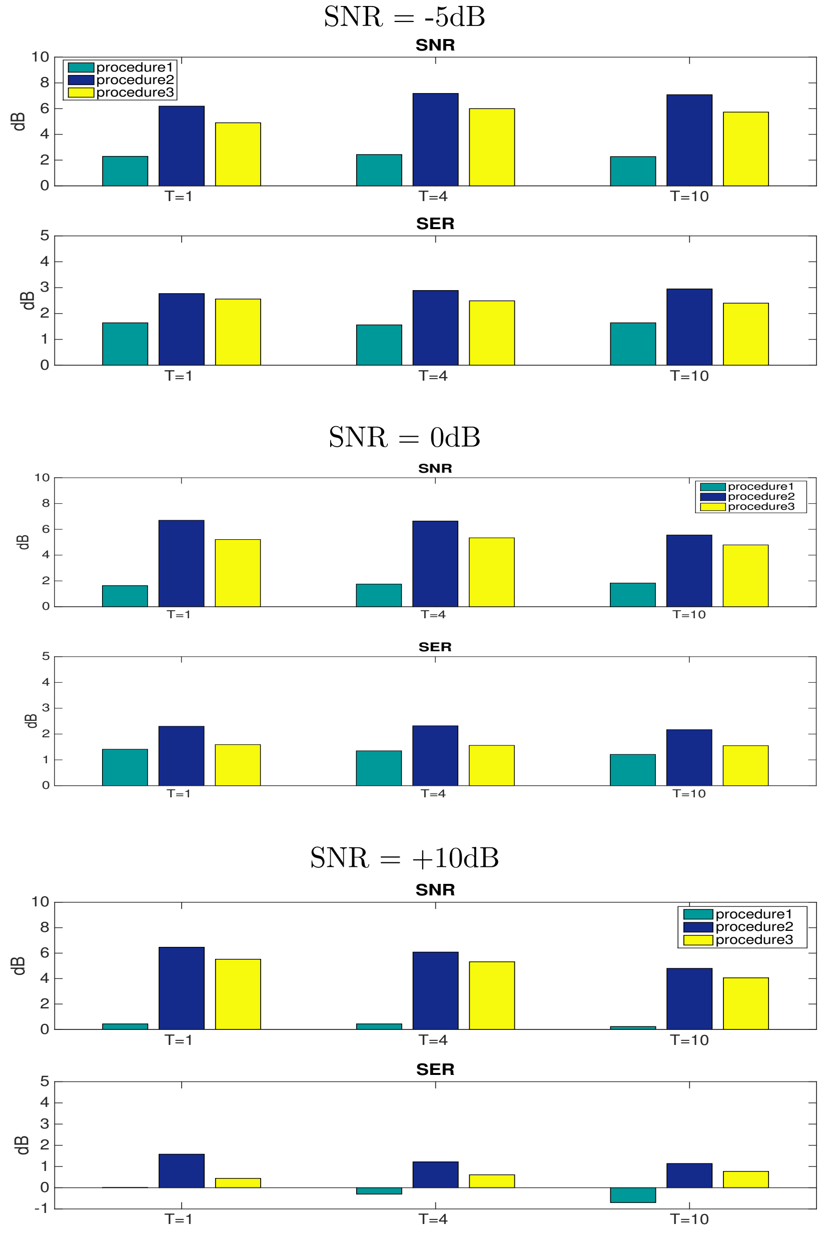
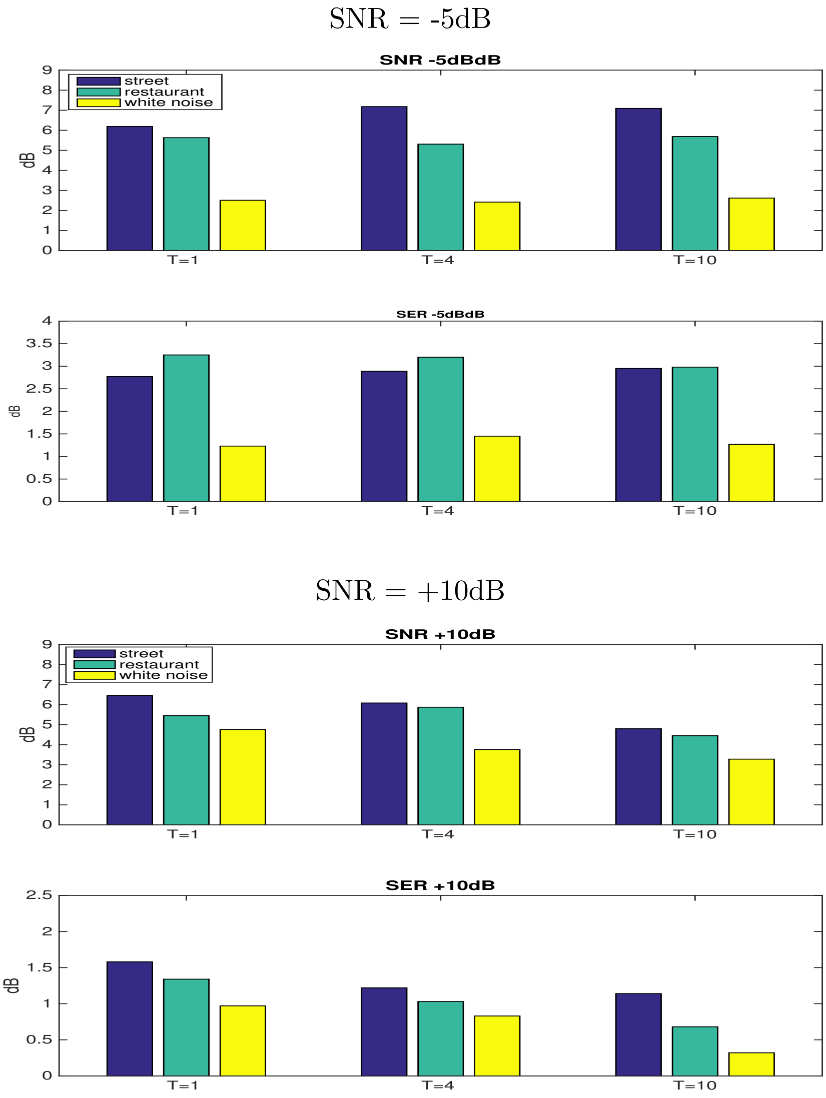


*Figure 4.2: SNR and SER outputs obtained with 4000 bases dictionaries for several size of T.*

cise in describing clean speech as a linear combination of such large exemplars and it does not provide any improvement. Similar results where founded in [78].

The main problem when using smaller size exemplars is that they


SNR = -5dB

SNR = +10dB

*Figure 4.3: SNR and SER results obtained for the procedure 2 with 4000 bases dictionaries with several T and different background noise degradation.*

create discontinuities at the concatenating point which results in a unpleasant jerky audio reconstruction. At the same time, there are two mainly drawbacks when using big exemplars dimension $T$: the first one is related with the first procedure, is the amount of noise that the NMF can wrongly consider as speech part during the factorization. The NMF recognize the weight of each bases during the factorization

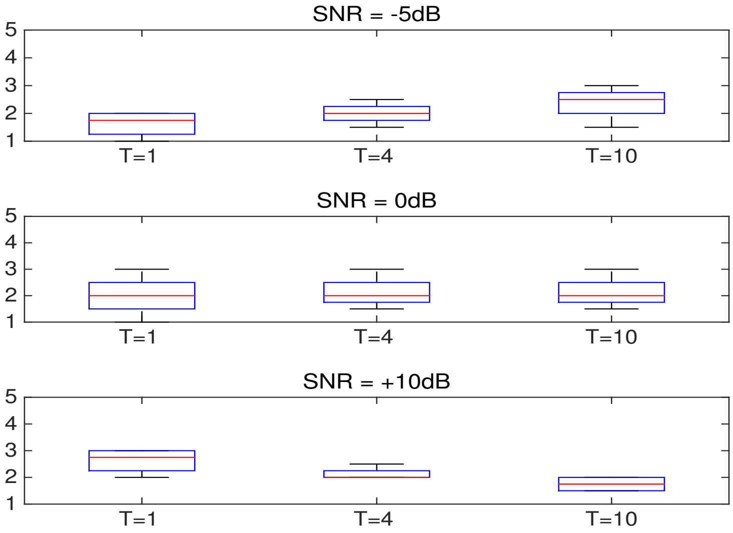


*Figure 4.4: MOS results for different T and different degradation noise.*

and, as it is very likely, in big speech exemplars are present some noise or silence part. This causes that some noise bases are recognize as speech and wrongly kept during the decomposition process. As we can see in the MOS results above (figure 4.4), with big $T$ the resulting quality perception decreases linearly with the SNR rather than continue increasing as expected. This particular problem occurs mainly in the first procedure, while the other two approaches do not seems to suffer for this issue, as they use clean speech dictionary,
The second one is the time complexity. In fact, the bigger the exemplar size $T$ is, the larger the time we spend into computing the NMF factorization. The execution time grows linearly with the dictionary dimension and, if we are dealing with very big dictionaries, this can be a problem. Resulting execution times are showed in the figure below (4.5).
Due to the results obtained in this work, and following the results provided in [3], we conclude that the best solution for a general case is to use a medium size for the exemplars (T = 4). In an unpredictable real noisy situation, this dimension provides the most robust results (in terms of noise suppression) in an acceptable amount of time. Anyway, another implementation, not developed in this work, may be use different exemplar sizes $T$: a larger one to model in a more accurately

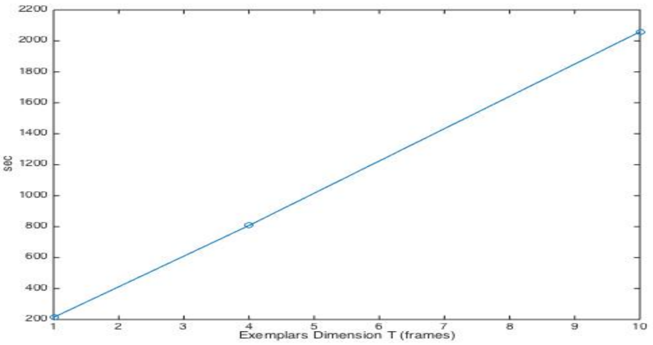


*Figure 4.5: Time expended for the NMF factorization of a 10 sec audio signal with speech and noise dictionaries dimensions of 4000 frames and variable T.*

way the bases of the already known sources and a smaller one unknown sources as the noise. This may lead to a better enhancement and suppress the problems described above at low and high SNR.

### 4.2.2 Dictionary Dimension

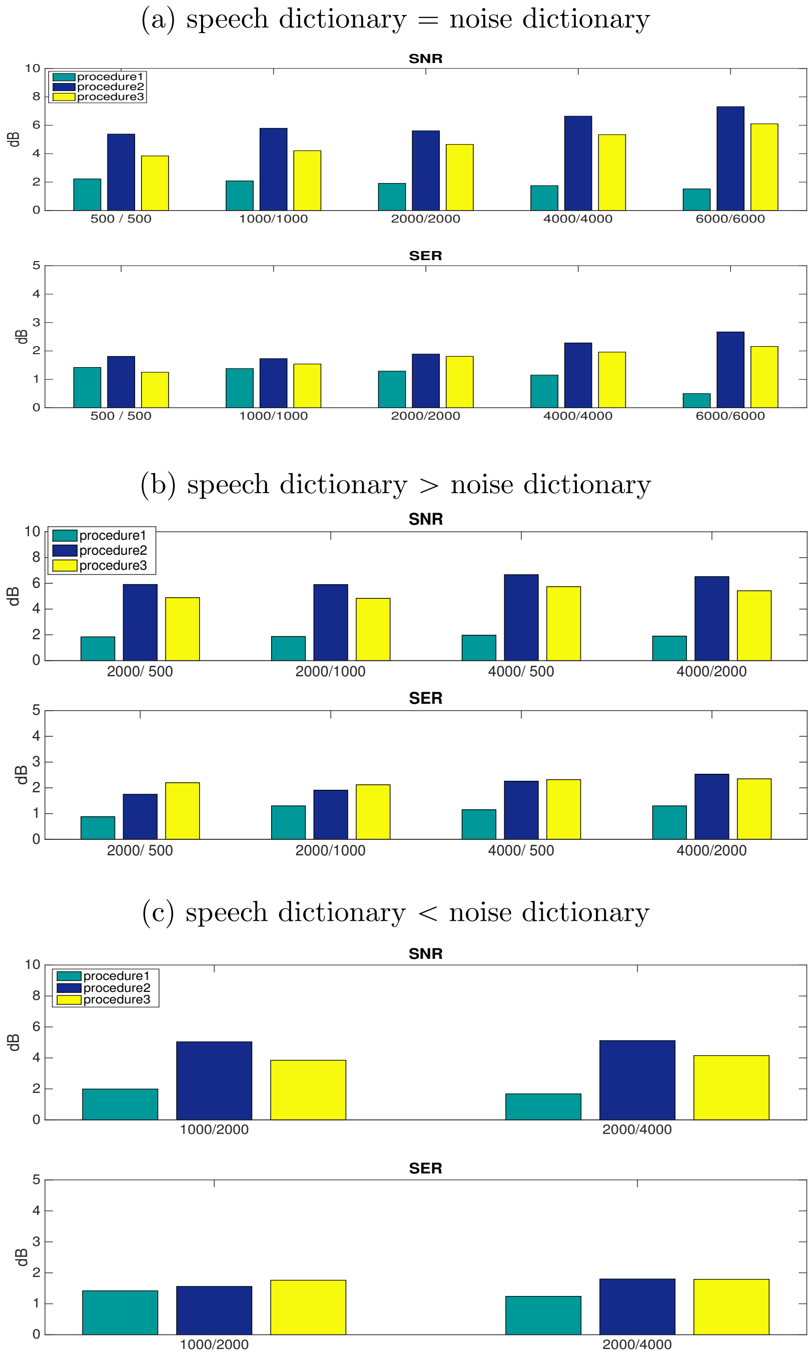
The speech and noise dictionaries dimension are the major problem in speech enhancement methods. There are many authors who are currently researching about this still open question [79–81]. We provide two way to set them: by using a percentage of the provided signal or by using a fixed number of frames.

The percentage method strongly depends on the provided signal dimension and therefore is not useful for comparative test. Given that, we focus our attention over the fixed frame number method.

As we can easily understand, the bigger the dictionaries are, the bigger number of exemplars that are available for a more precise estimation of noise and speech part in the test signal.

The increasing SNR y SER results for the second and the third procedure (figure 4.6) and the reconstruction error obtained (results not showed here) inversely proportional to the size of the dictionaries, encourage us to follow this direction. Anyway, this approach suffers of two important drawbacks.

First of all, big dimensions mean higher probability to having lots of very similar exemplars active at the same time, causing echoes and annoying noise in the reconstruction. To avoid this problem we can



(a) speech dictionary = noise dictionary

(b) speech dictionary > noise dictionary

(c) speech dictionary < noise dictionary

*Figure 4.6: SNR and SER results for different dimension speech and noise dictionaries (speech dimension / noise dimension).*

introduce sparsity constraints on the calculus of the activations matrix, but the time complexity will grow anyway. Another option, followed at the beginning of this work, was to select only the activation that passed some criterion. While this method turned out to be useful in the way which is adopted in the third procedure, in general does not lead to correct reconstruction due to the fact that the criterions may change depending on the context. While in a specific situation some thresholds can be good, in others they can drastically affect the reconstruction. As we are trying to build a general enhancement system, we discarded this approach and we continued our experiments using the whole activation matrix $H$ without forced activation decision. Given that, we focused our attention on the research of an optimal dictionary construction that gives the possibility of obtaining a better reconstruction without imposing thresholds on the activation. We also follow the approach of imposing a sparsity constraint in the calculus of the activation matrix as described in 2.8. The results for the sparsity constraint are depicted below in the next section 4.2.5.

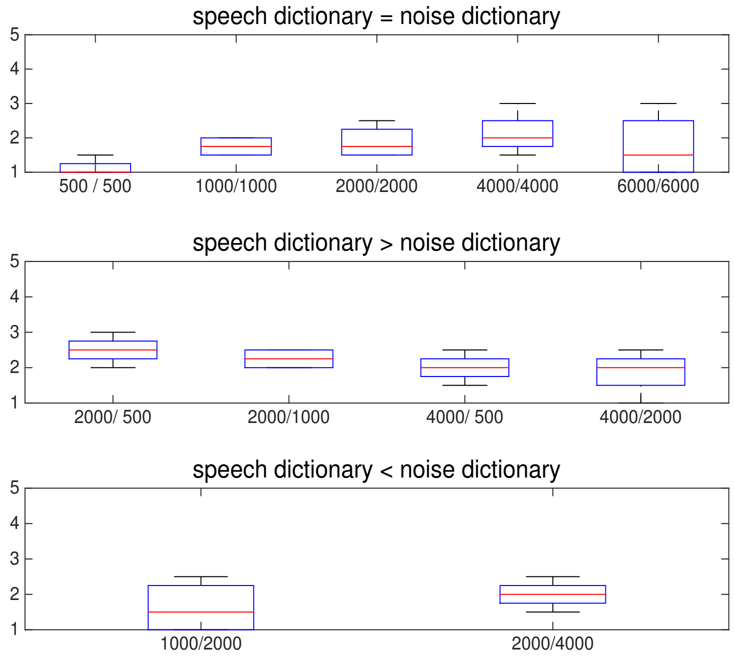


*Figure 4.7: MOS results for different dictionary dimensions.*

Another issue, occurring when we are dealing with dictionaries of different sizes, is that the NMF split the energy in a different way according with the dimension of the dictionaries. In fact, big dictionaries contain more bases that can better estimate an input segment compared with smaller dictionaries that unlikely perfectly match an input segment. For this higher amount of bases, bigger dictionaries will have more weight in comparison with the smaller ones. Moreover, the final reconstruction will be more influenced by the source which has been modeled by the bigger dictionary. Comparing these results with the ones resulting from the tests performed with the same noise and speech dictionaries dimension, we noticed that, in the case of noise dictionary bigger than speech dictionary, the results have a good noise suppression. However, they result in lower speech SNR and quality reconstruction. In the opposite case, the results obtain a better speech enhancement but still a quite noisy reconstruction due to the bad estimation of noise bases. It is possible to see this effect in the results showed in figure 4.6 (a), 4.6 (b) and in the MOS results (4.7).
As can be seen in the MOS results in figure (4.7), the reciprocal dimension of the dictionaries is equally important as the single dimension itself because it influences the human's perception.
Moreover, we can evidence that bigger dictionaries, despite the fact that they improve speech estimation, can also underperform the final quality perceived compared with smaller size dictionaries. This is caused by the huge number of bases active at the same time which introduce echoes and distortions.

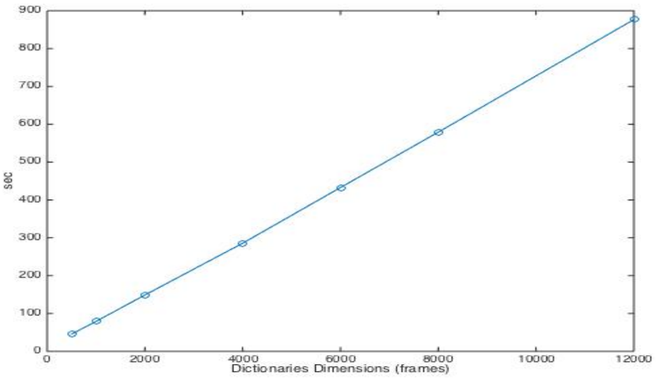


*Figure 4.8: Time expended for the NMF factorization of a 10 sec audio signal with $T = 1$ and the dictionaries dimensions variables.*

At last, with a big dictionary, the computational cost and time of execution are increased. We can observe this phenomenon in figure (4.8), showing some execution time, organized for each size. The execution time grows linearly according with the dictionary dimension.
Due to the trade-off between the time cost and the results obtained, the dictionaries dimension have to be chosen following the system's performances desired, if we want to reduce the execution time or if we want to obtain a better estimation. In this work we will continue our tests using the speech dictionary dimension of 4000 frames. This dimension is the one which gives the bests results in an fair amount of time as proved in this section and also following the results obtained by [3]. Moreover, comparing out outcomes with the same ones obtained with the same setting by [82, 83], we noticed that our system provide the same trend in the results without outperforming they. From now on, we limit our case of study to the second and the third procedure as they provide much better results compared with the first one.

### 4.2.3 Bases Selection Methods

The bases selection method represents the way we choose the exemplars to build the dictionaries. The noise dictionary bases are always randomly selected while for the speech dictionary we provide two methods: Randomly and a Maximum Energy approach (see section 2.6 and formula 2.24). This test is important to determine if, with the same general setting, Maximum Energy base selection can outperform Random base selection.

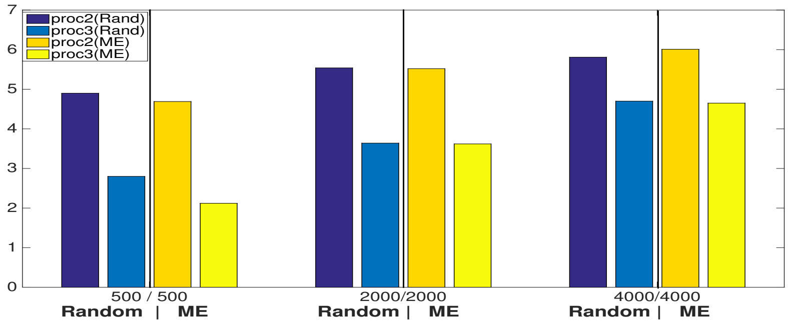


*Figure 4.9: SNR and SER results for Random and Maximum Energy (ME) base selection methods.*

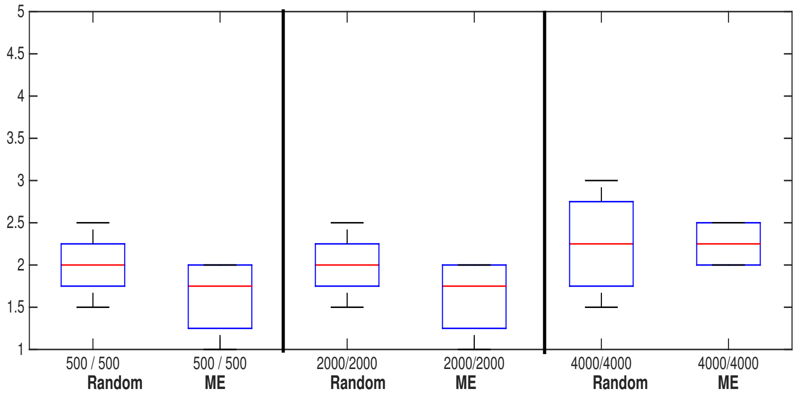


*Figure 4.10: MOS results for Random and Maximum Energy (ME) bases selection methods.*

As showed by the SNR and the MOS results (figure 4.10), this technique do not improve the Random approach for the choice of the exemplars. In fact it underperform or equals the results of the Random technique and has a lower MOS perception. This probably happens due to the fact that the Maximum Energy approach only models in a proper way the highest energy speech part, while the random exemplar method has more probability to model a wider range of exemplars with the same quality.

### 4.2.4 Additional Filtering

This test is made in order to prove that our system is equally reliable with strong band-limited communications signals as Voice over IP. We want to verify if the enhanced reconstruction also has the energy content in the frequencies that were not present in the test signal. To be able to simulate an harder communication condition, we use an additional filtering process that rejects the frequencies outside a determined range provided. We tested our system applying an extra band-pass filter between 300 Hz and 4 KHz. The results compare the SNR obtained with our system to a full-band version of the test signal. This test serves to demonstrate how much energy of a full-band test signal we lose or, in the opposite point of view, how much energy of the band-limited test signal we can reconstruct with our system.
As we can see in the figure 4.11 and in the picture representing the enhanced spectrogram 4.12 (b), our system can recreate great part of the

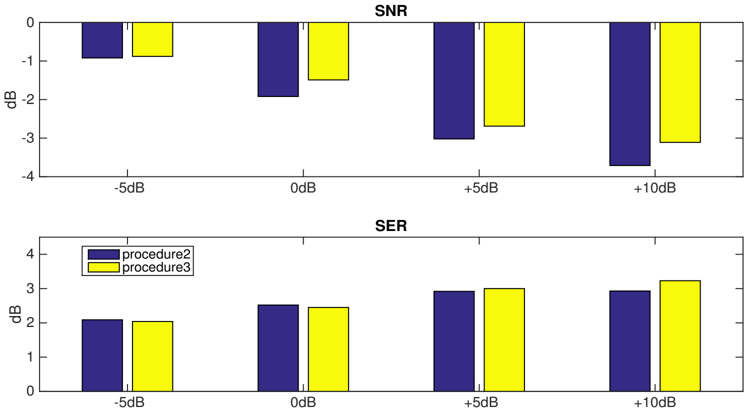
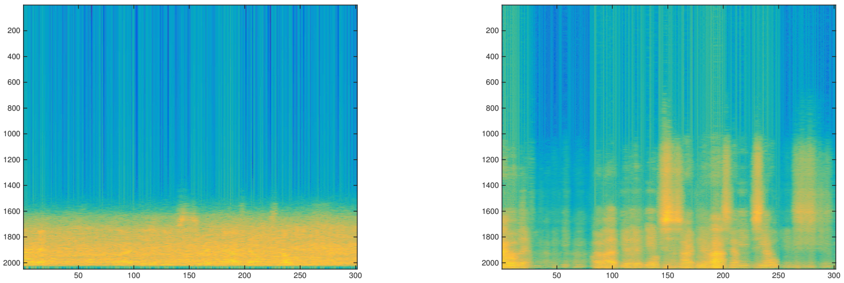


*Figure 4.11: SNR and SER results for the band-limited test signal.*



(a)

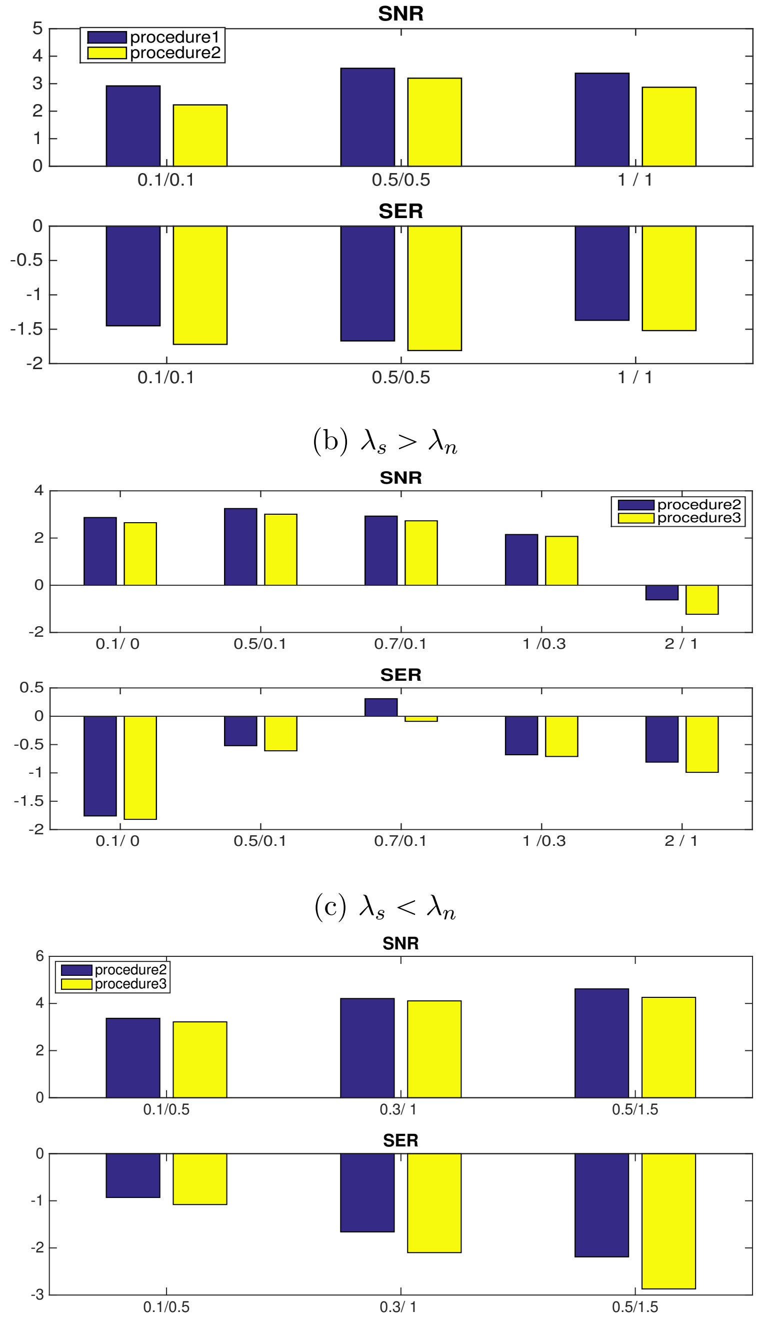
(b)

*Figure 4.12: Magnitudes of the band-limited input test signal (a) and the enhanced reconstruction (b).*

frequencies not present in the original band-limited test signal. Using the original filtered phase with the enhanced magnitude spectrogram we will obtain some distortion in the perception and the incorrect reconstruction of high frequencies. Usually, this problem can be solved by the G&L phase reconstruction technique, but observing our results, we can see that is not true. We tried to solve this problem with different settings of the system, as proved by the SER results obtained till now, but the phase reconstruction technique keeps providing worst results if compared with the one that use the original phase. This worst results are probably obtained by a badly phase reconstruction due to the presence of wrong bases in the reconstructed magnitude that alter the correct estimation. The resulting enhanced output with the original filtered phase contains some distortion in the frequencies not present in the test signal. However, it is still less unpleasant than the one obtained with disturb and echoes that wrong bases can introduce during the phase reconstruction. An improvement for the phase reconstruction can reside on estimating the frequencies' phase that are not present in the original signal while keeping the original phase for the rest of the magnitude spectrogram.

### 4.2.5 Sparsity Constraint

Of particular interest is the additional sparsity constraint that we can apply to the speech and noise activations detection. In order to avoid the limitative and dependent decision of which base make active in the reconstruction, we can apply an additional constraint during the factorization, as described in 2.8 and in 2.9. This constraint is an additional parameter that we can apply to each source to avoid the over-fitting of the activations matrix $H$ and to force the exemplars selected to be closer to the factorized speech part in the test speech signal. We test our system, using our previous best results, with the same sparsity values $\lambda$, as proposed in [6, 45], and with different sparsity values ($\lambda_s$ and $\lambda_n$) for each source, as proposed in [3].
The results delineated in figure 4.13 and also the MOS results, figure 4.14, describe how the selection of a good sparsity parameter can strongly influence the final enhancement, providing better results in terms of quality perceived. Analyzing these results, we can also see how the parameter $\lambda$ controls the trade off between sparseness and accurate reconstruction. In fact, low $\lambda$ can provides very sparse matrices, good to avoid multiple active bases, but also resulting in a quite low intelligibility reconstruction. From the analysis of the results, using this parameter, we conclude that the additional sparsity constraint greatly helps to obtain a better enhanced signal but has to be set depending on the situation. Anyway, from the outcomes in 4.14 we obtain the sparsity parameter produce better perceived results when $\lambda_s$ is at least bigger or equal to the sparsity parameter of the noise, ($\lambda_s >= \lambda_n$). In the opposite case, when the speech sparsity parameters is low, $\lambda$ provides a very sparse speech estimation, resulting in a still disturbed and low quality test signal reconstruction.

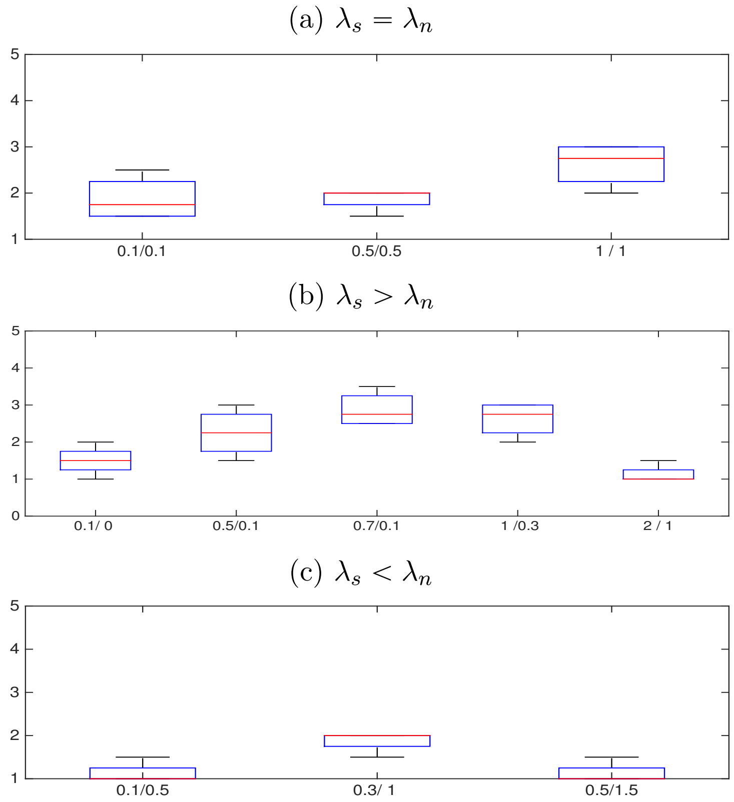


(a) $\lambda_s = \lambda_n$

(b) $\lambda_s > \lambda_n$

(c) $\lambda_s < \lambda_n$

*Figure 4.13: SNR and SER results for different $\lambda_s$ and $\lambda_n$.*



(a) $\lambda_s = \lambda_n$

(b) $\lambda_s > \lambda_n$

(c) $\lambda_s < \lambda_n$

*Figure 4.14: MOS results for different $\lambda_s$ and $\lambda_n$.*

### 4.2.6 Automatically Learned Bases

In our work we always used two sources: speech and noise. However, the NMF is also able to learn the dictionaries' bases by automatically updating a random initialized dictionary in order to approximate the input signal. In this section we want to explain the importance of providing the NMF some dictionaries capable of explaining each source that needs to be extracted. We performed three tests with: no fixed dictionaries provided (speech and noise automatically learned), fixed noise and learned speech and fixed speech and learned noise. We compared the outputs obtained with the results achieved with both fixed dictionaries as performed till now. The results are showed in figure (4.15).

As can be seen in figure 4.15, modeling the speech is fundamental in order to obtain an enhanced output. At the same time, due to the variability and unpredictability of the background degradation, the noise

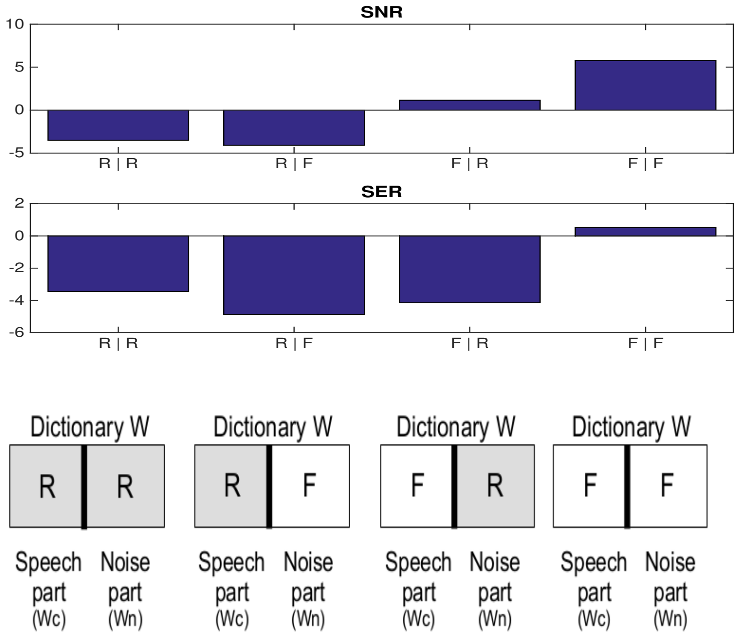


*Figure 4.15: SNR and SER results for fixed (F) and random initialized (R) dictionaries (speech dictionary type | noise dictionary type).*

bases can be learned automatically from the NMF while factorizing the input. Regardless, the performance obtained with this automatically learned bases underperformed the result obtained with a learned dictionary in the training phase.

We concluded that fixing the noise and the speech dictionaries is the best approach. A possible improvement would be provide the NMF factorization a fixed noise dictionaries with some bases that can be automatically learned to follow the changes of background noise through time. However, the influence of this automatically learned bases have to be studied in order to avoid wrong identification. In fact, the NMF can use these undefined bases to wrongly approximate speech or noise segment with them. In this case, the outputs will present lots of speech and noise estimation errors.

### 4.2.7 White Noise Dictionary

We did this test in order to verify if our system can be noise independent and assign the weight of real noises to white noise bases. As usual, the test signal was degraded with a real noise, while for the noise dictionary we use only white noise bases. We want to verify if this way our algorithm can outperform the results obtained with the automatically learned noise bases obtained in the previous test and recognize any kind of noise.

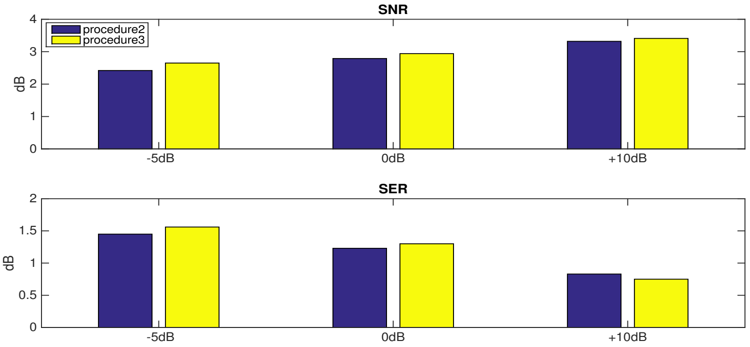


*Figure 4.16: SNR and SER results obtained using white noise as noise dictionary.*

As we can see in figure 4.16, the white noise dictionary outperform the results obtained for the automatic learned noise dictionary. However, listening to the reconstruction, it is possible to notice that it still sounds very noisy, with plenty background noise. This is explicated by the fact that the NMF, without a good matching case, approximate the noise with the speech bases. This make us unable to correctly remove the noise and this is not what we are looking for. So, we concluded that only using white noise bases outperform the automatic noise bases estimation but cannot suppress the background noise. In fact, if the NMF is forced to use only white noise bases, it cannot perform correctly the noise reduction. We can provide some white noise bases to NMF, but it must at least be free to automatically update the rest of bases if we want to eliminate the background noise.

# Chapter 5

# Conclusions and Future Works

In this thesis we continue the investigation on the exemplar-based NMF technique for speech enhancement. We studied the influence of different dictionaries in the NMF factorization and in the reconstruction. In particular, we focused on the enhancement of a noisy signal degraded with a real background noise using different speeches for the system training. The scope of this work was to give a proof-of-concept for developing a more general speech enhancement system. We analyzed the most important aspects in the NMF factorization such as: the dictionaries and the exemplars dimensions, the dictionary creation and the exemplars activation decision. We provide three approaches, with several settings for each one, to obtain the best enhanced version of the input speech signal.

Extrapolating the outputs obtained from each test, we prove that exemplars spanning multiple frames are only beneficial if the underlying sources are known. It also avoids discontinuities at concatenation points, typical of single frame exemplars. We also concluded that bigger dictionaries achieve better estimation. However, they have to be used jointly with a sparsity constraint factor in order to avoid the simultaneous activation of multiple exemplars. In any case, the NMF time of execution will increase linearly with the size of the dictionary. The sparsity constraint results prove that this parameter is beneficial and that it removes the problem of the activation decision. With our enhancement system we are also able to recreate the missing frequency parts in the filtered input signal. In return, the phase estimation algorithm cannot outperform the results obtained with the original test signal phase. Considering the influence of all these parameters, we conclude that the second approach is the best one over the three proposed. However, with more accurate study over speech features, the third approach can be faster and as good as the actual best. A great achievement is that we managed to obtain an enhanced test signal without needing parallel data, as it was required in the first method and in [8, 9]. In fact, by directly using a clean reference we are able to extract the speech part from a noisy signal and reconstruct its enhanced version.

Further research can be focused on introducing variable length exemplars size. In fact, larger exemplar dimensions are beneficial if the underlined source is known while shorter exemplars work better with unknown sources. Additional investigation needs to be done over the feature extraction and classification for voiced and unvoiced exemplars in the third procedure. This will lead to a more accurate classification and a better perceived enhanced reconstruction. Another important study arise from the application of a different distance for the NMF estimation. As described in section 2.5.1, a different divergences, like the Itakura-Saito (IS), can lead to faster results if we want to apply our system to a real-time scenario. Anyway, the enhancement obtained has to be studied and compared with the one obtained with the KL divergence. Related with the execution time there is also the problem of the phase estimation algorithm. Since we worked in batch processing, the time of execution was not an important issue. Now, if we want to get a faster and continuous phase estimation, we have to change the phase estimation algorithm and use a faster approach as the (RTISI-LA) proposed by Zhu [72]. Another improvement, related with the phase estimation in band-limited input signal, is the jointly use of the original phase and an estimated one. In order to obtain a better perceived output, we can reconstruct the frequencies not present, or not pertaining to the human frequency range, in the original signal while keeping the original phase for the other frequencies.

At last, in this thesis, we have restricted our study on just one individual male speaker and a minor number of degradations. Future experiments expect to increase the number and the gender of the speakers and the variety of background noises in order to examine the effectiveness of our methods.

# Bibliography

[1] Jacob Benesty, Shoji Makino, and Jingdong Chen. *Speech enhancement*. Springer, 2005.

[2] D D Lee and H S Seung. Algorithms for non-negative matrix factorization. *in Proc. Neural Information Processing System*, pages pp. 556–562, 2001.

[3] Jort F Gemmeke, Tuomas Virtanen, and Antti Hurmalainen. Exemplar-based sparse representations for noise robust automatic speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, pages 2067–2080, 2011.

[4] Paris Smaragdis and Judith C Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180. IEEE, 2003.

[5] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1066–1074, 2007.

[6] Mikkel Schmidt and Rasmus Olsson. Single-channel speech separation using sparse non-negative matrix factorization. 2006.

[7] Bhiksha Raj, Tuomas Virtanen, Sourish Chaudhuri, and Rita Singh. Non-negative matrix factorization based compensation of music for automatic speech recognition. In *INTERSPEECH*, pages 717–720, 2010.

[8] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Exemplar-based voice conversion in noisy environment. In *SLT*, pages 313–317, 2012.

[9] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Exemplar-based individuality-preserving voice conversion

for articulation disorders in noisy environments. In *INTERSPEECH*, pages 3637–3641, 2013.

[10] Simon Doclo and Marc Moonen. Gsvd-based optimal filtering for single and multimicrophone speech enhancement. *Signal Processing, IEEE Transactions on*, 50(9):2230–2244, 2002.

[11] Thomas Lotter, Christian Benien, and Peter Vary. Multichannel speech enhancement using bayesian spectral amplitude estimation. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–880. IEEE, 2003.

[12] Saeed V Vaseghi. *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008.

[13] Thomas Esch and Peter Vary. Efficient musical noise suppression for speech enhancement system. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4409–4412. IEEE, 2009.

[14] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*, volume 2. MIT press Cambridge, MA, 1949.

[15] Jalal Taghia, N Mohammadiha, Jinqiu Sang, V Bouse, and R Martin. An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4640–4643. IEEE, 2011.

[16] Yariv Ephraim. A bayesian estimation approach for speech enhancement using hidden markov models. *Signal processing, IEEE transactions on*, 40(4):725–735, 1992.

[17] Yariv Ephraim. Gain-adapted hidden markov models for recognition of clean and noisy speech. *Signal Processing, IEEE Transactions on*, 40(6):1303–1316, 1992.

[18] Yariv Ephraim, David Malah, and B-H Juang. On the application of hidden markov models for enhancing noisy speech. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 37(12):1846–1856, 1989.

[19] Hossein Sameti, Hamid Sheikhzadeh, Li Deng, and Robert L Brennan. Hmm-based strategies for enhancement of speech signals embedded in nonstationary noise. *Speech and Audio Processing, IEEE Transactions on*, 6(5):445–455, 1998.

[20] Hossein Sameti and Li Deng. Nonstationary-state hidden markov model representation of speech signals for speech enhancement. *Signal Processing*, 82(2):205–227, 2002.

[21] Nasser Mohammadiha, Rainer Martin, and Arne Leijon. Spectral domain speech enhancement using hmm state-dependent super-gaussian priors. *IEEE Signal Processing Letters*, 20(3):253–256, 2013.

[22] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[23] Jingu Kim and Haesun Park. Sparse nonnegative matrix factorization for clustering. 2008.

[24] Paris Smaragdis, Cedric Fevotte, G Mysore, Nasser Mohammadiha, and Matthew Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *Signal Processing Magazine, IEEE*, 31(3):66–75, 2014.

[25] Pentti Paatero. Least squares formulation of robust non-negative factor analysis. *Chemometrics and intelligent laboratory systems*, 37(1):23–35, 1997.

[26] Solomon Kullback. The kullback-leibler distance, 1987.

[27] Suvrit Sra and Inderjit S Dhillon. Generalized nonnegative matrix approximations with bregman divergences. In *Advances in neural information processing systems*, pages 283–290, 2005.

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

[29] Y Kenan Yilmaz. Generalized beta divergence. *arXiv preprint arXiv:1306.3530*, 2013.

[30] Shinto Eguchi and Yutaka Kano. Robustifing maximum likelihood estimation by psi-divergence. *ISM Research Memorandam*, 802, 2001.

[31] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the $\beta$-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

[32] Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, 2011.

[33] Fumitada Itakura and Shuzo Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proceedings of the 6th International Congress on Acoustics*, volume 17, pages C17–C20. pp. C17–C20, 1968.

[34] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830, 2009.

[35] Chih-Jen Lin. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *Neural Networks, IEEE Transactions on*, 18(6):1589–1596, 2007.

[36] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[37] Jingu Kim and Haesun Park. Fast nonnegative matrix factorization: An active-set-like method and comparisons. *SIAM Journal on Scientific Computing*, 33(6):3261–3281, 2011.

[38] Cho-Jui Hsieh and Inderjit S Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1064–1072. ACM, 2011.

[39] Paris Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Independent Component Analysis and Blind Signal Separation*, pages 494–499. Springer, 2004.

[40] Paris Smaragdis. Convolutive speech bases and their application to supervised speech separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):1–12, 2007.

[41] Rahim Saeidi, Antti Hurmalainen, Tuomas Virtanen, and David A van Leeuwen. Exemplar-based sparse representation and sparse discrimination for noise robust speaker identification. In *Odyssey speaker and language recognition workshop*, 2012.

[42] Antti Hurmalainen, Jort Gemmeke, and Tuomas Virtanen. Nonnegative matrix deconvolution in noise robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4588–4591. IEEE, 2011.

[43] Felix Weninger, Martin Wollmer, Jurgen Geiger, Bjorn Schuller, Jort F Gemmeke, Antti Hurmalainen, Tuomas Virtanen, and Ger-

hard Rigoll. Non-negative matrix factorization for highly noise-robust asr: to enhance or to recognize? In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4681–4684. IEEE, 2012.

[44] Michael Carlin, Nicolas Malyska, and Thomas F Quatieri. Speech enhancement using sparse convolutive non-negative matrix factorization with basis adaptation. In *INTERSPEECH*, 2012.

[45] Paul D O'grady and Barak A Pearlmutter. Convolutive non-negative matrix factorisation with a sparseness constraint. In *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*, pages 427–432. IEEE, 2006.

[46] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004.

[47] Lidan Miao and Hairong Qi. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(3):765–777, 2007.

[48] Guoxu Zhou, Shengli Xie, Zuyuan Yang, Jun-Mei Yang, and Zhaoshui He. Minimum-volume-constrained nonnegative matrix factorization: Enhanced ability of learning parts. *Neural Networks, IEEE Transactions on*, 22(10):1626–1637, 2011.

[49] Chris Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):45–55, 2010.

[50] Jort Gemmeke and Bert Cranen. Noise robust digit recognition using sparse representations. *Proceedings of ISCA 2008 ITRW? Speech Analysis and Processing for knowledge discovery*, 2008.

[51] Nicolas Gillis. Sparse and unique nonnegative matrix factorization through data preprocessing. *The Journal of Machine Learning Research*, 13(1):3349–3386, 2012.

[52] Michael Zibulevsky and Barak Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural computation*, 13(4):863–882, 2001.

[53] Tuomas Virtanen, Ali Taylan Cemgil, and Simon Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Acoustics, Speech and Signal Processing,*

*2008. ICASSP 2008. IEEE International Conference on*, pages 1825–1828. IEEE, 2008.

[54] David Field. What is the goal of sensory coding? *Neural computation*, 6(4):559–601, 1994.

[55] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.

[56] Jort Gemmeke, Louis ten Bosch, Lou Boves, and Bert Cranen. Using sparse representations for exemplar based continuous digit recognition. In *Proc. EUSIPCO*, pages 24–28. Citeseer, 2009.

[57] Yanping Zhao, Xiaohui Zhao, and Bo Wang. A speech enhancement method employing sparse representation of power spectral density?

[58] Li Shang, Yan Zhou, Jie Chen, and Wen-jun Huai. Nature image feature extraction using several sparse variants of non-negative matrix factorization algorithm. In *Advances in Neural Networks–ISNN 2012*, pages 274–281. Springer, 2012.

[59] Grindlay. Nmflib. http://www.ee.columbia.edu/∼ *grindlay/code.html*, 2010.

[60] Matthias Mauch and Sebastian Ewert. The audio degradation toolbox and its application to robustness evaluation. https://code.soundsoftware.ac.uk/projects/audio-degradation-toolbox/, 2013.

[61] Waleed H Abdulla, David Chow, and Gary Sin. Cross-words reference template for dtw-based speech recognition systems. In *TENCON 2003. Conference on Convergent Technologies for the Asia-Pacific Region*, volume 4, pages 1576–1579. IEEE, 2003.

[62] D. Ellis. Dynamic time warp (dtw) in matlab. http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/, 2003.

[63] Mike Brookes et al. Voicebox: Speech processing toolbox for matlab. *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, 1997.

[64] Bojana Gajic and Kuldip K Paliwal. Robust speech recognition using features based on zero crossings with peak amplitudes. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 1, pages I–64. IEEE, 2003.

[65] Doh-Suk Kim, Soo-Young Lee, and Rhee Man Kil. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *Speech and Audio Processing, IEEE Transactions on*, 7(1):55–69, 1999.

[66] Yingyong Qi and Bobby R Hunt. Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier. *IEEE Transactions on Speech and Audio Processing*, 1(2):250–255, 1993.

[67] RG Bachu, S Kopparthi, B Adapa, and BD Barkana. Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In *American Society for Engineering Education (ASEE) Zone Conference Proceedings*, pages 1–7, 2008.

[68] DS Shete and SB Patil. Zero crossing rate and energy of the speech signal of devanagari script.

[69] Olivier Lartillot and Petri Toiviainen. A matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*, pages 237–244, 2007.

[70] Nicolas Sturmel and Laurent Daudet. Signal reconstruction from stft magnitude: a state of the art. In *International Conference on Digital Audio Effects (DAFx)*, pages 375–386, 2011.

[71] Daniel Griffin and Jae S Lim. Signal estimation from modified short-time fourier transform. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(2):236–243, 1984.

[72] Xinglei Zhu, Gerald T Beauregard, and Lonce Wyse. Real-time iterative spectrum inversion with look-ahead. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 229–232. IEEE, 2006.

[73] Vincent Akkermans, Frederic Font, Jordi Funollet, Bram De Jong, Gerard Roma, Stelios Togias, and Xavier Serra. Freesound 2: An improved platform for sharing audio clips. In *Late-braking demo abstract of the Int. Soc. for Music Information Retrieval Conf*, 2011.

[74] D Griffin, D Deadrick, and Jae S Lim. Speech synthesis from short-time fourier transform magnitude and its application to speech processing. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, volume 9, pages 61–64. IEEE, 1984.

[75] Xinglei Zhu, Gerald T Beauregard, and L Wyse. Real-time signal estimation from modified short-time fourier transform magnitude spectra. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(5):1645–1653, 2007.

[76] Gerald T Beauregard, Xinglei Zhu, and Lonce Wyse. An efficient algorithm for real-time spectrogram inversion. In *Proceedings of the 8th International Conference on Digital Audio Effects*, pages 116–118, 2005.

[77] Mouhcine Chami, Joseph Di Martino, Laurent Pierron, et al. Real-time signal reconstruction from short-time fourier transform magnitude spectra using fpgas. In *5th. International Conference on Information Systems and Economic Intelligence-SIIE ? 2012*, 2012.

[78] Jort Florent Gemmeke, Bert Cranen, and Ulpu Remes. Sparse imputation for large vocabulary noise robust asr. *Computer Speech & Language*, 25(2):462–479, 2011.

[79] Ivan Ivek. Supervised dictionary learning by a variational bayesian group sparse nonnegative matrix factorization. *arXiv preprint arXiv:1405.6914*, 2014.

[80] Yifeng Li and Alioune Ngom. Supervised dictionary learning via non-negative matrix factorization for classification. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 1, pages 439–443. IEEE, 2012.

[81] Zunyi Tang, Shuxue Ding, Zhenni Li, and Linlin Jiang. Dictionary learning based on nonnegative matrix factorization using parallel coordinate descent. In *Abstract and Applied Analysis*, volume 2013. Hindawi Publishing Corporation, 2013.

[82] Deepak Baby, Tuomas Virtanen, Tom Barker, et al. Coupled dictionary training for exemplar-based speech enhancement. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2883–2887. IEEE, 2014.

[83] Nikolay Lyubimov and Mikhail Kotov. Non-negative matrix factorization with linear constraints for single-channel speech enhancement. *arXiv preprint arXiv:1309.6047*, 2013.