Low Latency Audio Source Separation for Speech Enhancement in Cochlear Implants

Jordi Hidalgo Gomez

MASTER THESIS UPF/ YEAR 2012
Master in Sound and Music Computing

Master Thesis Supervisor
Dr. Waldo Nogueira

Master Thesis Co-Supervisor
Dr. Jordi Janer

Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona

Abstract

This master thesis is a combination of two areas of Sound and Music Computing, Blind Source Separation and Cochlear Implants. The research focuses in the evaluation of existing source separation algorithms in order to improve noise reduction strategies in the context of cochlear implants. The modification and adaptation of a low latency algorithm is the point of start for the evaluation based in the requirements of speech signals in cochlear implants. The evaluation consists in a different set of objective and subjective experiments to determine the speech intelligibility enhancement produced by the separation process. Objective evaluation has revealed that a very good performance level is achieved with low latency algorithms compared to NMF which take considerably higher computation time. A series of subjective tests have been conducted with cochlear implant patients in order to compare the objective results and determine the real speech intelligibility level. The low latency algorithm showed only improvements in few situations where the noise reduction algorithm outperforms in most of the cases. Accurate analysis determined that the main reason of the speech degradation caused by low latency algorithm is because of the algorithm is not designed to detect unvoiced consonants and a lot of speech content is missing. But last experiments revealed that is possible to recover this consonants which can considerably improve the performance and later speech intelligibility of LLIS.

Acknowledgements

I would like to thank to dr. Xavier Serra and the people from the MTG to give me the opportunity of join the master where I have learned a lot concepts and discovered a lot of new exciting areas of knowledge. Then I would like to mention also the SMC collages which have made this experience even more grateful.

Thanks also to my cosupervisor Jordi Janer who have helped in many aspects of the thesis. As well as Ricard Marxer which has advised in many occasions regarding the algorithm and many concepts about source separation. Specially I am so thankful to my supervisor Waldo Nogueira who has spent many ours helping me in building, designing and evaluating the system as well as his patience for many of the times when we realized the tests.

Then another people who without them this thesis had not been carried out are the participants in the tests. First the volunteers Miquel Matas, Maria Boada, Aleix Bou, Marc Freixes, Telma LLos and Miguel Angel Betoret. Then the CI patients Jose Ma Folch, Mario Millan, Gloria Salvado, Jordi A. and Juan Ibanyez. Thanks for your great patience. Also the people from the CBCLaB for letting us make the tests in one of their rooms.

Finally I would like to thank my family, my girlfriend Maria and all my friends for all his support over the years.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Motivation and Goals

Cochlear implants are electronic devices that are implanted in people suffering from hearing loss. These devices are implanted surgically inside the patient head directly connected to the cochlea. By means of electric pulses, the implant stimulates the cochlea with a set of small electrodes corresponding to different frequency bands. The sounds are transduced by a microphone located in the processor part, out of the head. This has the function of process, adapt and after transmitted to the internal part, map the sound in the different frequency bands for the stimulation.

Ideally patients are implanted at early stages of the childhood, which is perfect to adapt to the behavior and perception obtained with the implant. This really facilitates the speech learning process. Another successful case of implantation is for post locutive hearing people i.e. for people that suffered an accident or disease that provoked hearing loss but they learned to speech years ago.

Once implanted, there is a whole process of adaption which is regularly calibrated. The patients need to adjust the intensity of the currents in the electrodes and other configuration issues regarding the transmission between the processor and the internal part. Ideally when the implant is working, patients can perceive sounds in a very similar way than normal people do. There are some limitations like the frequency resolution that is lower or the dynamic range which is narrower than for normal people. But real problems begin in the presence of noisy environments. The noise reduction is one of the big challenges of the audio processing in cochlear implants. There is a large set of successful noise reduction algorithms and all the implants have several strategies implemented in their processors. Patients can dynamically select different programs depending on the nature of the noise and the scenario. But the most conflict situations are in the presence of non stationary noises, where most noise reduction algorithms

reduce their performance. The problem is that in real life, most of the noises are non stationary noises, like could be a cafeteria, a train station or simply the traffic sound. In these situations if the SNR of speech is not considerably higher the intelligibility becomes harder and sometimes impossible.

In some way blind source separation can be seen as a noise reduction method. If we have a mixture of sounds composed by a speech signal and a set of other components, if we detect the speech signal and we separate them from the rest of components of the mixture we can considerate it as a noise reduction strategy. Blind audio source separation is a well known technique specially for music signals. Retrieving the isolated components in a given mixture like the instruments or the lead voice have been demonstrated to be successfully achievable.

Most well known algorithms for blind audio source separation are tho so called offline algorithms. The term offline means that can not operate in real time due the computation time required to separate the sources. These normally are implemented with a popular technique called Non Negative Matrix Factorization (NMF). This technique is very useful for the factorization of audio data such as spectrograms which values are all positive. This method reported some of the best quality results in the recent years. But unfortunately the computation cost of the iteration algorithms used for NMF makes this algorithms unable to use it real time. In contrast to NMF, time frequency masking is a technique that has some online implementations which has obtained really good results. This technique is principally focused on the creation of binary masks corresponding to the target frequency components of the sources.

The main goal of this thesis is to determine the state of the art of blind source separation and then evaluate how these techniques can improve speech intelligibility in the context of cochlear implants. More concretely we want to evaluate the performance of an online or low latency method in contrast to offline strategies. The reason is that if an algorithm presents good results the next step could be include the strategy to the cochlear implant processor and this must be obviously with the low latency as possible. Derived objectives of this work is to evaluate the low latency source separation techniques with different kind of noise, specially non stationary noises which are the most conflict for the current noise reduction algorithms implemented in the implants. Also compare the performance of other source separation algorithms is contemplated to determine main differences and problems.

1.2 Structure of the document

This document is organized in six chapters. First chapter is this same chapter which represents an introduction to the both fields treated in the documents and also the motivation and goals are defined. Second chapter is a literature review of the most relevant research done in the last years within the context of source separation and cochlear implants. The third chapter is basically a quick description of the algorithm selected for source separation and some of his requirements regarding speech. Then the evaluation process is explained in the fourth chapter with the description of all the objective and subjective measures used. Results are detailed in the fifth chapter where also a discussion related is made. Finally, the last chapter describes the main contributions and conclusions followed of some future work strategies.

Chapter 2

STATE OF THE ART

This chapter presents the present state of the art of Source Separation and Cochlear implants. In section 2.1 we define the theoretical and basic principles of Source Separation. We define also the most common approaches in order to deal with speech and music. Finally we describe some of the existing Frameworks and its characteristics.

In section 2.2 we will explain the main features of the hearing with Cochlear Implants (CI) and then the used signal processing strategies in CI. At the end, Some Source separation methods applied to CI are reviewed.

2.1 Source Separation

Source separation is the task of extracting individual signals from an observed mixture by computational means. This technique is used in different fields of research like Audio and Image processing, Biology or Data Mining. We will be focused on the separation of Audio signals. This problem is always compared with the *cocktail party problem* which describes a cocktail party situation, where many sound sources are mixed like voices, music and noise in general. Humans have the ability of naturally focus on a specific source and isolate it from the rest. This is a relatively easy task for the human auditory system but it becomes difficult when we attempt to simulate the problem in a computational way.

One of the responsible of the popularization of source separation is [Bregman, 1990]. He proposed a cognitive approach for describing how humans perceive and understand sound objects in mixtures as independent components. This process was called *Auditory Scene Analysis* (ASA). He proposed five grouping principles that are used by the brain to separate and isolate sound sources. These are proximity, similarity, good continuation, closure and common fate. These ideas correspond to how we perceive sounds in time and frequency and are very close to the *Gestalt Psychology*. In order to simu-

late the ASA principles, it was raised *Computational Auditory Analysis* (CASA) which proposes algorithms focused on extract the individual sources of an audio mixture in a similar way as the human auditory system does. Although most of the methods used for source separation means in the field of signal processing are more focused on the mathematical way to solve the problem, it is always good to consider the cognitive approach by Bergman in order to build a more complete and robust strategy.

Different names are used in the literature to talk about source separation, mainly because the difference between blind and non-blind source separation. This could induce sometimes to the confusion of what is blind and what is non-blind. The requisite to be blind is to have any prior information about the sources that are going to be separated. But this is a little bit ambiguous, is difficult(or impossible) to have absolutely any prior information of the sources, we normally assume at least statistical independence or even if we are considering to separate speech from music, this can be itself considerate as prior information.

[Burred, 2009] gives a good approach for classify the mentioned types of source separation. *Blind Source Separation* (BSS) is defined as the situations where there is a little or any previous knowledge of the sources. We talk about *Semi Blind Source Separation* (SBSS) when advanced models like sinusoidal are took into account in order to separate the sources. This method can be considerate as a supervised separation, where the system is trained with sound/music examples. Finally we can consider that the system is *Non-Blind Source Separation* when the an input is required to perform the separation, like could be a score or other high-detailed information about the sources. In the rest of the document, if is not specified (blindness), we will be talking about source separation referring to a general approach.

A good distinction between the applications of source separation could be the made by [Vincent et al., 2003]. They distinguished two main types of applications, Audio Quality Oriented (AQO) and Significance Oriented(SO) applications. The main difference between AQO and SO is the quality of the output signal. AQO applications deals with the quality of the extracted sound, and the finality of his application is to listen the separated signal. In contrast, SO applications are interested more in the content and information of the extracted signal. This is very well suited with Music Information Retrieval (MIR) where the intrinsic information about the source in form of features and specific descriptors are required. As we will see in section 2.2, speech and music intelligibility deals with the maximization of audio quality, so the work done in this thesis will be more related to AQO applications.

2.1.1 Overview and principles

Mixing Models

[Burred, 2009] describes the concept of sound as mixtures of signals. When a sound wave incises into a microphone it starts the process of transduction form acoustic sound to electrical oscillation. The information about propagation directions gets lost and is translated into vibrations that at most, can simulate the effect of propagation in function of the amplitude and depending on the directionality of the given microphone. Once the sound is transduced, it will become unidimensional so the different sources that are present at this moment will be merged into one. We can define our sound $x(t)$ as a mixture of N signals $y(t), n = 1, ..., N$ where each signal has its corresponding instantaneous amplitude.

$$x(t) = \sum_{n=1}^{N} y(t) \tag{2.1}$$

The different mixing conditions are these represented by the source signals $s_n(t)$ that will be transformed into the source images $y_n(t)$.

Instantaneous mixing model The most simple mixing model is the called linear or instantaneous mixing model, and only assumes that the source signals have been modified by an amplitude scaling:

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t), \qquad m = 1, \cdots, M \tag{2.2}$$

Equation 2.2 can be traduced as a linear equations system:

$$\left\{ \begin{array}{rcl} x_1(t) & = & a_{11}s_1(t) + a_{12}s_2(t) + ... + a_{1N}s_N(t) \\ \vdots & & \\ x_M(t) & = & a_{M1}s_M(t) + a_{M2}s_M(t) + ... + a_{MN}s_N(t) \end{array} \right) \tag{2.3}$$

We can express the whole system as Mx1 vector mixtures $x = (x_1(t), ..., x_M(t))^T$ and Nx1 vector of sources $s = (s_1(t), ..., s_N(t))^T$:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_M(t) \end{pmatrix} = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1N}(t) \\ a_{21}(t) & a_{22}(t) & \cdots & a_{2N}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1}(t) & a_{M2}(t) & \cdots & a_{MN}(t) \end{pmatrix} \cdot \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{pmatrix}, \tag{2.4}$$

what can be expressed as:

$$\mathbf{x} = \mathbf{As} \tag{2.5}$$

The goal of linear source separation is, given the observed set of mixtures x, to solve such a set of linear equations towards the unknown s. However, in contrast to basic linear algebra problems, the coefficients a_{mn} are also unknown. As in linear algebra, a system with more equations than unknowns (M mixtures $> N$ sources) is called overdetermined, less equations than unknowns (M mixtures $< N$ sources) undetermined and finally the same number of equations than unknowns (M mixtures $= N$ sources) even-determined.

Delayed mixing model The delayed model is referred for this cases where sounds reach each of the existing sensors at different times giving a source-to-sensor delay $\delta_m n$:

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - \delta_{mn}), \qquad m = 1, \cdots, M \tag{2.6}$$

Convolutive mixing model The convolutive model appears when a source is filtered between the sensor. We can define the impulse response filter is:

$$h_{mn}(t) = \sum_{mn}^{Kmn} a_{mnk} \delta(t - \delta_{mnk}), \tag{2.7}$$

where K_{mn} is the length of that particular impulse response. Then the mixture received is:

$$x_m(t) = \sum_{n=1} h_{mn} * s_n(t) = \sum_{n=1}^{N} \sum_{mn}^{Kmn} s_n(t - \delta_{mn}), \qquad m = 1, \cdots, M \tag{2.8}$$

This model is often referred to reverberant scenarios, where K_{mn} is defined by the number of possible reverberant paths that the sources need to do to reach the sensors. The delayed model is the same than a convolutive model for $K_{mn} = 1$ for all m,n.

PCA, ICA and ISA

Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are methods widely used for BSS. The main objective of PCA[Jolliffe, 2002] is to reduce the dimensionality of a data set consisting of a large number or interrelated variable, maintaining as much as possible the variance. This is achieved transforming the data into new variables, the principal components, which are uncorrelated.

Independent component analysis (ICA) is an extension of PCA which estimates [Itoyama, 2011] source data from observed data which have instantaneous or convolutive mixture representation based only on the assumption that sources are statistically independent. Basically [Jutten and Comon, 2010], for linear instantaneous mixtures, ICA methods aim at estimating the demixing matrix A yielding the estimated sources $x(t) = As(t)$, which are statistically independent.

ICA[Itoyama, 2011] has the advantage that can by applied to arbitrary time-series signals by assuming statistical independence of the sources. Although this assumption is correct for speech signals, is unsuitable for separating complex musical signal where the independence is not correct.

Finally, another used method for source separation related to PCA and ICA is Independent Subspace Analysis (ISA). ISA[Casey, 2000] extends ICA by identifying mutiple independent "spaces" from a given data. A typical[Vincent, 2006] decomposition in sound could be the power spectrogram as two independent subspaces.

NMF

Nonnegative matrix factorization(NMF) is another common method used for dimensionality reduction and one of the most used methods for source separation. This method is used to separate music signals, but is also very common in other fields like image processing or text mining[Jutten and Comon, 2010]. One of the main reasons why NMF is due it is a representation of only non-negative data, this is very useful to represent certain amounts of data like for example could be an spectrogram where time and frequency are factorized in two independent matrices.

We can approximate our mixture as[Lee and Seung, 1999][Hoyer, 2004]:

$$\mathbf{v}^t \approx \sum_{i=1}^{M} \mathbf{w}_i h_i^t = \mathbf{W}\mathbf{h}^t \tag{2.9}$$

$\mathbf{W}$ is a $\mathbf{N}$x$\mathbf{M}$ matrix containing the basis vectors of our mixture, so if we consider that we have T basis vectors $\mathbf{w}_i$ then we can assume that our system can be expressed as:

$$\mathbf{V} \approx \mathbf{W}\mathbf{H}, \tag{2.10}$$

where $\mathbf{H}$ contains the coefficients $\mathbf{h}^t$ corresponding to each measurement vector $\mathbf{v}^t$. PCA, ICA and NMF all can be seen as a matrix factorization of the data that we want to separate, the difference is that NMF imposes that all elements will be positives. So any component will be subtractive.

Once the factorization equation is defined, the goal is to solve the minimization problem given by the following distance[Févotte et al., 2009]:

$$\min \quad D(\mathbf{V}|\mathbf{WH}) \qquad W, H \geq 0, \tag{2.11}$$

where $D(\mathbf{V}|\mathbf{WH})$ is a cost function defined by:

$$\min \quad D(\mathbf{V}|\mathbf{WH}) = \sum_{f=1}^{N} \sum_{n=1}^{M} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}), \tag{2.12}$$

The most popular cost functions are the Euclidean distance in equation 2.13 and the Kullback-Leibler divergence in equation 2.14.

$$d_{EUC}(x|y) = \frac{1}{2}(x-y)^2 \tag{2.13}$$

$$d_{KL}(x|y) = x \log \frac{x}{y} - x + y \tag{2.14}$$

[Lee and Seung, 2001] proposed a series of "Multiplicative update rules" in order to solve more efficiently the gradient step algorithms raised from the cost functions. Here, they proved that convergence is considerably faster using the mentioned multiplicative update rules.

2.1.2 Source Separation Methods for Speech and Music

Many of the different approaches are studied during these years have been proposed for extracting the independent sources of mixed audio signals. Some of them are designed to extract a very specific source like could be speech/singing or percussive sounds. But most of the proposed algorithms are thought to work in a more general mixture of signals, where voice, instruments and noise are all mixed together. Most of the approaches are different combinations or extensions of the methods presented in the following sections.

Source Separation with PCA, ICA and ISA

As we have mentioned in section 2.1.1 PCA, ICA and ISA is a widely used method for BSS. [Casey, 2000] presented a method for extracting independent audio sources from a single-channel mixture using ISA. An example of how the audio is separated is the called "Spectrogram subspace separation". In this example the audio signal is decomposed as an spectrogram separated by its spectral components as independent subspaces.

Another interesting approach is presented by [Burred, 2009] where the core of the separation tasks is PCA. This separation approach is based on timbre models. A mixture signal is sinusoidally modeled. After an onset detection, the source separation task is made by comparing the timbres of each signal with a predefined timbre model library. Finally the tracks are extracted by matching the timbres.

But ICA[Vincent, 2006] is more effective when the system is overdetermined, there are more mixtures than sources and also his performance decreases when the reverberation increases.

Source Separation with NMF

In ISA, ICA and PCA the estimation is based on the independence of the spectral components, if those components are magnitude or power is a good approach to use nonnegative constraints. Currently there are not effective algorithms for non-negative PCA. This is the reason why in the recent years NMF has become more popular. As we have seen in section 2.1.1, the estimation is based on a series of iteration algorithms where the cost function of a distance between the mixture X and the sources and its coefficients AS is attempted to be minimized. The algorithms proposed by [Févotte et al., 2009] are based on the Itakura-Saito divergence denoted by:

$$d_{IS}(x|y) = \frac{x}{y} - log\frac{x}{y} - 1 \tag{2.15}$$

The use of IS-divergence has multiple advantages in front of other cost functions like Euclidean distance or Kullback-Leiber. For example the scale invariance is very useful for the decomposition of audio spectra and also has a faster convergence. The performance of the IS cost function is demonstrated with a piano excerpt. Here the six individual components have been successfully separated, the first four as the individual pitches, the fifth with the content of the hammer hits and pedal releases. Finally the sixth component contents the residual noise.

[Durrieu et al., 2009a] presented a system for separating the main instrument in stereophonic mixtures. His algorithm is an extension of the algorithm presented in [Durrieu et al., 2009b], where the same methodology is applied for mono mixtures. The operation of the algorithm has two very important steps for the separation of the solo or main instrument and the accompaniment. These are the Melody tracking and the Wiener filters. The melody tracking is made with the Viterbi smoothing algorithm which estimates the fundamental frequency of the solo instrument at each frame. The final separation is based on Weiner filtering which is applied in the frequency domain and allows separate the solo from the accompaniment. The algorithm is complemented

with different steps of parameter estimation through multiplicative update rules. The overall operation of solo/accompaniment separation for mono signals is depicted in figure 2.1

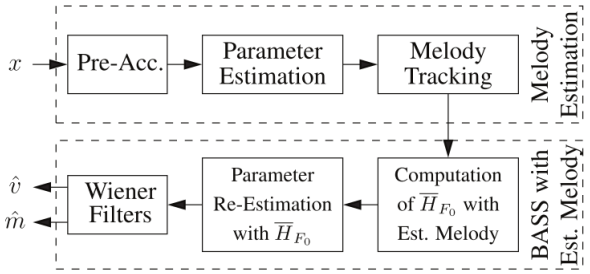

Figure 2.1: Solo/Accompaniment Separation algorithm [Durrieu et al., 2009b]

Another interesting example of the use of NMF is the study done by [Virtanen, 2005]. Virtanen defines his method as "Weighted Non-Negative Matrix Factorization". As in other cases the observed signal is transformed to the frequency domain to obtain the spectrogram and factorize the sources and its coefficients. But he introduces the concept of weight for the divergence in contrast to other methods. He defines the weighted divergence as:

$$D(\mathbf{X}|\mathbf{AS};\mathbf{W}) = D(\mathbf{W}.*\mathbf{X}|\mathbf{W}.*(\mathbf{AS})), \tag{2.16}$$

where W is a positive TxF weight matrix and .* is the element-wise multiplication. This method is described as perceptual weight, because the weighted sum of the spectrum bins are selected to be equal to the estimated loudness. The algorithm was tested with mixed pitched and percussive sounds and it was found that a very high perceptual quality of the separated sources is achieved. In the other hand, some problems were encountered in order to separate two note mixtures of pitched instruments.

[Paulus and Virtanen, 2005] describes a method fully focused on percussive sounds. They start by obtain the instrument spectra from a trained dataset. This dataset is composed of different percussive instrument sounds(bass drum, snare and hi-hat) recordings. These sounds are factorized by NMF to extract the spectrogram. The estimation of the time-varying gains is done by minimizing the cost function between the observed spectrum and the model **AS**. Finally an onset detection procedure is applied with the time-varying gains and finally the training signals are separated according to the onset locations.

Usually NMF-based algorithms are designated as off-line. This is due in most of the algorithms the estimation requires that the whole signal has to be known. The problem of off-line methods is that can not be used for real-time applications where

only the present and past frames of the audio information are known. But recently some approaches for real-time NMF have been proposed. An example could be the approach presented by [Joder et al., 2012]. Here they define an "On-Line NMF" as a sliding window method for decomposing the recent and past spectrum. The matrices are updated using a fixed number of iterations. With the sliding window approach, a low number of iterations is enough. Although this system is outperformed by off-line algorithms, the results obtained are successfully achieved for speech separation.

Time-Frequency Masking

In contrast to NMF, Time-Frequency Masking is a method very extended for real-time applications. This is mainly due the simplicity and inexpensiveness of computation of his algorithms. Time-frequency masks assumes that the sources to separate are orthogonal in the time-frequency domain. One of the most common binary mask in stereo music is based on panning information.

An approach for separation of speech mixtures with time-frequency masking is described in [Rickard and Yilmaz, 2004]. They demonstrate that through the use of *W-disjoint orthogonality* a perfect demixing of not overlapping sources is possible with binary time-frequency masks. His results demonstrate that an ideal binary time-frequency mask allow separate speech mixtures.

[Vinyes et al., 2006] proposes a method where is combined the use of time-frequency masks with stereophonic information of the audio tracks. This means using the panning of each channel of the stereophonic track. Thanks to this information they can extract inter-channel phase difference (IPD) and distinguish the different sources thanks to phase information. Moreover, this approach is implemented with a visual interface where the user can select dynamically the inter-channel magnitude ratio and IPD to adapt the separation.

Another example of on-line source separation approach related to time-frequency masking is presented by [Marxer et al., 2012]. They propose a method for real-time applications that unless not the same performance than in off-line approaches like NMF algorithms is achieved, they demonstrate how it is better than other existing real time algorithms with only a low-latency of 232 ms. Their method is designed for harmonic sources, specially for singing voice, although can be extended to other instruments.

2.1.3 Existing Frameworks

Several studies and approaches for source separation tend to include his implemented algorithms in his own websites, which is very useful in order to evaluate the performance

of them not only by reading statistics. But some of the authors have contributed with more than simple algorithm implementations. There exist some frameworks that are the most used and extended. This frameworks are more complete than the simple algorithms, in the sense that allow to configure the source separation according to the user preferred strategy and methods.

FASST, the Flexible Audio Source Separation Toolbox [Ozerov and Vincent, 2010] is the most extended and used framework. This framework is implemented in MATLAB and it can be downloaded for free. His main advantages are that is a general, flexible and modular framework that generalizes some of the most used source separation methods and it merges all of them in the same framework. To good point of that is that the user can design new sources separation strategies based on the delivered methods which can induce in more effective methods.

VIMM or VUIMM is another well known framework based on the work done by [Durrieu et al., 2011]. VIMM and VUIMM aim to Voiced and Unvoiced Instantaneous Mixture Model (IMM). This software is also free and is implemented in Python/NumPy. This framework is not that general than FASST, due it is based on the source/filter model described in [Durrieu et al., 2011], but is still very useful in order to separate instruments with solo and accompaniment.

Another existing framework is Low-Latency Instrument Separation in Polyphonic Audio Using Timbre Models. This is developed by [Marxer et al., 2012] and is still in development phase. This is done in collaboration between MTG-UPF and Yamaha Corp. As we seen in section 2.1.2 this framework is designed to work with singing voice mainly, but can be easily extended to work with other sounds.

2.2 Source Separation in Cochlear Implants

2.2.1 Hearing with CIs

Sound

Sound [Nogueira, 2008] is the perception of pressure waves by compressing the air molecules. The unit for sound pressure is the sound pressure level(SPL) in db:

$$SPL(dB) = 20 \log_{10} \frac{P}{P_0}, \tag{2.17}$$

where P is the absolute pressure in μbar and P_0=0.0002 μbar. This value is the reference 0 dB that corresponds to the threshold of hearing.

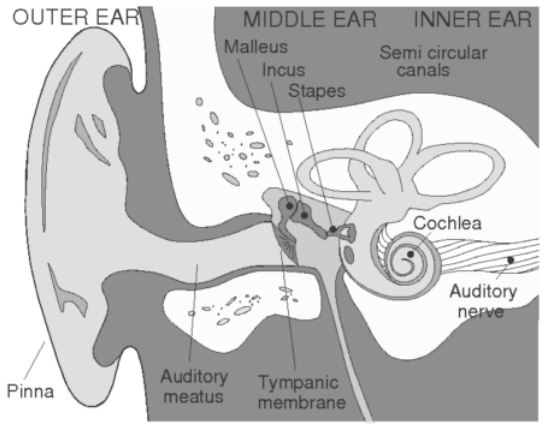

Figure 2.2: Outer, Middle and Inner Ear [Nogueira, 2008]

Anatomy and physiology of the auditory system

Sound is transformed into neural codes once pressure waves incises in the Ear. In figure 2.2 we can see a graphical representation of the outer, middle and inner ear. The part where sounds are coded to the brain is the cochlea, in the inner ear. The outer and middle ear with its components is the responsible of amplify pressure waves and modify the timbre of sound depending on the position of the source.

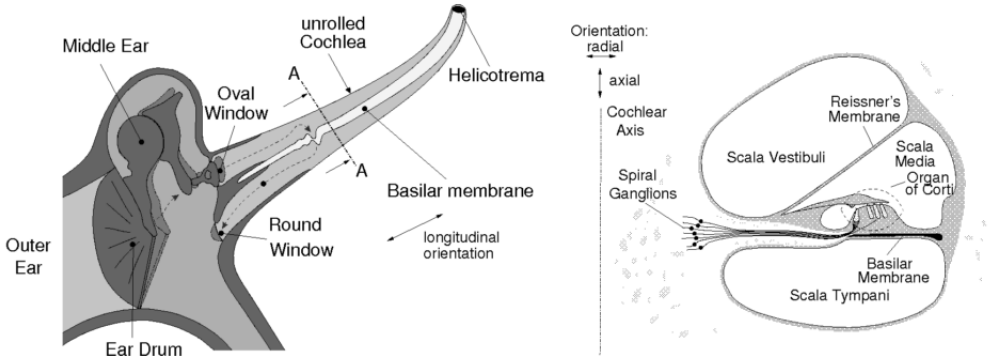

Figure 2.3: Middle and Inner Ear [Nogueira, 2008]

In figure 2.3 we can see both middle ear and the unrolled cochlea at the left side. In the other side we can distinguish also a representation of a transversal cut of the cochlea. Sound vibrations incoming from the middle ear travel along the cochlea fluids and is traduced in perpendicular pressure into the basilar membrane (BM). The frequency at which the BM is most sensitive to sound vibrations is the characteristic frequency (CF). The BM is narrow at the base, which makes it more sensitive to high frequency vibrations while it is three times wider at the apex where more low

frequency vibrations happen. The BM can be considered as a overlapping filter bank with a bandwidth equal to the critical bandwidth. This critical bandwidth is usually to be constant from 100Hz up to 500Hz.

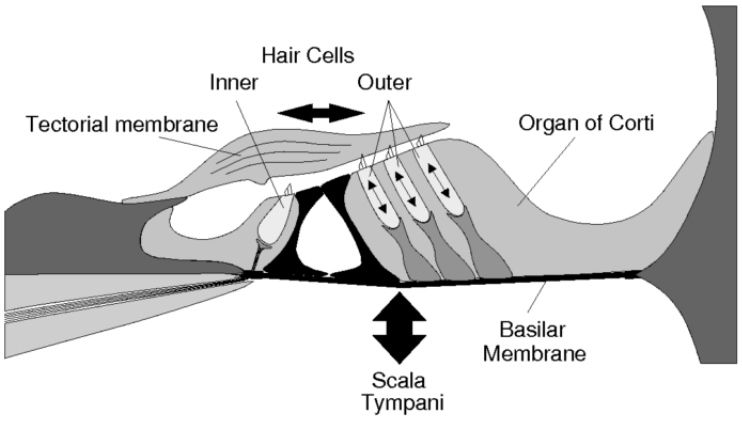

Figure 2.4: Organ of Corti [Nogueira, 2008]

The organ of Corti, in figure 2.4 is located on the top of the BM and comprises the sound receptors or hair cells and the tectorial membrane. There are two kind of hair cells, the Inner Hair Cells (IHC) and the Outer Hair Cells (OHC). IHC are the responsible of transmit the electric current to the auditory nerve. This is made by the movement of the tectorial membrane in contact with the IHC. OHCs are in charge of the mechanics of the organ of Corti by influencing the response of the BM.

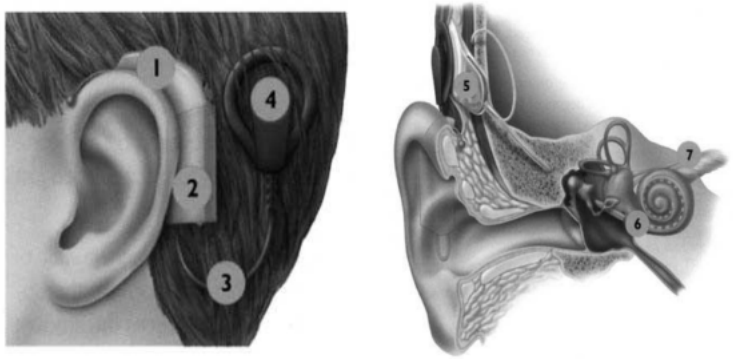

Figure 2.5: External and Internal part of a CI device [Nogueira, 2008]

Cochlear Implants

People suffering from hearing loss have a damage that hinders the sound path between the outer ear and the auditory nerve. The optimal case for a CI treatment is when this

damage is caused directly in the hair cells which stimulate electrically the auditory nerve. The main function of CI is to stimulate directly the auditory nerve with small currents and voltages in the order of milivolts.

All the currently implanted CIs are composed by two main parts, the internal and external part. The external part can be Behind-The-Ear (BTE) or a body-worn processor and both are connected to a transmitting coil through a cable. The internal part is surgically implanted in the patient head, and consists of a receiving coil a decode and stimulator, and the electrode array. The first carrier is composed by 16-22 electrode contacts and is inserted into the scala tympani of the cochlea. The second carrier is inserted beneath the skin and can be used as return electrode. In figure 2.5 we can observe a representation of an implanted CI. In the external part, 1 and 2 corresponds to the microphone and the sound processor respectively, 3 is the small cable that connects the transmitting coil represented as 4. In the other side, 5 is the receiving coil that is connected to 6, the electrodes array, and finally 7 represents the path to the brain.

Figure 2.6: CI device [Nogueira, 2008]

Almost all the existing CIs are manufactured by four companies: Clarion implants from Advanced Bionics (U.S.), the Nucleus implants by Cochlear Ltd.(Australia), the Pulsar implants from Med-El(Austria) and the Digisonic implants from Neurelec (France). In figure 2.6 we can see the Harmony BTE speech processor at the left side and the HiRes90k electrode array of Advanced Bionics.

Speech and music perception with CIs

Speech intelligibility in noise and reverberation can be predicted from how well envelope information from different frequency bands is preserved. For normal hearing listeners, envelope information for 3-4 frequency bands is enough for speech intelligibility and are able to understand sentences when noise is around 5 dB louder than speech. Implant

patients needs a higher Signal-to-Noise Ratio (SNR) to understand speech. There is a 10 dB of difference between implant patients and normal hearing listeners for speech intelligibility. A very important factor in the success of the CI is the age of the patients and whether the deafness is produced before(pre-lingual) or after(post-lingual) learning speech and language. This last are the patients that perform better with CI.

It has been demonstrated that CI patients can enjoy music. However, the enjoyment is poorer than before their haring loss. One of the main problem in music perception is the melody recognition. Normal hearing listeners can achieve up to 0.2 semitones while CI patients only achieve between 2 to 7 semitones. This bad perception of pitch is what induces to bad melody recognition. In the other hand, in terms of timber perception, CIs are very similar to normal hearing.

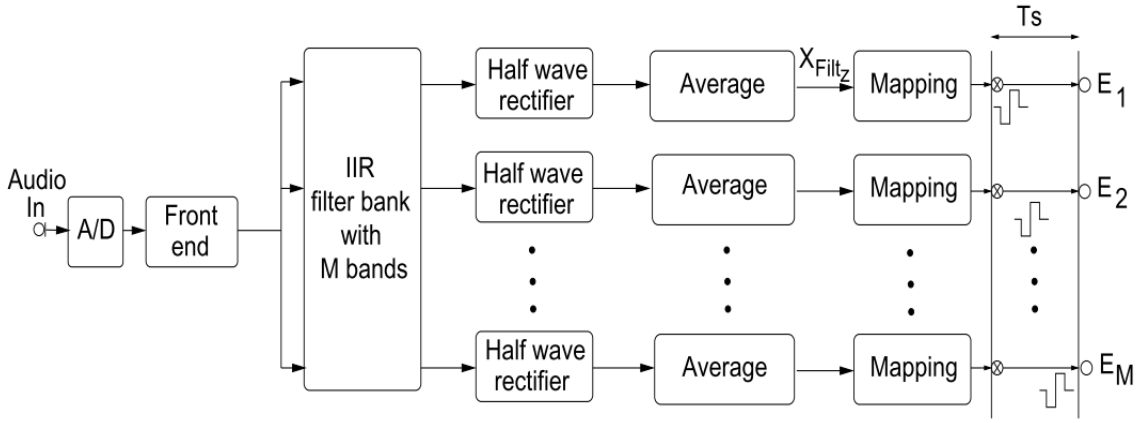

Figure 2.7: HiRes Block diagram [Nogueira, 2008]

2.2.2 Audio signal processing for CIs

Each CI manufacturer has its own algorithms and signal processing strategy adapted for his specific CI. We will be focused on the strategy followed by Clarion, Auria and Harmony devices by Advanced Bionics which is the High Resolution strategy (HiRes). The basic block diagram of the HiRes strategy is the presented in figure 2.7. In the first stage, the audio is sampled at 17400 Hz, pre-emphasized by the microphone and then digitalized. Then an Adaptive Gain Control (AGC) is applied digitally. After that the signal is divided in frequency bands using IIR Butterworth filters of order 6. Each of the frequency bands is associated with one electrode of the implant. Then each filtered signal is half-wave rectified (only positive amplitudes) and averaged for the duration T_s of the stimulation cycle. Finally the last block maps the acoustic signal for each

band. The output of each band will be in a dynamic range set by the clinician. In each cycle the HiRes stimulates the 16 electrodes sequentially to avoid channel interactions.

2.2.3 Speech recognition with BSS in CIs

As we have discussed in section 2.2.1 the speech intelligibility is as task that has still room for improvement. Independently of the many speech coding and signal processing techniques developed in the past years, BSS is a new approach in order to attempt for enhance speech recognition in CIs. The main objective of BSS in CIs is to recover and improve independently the desired source from a set of mixed signals.

The research done in BSS applied to CIs is not so much extensive yet but some studies have demonstrated that it is a very good strategy for improving speech recognition. In the other hand, the inclusion of BSS in speech recognition algorithms for Hearing Aids (HA) it has been considerably active in the recent years. HA are devices very much simpler than CI, his main operation is based on filtering and amplification of the input audio signal. But at the end, the objective of HA is the same, improve speech recognition and perception. The research done by [Reindl et al., 2010] is an example of the use of BSS for speech enhancement with HA. Here a combination of different source separation techniques is proposed, Directional BSS which estimates the interfering point sources and Weiner filtering which enhances the independent sources. [Han et al., 2009] proposed a method for post-processing of audio signals based on convolutive blind source separation in reverberant scenarios. The main separation strategy is achieved through Binary masks.

As we have said there is not a huge amount of literature about research in BSS and CIs. One of the most relevant studies, is the work done by [Kokkinakis and Loizou, 2008]. His approach is based on BSS applied to bilateral CIs. The core separation algorithm is based on ICA and he assumed a cascaded mixing and unmixing system with two sources and two sensors as can be shown in figure 2.8. Here the experiments were done with patients with bilateral implants and different configuration for the sources. A fixed source with the target speech sound situated at a fixed phase of $0\,^\circ$ and a noise masker with variable phase between $0\,^\circ$ and $90\,^\circ$. The experiments were ran in two different scenarios, first in an anechoic room and in then in two reverberant rooms with different reverberation time. The results obtained for the both scenarios are found to improve speech recognition with the use of BSS, specially for the non-reverberant case and low-reverberant time case.

Another example of the application of source separation in CI was proposed by [Suhail and Oweiss, 2006]. This approach proposes a not very common method as is Subband Decomposition. This method uses the Discrete Wavelet Packet Decomposi-

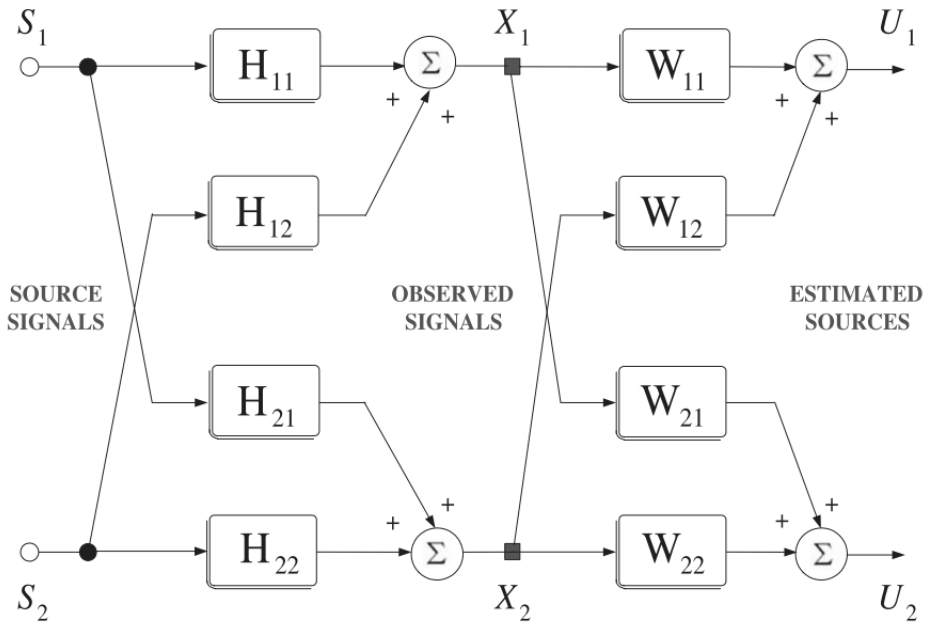

Figure 2.8: Cascaded mixing and unmixing MBD system configuration with two-source two-sensor scenario [Kokkinakis and Loizou, 2008]

tion (DWPT) to separate the desired signals. This study was tested with mixtures of multiple speech voices and the target speech signal was extracted as expected, but it was not tested with CI patients.

Chapter 3

METHODOLOGY

The methodology followed in this master thesis is relatively simple. Since the primary objective of the thesis is to enchance speech inteligibility by means of source separation, a decisive point is how the separation is made. As we commented, the development and implementation of the source separation algorithm is not the goal of this thesis but we will be more focused on the evaluation. This is the reason why the methodology is simple. After selecting the desired algorithm, most of the work will be on the evaluation and analysis of the results.

A very important aspect, like in all research, is to make an accurate literature review. Maybe in this case even more important due we will be using an existing framework, and we need to ensure that this will fit our requirements and needs. As we have seen in the state of the art section, we can divide source separation methods as online and offline. So this is a decisive aspect to have into account if we want to work with CI. The sound processors of the CI work by capturing the incoming signal directly from the microphones and after the corresponding processing, the information is translated in stimulation pulses in the cochlea. If we take an offline algorithm in order to pre-process the signal in a CI this maybe will be translated in a higher quality of the signal. But the stimulation will be produced after some seconds than the sound arrives the microphone or even minutes. The patient will be listening continuously with a considerable delay in the perceived sound. This is the reason why we encouraged to use an online method in order to separate the sources.

Although we will evaluate also an offline method to compare the performance, we will be focused in the use of an online method. Specifically we used the Low-Latency Instrument Separation in Polyphonic Audio Using Timbre Models (LLIS) by [Marxer et al., 2012]. This algorithm is designed to separate music signals and more specifically to remove the lead singing voice of the audio track. These are not a priory requirements for improve speech intelligibility, and this is why we have modified some

aspects of the algorithm.

3.1 Baseline Algorithm: LLIS

Low-Latency Instrument Separation in Polyphonic Audio Using Timbre Models represents the starting point of the research done in this Master Thesis. As we have commented in section 2.1.2 some of the available online source separation solutions are designed used Time Frequency Masking. LLIS is one example of using Time Frequency Masking, using the binary masks is achieved a good quality level while maintaining the low-latency.

A useful method to decompose sources with Time Frequency Masking is by using panning information and IPD (inter-channel phase difference)[Vinyes et al., 2006]. Using the pan and the frequency parameters we can create the desired frequency mask. But in some cases is not enough or effective due reverberations or monophonic recordings that lack of pan information. This method is incompatible for our case because we need to work with monophonic recordings to be compatible with the monaural CI of our experiments.

3.1.1 Harmonic Mask

The base of this algorithm is the creation of harmonic masks. The creation of this masks starts with the assumption that the vocal component is localized around the partials of the singing voice so the optimal mask to remove the voice, consists on zeros around the partials and ones elsewhere.

But many parts of the vocal component such as consonants, fricatives or breath that are very important for speech intelligibility, normally are not harmonic components and thus will not be neither detected nor separated using this algorithm. For the voiced components, the harmonic mask is created considering the whole components of the vocal part. To do that the $f0_i$ of the source must be estimated. This is made in three steps: pitch likelihood estimation, timbre classification and pitch tracking.

Pitch Likelihood Estimation

This step is similar to NMF in the sense that a linear signal decomposition is made in order to do the pitch likelihood estimation. As in other linear decompositions, here is assumed that the spectrum at each frame is a linear combination of the elementary spectra or the so-called basis components. One of the peculiarities of this method is that the Tikhonov regularization is used to estimate the components. This method

has the advantages in front of NMF of the low-latency and his simple implementation but in contrast the gains can take negative values. To have an effective likelihood, all the negative values are set to 0.

Timbre Classification

In order to estimate the pitch of the present instruments in the mixture it is needed to select the right values from the pitch likelihood estimation algorithm corresponding to each instrument. With the pitch candidates it is created a group of features vectors which is classified using Support Vector Machines (SVM). The envelopes are calculated with an interpolation on the magnitude of the spectrum at the harmonic frequency bins. The feature vectors are a variant of the Mel-Frequency Cepstrum Coefficients (MFCC). In the figure 3.1 we can see the spectrum (black line) and the pitch candidates as the different envelopes (colored dashed lines).

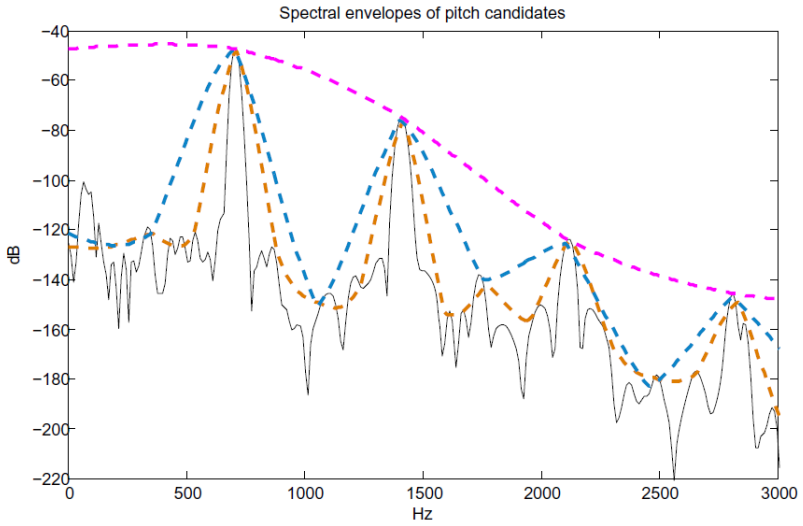

Figure 3.1: Spectrum magnitude (solid black line) and the harmonic spectral envelopes (colored dashed lines) of three pitch candidates [Marxer et al., 2012]

Figure 3.2 shows a diagram of how the voice model is created based on the pitch estimations and the annotations. Based on the pitch information, the envelopes are created and the timbre features are extracted. Finally, the classification is made comparing with a test dataset and the model is generated.

Instrument Pitch Tracking

This step is a dynamic programming algorithm that is composed of two processes. First a Viterbi allows determine the optimal pitch track and the second determines

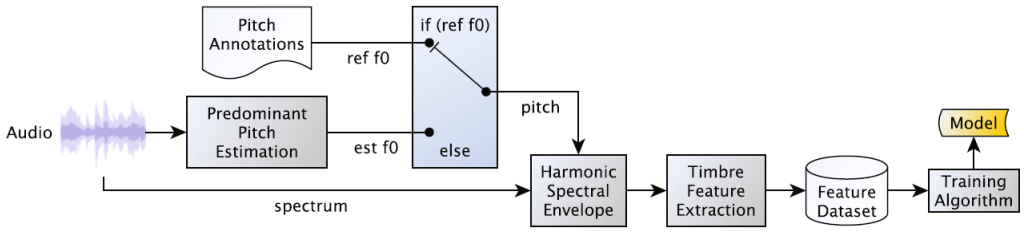

Figure 3.2: Workflow of the supervised training method[Marxer et al., 2012]

the voiced and unvoiced frames. This step has a latency of 20 frames (232 ms) what makes possible to use it online.

3.2 LLIS Requeriments and modifications

As explained, LLIS is an existing algorithm designed to deal with musical signals and one of the main objectives is to remove the main singing voice. This is not the goal of this research, but since singing voice signals are very similar to a normal speech signal, we decided to use this algorithm and try to adapt as maximum as possible to detect speech signals.

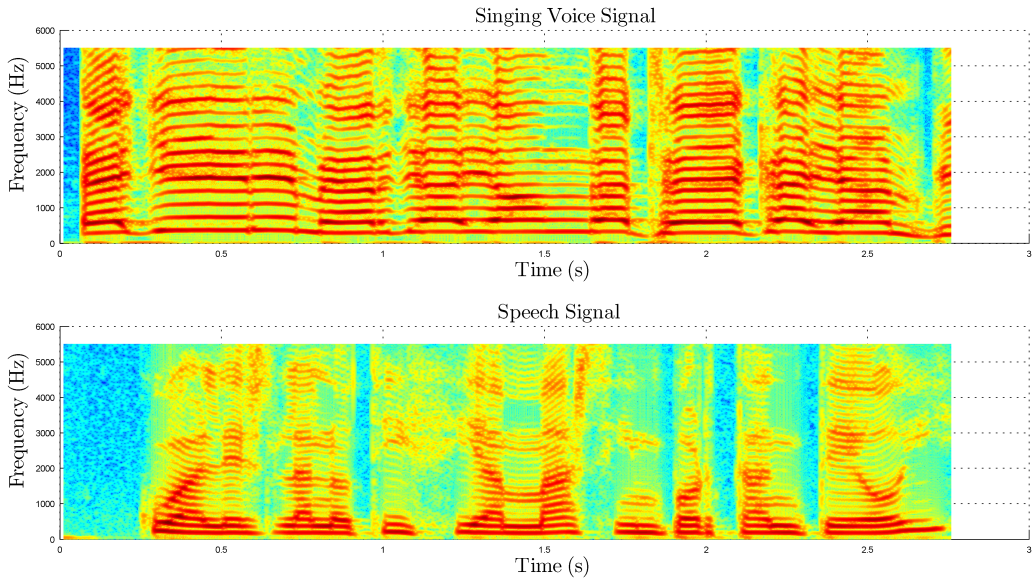

Figure 3.3: Singing voice and Speech spectrograms

In figure 3.3 we have the spectrograms of a **Singing Voice signal** and a **Speech signal**. In order to see the differences we have chosen two random signals, different

languages and male and female. We can distinguish easily the two spectrograms, one considerable difference is that the singing voice is more continuous in time while the speech signal have the words, even syllables well separated with short silences. Another important difference is that in the singing voice most of the partials have similar energy than the fundamental while in the speech signal only the first partials have considerable energy. These are simple examples of the differences in both signals, but still the behavior is similar so we can consider that if the system can recognize the singing voice it will be capable of detect the speech signal.

The first requirement is to restrict the system to only monophonic recordings due the aforementioned problem that all the CI patients have only one implant. Then an important but simple modification is to remove the components except the speech. As we have seen in the algorithm description, the algorithm is designed to remove the lead singing voice. Then to get only the speech part we need only to invert the harmonic mask. The algorithm does not allow a big number of modifications, most of them are frame size, hop size, window size or type and none of them affects directly speech intelligibility or to improve speech separation. The only parameters that seemed to improve speech separation in terms of Signal to Distortion Ration (SDR) is the limits of the Viterbi algorithm, which have been improved (about 0,3-0,5 dB) by reducing the low boundary. One of the important aspects of the configuration of the system is the selected voice model. The voice model is what determines how the pitch candidates of the envelopes are selected. So the used models are trained for singing voice, what means that the pitch tracking will perform better for singing voice. The optimal case will be a model totally trained with speech signals. This should be one of the first steps in the future work out of the scope of this thesis.

It is worth to comment that LLIS is not designed to detect unvoiced consonants such as stop, affricate, fricative consonants. This mean that these kind of consonants like “s” or “p” the type of consonant will never strictly detected. In some cases some content of these consonants will remain in the estimated speech due it can be mixed with the harmonic component. In order to solve this missing capability, is being developed a new strategy complementary to LLIS with the goal of detect unvoiced consonants. This new strategy is originally designed to detect transient signals in music like percussive instruments. But this can be adapted to detect some unvoiced consonants. Specially consonants from the beginning of the word or syllable which in many cases act as transient signals.

Chapter 4

EVALUATION

Typically, the evaluation of SS algorithms is a simple task, computing a number of objective measures (SDR,..) given a the original separated sources. But when we need to evaluate speech intelligibility it becomes harder. None of the objective methods used to evaluate the algorithm is strictly reliable in order to detect an speech intelligibility enhance. Moreover if we add the difficulty that we need to test our system with CI patients, it becomes even more unpredictable.

A goal of this research is also to determine the performance of the used source separation system in order to separate speech signals. Taking this assumption, most of the objective measures are good indicatives. Then we decided that the best way to evaluate speech intelligibility is to realize tests directly with the patients and volunteers and trust in this results.

Another important aspect in the evaluation of the system is the selected dataset. Previously, LLIS have been tested mostly with music samples which with the multitrack recordings we can evaluate how well each of the individual sources have been separated. In our case we are more interested in separate speech signals from different kind of noises. For this reason we have created our own dataset from different speech and commonly used noise audio signals. Furthermore, we have had to use the appropriate speech and noise samples to run the experiments with the CI patients.

4.1 Dataset

As mentioned, normally when evaluating source separation algorithms, multitrack recordings are used in order to have a reference of the real sources of the mixtures. In our case we do not need strictly to use music samples. Concretely we used speech samples mixed with different noise samples. As the main goal is to separate the speech from the rest of sources, we assumed that our mixtures are composed by only two

components: speech plus noise, where the noise could be more than one source itself. So first we need to define which kind of speech and noise samples we are going to use.

4.1.1 Noise

We can see this system more than a source separation, as a speech enhancement or noise reduction tool since the key point is to isolate the speech signal from the rest of components and try to keep it unchanged. In order to evaluate real or similar situations that CI patients can suffer we decided to try different kind of noises. We have divided the noise samples in three cathegories accordint to their structure or behavior:

- **Stationary Noise**
  The main property of stationary noises is that the probability distribution does not changes along time. This kind of noise is often used in order to evaluate noise reduction algorithms and is with this kind of noise when normally these algorithms perform better. Although this kind of noise is not very "real" in the sense that we normally do not perceive sounds with such invariability among time. So normally these are artificially generated sounds.

- **Non-stationary Noise**
  In contrast to the stationary, the non-stationary noise as expected changes the probability function along time. These noises are very commonly found in real life like could be any kind of background noise such a fabric, traffic or a cafeteria situation. It is very interesting to do the evaluation using this kind of noise due normally noise reduction algorithm does not work at all since these are very unpredictable and are an heterogeneous mix of sounds which his spectral content could be very different or very similar.

- **Music Noise**
  We decided to include also a simple music sample because of the nature of LLIS algorithm and in some cases can be perceived as a noise that is interfering with the speech. As this is originally designed to separate music signals like instruments and singing voice it is also a good opportunity to see how well speech can be separated from music sounds.

Noise Selection

- **CCITT**
  CCIT is the chosen noise for the stationary type. CCITT is a well known noise

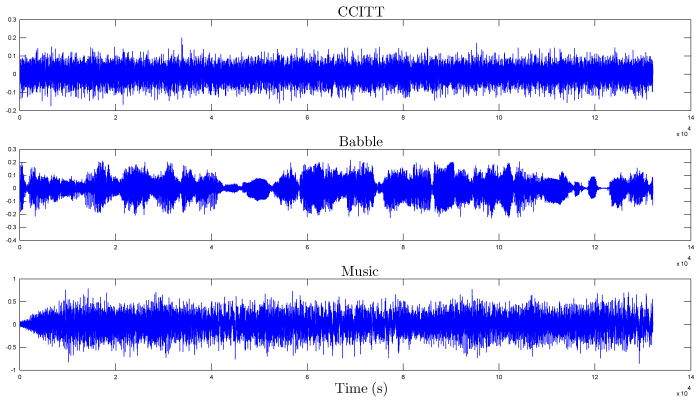

Figure 4.1: Three example waveforms of the different selected noises

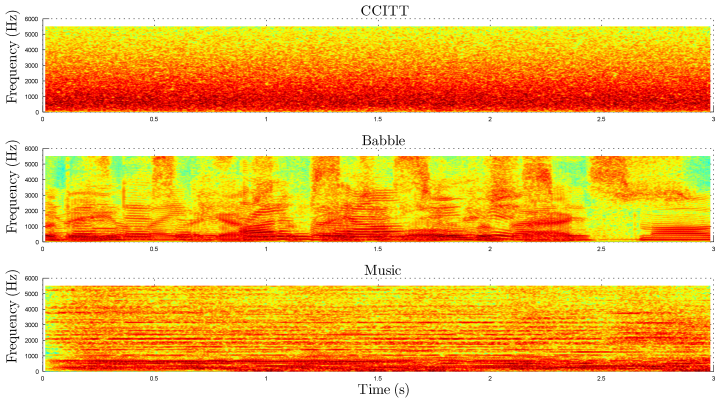

Figure 4.2: Three example spectrograms of the different selected noises

and commonly used in tests for speech enhancement that was created by the [International Telecommunication Union,]. As we can see in figure 4.1 the amplitude of the waveform varies between about 0.1 and -0.1 but the envelope is very stable compared to the other two. About the spectrogram in figure 4.2 we can see that the energy distribution is equally distributed along time with a huge precence in frequencies from 0 up to 4-5 kHz, specially up to 1 kHz.

- **Auditech 4T Babble**
  As we commented we can found non-stationary noise in a lot of quotidian situations. This is why we selected a babble noise composed by 4 talkers. We thought about this noise as a very common situation like could be a cafeteria where different people talks at the same time. This is very difficult noise to deal with because of the problems that we commented about the non-stationary noises. Moreover if we consider that we will be adding more voices to the speech sig-

nal this complicates even more the task of differentiate and separate the desired speech signal. Mainly because the noise itself it is a mixture of different speech signals that can easily be confused with the target signal. In figure 4.1 we can see clearly that this is the one with more changes in time. Furthermore figure 4.2 reveals how the different speech signals are mixed together most of the time very close in frequency.

- **Mazoni**
  The selected music sample is an excerpt of a pop rock song. We did not have many requisites for the music sample selection. Principally we chose this song because we wanted the presence of typical instruments like guitar, piano, drums, bass-guitar and singing voice. The idea was to simulate a real situation similar to a concert or a place where the music is considerably higher. Tho reinforce this situation we added a slightly reverberation effect in order to simulate the acoustics of a standard hall. This signal can be considerered as non-stationary due different instruments are playing at same time changing arbitrarily during this time. In figure 4.2 the spectrogram of the music sample reveals that the signal is more constant than expected. This could obviously change with each music sample that we choose but in this case each instrument is very constant in time and this effect is reinforced thanks to the reverberation effect.

4.1.2 Speech

The set of speech samples selected is composed by two collections of words and sentences. Both of this collections have been elaborated and recorded by spanish linguists in order to have an standarized speech dataset. The only requisite that we had in order to select the speech dataset is that this had to be in spanish basically because the experiments with CI patients have to be in spanish due it is the native language of all of them.

As we can see in figure 4.3 although both signals are similar, there are some differences. Mainly the word has a higher spectral content what means that each fundamental frequency has more partials and the non pitched components such us some consonants or other sounds produced by the speech are more defined. In the sentence spectrogram we can see the silence spaces between words but moreover, there are some frequency regions between the words that are not very energy dense. This behavior gives the word a higher intelligibility level due each phoneme is more defined and makes it highly noticeable.

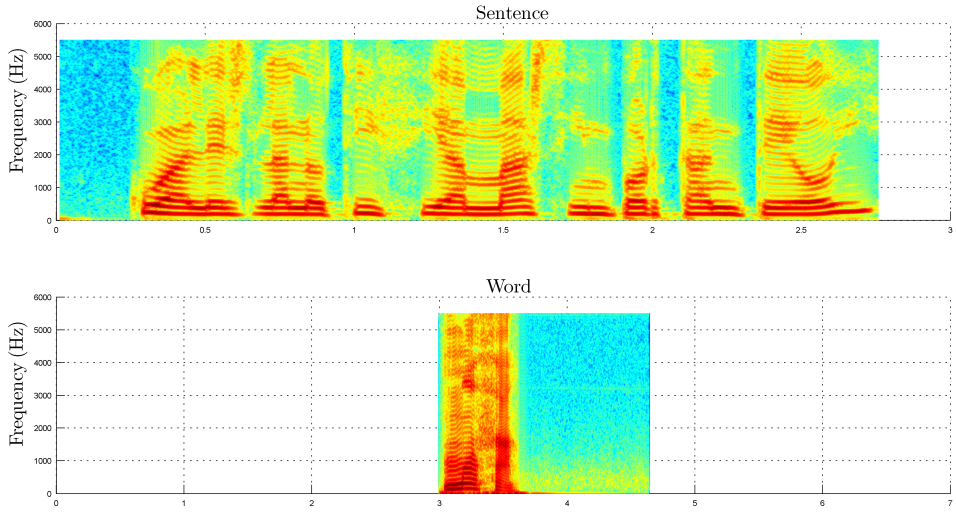

Figure 4.3: Sentence and word spectrograms

4.1.3 Calibration and Mixing

The final database that he used to evaluate LLIS is basically a set of mixed speech plus noise signals. The database is composed different lists of words and sentences and each one has his own different SNR levels with each one of the used noise samples. The procedure followed for mixing in each case is described in figure 4.4. We can distinguish basically two main steps, first the calibration and second the mixing step. In order to compute the SNR for each signal we decided to use the Root Mean Square (RMS) measure. The RMS is basically a measurement of the magnitude of a set of values in continuous change over time. It is very useful in electronic signals like audio which are constantly changing their positive and negative values. Usually RMS is calculated as is described in equation 4.1. But in order to have more accurate values on the RMS we decide to average the squared signal by low pass filtering. This method is normally used to remove the ripple frequency noise. So in equation 4.2 $H(e^{j\omega})$ represents a low pass filter, and $\mu[n]$ is the average squared signal. Finally $x_{rms}[n]$ represent the magnitude of the values.

$$x_{rms}[n] = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x^2[n]} \tag{4.1}$$

$$\mu[n] = x^2[n]H(e^{j\omega}) \tag{4.2}$$

$$x_{rms}[n] = 10\log(\mu[n]) \tag{4.3}$$

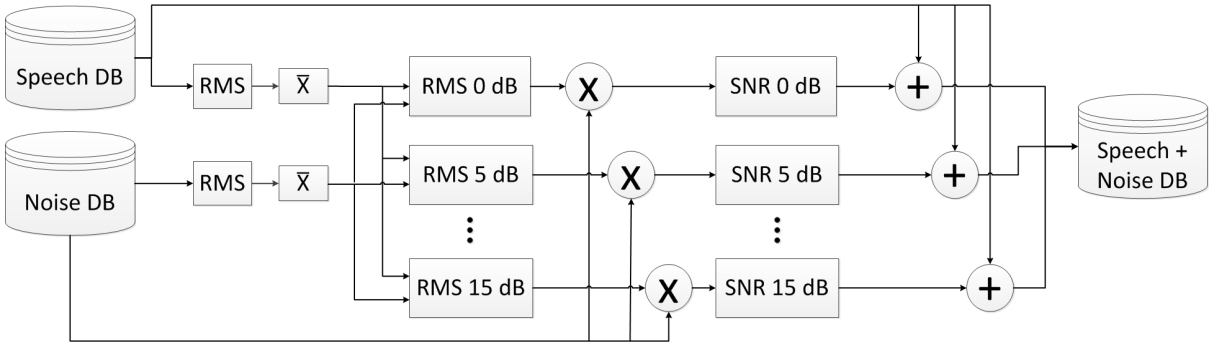

Figure 4.4: Speech + Noise Database creation through the calibration and mixing process

After computing the RMS value of both speech and noise signals, we calibrate the noise signal by adjusting the desired dB level. We simply compute the mean of the RMS values and we adjust 0, 5, 10 or the desired dB level in each case. We think that this is a good method to compute the SNR because it depends on each case how the word or the sentences are composed. For example in figure 4.5 we have an example of a word mixed with noise at 4 different levels. If we observe to the first plot at 0 dB SNR, in the middle of the word the magnitude of the noise is higher than the noise but at the beginning is a little bit higher the word. In this example is shown how the average value of the RMS is appropriate to compute the SNR.
Finally the mixing process it is a very simple procedure where each speech signal is joint with the corresponding speech signals at each SNR level. Just some other issues are considered like adjust the duration of the noise to the same duration of the speech or re-sample all the samples to 44100.

4.2 Objective Evaluation

For the objective evaluation we selected two type of measures. First the measures in charge of evaluate the performance of the separation i.e. how well the sources are separated or basically how well the speech is separated from the rest due we only consider a mixture of two sources. Then we use other measures in order to evaluate the intelligibility or speech degradation caused by the separation where some speech components are lost and some noise artifacts are added.

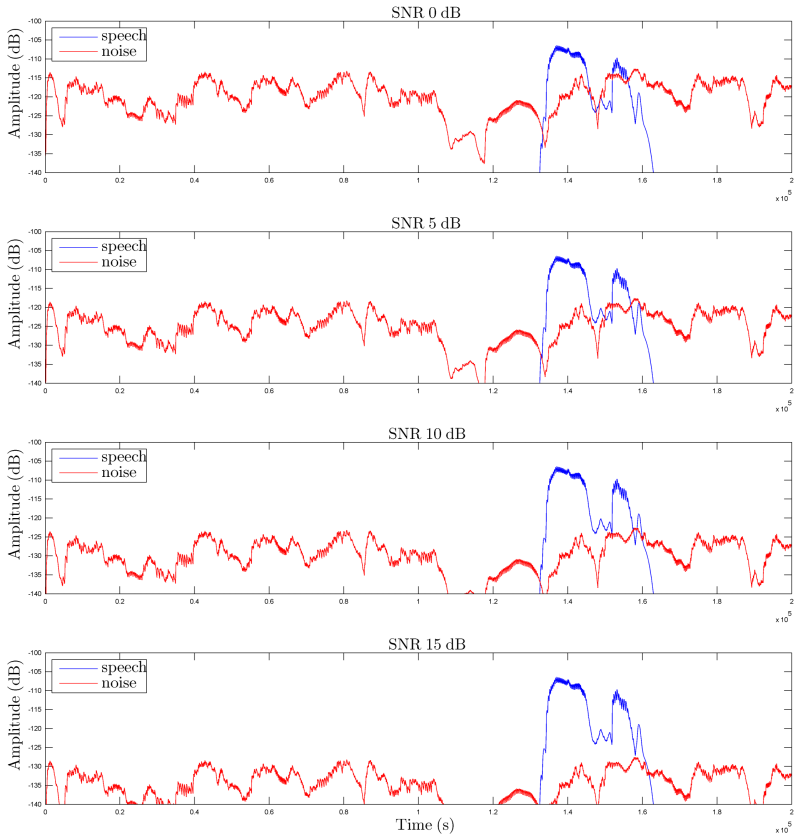

Figure 4.5: RMS values from both noise and speech signals at **0**, **5**, **10** and **15** dB SNR levels

In figure 4.6 is depicted a representation of the objective evaluation procedure. As we can see is very simple, after the speech and noise samples are mixed, each mixed sample is passed to the algorithms in order to estimate the sources. Then with the estimates and the original sources is evaluated the performance with a given set of measures. We have to comment also that we used another algorithm a part from LLIS to estimate the sources called IMM developed by [Durrieu et al., 2011] which is based in NMF. We used that algorithm in order to compare the performance between online and offline methods.

The second kind of measures used in the evaluation are those more focused on a perceptive point of view. These measures are centered in the evaluation of speech intelligibility in contrast to most of the objective measures that are focused mainly in the separation performance.

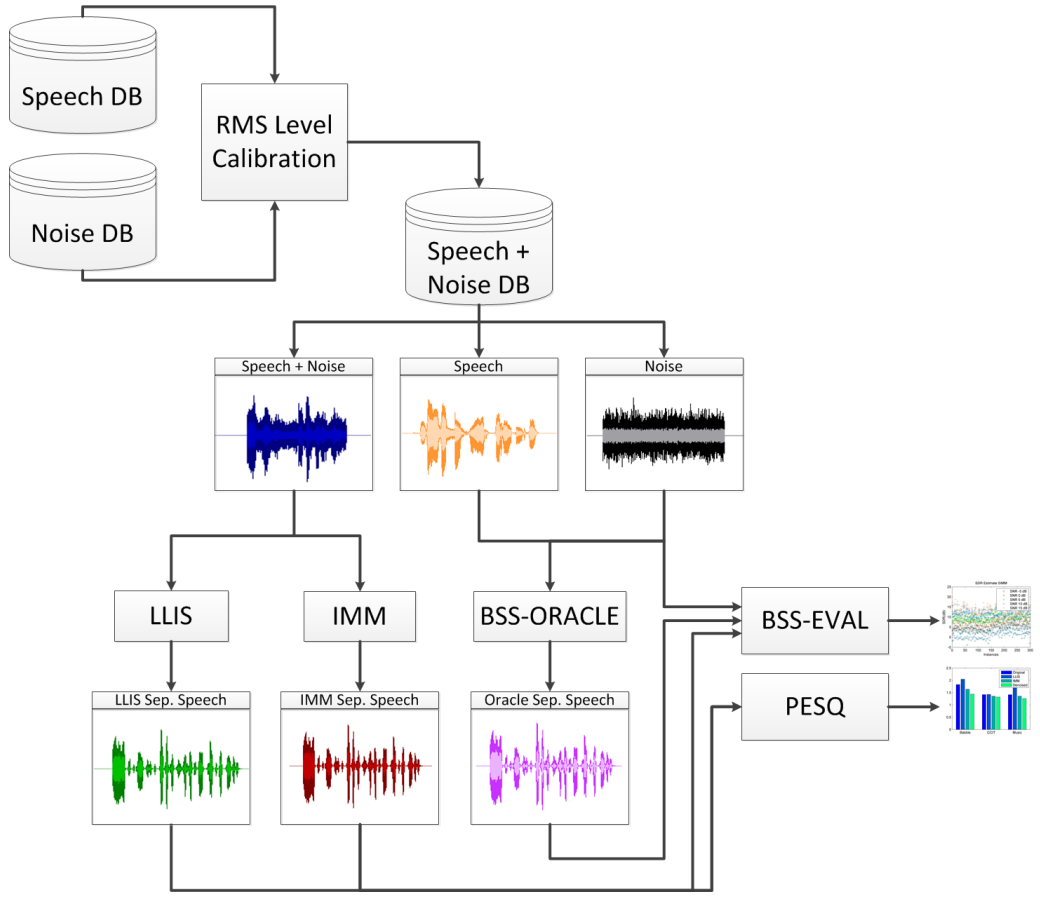

Figure 4.6: Objective Evaluation Diagram

4.2.1 Evaluation Measures

BSS Measures

Signal Separation Evaluation Campaign (SISEC)[1] is contest oriented for scientific evaluation where speech and music datasets are evaluated and standarized measures are used. Within the context of SISEC was created BSS Eval[2] [Févotte et al., 2005] that is a MATLAB toolbox specially implemented for evaluate the performance of source separation. Once the estimate singnals $\hat{s}_{ij}^{img}(t)$ are obtained, these must be compared with the source images $s_{ij}^{img}(t)$. The criteria[Vincent et al., 2007b] to evaluate this comparison is by express the estimate $\hat{s}_{ij}^{img}(t)$ as:

$$\hat{s}_{ij}^{img}(t) = s_{ij}^{img}(t) + e_{ij}^{spat}(t) + e_{ij}^{interf}(t) + e_{ij}^{artif}(t) \tag{4.4}$$

[1] http://sisec.wiki.irisa.fr/tiki-index.php
[2] http://bass-db.gforge.inria.fr/bss_eval/

where $s_{ij}^{img}(t)$ is the true image and $e_{ij}^{spat}(t)$, $e_{ij}^{interf}(t)$ and $e_{ij}^{artif}(t)$ are error components representing spatial distortion, Interference and artifacts. The relative amounts of this error components are represented by the Source Image to Spatial Distortion Ratio (ISR_j), the Source to Interference Ratio (SIR_j) and the Sources to Artifacts Ratio (SAR_j), defined by

$$ISR_j = 10 \cdot \log_{10} \frac{\sum_i^I \sum_t s_{ij}^{img}(t)^2}{\sum_i^I \sum_t e_{ij}^{spat}(t)^2} \tag{4.5}$$

$$SIR_j = 10 \cdot \log_{10} \frac{\sum_i^I \sum_t (s_{ij}^{img}(t) + e_{ij}^{spat}(t))^2}{\sum_i^I \sum_t e_{ij}^{interf}(t)^2} \tag{4.6}$$

$$SAR_j = 10 \cdot \log_{10} \frac{\sum_i^I \sum_t (s_{ij}^{img}(t) + e_{ij}^{spat}(t) + e_{ij}^{interf}(t))^2}{\sum_i^I \sum_t e_{ij}^{artif}(t)^2} \tag{4.7}$$

Finally a combination of the previous measures, the Signal to Distortion Ratio (SDR_j)

$$SDR_j = 10 \cdot \log_{10} \frac{\sum_i^I \sum_t s_{ij}^{img}(t)^2}{\sum_i^I \sum_t (e_{ij}^{spat}(t) + e_{ij}^{interf}(t) + e_{ij}^{artif}(t))^2} \tag{4.8}$$

Sometimes comparing different evaluation measures from the mentioned above can be complicated. This could happen when we compare two different samples which their in dB differ substantially. To solve this, BSS Oracle[1][Vincent et al., 2007a] was created. BSS Oracle, like BSS Eval, is a MATLAB Toolbox which has a set of algorithms dealing with the mentioned problem. Principally, the objective is to define oracle estimators which compute the best performance achievable by the separation algorithms. With this estimators we can have an accurate version of the BSS measures. Instead of comparing the performance of the BSS measures, we can evaluate the performance with the error, or the difference between the estimates and the oracle estimates. If

[1] http://bass-db.gforge.inria.fr/bss_oracle/

$SDROracle_j$ is the maximum SDR achievable level and SDR_j is the estimated SDR, the error is defined as

$$SDRErr_j = SDROracle_j - SDR_j \tag{4.9}$$

PESQ

Perceptual Evaluation of Speech Quality (PESQ) is a measure to evaluate speech enhancement defined by [a.W. Rix et al.,] and also used by [Hu and Loizou, 2008]. PESQ measure is recommended by ITU-T for speech quality assessment of 3.2 kHz(narrow-band) handset telephony and narrow-band speech codecs. PESQ aims to compare a reference signal with the same signal with a quality degradation. To do that, first each signal is aligned to a standard listening level. Then, they are filtered using the FFT. After that, the signals are aligned temporally and processed into an Auditory transform. Finally thanks a disturbance processing, errors are extracted and aggregated in frequency and time to be evaluated with a subjective mean score (MOS). Some measures taken into account in order to do the transformations and quantize errors are the Bark Spectrum, Frequency Equalization, Gain Variation or Loudness Mapping.

4.3 Subjective Evaluation

In order to have a subjective point of view we have evaluated the system with different tests which principally consist in listening and evaluate the speech intelligibility of the samples produced by the LLIS system.

4.3.1 Word Identification Test

The used method for the subjective evaluation is a word identification test. In this test, the user is asked to listen to different speech plus noise samples and answer the understood words. The test is made by combining different SNR levels and several algorithms which produce different intelligibility levels depending on the case. Each test is evaluated by simply counting the percent of correct words of each list.This method is very simple, but is a very effective way to detect when and when not the user identifies correct words and sentences.

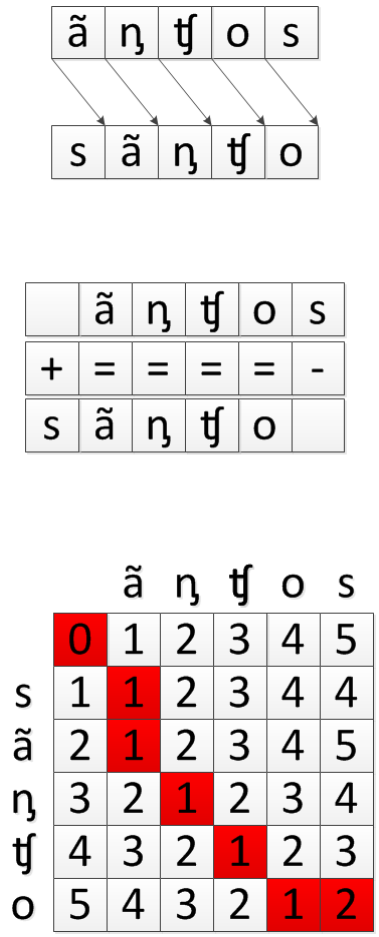

Figure 4.7: Levenshtein Distance operation examples. + = Insertion and - = Deletion

4.3.2 Phoneme Error Rate

Parallel to the word identification test, each test is recorded and transcribed with the goal of having all the answered words transcribed to be able to evaluate them. Thanks to that we can then observe and have an idea of wich words are more typically failed. Specially we found very interesting to go deeper and make the evaluation phoneme errors wich are the responisbles of the intelligibility. To do this evaluation we need to calculate the phoneme errors between the original and the answered word. We selected the Levenshtein Distance[1][Apostolico and Galil, 1997] also called Edit Distance to determine this error. The Levenshtein distance is a metric tha measures the similarity of two strings. We used the simple form where the Levenshtein Distance, is the minimum number of operations necessary to modify a srting to another. The possible operations are substitution, deletion or insertion. This is normally implemented with a dynamic

[1] http://www.levenshtein.net/index.html

programming algorithm wich checks both strings character by character.

In figure 4.7 there is represented the Levenshtein edit operations. The example takes the original and observed strings “anchos” and “sancho” in their phonetic transcription. The advantage of using this algorithm in front of simply align the strings and compare character by character is that matches the best similarity between the strings. In the example if each character is simply aligned, each character will be confused by substitution errors what it is erroneous. As can be seen, the algorithm computes the minimum value from the edit operations substitution, insertion or deletion. The matrix with both strings contain the values of the distances for each possibility. Each jump horizontally, vertically or diagonally represents a cost of 1 and if the characters match then the cost is 0. Following the path from the upper left corner to the lower right we can find the Levenshtein distance. In this case we have only 1 deletion and 1 substitution, so the Levenshtein distance is 2.

Finally we compute the Phoneme Error Rate (PER) as follows

$$PER = \frac{LevDist(S_{original}, S_{observed})}{length(S_{original})} \times 100\% \tag{4.10}$$

Chapter 5

RESULTS

After a clear evaluation structure definition, a big set of date is obtained from the simulations and tests that must be accurately analyzed. The procedure is as follows, first each list of words and sentences from the database is processed with LLIS to obtain the audio estimates of the isolated speech. With these samples we can compute the evaluation measures for the objective evaluation. After the objective evaluation, the estimated speech is used in the tests with CI patients to compare the speech intelligibility obtained in front of other algorithms. Some audio examples obtained with the source separation algorithms can be found in tables 5.1,5.2 and 5.3.

Babble Noise			
Algorithm	**SNR 0 dB**	**SNR 5 dB**	**SNR 10 dB**
Original Mix	Babble SNR 0	Babble SNR 5	Babble SNR 10
IMM	Babble IMM SNR 0	Babble IMM SNR 5	Babble IMM SNR 10
LLIS	Babble LLIS SNR 0	Babble LLIS SNR 5	Babble LLIS SNR 10

Table 5.1: Audio Examples of Babble noise with the different condition

Music Noise			
Algorithm	**SNR 0 dB**	**SNR 5 dB**	**SNR 10 dB**
Original Mix	Music SNR 0	Music SNR 5	Music SNR 10
IMM	Music IMM SNR 0	Music IMM SNR 5	Music IMM SNR 10
LLIS	Music LLIS SNR 0	Music LLIS SNR 5	Music LLIS SNR 10

Table 5.2: Audio Examples of Music noise with the different condition

CCITT Noise			
Algorithm	**SNR 0 dB**	**SNR 5 dB**	**SNR 10 dB**
Original Mix	CCITT SNR 0	CCITT SNR 5	CCITT SNR 10
IMM	CCITT IMM SNR 0	CCITT IMM SNR 5	CCITT IMM SNR 10
LLIS	CCITT LLIS SNR 0	CCITT LLIS SNR 5	CCITT LLIS SNR 10

Table 5.3: Audio Examples of CCITT noise with the different condition

5.1 Objective Evaluation

We have detailed in chapter 4 the differences of how we have evaluated the results obtained. The objective evaluation is all based in the simulations done with MATLAB using the target algorithms.

5.1.1 BSS-EVAL Results

BSS-EVAL Toolbox provides a set of measures which all of them are used to evaluate the performance of the obtained separation from a given mixture of sources. As we have seen in section 4.2.1, ISR, SIR, SAR and SDR are these main evaluation measures. For the analysis of the results we found more interesting to focus on SDR since this is a combination of the rest of measures. Figure 5.1 depict the behavior of SDR for a set of 20 instances of speech estimates. These instances are composed by estimates obtained through mixtures of different noises. In this case the used noises are the music, babble and CCITT noise. As can be seen, each SNR level is represented by a different color. The plot of SDR Oracle shows the ideal SDR level which is the maximum that can be achieved by separating the mixtures. In this case, it is very clear the curve that follows each instance at the same level which are clearly separated increasing the SDR while increasing the SNR level. In the case of the estimates for IMM and LLIS algorithms, we can still appreciate the differences between SNR levels but now some values are mixed.

In order to have a clear idea of which instances perform better or wrong in contrast of the rest, figures 5.2 and 5.3 shows the obtained SDR Error for the cases IMM and LLIS. Contrary as expected the higher error values are obtained at 15 dB SNR. We explain this effect because of the oracle levels obtained at this SNR are very high and the algorithms can not reach such level by estimating the sources. Then we can distinguish between the effect of the noise and the SNR to see which one has a bigger impact in the SDR. In figure 5.4 we an see the different average SDR values obtained by both algorithms at 5 different SNR levels. As commented before, at SNR = 15 dB

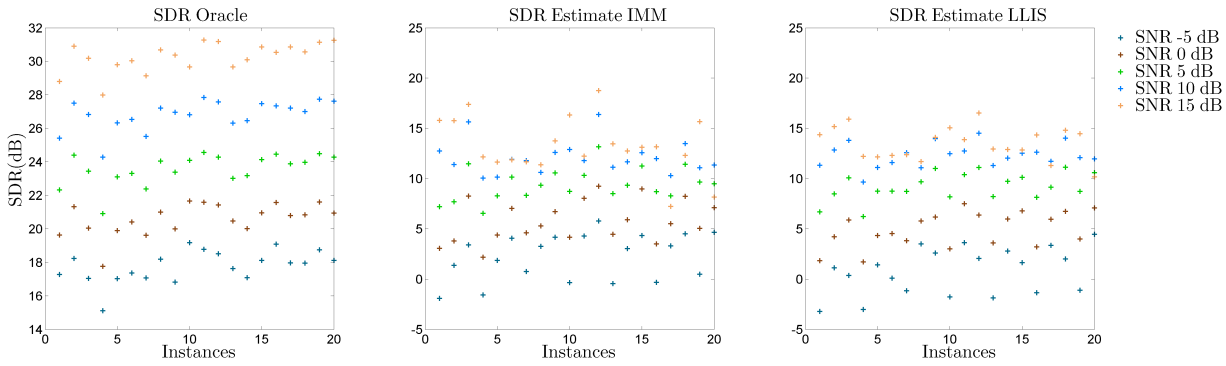

Figure 5.1: Speech SDR values for Oracle, IMM and LLIS estimates

is reached the maximum SDR error level for both cases IMM and LLIS. At levels 0 and 5 dB seems to be the best results obtained below 15 dB. There is not a significance difference in the results obtained by IMM comparing with LLIS($\approx$ 1 - 1.5 dB), but IMM seems to work better at 10 and 15 dB while LLIS at -5 and 0 dB. Regarding the results by noise, figure 5.5 shows the average values for a fixed SNR level at 10 dB. Again, both algorithms show similar SDR error levels where the music noise is the one with more difference where IMM obtains slightly better results.

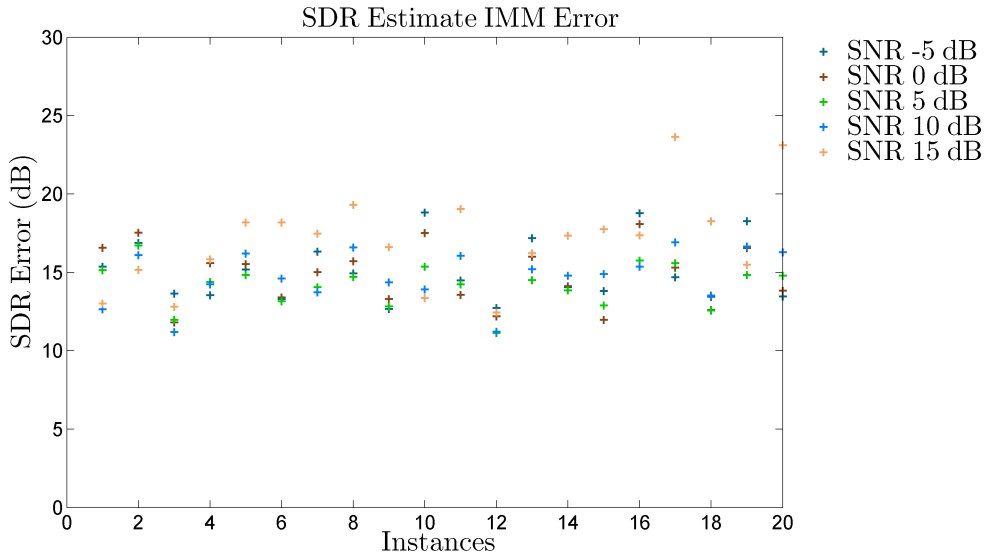

Figure 5.2: Speech SDR Error values for IMM estimate

Comparing these results to other cases like in [Marxer et al., 2012] where the difference on average SDR error obtained for IMM and LLIS is between 5-10 dB, it is clear that the algorithms do not work with the same accuracy. But this could be produced by different factors, like the data used in the evaluation that in their case are music

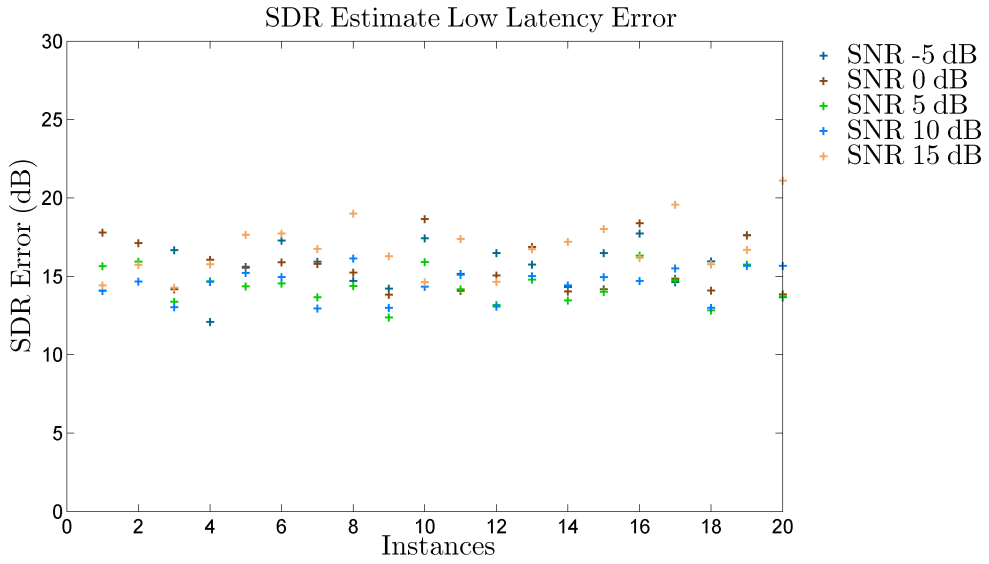

Figure 5.3: Speech SDR Error values for LLIS estimate

samples which have a different SNR between the lead voice and the rest of components than in our mixtures. Another example could be the effect of the training, which in our case the system trained with singing voice could be not detecting the speech signal with the same precision.

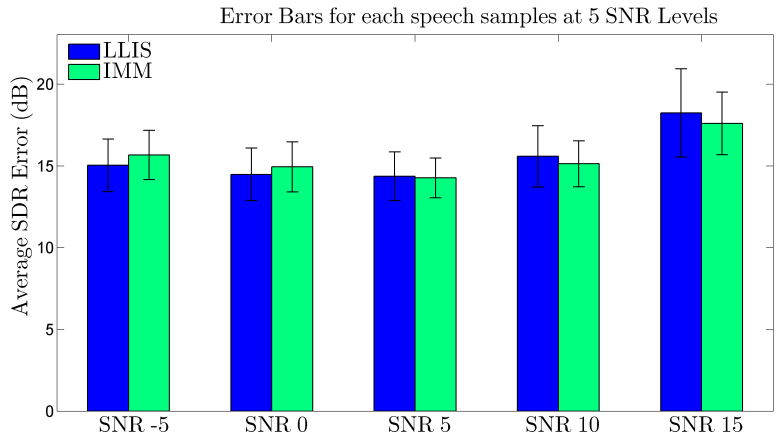

Figure 5.4: Speech average SDR error for each noise SNR level with different noise mixed

5.1.2 PESQ Results

The other side of the objective results is the Perceptual Evaluation of Speech Quality. In these experiments we have used an additional algorithm in order to contrast the

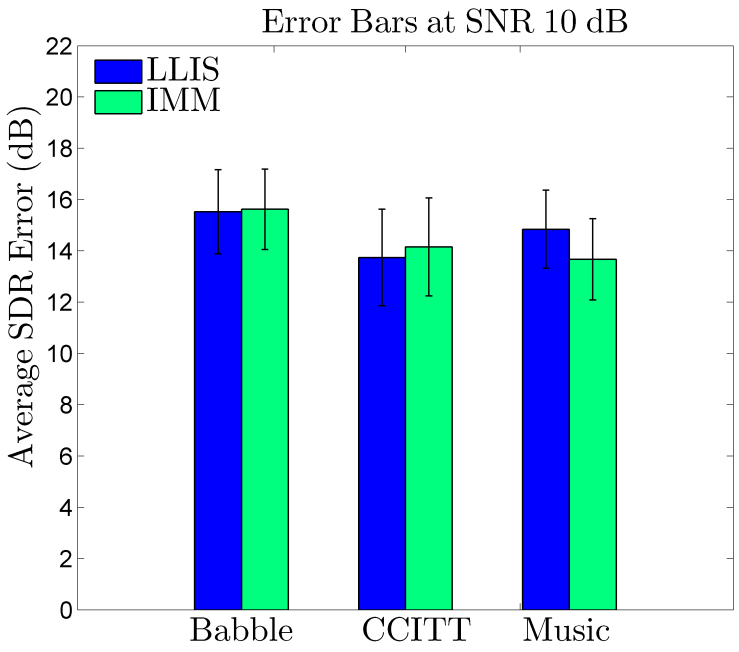

Figure 5.5: Speech average SDR error for each noise Babble, CCITT and Music at SNR 10 dB

results. This is a noise reduction algorithm [J. Benesty,][M. M. Sondhi and Rabiner,] [J. Chen and Huang,][Diethorn,] which basically reduces the background noise. The algorithm works perfectly with stationary noises but has not very good effect with non stationary noises. In figures 5.6 and 5.7 we can see the results obtained with the pesq evaluation. As can be observed, in most of the cases the original mixture with speech plus noise and the results obtained with the noise reduction algorithm are significantly better than both algorithms LLIS and IMM. The most optimistic case is the babble noise, where LLIS reaches a similar level than the original and the denoised. The good point is that the results obtained with LLIS for the babble noise is considerably better than the IMM.

5.2 Subjective Evaluation

The tests realized for the subjective evaluation are the decisive exercise to determine how well the system works for our main objective, improve speech intelligibility. After evaluating the objective results we can have a preliminary idea of how the algorithm will work in the different scenarios. But unfortunately objective measures are not always reliable specially with speech intelligibility. This is why we decided to do some

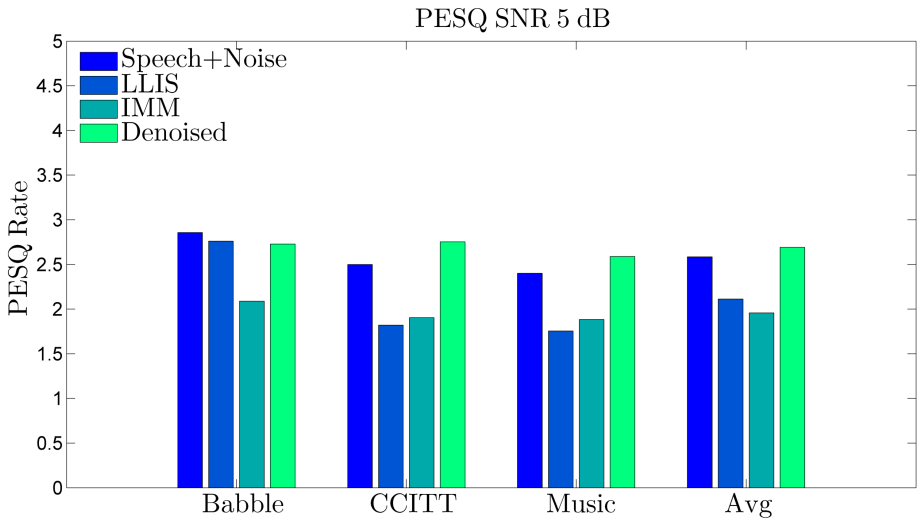

Figure 5.6: PESQ measure for each noise Babble, CCITT and Music at SNR 5 dB obtained with the different algorithms

tests with volunteers before the definitive tests with CI patients.

With the goal of have coherent and useful results, we tried to have a correct configuration reagarding the environment and the procedure. The first requirement was to have a quiet and confortable place to run the experiments essentially because the tests need a considerable level of concentration specially for the CI patients which any kind of background noise can disturb their perception. We used a small room from the CBCLaB[1], a neuroscience laboratories used by some brain and cognition research groups of the UPF. In figure 5.8 is depicted a simple representation of the room. The room is considerably good isolated acoustically and has no reberveration. The speaker is placed 1.5 meters from the subject and the audio level is calibrated each time at 60 dBSPL. We were very restricted with the test structure because with 6 conditions we had 9 possible combinations, so 9 lists of 25 words or 20 sentences plus a training phase took about 1.5 hours. It is important to mention that all the tests were not realized at the same SNR level. This is because each CI patient has very different hearing levels. So the procedure was during the training stage, different levels were presented to the CI patient and we selected the optimal level which we considered that is at 50% of correct words independently of the SNR.

[1]http://lnucc.upf.edu/?q=en

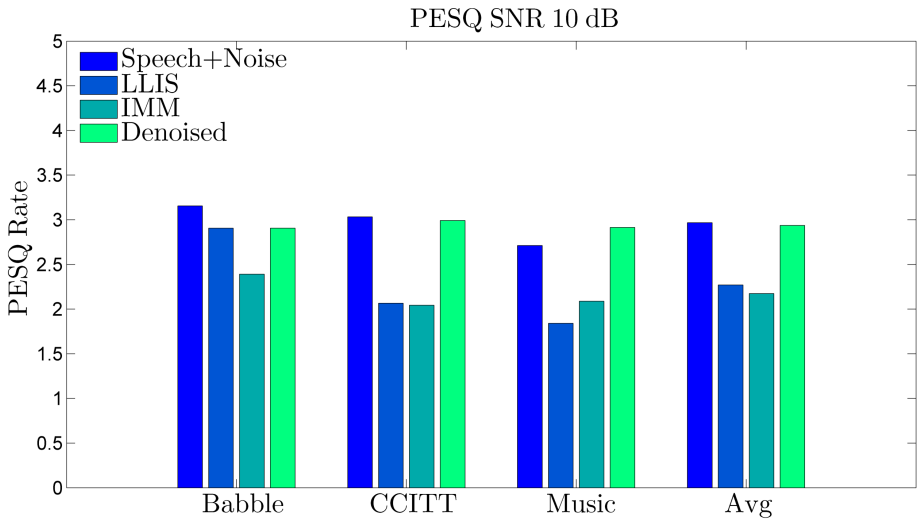

Figure 5.7: PESQ measure for each noise Babble, CCITT and Music at SNR 10 dB obtained with the different algorithms

5.2.1 Experiments with normal hearing people

The tests done with volunteers are not 100% reliable mainly by the fact that the volunteers had not any known hearing disease. So we assumed that they can hear perfectly good. In order to adapt this situation to be similar to the CI patients we used a processed version of the estimated audios. We can see that as a CI simulator but essentially it is not. The objective is not simulate how CI patients hear due it is very difficult to approximate to model how they internally perceive. So what we have done is degrade the signal to an inferior level similar to the degradation level that CI have. An example of these measures can be the lower frequency bandwidth used in the implants, where is mapped the original frequency bands into 16-22 frequency bands corresponding to the number of electrodes used to stimulate the cochlea.

In these preliminary tests we had 6 volunteers which 2 of them are not reflected in the results due in the first tests were more introductory and we needed to test different combinations of noise plus SNR in order to have a good test structure. In these tests we evaluated both algorithms IMM and LLIS compared to the original mix. These conditions where tested for a sentence plus babble noise, and a word mixed with music and babble noise. We chose babble noise because was one of the only scenarios where the algorithms obtained good results. Then we selected the music noise to compare it. As can be seen in figure 5.9 the obtained results are quite different for each scenario. For the sentence case, the results obtained by the LLIS algorithm are considerably good. In all the cases outperforms IMM and also the average is higher than the original.

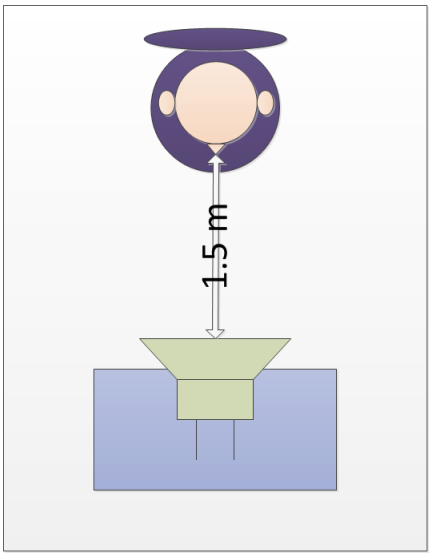

Figure 5.8: CBCLaB Room

Although in this case the subject number 3 makes elevate the mean considerably. In the word plus babble noise case the original obtains the best results in most of the situations, specially in the word plus babble noise. The case of word plus music is favorable to the IMM algorithm if we consider that with the first subject where some problems and these conditions were unable to be realized. The differences obtained with the sentence and word with babble noise are introduced by the differences between the sentence and the word. As the sentence have a context, it is easier to deduce more words while in the test with words it becomes harder to deduce it. This demonstrates that the word test is more reliable in terms of intelligibility.

5.2.2 Experiments with CI patients

After the tests with volunteers we built the definitive test configuration. First of all we focused on evaluate only lists of words due the reliability in order to evaluate intelligibility. Then we focused on same music noise from the previous tests and we added the CCITT and a new music noise. We discarted the babble noise because, even though LLIS showed significant good results with in the previous tests, we tested some lists with the first patient and it was impossible to recognize any word. We introduced the sationary noise CCITT to compare LLIS with the noise reduction algorithm. Finally the new music nois it is composed by only cello playing some notes at fixed frequency.

In table 5.4 we can see a summary of some characteristics regarding their experience with Hearing loss and CI. Is important to consider these set of characteristics in order to determine possible differences in the obtained results.

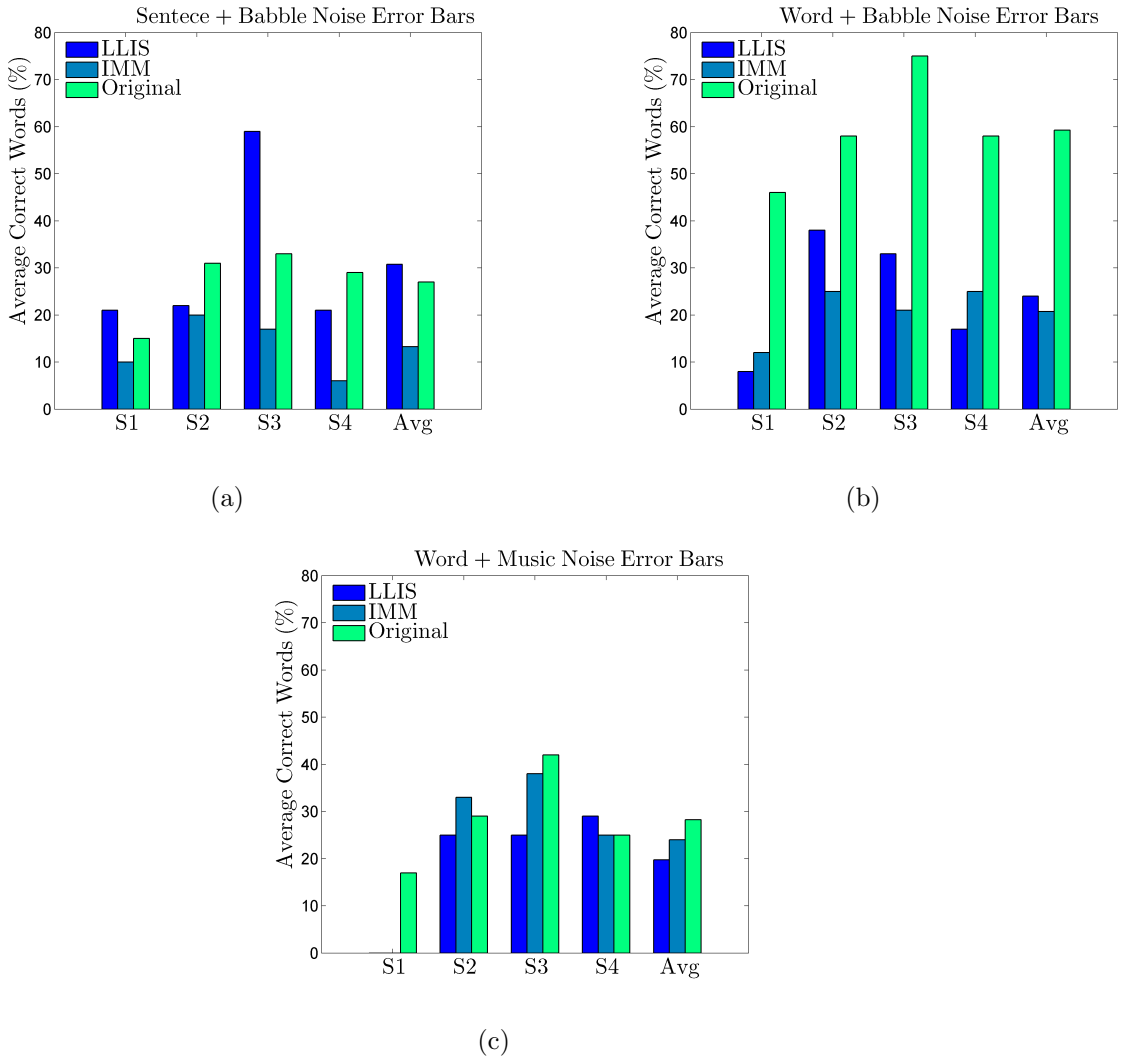

Figure 5.9: Test realized with 4 volunteers with mixed words and noise and with the different algorithms

Pacient	Age	Cause of H.loss	Device	Duration of H.loss	Experience with CI	Pre/Postlocutive
P1	57	Cofosis	Nucleus5	1.5 years	8 years	Postlocutive
P2	34	Accident	Nucleus5	10 years	3 years	Postlocutive
P3	49	Congenital	Esprit	32 years	17 years	Prelocutive
P4	59	Infection	Nucleus5	2 years	1 years	Postlocutive
P5	56	Infection	Nucleus5	15 days	2 years	Postlocutive

Table 5.4: CI patients main characteristics

After realizing the tests, we had the results in percentage of correct words per list, but we decided to go deeper and analyze it phoneme by phoneme as explained in section 4.3.2. The results are practically the same that if we count the percentage of correct words but during the tests we encountered many situations where the patient answered very similar words but with some changed phonemes which sometimes can

be a wrong word but with only one wrong phoneme. Even this way, the overall results for most of the cases is favorable to the noise reduction algorithm and for the original speech mixed with noise. As can be seen in figure 5.10 the PER error for LLIS is considerable higher than the rest. Only in some sporadic cases LLIS obtained good results like in the music case for the patient 1 which improved significantly the PER error of the original mixture but still very similar to the noise reduction.

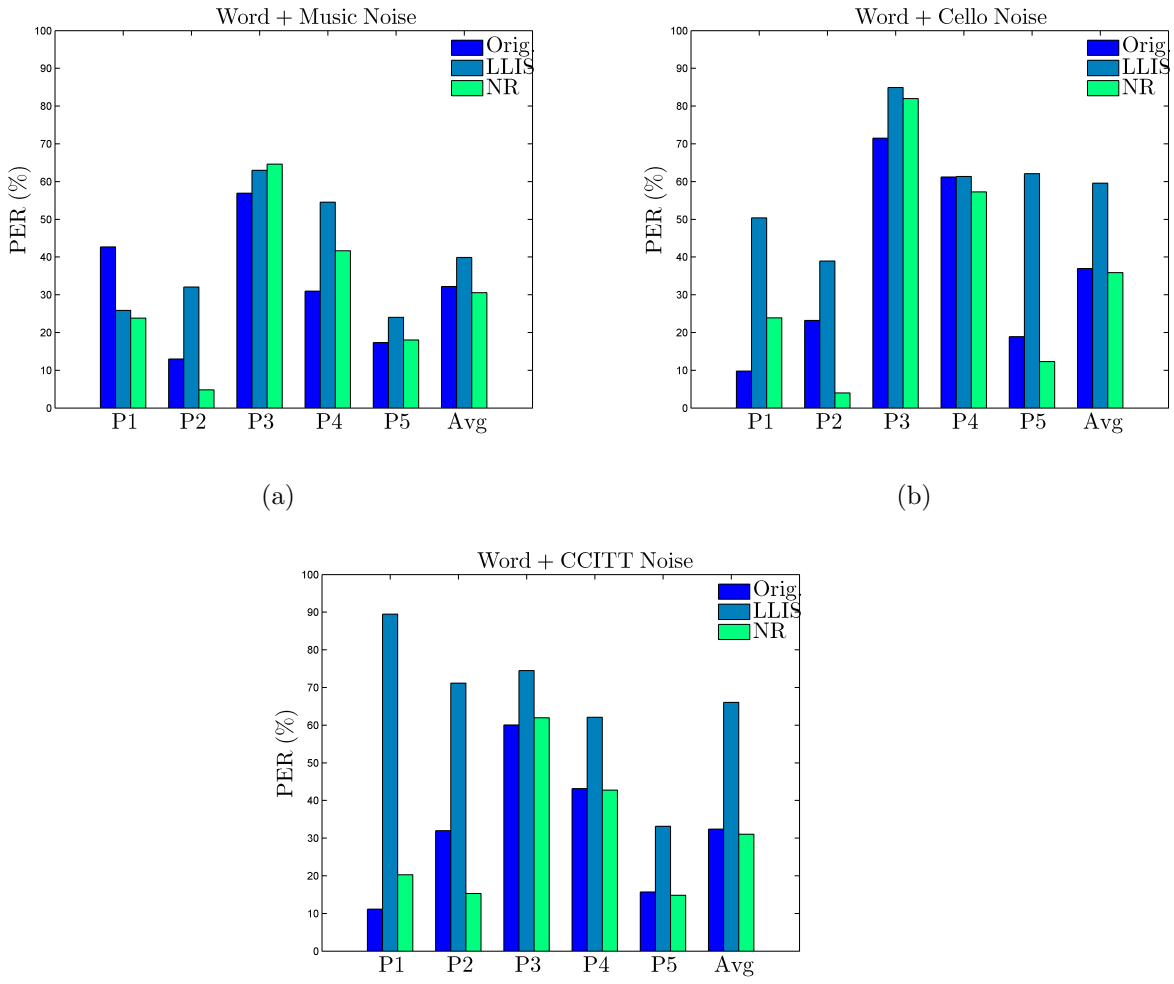

Figure 5.10: Test realized with 5 CI Patients with mixed words and noise and with the different algorithms

With the goal of analyse where specifically LLIS fails, we have made a re-count of the errors made by the patients. Concretely we have gathered all the substitution errors to determine which phonemes are more commonly confused. In figure 5.11 is showed an histogram of the error phonemes ordered by repetition times. In the first

4 positions we have the phonemes “e”, “s”, “a” and “o” which have been confused about 50 or more times. It is normal that the most confused phonemes are vowels due are more common in each word (at least in Spanish) and every time that a word is confused, some vowels will be confused. And considering that there are 5 vowels the probability of fail is higher. But the interesting point is that in the second position it is the phoneme s. We can assume with these results that the fact that LLIS does not recognize some consonants is a problem, and s is a clear example.

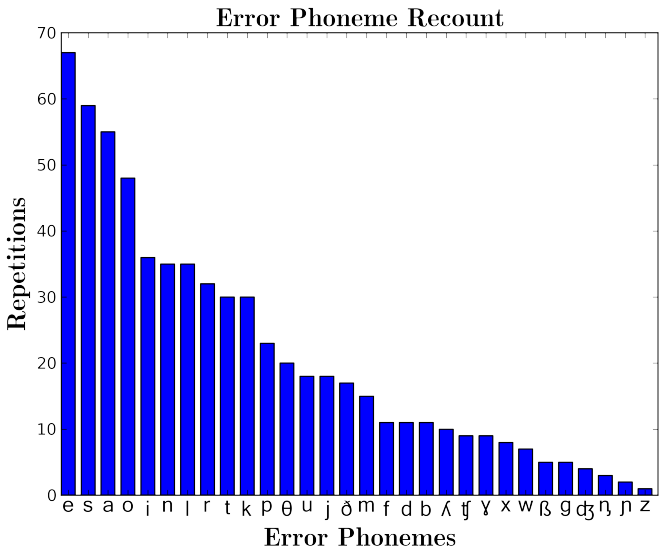

Figure 5.11: Phoneme error recount

5.3 Discussion

From the results obtained with the 5 CI patients we did not need plots to determine that there is some problems with LLIS regarding to the speech intelligibility. After generating all the estimates, in most cases LLIS successfully separated the speech signals from the rest of components. But we encountered another problem that is worse for speech intelligibility that is speech degradation. Although it only affects to some parts of the words, degrading the speech signal makes reduce abruptly the intelligibility. We found that consonants are fundamental components of the speech in order to understand. As commented in the methodology section, LLIS does not detects unvoiced consonants. There are two type of consonants: obstruents and sonorants. Sonorants are those type of phonemes that have a certain pitch like could be “n” or “m”. Vowels are also sonorants. But obstruent consonants are those sounds produced by obstructing the airflow with no interaction of the vocal folds. So obstruent consonants

have no pitch which is the main problem of dealing with LLIS wich does not detect these certain consonants.

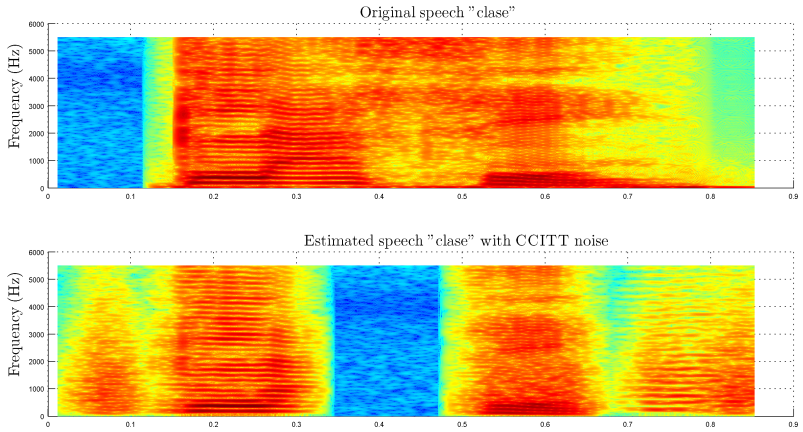

Figure 5.12: Spectrograms of a original word and a word mixed with CCITT noise with consonant "s" missing

We can return to figure 5.11 and the case of the phoneme "s". This phoneme is a fricative obstruent so it has no specific pitch. Figure 5.12 shows the spectrograms of a **original** word and **estimated** from a mixture with CCITT noise. It is a perfect example of how LLIS remove the consonants. In this case there is practically absolute silence between phonemes "a" and "e". In the original spectrogram we can see how the spectral content of the phoneme "s" is not harmonic at all. We can see a concentration of energy randomly distributed in time specially around 3.5 kHz and 5 kHz. There is some problems also with the first phoneme "k" which is a stop consonant or occlusive and it behaves like a transient with a huge concentration of energy suddenly when starts. Here the estimates also loose a lot of content but still there are some parts that make the phoneme recognizable.

Another phenomena could be the observed in figure 5.13 where now the **estimated** is a mix with cello noise. In this case LLIS neither detects correctly "s". But now in contrast to the anterior case, when the algorithm does not detect any pitch in the place of the consonant, some harmonics of the cello are present and this makes that also some spectral content of phoneme "s" remain present. The fact that the cello shares some partials makes that this is not removed totally and this is also a problem for intelligibility.

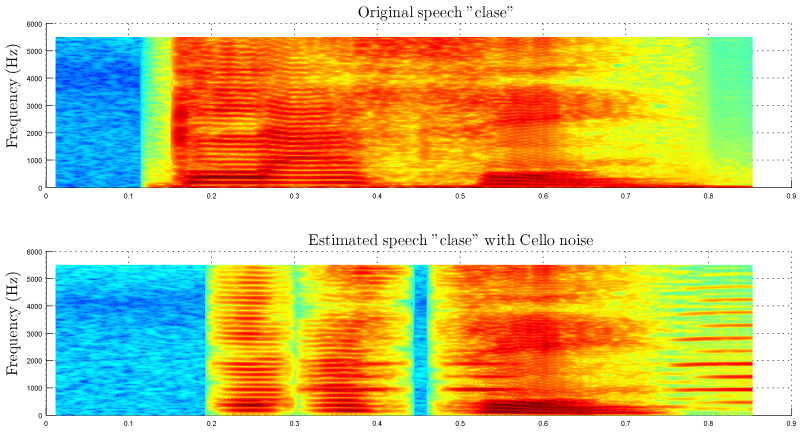

Figure 5.13: Spectrograms of a original word and a word mixed with Cello noise with consonant "s" replaced or degradated

5.3.1 Transient detection

The problem related to the consonants was expected before running the experiments since LLIS was designed without considering unvoiced consonants detection. As commented in section 3.2 a new approach with the goal of solve the consonants detection is being carried out. Following with the example of the word "clase", we used this new strategy to have preliminary results of the consonant detection. In figure 5.14 we have the same estimated speech signal from the previous example compared with the transient detection.

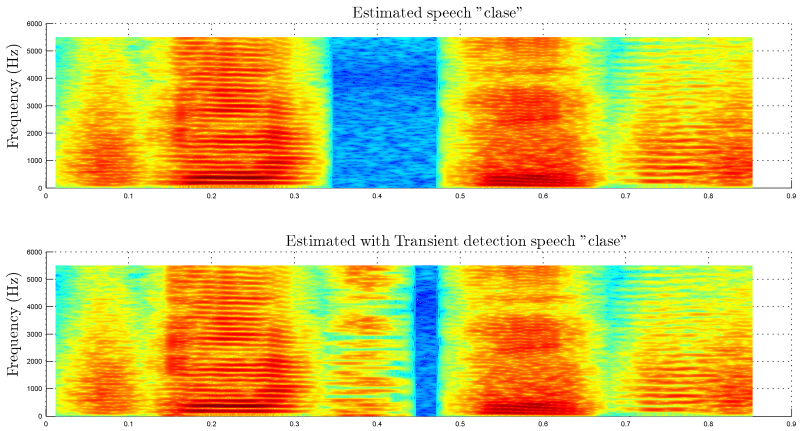

Figure 5.14: Spectrograms of estimated speech and estimated speech with transient detection strategy

The obtained **consonants signal** is extracted from the same noise plus speech signal of the previous example. As can be noticed it is not completely a consonant recovery but some content of the phonemes “k” and “s” is detected. But this few content of the consonants is enough to make change the intelligibility of the word. The result of the **added consonants to the estimated signal** can be easily perceived audibly and also in the spectrogram we can observe how some of the missing component in the middle of the word is filled with part of the recovered phoneme “s” as well as “k”. This simple example shows how this strategy can successfully improve speech intelligibility working with LLIS.

Chapter 6

CONCLUSIONS AND FUTURE WORK

The research done in this master thesis is a relatively new contribution for both areas blind source separation and cochlear implants. Perhaps the contribution has more impact in the field of cochlear implant as the main goal was to improve speech intelligibility in the context of cochlear implants. While we used an existing source separation framework with not so much emphasis in the modification but with an accurate evaluation process.

6.0.2 Contributions

Regarding both topics involved in this master thesis we can extract the following contributions:

- The literature review done in the related areas of research have determined that there is a large amount of research done in source separation and his different methods and strategies, but the application of this methods is still very new to the field of cochlear implants which only a few contributions have been done.
- It has been demonstrated how a low latency source separation algorithm can reach similar quality levels to an offline algorithm in many scenarios with different kind of noises and SNR levels.
- We showed how the low latency algorithm can reach considerable good results in tests with normal hearing people
- The realization of tests with volunteers and CI patients revealed some critical aspects of LLIS algorithm regarding to speech intelligibility

- The proposed solution can be a very effective solution to noise reduction algorithms which normally work only with stationary noise

6.0.3 Conclusions

Considering the motivations and goals presented in the introduction and having into account the contributions commented in section 6.0.2 we can obtain some conclusions.

First we can consider the presented work as an innovative contribution due we have encountered only a few contributions dealing with cochlear implants and source separation. It is worth to mention that in contrast to cochlear implants, the use of source separation in hearing aids it is quite extensive with successful results.

Then, in respect to source separation we can comment that even using a framework focused in singing voice and music, this have obtained relatively good results in the task of separating speech and other components different from musical instruments. It is very important to have into account that only obtaining similar results to an offline algorithm this is very positive, but in some cases LLIS even outperforms IMM. We observed that still the estimation error is considerable high compared to other examples. This can be improved adapting more the system with a speech training and consonant detection strategies which is part of the future work. For the case of the transient detection, some preliminary tests were conducted, but unfortunately was not possible to be evaluated with the subjective tests. However the first obtained results are promising.

The results obtained with normal hearing people in the preliminary tests were promisingly due in some scenarios like the music noise the results obtained with the source separation algorithms have reported some good results. In contrast, babble noise needs improvement and was not evalueted with the cochlear implants test.

Tests realized with CI patients have been the final point in this evaluation process. The results obtained with the tests showed that there is not so much correlation between separation measures and speech intelligibility. Even in cases where the separation is considerably good, if some parts of the speech components are degraded this becomes a huge problem for intelligibility. In most of the cases the original audio samples with the speech and the noise is best understood together with the de-noised sample through the noise reduction algorithm. Another good point of making the subjective evaluation directly with the patients is that a part from the data strictly from the test, we were constantly receiving feedback about little details of the algorithms. This has a lot of advantages in order to analyze the performance of the system. For example in some cases the intelligibility obtained with the LLIS algorithm was similar or equal to the original mixture. This could be a lack of improvement, but if the

patient reveals that in this case the noise was totally removed and he felt more comfortable, although we are not enhancing speech intelligibility, we can assume that we are improving his quality of life.

We observed also in the tests that the noise reduction algorithm, the most effective, does not work at all with non stationary noise like the babble or the cello. This gives a good opportunity for LLIS due can perfectly work with non stationary noise. It is also obvious that the problem of the consonants needs to be solved in order to be a serious solution.

6.0.4 Futrue Work

Transient detection

Some preliminary tests have been done with an improved version of LLIS with the goal of detect transient signals as it have been demonstrated in section 5.3.1. This is mostly designed to track percussive instruments or similar. But as we have seen in some of the spectrograms exposed in this work, some consonants behaves similar to transients, like the case of the word "clase", the phoneme "k" it can be seen as a transient signal. Specially occlusive consonants which are a quick duration in time and a sudden change of energy. This could work for some cases where LLIS does not detect unvoiced consonants by applying the transient detection and then adding the found consonants to the estimated speech signal.

Speech training

One of the commented problems of LLIS when working with speech signals is the mentioned training strategy. As the system is trained to remove the lead singing voice, the training strategy have been designed with singing voice signals which makes that the model generated is not the best. A good improvement of the system could be the generation of a new model fully trained with speech samples. A very special and restricted scenario could be the possibility of restricting the speech training to real recordings of the CI patients familiar or friends. This can be seen as a customized program from the amount of strategies installed in the implants with the possibility that the user selects the program in presence of the person whose voice has been trained.

Bibliography

[Apostolico and Galil, 1997] Apostolico, A. and Galil, Z. (1997). *Pattern Matching Algorithms*.

[a.W. Rix et al.,] a.W. Rix, Beerends, J., Hollier, M., and a.P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, 2:749–752.

[Bregman, 1990] Bregman, A. (1990). *Auditory Scene Analysis, Second Edition*.

[Burred, 2009] Burred, J. J. (2009). *From Sparse Models to Timbre Learning: New Methods for Musical Source Separation*. PhD thesis.

[Casey, 2000] Casey, M. A. (2000). Separation of Mixed Audio Sources by Independent Subspace Analysis. *Proceedings of the International Computer Music Conference*.

[Diethorn,] Diethorn, E. J. Subband Noise Reduction Methods For Speech Enhancement.

[Durrieu et al., 2011] Durrieu, J.-l., David, B., and Member, S. (2011). A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal on Selected Topics on Signal Processing, Music Signal Processing*, pages 1–12.

[Durrieu et al., 2009a] Durrieu, J.-L., Ozerov, A., Févotte, C., Richard, G., and David, B. (2009a). Main instrument separation from stereophonic audio signals using a source/filter model. *European Signal Processing Conference (EUSIPCO)*, (1).

[Durrieu et al., 2009b] Durrieu, J.-L., Richard, G., and David, B. (2009b). An iterative approach to monaural musical mixture de-soloing. *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (1):105–108.

[Févotte et al., 2009] Févotte, C., Bertin, N., and Durrieu, J.-L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. *Neural computation*, 21(3):793–830.

[Févotte et al., 2005] Févotte, C., Gribonval, R., and Vincent, E. (2005). BSS EVAL Toolbox User Guide Revision 2.0. *Technical Report 1706, IRISA*.

[Han et al., 2009] Han, S., Cui, J., Academy, C., and Beijing, S. (2009). Post-processing for Frequency-domain Blind Source Separation in Hearing Aids. *7th International Conference on Information, Communications and Signal Processing (ICICS)*, pages 1–5.

[Hoyer, 2004] Hoyer, P. O. (2004). Non-negative Matrix Factorization with Sparseness Constraints. *Journal of Machine Learning Research*, 5:1457–1469.

[Hu and Loizou, 2008] Hu, Y. and Loizou, P. C. (2008). Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238.

[International Telecommunication Union,] International Telecommunication Union. ITU-T Recommendation G.227.

[Itoyama, 2011] Itoyama, K. (2011). *Source Separation of Musical Instrument Sounds in Polyphonic Musical Audio Signal and Its Application*. PhD thesis.

[J. Benesty,] J. Benesty, J. Chen, Y. C. I. C.

[J. Chen and Huang,] J. Chen, J. B. and Huang, Y. *IEEE Trans Speech Audio and Sig. Proc.*

[Joder et al., 2012] Joder, C., Weninger, F., Eyben, F., Virette, D., and Schuller, B. (2012). Real-Time Speech Separation by Semi-supervised Nonnegative Matrix Factorization. *International Conference on Latent Variable Analysis and Source Separation (LVA/ICA)*, 2010:322–329.

[Jolliffe, 2002] Jolliffe, I. T. (2002). *Principal Component Analysis, Second Edition*.

[Jutten and Comon, 2010] Jutten, C. and Comon, P. (2010). *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Elsevier Ltd., first edit edition.

[Kokkinakis and Loizou, 2008] Kokkinakis, K. and Loizou, P. C. (2008). Using blind source separation techniques to improve speech recognition in bilateral cochlear implant patients. *The Journal of the Acoustical Society of America*, 123(4):2379–90.

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91.

[Lee and Seung, 2001] Lee, D. D. and Seung, H. S. (2001). Algorithms for Non-negative Matrix Factorization. *Advances in Neural Information Processing Systems*, 13(1):556–562.

[M. M. Sondhi and Rabiner,] M. M. Sondhi, C. E. S. and Rabiner, L. R. *Bell Syst. Techn, J.*

[Marxer et al., 2012] Marxer, R., Janer, J., and Bonada, J. (2012). Low-Latency Instrument Separation in Polyphonic Audio Using Timbre Models. *International Conference on Latent Variable Analysis and Source Separation (LVA/ICA)*, pages 314–321.

[Nogueira, 2008] Nogueira, W. (2008). *Design and Evaluation of Signal Processing Strategies for Cochlear Implants based on Psychoacoustic Masking and Current Steering.* PhD thesis.

[Ozerov and Vincent, 2010] Ozerov, A. and Vincent, E. (2010). A General Modular Framework for Audio Source Separation. *International Conference on Latent Variable Analysis and Source Separation (LVA/ICA)*, pages 33–40.

[Paulus and Virtanen, 2005] Paulus, J. and Virtanen, T. O. (2005). Drum transcription with non-negative spectrogram factorisation. *European Signal Processing Conference (EUSIPCO).*

[Reindl et al., 2010] Reindl, K., Zheng, Y., and Kellermannt, W. (2010). Speech Enhancement for Binaural Hearing Aids based on Blind Source Separation. *Proceedings of the 4th International Symposium on Communications, Control and Signal Processing*, (March):3–5.

[Rickard and Yilmaz, 2004] Rickard, S. and Yilmaz, O. (2004). Blind Separation of Speech Mixtures via Time-Frequency Masking. *IEEE Transactions On Signal Processing*, 52(7):1830–1847.

[Suhail and Oweiss, 2006] Suhail, Y. and Oweiss, K. G. (2006). Augmenting information channels in hearing aids and cochlear implants under adverse conditions.

Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2):889–892.

[Vincent, 2006] Vincent, E. (2006). Musical Source Separation Using Time-Frequency Source Priors. *IEEE Transactions on Speech and Audio Processing Special Issue on Statistical and Perceptual Audio Processing*, pages 1–8.

[Vincent et al., 2003] Vincent, E., Févotte, C., and Gribonval, R. (2003). A tentative typology of audio source separation tasks. *International Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, (April):715–720.

[Vincent et al., 2007a] Vincent, E., Gribonval, R., and Plumbley, M. D. (2007a). Oracle estimators for the benchmarking of source separation algorithms. *Signal Processing*, 87(8):1933–1950.

[Vincent et al., 2007b] Vincent, E., Sawada, H., Bofill, P., Makino, S., and Rosca, J. P. (2007b). First Stereo Audio Source Separation Evaluation Campaign : Data , Algorithms and Results. *Proc. 7th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, pages 8–15.

[Vinyes et al., 2006] Vinyes, M., Bonada, J., and Loscos, A. (2006). Demixing Commercial Music Productions via Human-Assisted Time-Frequency Masking. *Audio Engineering Society*.

[Virtanen, 2005] Virtanen, T. O. (2005). Monaural Sound Source Separation by Perceptually Weighted Non-Negative Matrix Factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1–9.