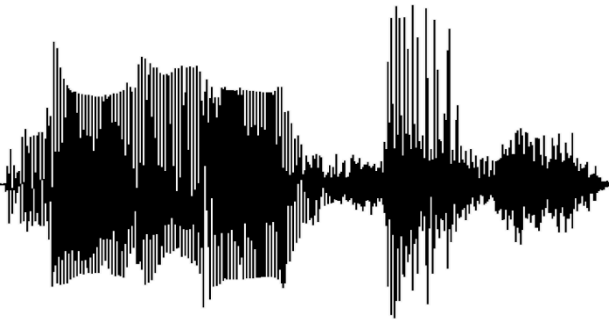
[Bachelor Thesis]

Noise cancelling in acoustic voice signals



with spectral subtraction

Grado en Ingeniería de Sistemas Audiovisuales.

Author: Andrea García González.

Tutor: Emilio Parrado Hernández.

Universidad Carlos III de Madrid

Title: Noise cancelling in acoustic voice signals with spectral subtraction.

Author: Andrea García González

Tutor: Emilio Parrado Hernández

Date: February 2014

Acknowledgements

Quiero dedicar este proyecto a toda la gente que ha acompañado, me ha ayudado y enseñado a crecer en estos años de Universidad.

A todos aquellos, que aunque sea con un pequeño detalle han hecho que estos 24 años hayan sido increíbles. Y a los que después de este tiempo siguen a mi lado, esperando muchas más historias que luego poder contar.

Gracias…

...a todos los que me han ayudado en este proyecto, y a los que cuando lo veía difícil, me han animado a acabar.

… a todos los que han hecho que yo sea como soy. A esa persona que me ha hecho disfrutar como una *groupie* en un concierto, a la que me hace pensar en sapos y princesas, en definitiva, a los que me hacen soñar.

… a mis amigos que tanto me conocen y que saben que pueden sacarme una sonrisa por el centro de Madrid, ya sea con un café, con un helado, o con una charla interminable.

... a mis queridas vecinas.

...a los nuevos amigos que me llevo de Alemania, en especial a mis chicas.

...a mis padres, gracias por enseñarme a luchar por lo que quiero, por darme la oportunidad de estudiar algo que me gusta.

… a mi hermana, por todo.

… a mis abuelos, a mis tíos y a mis primos… todos muy importantes para mí, gracias por aconsejarme siempre y animarme a aprovechar las oportunidades.

..Y a mi tutor, por ayudarme y guiarme… y por supuesto, ¡por hacer las revisiones más amenas con sus comentarios!

Abstract

The main purpose of study throughout this entire End of Degree Project would be the noise removal within speech signals, focusing on the diverse amount of algorithms using the spectral subtraction method. A Matlab application has been designed and created. The application main goal is to remove any meaningless thing considered as a disturb element when trying to perceive a voice; that is, anything considered as a noise.

Noise removal is the basis for any voice processing that the user wants to do later, as speech recognition, save the clean audio, voice analysis, etc.

A studio on four algorithms has been executed, in order to perform the spectral subtraction: Boll, Berouti, Lockwood & Boudy, and Multiband. This document presents a theoretical study and its implementation.

Moreover, in order to have ready for the user a suitable implementation of an application, an intuitive and simple interface has been designed. This document shows how the different algorithms work in some voices and with various types of noise. A few amounts of noises are ideal, used by its mathematical characteristics, while others, are quite common and presented in daily routine, it is presented as for example, the noise of a bus.

To apply the method of spectral subtraction is necessary the implementation of a Vocal Activity Detector, able to recognize in which precise moments of the audio there is voice or not. Two types have been studied and implemented: the first one establishes the meaning of voice according to a threshold which is adequate to this record, while the second one is the combination of Zero Crossing Rate and energy.

In the end, once the application is implemented, evaluating its performances was the next process, either in an objective and a subjective form. People stand point was considered and asked, in order to obtain the proper functioning of the application along different types of noise, voice, variables, algorithm, etc.

Resumen

Este Trabajo de Fin de Grado, consiste en el estudio de la eliminación de ruido en voces; en concreto en el estudio de distintos algoritmos para el método de la resta espectral. Se ha creado una aplicación en el programa de cálculo Matlab cuyo uso es la eliminación de todo aquello que nos pueda molestar a la hora de escuchar una voz, es decir, lo que se considera ruido.

La eliminación de ruido es la base de cualquier tratamiento de voz que se quiera aplicar posteriormente; desde reconocimiento de voz, el análisis de la misma, la conservación de la grabación limpia. etc.

Se ha hecho un estudio de cuatro algoritmos para llevar a cabo esta resta espectral: Boll, Berouti, Lockwood & Boudy y Multibanda. En este documento se encuentra tanto un estudio teórico, así como su implementación.

Para la implementación de una aplicación que pueda ser usada por un usuario, se ha diseñado una interfaz fácil e intuitiva de usar, en ésta se muestra cómo funcionan los distintos algoritmos en distintas voces y con distintos tipos de ruido, algunos ideales, usados en las medidas oficiales de ruido por sus concretas características matemáticas, y otros, los de la vida cotidiana como el ruido de un autobús.

Para aplicar el método de la resta espectral es necesario la implementación de un Detector de Actividad Vocal (VAD) que reconozca en qué momentos del audio hay voz o no. Se han estudiado e implementado dos: Uno de ellos establece qué es voz según un límite adecuado a esa grabación y el otro es la combinación de la Tasa de Cruces por Cero (ZCR) y la energía.

Por último, una vez implementada esta aplicación se ha procedido a evaluar su funcionamiento, tanto de una forma objetiva como subjetiva, a través de la escucha de distintas personas, las cuales dan su opinión, para poder obtener el comportamiento de la aplicación con distintos tipos de ruidos, voces, variables, algoritmos, etc.

Index

Figures index

Figures index

Chapter 1

1. Introduction and objectives

Throughout the following chapter the reader would encounter the diverse arguments and the multiple goals for which the Bachelor Thesis has been carried out.

1.1. Setting and motivation

In our present world, the recording and audio treatment is very common. And it would be possible the emergence of a discipline linked to this field. A necessary discipline able to take into consideration what many of the devices that usually use the speech need to run audio processing subsystems. Common examples are telephony and music applications.

The operation interface through the human voice is increasingly common. This way of working with the devices makes it easier to interact with computers, phones, etc. The control of these devices by voices is a way for old people and people with disabilities can access information. Moreover, this development allows disabled people to interact with the world around them trough these devices working voice.

We also work with the conversion of analogue audio into digital audio in order to have more freedom to work with it. However, not all systems or devices seem to be perfect, since the process of working with audio most of times registers the presence of noise, something that can contaminate the signal of interest.

Nowadays, noise removal techniques are pretty important, and must be taken as necessary first step. Consequently, the first thing that needs to be done is to clean the signal, removing unwanted and background noises.

When performing voice activity detection, the quality of this cleaning may limit the voice recognition performance. In addition, when recording music, listeners do not want to be disturbed with other noises that later would have to be heard. When talking about speech, the sound is even more important because that voice cannot be heard well.

Therefore, this Bachelor Thesis would present different chapters, one of several forms of audio noise removal, particularly the one dedicated to eliminate noise at voice recordings. Spectral subtraction algorithm has been chosen because it is the basis of many other noise removal techniques. Moreover, this technique implies low computational cost.

1.2. Objectives

The objectives of this Bachelor Thesis are to study different algorithms which one carries on the spectral subtraction and to study the results. Another goal consists in to analyze their performance in different scenarios as well as to determinate according to objective and subjective criteria about which is the most suitable for our main objective, to clean an audio signal. How these algorithms operate for each type of noise would be another subject to analyse in deep way.

Furthermore, another objective of this Bachelor Thesis is to study different types of voice activity detectors and to implement the one that best fits the subsequent spectral subtraction. I mean, this module has to be able to estimate the noise well and not cut the words.

This project is formed by several modules. The first one is devoted to data acquisition. It involves the recording of speech of different types under different ambient noise conditions.

The second module examines the voice activity detector (VAD). Studying and creating a suitable and proper one which would act at the base of the spectral subtraction algorithm. A high classification error rate in the VAD will lead to poor estimations of the noise spectra. This fact will bring out a severe deterioration of the performance of noise canceller.

The third module focuses on the spectral subtraction, studying the different algorithm chosen. Its implementation has been done and all the appropriate improvements have been inserted.

The last module is the testing. Obtaining a subjective and an objective quality measures so as to assess the performance of the analyzed algorithms is its main target.

1.3. Reading Guide

This project report is divided into 8 chapters, which are described below.

- *Chapter 1: Introduction and objectives.*
  This chapter includes a brief introduction about one of several form of audio noise removal is made, particularly focusing in eliminating noise of voice recordings. A description of the motivations and the objectives of this End of Degree Project are also provided.

- *Chapter 2: State of the art.*
  A general theoretical basis of audio signal processing is explained in this section and a study about different analysis techniques is done. As well as the problem that there is necessary to eliminate: the noise in voice signal.

- *Chapter 3: Technical solution design.*
  The technical solution design is explained from the point of view of the theory. I explain the different algorithms, the problem with noise and distortions, and the necessary algorithm to carry out the spectral subtraction.

- *Chapter 4: Implementation.*
  This section explains the development of the spectral subtraction application in Matlab.

- *Chapter 5: Results.*
  This chapter describes the experimental work carried out to validate the theoretical analysis presented in the previous chapters. The experiments include noise cancellation in different environments (bus noise, restaurant noise, etc) and with different target signals (male, female).

- *Chapter 6: Application uses and future lines.*
  This chapter explains the futures steps which are not carried out in this project and its uses.

Finally, the last two chapters show the conclusions of this project and the budget.

Chapter 2

2. State of the art

2.1. Introduction to digital audio signals

A digital audio signal is a representation of sound signals through a stream of binary data.

Usually a digital audio system originates from a transducer (a microphone) which converts the sound pressure wave into an analogue electrical signal.

This analogue signal passes through a processing system which has the ability to apply different treatments as audio frequency equalization, amplification and other processes.

As an example of what has been previously said, the equalization counteracts the frequency response of the transducer used to form the analogue signal so that the final signal closely matches the original audio signal.

The digital audio signal is finally obtained after sampling, quantizing and coding, in order to be converted into a digital audio signal. Sampling would involve the process of taking a number of discrete analogue signal values per second (sampling frequency), using and discrete number of values so as to codify the value. That procedure implies a loss of information, since the values of the signal are approximated to the nearest code value. The digital signal consists of the coding sequence of bits assigned to each discrete analogue value.

The quality of this process depends primarily on two values:

- *Sampling rate:*
  The sampling frequency is the number of samples per time unit taken from a continuous signal to produce a discrete signal. It is expressed in Hz per time unit. The Nyquist-Shannon theorem states that for a signal to be properly sampled, the sampling rate must be at least greater than twice the highest frequency to be sampled.

Although the audible frequencies for humans are between 20 and 20,000 Hz, the human voice is almost always below 10,000 Hz. The sampling frequency must be chosen in accordance with the use to which the signal and the target are intended.

- *Bit depth:*
  The term bit depth describes the number of bits with which each sample is recorded. It corresponds to the resolution at which each sample is quantified. When the value of bit depth values is higher, the result obtained would be more close to the reality. This loss of information is called quantization error, and it is the difference between the real value and the assigned one.

An example of use of sampling frequency and bit depth is:

Application	Sampling rate	Bit depth
Telephone	8 kHz	8
Compact Disc (CD)	44.1 kHz	16
DVD Audio	96 kHz	24

Table 2.1 Example of uses of sampling frequency and bit depth.

2.2. Analysis techniques

As mentioned in the preceding paragraph, in order to signal, it is necessary to summarize along this section the most important signal analysis methods that have been used, such as windowing and frequency analysis.

2.2.1. Window functions

The mathematical functions called "windows" are used in the analysis and signal processing to alleviate the problem in which the audio signal is not stationary. The windowing allows the analysis of stationary stages of the audio signal, and consists in grouping a number of consecutive samples in a segment, so as to process lately each segment individually. Taking small pieces of the signal, those sections could be considered stationary when the signal is processed, and thus avoiding the problem of non-stationary voice signal.

There are a lot of types of windowing, for example: rectangular windowing, Hanning, Hamming, Gauss, Triangular, etc **[1]**.

In this case the Hanning window has been chosen. It has the shape of a cycle of a cosine wave plus an offset so it is always positive.

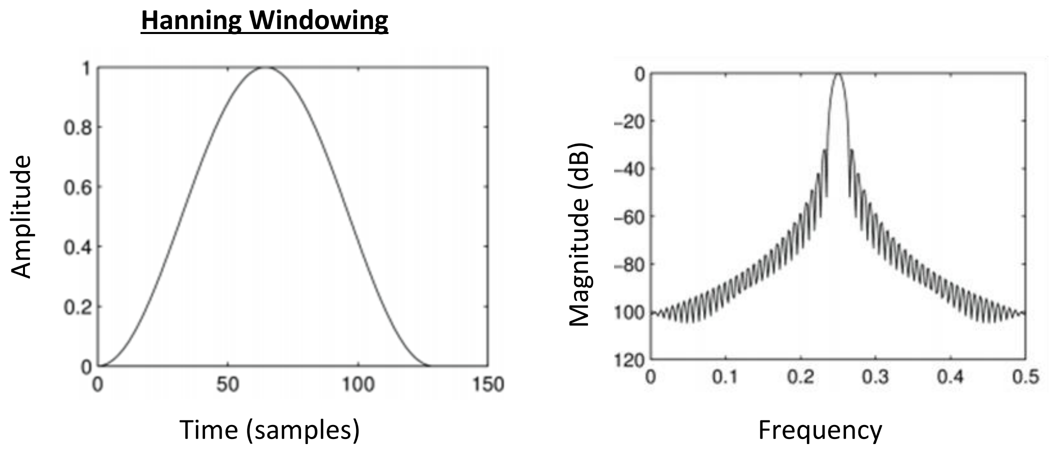


Figure 2.1 Example of Hanning Windowing.

By multiplying the signal by the Hanning window the beginning and the end tend to zero, avoiding problems when moving to infinite frequency signal.

By contrast, as it also adds distortion to both ends, forcing the area to zero modifies the signal.

2.2.2. Frequency analysis

Converting a signal to the frequency domain is the result of decomposing the signal into sinusoidal components. For this purpose, two mathematical tools are used, depending on the continuous or discrete nature of the signal.

For continuous signals, the Fast Fourier Transform (FFT) is employed, while for discrete signals or sequences the tool must be the Discrete Fourier Transform (DFT). This project focuses on the DFT, as the digital audio signals are discrete signals or sequences.

2.2.2.1. Discrete Fourier Transform (DFT)

The Discrete Fourier Transform, or the Fourier transform of a sequence x [n], is a function $X(e^{jw})$ [eq.2.1], continuous and periodic, with period 2π. It can be computed with the following expression:

$$X\left(e^{jw}\right) = \sum_{n=-\infty}^{\infty} x[n]e^{-jwn}$$

[2.1]

The inverse Fourier transform to sequences (IDFT) will return the original sequence, being its expression called synthesis:

$$x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X\left(e^{jw}\right)e^{jwn}dw$$

[2.2]

To synthesize again x[n] the inverse Fourier transform must be used.

As $X\left(e^{jw}\right)$ is a complex function with real and imaginary components, it can be represented as:

$$X\left(e^{jw}\right) = X_R\left(e^{jw}\right) + jX_I\left(e^{jw}\right)$$

[2.3]

Or, in polar form, the module and phase decomposition:

$$X\left(e^{jw}\right) = \left| X\left(e^{jw}\right)\right| \cdot \ e^{j\angle X\left(e^{jw}\right)}$$

[2.4]

Where $\left| X\left(e^{jw}\right)\right|$ is the module and $\angle X\left(e^{jw}\right)$ the phase.

2.3. The noise

When talking about noise in the field of acoustics sing, there is a necessity to define the concept of "noise" as any unwanted and annoying sound that can interfere with the reception of the sound wanted, or in a processing with it.

In general terms the noise can be classified into:

- *Additive noise:*
  The additive noise can be seen as the noise from various sources that coexist in the same acoustic environment.

- *Interfering signals:*
  In the case of voice signal, the interference signals come from other speakers than those of interest.

- *Reverberation:*
  This effect is produced by multipath propagation. It occurs in enclosed or semi-enclosed acoustical environments and it is a form of distortion.

- *Echo:*
  Usually it is produced by the coupling between the microphones and speakers. It is another form of distortion.

There are multiple studies devoted to each of these classes of noise or distortion, and these have led to different speech processing techniques designed to eliminate one of these types.

In this project there would be a focus on additive noise and possible ways of eliminating it.

2.3.1. Additive noise

The noise additive is considerate when the signal is formed by the addition of clean speech and noise. Thus, the noise reduction carries out the task of separating these two signals in the most optimal way.

An initial proposal would be to treat the noise elimination as a parameter estimation problem, where the optimal estimate of the clean speech can be carried out under the criteria of certain factors. For example, the factor MSE (Mean Squared Error) or the SNR (Signal to Noise Ratio) of the estimation of the clean speech versus the original audio will be an example.

In some cases, the results obtained using a method of noise removal are not considered optimal from the stand point of the listener, and here arises the subjective perception.

For this reason, subjective and objective aspects should be taken into account.

Upon the objectives to be achieved are:

- Optimize the objective criteria, MSE, SNR, etc.
- Optimize the perceived quality of the reconstructed signal (subjective criteria).

It can be considered as a pre-processing step to further voice processing and it can lead to an increase in the robustness of other systems (speech coding, speech recognition, etc...) against the noise.

The compliance objective determines the complexity and difficulty of the filtering method and hardware. An example of the variation of the number of hardware is the number of microphone or channels to be used. The more available channels, the more options for improved voice quality.

Although the situation of multiple microphones (or microphone “array”) is not the most common. An example of this is a mobile phone that only has a microphone through which voice and ambient noise are equally taken. This case consists of a single channel system.

The monochannel and multichannel techniques will be discussed later in detail.

When focusing in monochannel systems, Professor Schroeder was one of the firsts to propose an implementation of spectral subtraction, about 1958. Fifteen years later this model was applied to the field of digital signals. Around 1979, researchers like Jae S. Lim and Alan V. Oppenheim **[21]**, performed an analysis of the technique that existed in that moment, within the field of enhancing speech signals, and concluded that the reduction of noise was not only beneficial to the quality of the recovered voice, but also to the quality and intelligibility of linear predictive coding (LPC), broadly used in coding and voice recognition systems.

The techniques developed so far can be classified into three groups, depending on the noise reduction method: Linear adaptive filtering, Spectral subtraction and model based **[21]**:

- *Linear adaptive filtering:*
  The basis of adaptive filtering is to process the noisy signal with a linear filter that is adapted to remove noise, thus reducing the noise component and leaving the speech signal with the less possible amount of distortion.
  An example of this is the RLS estimation (Recursive least squares filter).

- *Spectral subtraction:*
  Spectral subtraction methods reduce noise through a spectrum estimation of the speech signal from the original noisy signal.
  An example would be the MMSE algorithm (Minimum-Mean-Squared-Error).

- *Based on model:*
  Reduction methods based on models reduce the Noise Ratio as a parameter estimation problem, which use mathematical models of voice generation.
  An example is the LP-Kalman technique (linear prediction).

2.3.2. Psychological and physiological effects of noise

The noise can be quite annoying in everyday life, as it interferes in activities such as study, work, sleep or even leisure time. It causes fatigue, forces us to make an extra effort and can cause irritation and headache. The noise with many decibels can cause temporary or even permanent deafness.

Psychologically, it has negative effects on worker productivity and efficiency because of a decrease in the concentration.

An example of this is the white noise that can be used to mislead people or as a sensory deprivation technique. Depriving the human from the other sounds, with white noise, can also be used to promote relaxation and sleep, or to mask other sudden and distressing noises.

2.3.3. Classification of the types of noise depending on their nature.

According to its nature, noise can be classified into different types:

- *Acoustic noise:*

  It emanates from moving, vibrating, or colliding sources. This type of noise is the most typical in everyday life. It is generated by sources such as computer fans, air-conditioners, traffic, wind, people talking, rain, etc.

- *Electrostatic noise:*

  This kind of noise is generated by the presence of a voltage with or without current flow.

  Fluorescent lighting is one of the most common sources of electrostatic noise.

- *Electromagnetic noise:*

  Present at all frequencies and in particular at radio frequencies. All the electrical devices used to transmit and receive signal generate electromagnetic noise (radio, television, etc.).

- *Processing noise:*

  It is the noise that results from the digital or analogue processing of signals. The quantization noise in the digital coding of speech or images would be an example, or noise due to packet loss in digital communication systems.

- *Channel distortions, echo and fading:*

  This noise is the result of non-ideal characteristics of communication channels. The mobile phone communications are particularly sensitive to the propagation characteristics of the channel.

2.3.4. Classification depending on its frequency or time characteristics

Depending on its frequency or time characteristics, a noise process can be classified into one of several categories:

- *Narrowband noise:*
  A noise process with a narrow bandwidth such as a 'hum' from the electric network (50/60 Hz).

- *White noise:*
  It is a purely random noise with a flat power spectrum. White noise contains theoretically all frequencies with equal intensity.

- *Band-limited noise:*
  It is a noise with a flat spectrum and band-limited, which usually covers the limited range of the device or the signal of interest.

- *Colored noise:*
  Non-white noise or any wideband noise whose spectrum has a non-flat shape. Examples are pink noise, brown noise and autoregressive noise.

- *Impulsive noise:*
  It consists of short-duration pulses of random amplitude and random duration.

- *Transient noise pulses:*
  This consists of pulses of relatively large length.

2.3.4.1. White noise

The white noise is defined as an uncorrelated noise process with equal power at all frequencies. The values of the signal at two different times are statistically uncorrelated. As a result, the power spectral density (PSD) is constant, presenting a flat graph [figure 2.2].

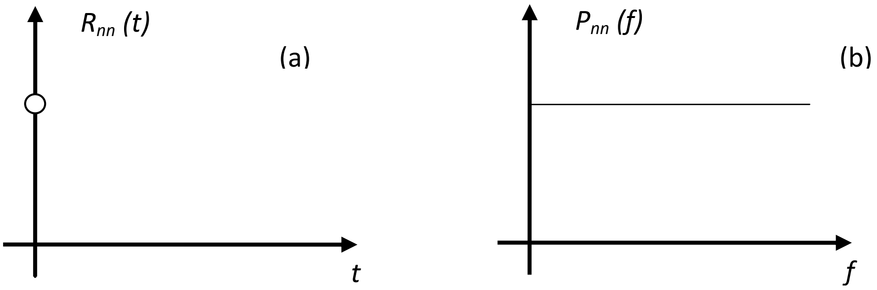


Figure 2.2. White noise Autocorrelation (a) and its power spectrum (b).

A noise that has equal power in all frequencies in the range of ± π must have an infinite power, although this is only a theoretical concept.

The shape of white noise is as follows:

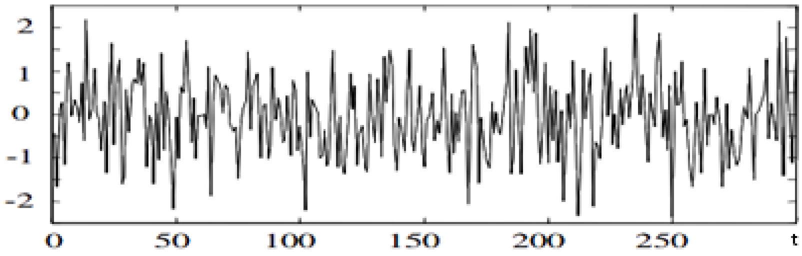


Figure 2.3. White noise.

This type of noise is not always annoying, since its features can even make it useful.

The following uses can be highlighted:

- In linear time invariant systems, it is used to determine the transfer function. In architectural acoustics, the transfer function is used to measure the acoustic insulation and room reverberation.

- In audio synthesis (electronic music), it is used to synthesize the sound of percussion instruments, deaf or speech phonemes.

2.3.4.2. Colored Noise

Although the concept of white noise is very important within the field of telecommunication systems, many other noise processes are considered non-white.

Therefore, the term of colored noise includes any broadband noise with a spectrum that is not flat. Examples of coloured noise are mostly audio frequency noise, caused by cars movement, the fan noise from computers or the background noise of people talking. All these noises have a spectrum that is not white, and which have predominantly low frequencies.

If white noise goes through certain channel, and changes its characteristics, it becomes colored noise. The two most popular classes of colored noise are pink and brown noise.

In **brown noise**, spectral density is inversely proportional to frequency [figure 2.4].

Hence, lower frequencies have more energy. It decreases in power by 6 dB per octave (20 dB per decade).

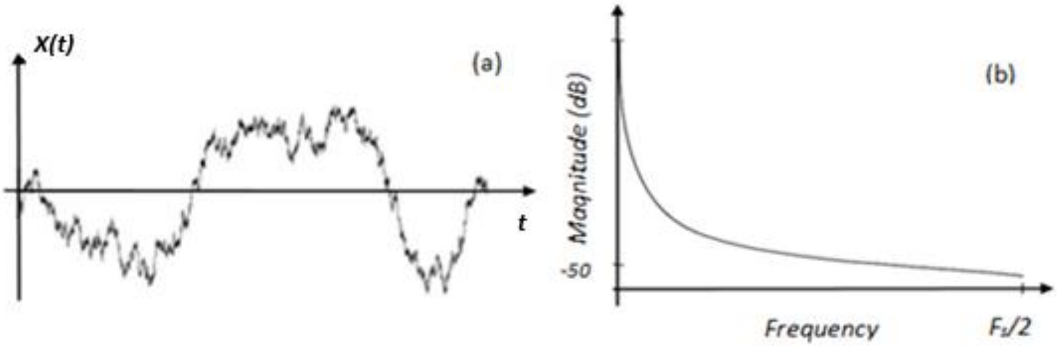


Figure 2.4 . A brown noise signal (a) and its magnitude spectrum (b).

The **Pink noise** is used to make acoustic measurements, and in the practice it is used to equalize room acoustics and to perform audio calibration.

This noise is also characterized by a spectral density, which is inversely proportional to frequency [Figure 2.5]. Pink noise shows a very particular characteristic, and this is the main reason why the noise level would be constant when processed through an octave band filter. Octave band filters, as well as one-third of octave band filters, are proportional between them and thus whenever we lower an octave, we double the bandwidth. Hence, the pink noise decreases 3dB per octave, just the rate at which the width increases band, correcting the integrated level of total noise. In conclusion, pink noise has a constant noise level in all octave bands.

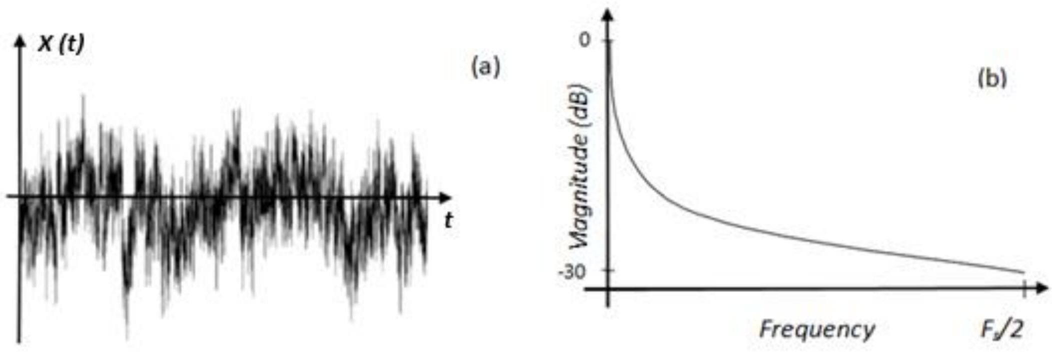


Figure 2.5. A pink noise signal (a) and its magnitude spectrum (b).

2.3.4.3. Impulsive noise

The impulsive noise consists of short duration pulses, due to a variety of sources such as noise switches, grooves or surface degradation of audio recordings, "clicks" of computer keyboards, etc.

One of the most significant characteristic of this kind of noise is that, supposing that the signal is ideal, in the time domain the signal is a delta however in the frequency domain it’s a constant. This characteristic is appreciated in the next figure:

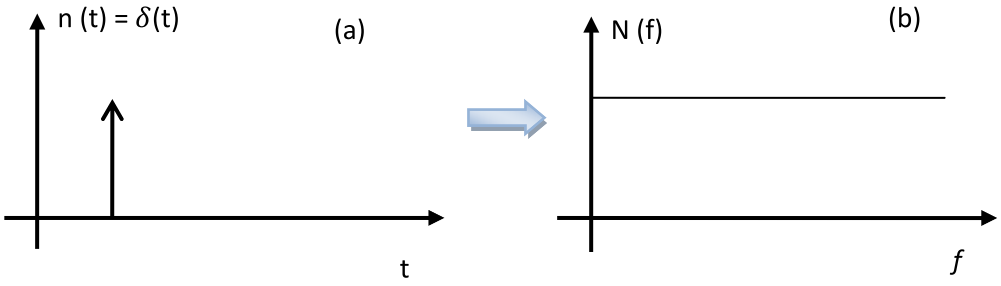


Figure 2.6. Ideal pulse in the time domain (a) and in frequency domain (b).

In communication systems, a real boost noise lasts more than one sample. For example, in the context of audio signals, one sharp pulse with short-duration of up to 3 milliseconds (60 samples at a 20 KHz sampling rate) may be considered as impulsive noise.

In communication system, an impulsive noise originates at some point in time and space, and then propagates through the channel to the receiver. This leads to discomfort in speaking. The figure 2.7 is the a real signal where we can appreciate the differences with the ideal signal [figure 2.6]

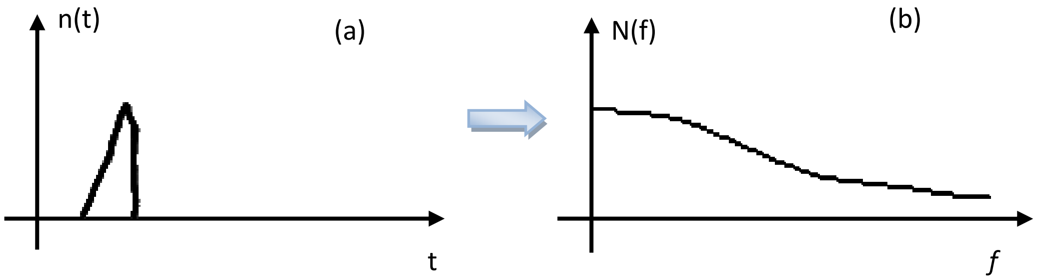


Figure 2.7. Real pulse in the time domain (a) and in frequency domain (b).

Yet often, this kind of noise is not considered as impulsive noise. The noise received is in turn dispersed in time and coloured by the channel itself, so it can be considered as the impulse response of the channel.

In general terms, the characteristics of communication channels can be linear or non-linear, stationary or time-varying. Indeed, most of the channels present non-linear responses to large amplitude pulses [figure 2.8].

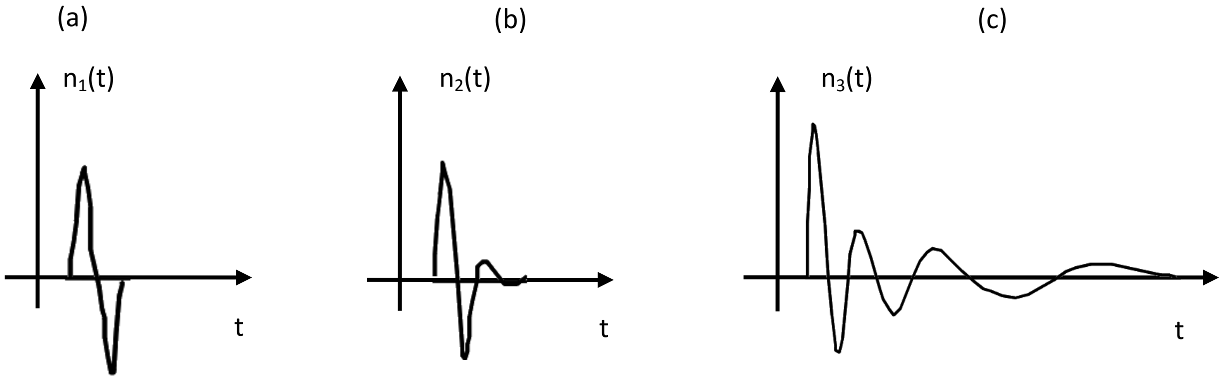


Figure 2.8. Variation of the impulse response of a non-linear system with the increasing amplitude impulse.

2.3.4.4. Transient Noise Pulses

These usually present sharp pulse profiles, which are followed by a drop formed by low frequency oscillations [Figure 2.9]. The initial pulse is usually the result of some impulsive interference (external or internal), and the fluctuations are usually caused by the communication channel resonance, which is excited by the initial pulse. Therefore, it would be considered as the response of the channel to the initial pulse.

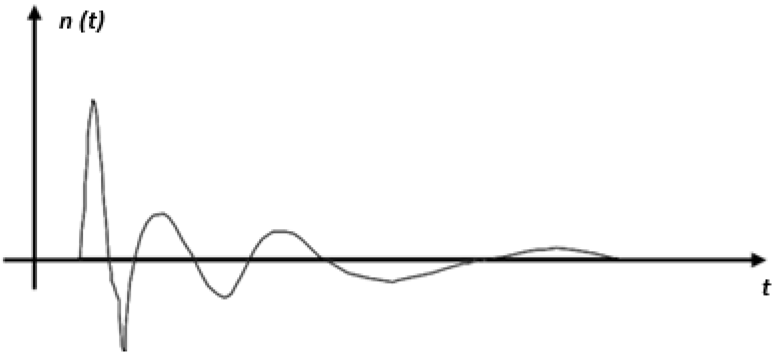


Figure 2.9. Example of transient pulse.

Thermal noise is based on thermodynamic concepts, and it is associated to the particles random motion, dependent on the temperature, as for example gas molecules in a container or electrons in a conductor.

The average of these random movements tends to zero, although the problem is the fluctuations over this average, which cause thermal noise. For example, the movements and collisions of gas molecules in a closed space, which generate random fluctuations, above the average pressure.

As temperature rises, the kinetic energy of the molecules and the thermal noise increases.

Similarly, if an electrical conductor has a large number of free electrons, ions along the conductor randomly vibrate around their equilibrium positions obstructing the movements of electrons. The free movement of electron flow forms spontaneously random noise or thermal noise. The electrons move to higher energy states because the conductor temperature increases, thus increasing random stream flows.

2.3.4.5. Electromagnetic noise

Every electrical device that generates, uses, or transmits energy is a potential source of electromagnetic noise and interference to other systems. High voltage or current levels, together with the proximity to electric circuits or devices, produce most of the induced noise.

The most common sources of electromagnetic noise are radio receivers, televisions, microwave transmitters, transformers, cellular phones, motors, generators, fluorescent tubes and thunderstorms.

The sources can be divided into two types:

- *Electrostatic noise.*
- *Magnetic noise.*

These two types are different and they need different shielding measures. But the problem is that most of the noise sources produce both types together.

Electrostatic fields are generated by the presence of voltage, with or without passage of current. Fluorescent lights are one of the most common sources of electrostatic noise.

For magnetic noise, motors and transformers are current driven examples, and without current, one example is the Earth's magnetic field.

2.3.4.6. Channel Distortions

When propagating through a channel, the signals are shaped and distorted by the frequency response and attenuation characteristics of the channel.

In the case of the analogue audio, there are two types of distortion in the channel: module and phase distortion.

Furthermore, in radio communications, there is the effect that produces a signal that is transmitted through different paths to the receiver. This results in multiple versions of the signal with different delays and attenuations. Channel distortions can degrade the signal or even interrupt the communication process.

2.3.5. Denoising techniques

Voice communication under noisy conditions implies a great effort. Certain sounds are masked with noise, making it difficult to hear audio and making speech intelligible.

Other forms of degradation of speech are the reverb and channel distortions due to multiple factors, such as the quality of the recording and reception equipment, together with the characteristics of the transmission channel and effects, due to different types of digital signal encoding used for transmission.

The purpose of all these techniques to improve speech intelligibility is to increase the speech audio signal, so that those parts that were incomprehensible, after this process, become clear. For this purpose, programs usually remove noise signals as much as possible, trying not to distort the audio signal of interest.

We can establish a division between the processing techniques by the number of channels employed for audio input: monochannel and multichannel.

2.3.5.1. Monochannel techniques

The first monochannel techniques consisted in to use the Wiener filter and in other techniques based on the periodicity of voiced speech, as the adaptative comb filter or the harmonic selection **[21]**.

The main monochannel techniques are those based on a direct estimation of the spectral amplitude in a short time period and, and which are named “Spectral Subtraction”.

The basic principle of the spectral subtraction is the Boll method **[5]**, created by Steven F. Boll. The main propose is to obtain, during segments without speech, a noise spectrum estimator of the contaminating noise, for later subtraction in the frequency domain of the instantaneous spectrum of the input signal in each moment.

Spectral subtraction is not perfect, and this aspect must be taken into account. Subtracting an estimation of noise, rather than the actual noise spectrum at each instant, can create spectral peaks that do not belong to the original signal. When returning to the time domain, these results in very short duration tones whose frequencies vary from frame to frame. This effect is called *musical* noise and it should be avoided.

However, the study of Berouti **[3]** introduced a possible solution to this musical noise. He established a minimum threshold below which spectral subtraction cannot be applied, attenuating musical noise but increasing the mean noise level.

Later, Lockwood and Boudy **[12]** proposed another solution: non-linear spectral subtraction. The thresholds and the factors are not constant because they will apply in less or more measure the spectral subtraction depends on the frequency. This optimizes the amount of noise that we can subtract without the appearance of *musical* noise.

Another option is to take into account the perception of people using auditory models and including the characteristics of human hearing and masking ability of certain sounds. The overall of this procedure is too complicated and it does not work well in general terms.

The last method presented is called multiband subtraction **[18]**. This considers that the noise is coloured and does not affect equally the whole spectrum, since it is not linear.

2.3.5.2. Multichannel techniques

There are three primary techniques according to the number of audio inputs:

- *Two channels:*
  This technique requires two input channels to make use of adaptive filtering so as to improve the signal contaminated. The limitation of this technique is the necessity to take in one of the channels as a good reference of the input noise. It is widely used in aviation.

- *Multiple input channel. Microphone arrays:*
  It takes into account two factors:

  - Additive acoustic noise, which is the one that reaches the receiver from unwanted sources.
  - The reverberation due to the transmission of signals between two points in the same room.

  These factors will depend on the characteristics of the room and the amount, type and position of the sound sources. By using microphone arrays, achieving a receive beam which steers its direction is possible, so as to get the desired signal by combining the outputs of the microphones while attenuating most of the other signals or noise. The main drawback is the need of a specific hardware.

- *Binaural process:*

  This is the best system because it is similar to the human ear's system.

  The human brain is capable of focusing on a conversation formed by two different signals that reach each ear, ignoring all others sounds. For example, it is able to distinguish one instrument between several other instruments sounding simultaneously.

  This ability is the result of the combination of two phenomena. First, the "binaural processing" based on the use of both ears to improve human hearing capabilities (e.g. the discrimination sources or spatial localization). Second, the analysis of the auditory scene, whereby the brain reconstructs the outside world with the sound signals it receives.

  For years scientists have done multiple and diverse experiments in which they have studied the functioning of the human ear. This takes into account many variables, since hearing depends not only on the ears, but also on the rest of the environment, such as the hair and the whole head, being also very influential.

Chapter 3

3. Technical solution design: Spectral subtraction

3.1. Introduction and explanation of the algorithm

From the total amount of options that have been studied, the Spectral Subtraction has been selected the best choice for diverse reasons. Primarily, there will be usually only one signal being recorded at a time with no reference noise. Furthermore, spectral subtraction requires only one signal, and consequently, employing a monochannel technique. Throughout this project, the investigation and implementation of the spectral subtraction will be faced, since it is the base of the other algorithm for denoising.

The spectral subtraction is a technique which implies low computational cost and consists in a simple concept.

There are three fundamental requirements involved in an audio signal noise removal method:

- Improved signal to noise ratio (SNR).
- Intelligibility and naturalness of the improved signal.
- Computational simplicity.

This method assumes that speech and noise are incorrelated, and the noise is added in the time domain. Therefore, the power spectrum of the noise signal is the sum of the power spectrums of the speech and noise. It is also important to mention that the noise characteristics will vary slowly with respect to the voice signal. Hence, there is a need to assume the fact that the noise is stationary, and with constant variance. In consequence, to suppress noise from the contaminated signal, its spectrum can be estimated during a voiceless segment. Nevertheless, unwanted effects would appear.

The noisy signal can be defined in the time domain as the sum of two components [eq.3.1]:

$$y(t) = x(t) + n(t)$$

[3.1]

Where: $y\ (t)$ Is the noisy signal.

$x(t)$ Is the original signal without noise.

$n(t)$ Is the noise.

In the frequency domain the equation 3.1 is:

$$Y(f) = X(f) + N(f)$$

[3.2]

The input signal is windowed in segments of data, containing a predefined number of samples, being these subsets named frames. In this case, the Hanning window is applied. Then, it is transformed to the frequency domain by means of the DFT. The windowing compensates the effects caused by discontinuities at the edges of each data frame.

The process of the subtraction, in a very general way, is:

$$\left|\hat{X}(f)\right|^b = |Y(f)|^b - \alpha\left|\overline{N(f)}\right|^b$$

[3.3]

Where: $\left|\hat{X}(f)\right|^b$ is the estimate of the original signal spectrum$|X(f)|^b$.

$|Y(f)|^b$ is the contaminated signal.

$\left|\overline{N(f)}\right|^b$is the noise spectrum averaged over time.

α is the variable that controls the amount of noise subtracted, where $\alpha = 1$ means that the complete calculated noise is subtracted. If $\alpha > 1$, the noise signal is amplified for an over-subtraction of the noise signal. The spectral subtraction can be done in power or in magnitude.

In the periods of absence of voice, the noise is averaged as follows:

$$\left|\overline{N(f)}\right|^{b} = \frac{1}{K}\sum_{i=0}^{k-1} |N_i(f)|^{\mathrm{b}}$$

[3.4]

If b=1, magnitudes are subtracted, while if b=2, we are working with power. The *i* sub index indicates the frame number and k frames which are expected during the averaging period. The average spectrum of the noise can be obtained as the output of a digital low pass first order filter:

$$\left|\overline{N_{\iota}(f)}\right|^{b} = \rho\left|\overline{N_{\iota-1}(f)^{b}}\right| + \ (1-\rho)\ |N_i(f)|^{b}$$

[3.5]

For a typical filter, ρ is usually a value between 0.85 and 0.99 **[6].**

My choice has been working in the frequency domain with the power spectrum subtraction. So, the value takes in the [eq. 3.6] for b is 2. The equation will be there:

$$\left|\hat{X}(f)\right|^{2} = \ |Y(f)|^{2} - \overline{|N(f)|^{2}}$$

[3.6]

In order to return to the time domain, the estimated magnitude spectrum $\left|\hat{X}(f)\right|$ is combined with the phase of the noisy signal, and then transformed into the time domain through IDFT (Inverse Discrete Fourier Transform), the inverse DFT process (Discrete Fourier Transform) [eq. 3.7].

$$\hat{x}(t) = \sum_{k-0}^{N-1} \left|\hat{X}(k)\right| e^{j\theta_{y}(k)} - e^{-j\frac{2\pi}{N}km}$$

[3.7]

In the previous equation, $\theta_y(k)$ is the phase of the signal with noise $Y_k(f)$ that was mentioned earlier in the process.

The following equation is used to avoid negative results:

$$T[|\hat{X}(f)|] = \begin{cases} |\hat{X}(f)| & |\hat{X}(f)| > \beta|Y(f)| \\ f_n[|Y(f)|] & otherwise \end{cases}$$

[3.8]

The figure 3.1 shows a block diagram of the spectral subtraction algorithm.

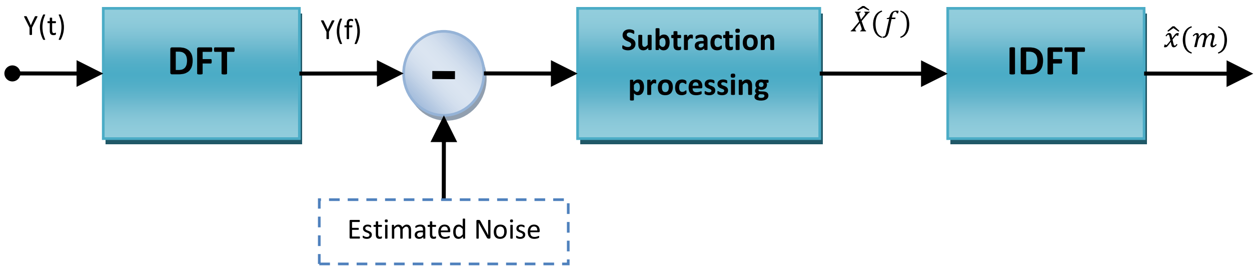


Figure 3.1. Block diagram of spectral subtraction.

If we show it graphically, the following would happen:

On one side, we have the original signal is presented [Figure 3.2], from which the noise signal is estimated [Figure 3.3].

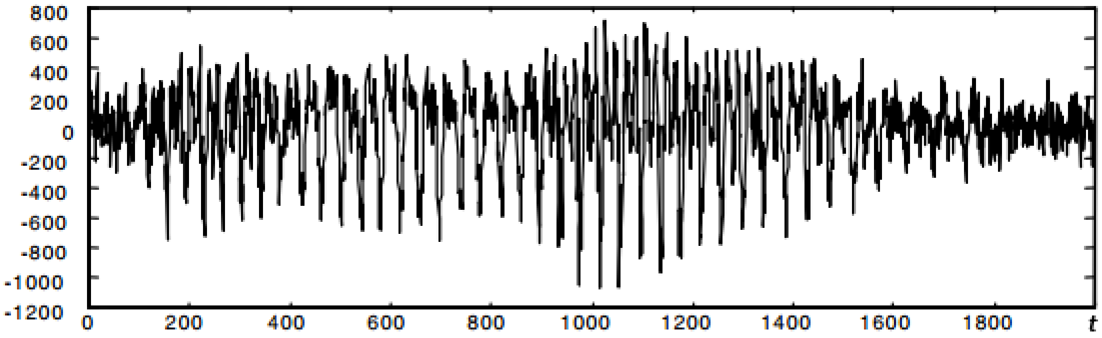


Figure 3.2. Original noise signal.

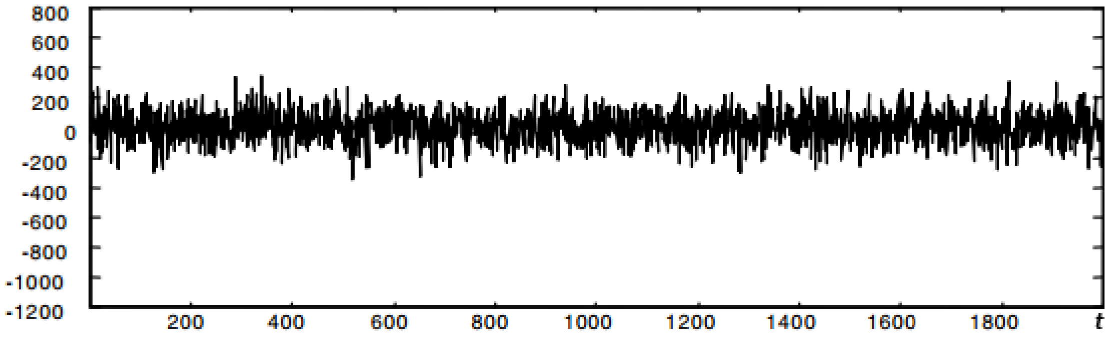


Figure 3.3. Noise estimate obtained from the subtraction of Figure 3.2 and Figure 3.4.

When subtracting the estimated noise [Figure 3.3] to the original signal [Figure 3.2], the reconstructed signal is obtained after applying the spectral subtraction [Figure 3.4].

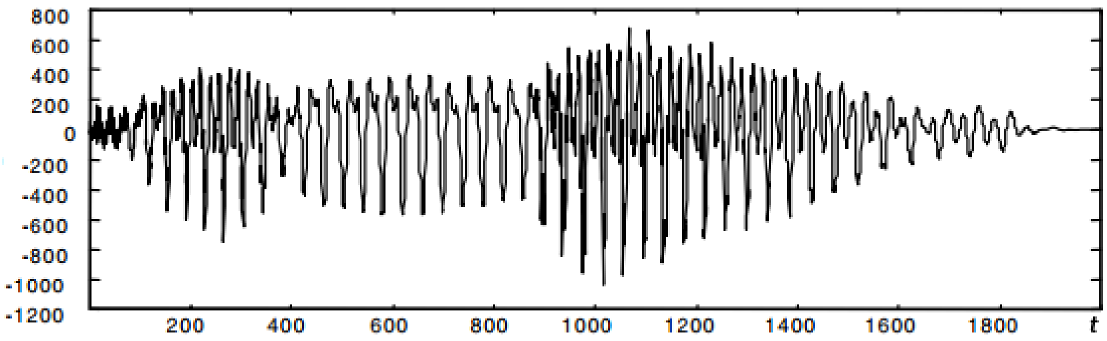


Figure 3.4. The restored signal after applying the spectral subtraction (c).

The [Figure 3.4] is the cleaned signal, i.e., this is the result of apply the spectral subtraction to a signal with noise.

3.2. Problems with musical noise and distortions

The main problem of the spectral subtraction is the non-linear distortion of the processing, caused by the nature of random noise.

There are three sources of distortion:

- Variations of instantaneous power of noise.
- Signal and noise cross-product terms.
- Non-linear mapping of the estimated spectrum which falls below a threshold.

The distortion that appears most often is the one due to non-linear mapping of the negative estimations or small valued, in the estimates. This distortion produces a metallic sounding noise called *musical tone noise,* as a result of its narrow spectral band.

The success of the spectral subtraction depends on the ability of the algorithm to reduce variations of noise and compensate the processing distortions. In the worst case, the residual noise may have the following forms:

- A sharp though peak in the spectral signal. [Figure 3.5].
- Isolated narrow frequency bands.[Figure 3.5].

Near a frequency of high amplitude, the noise mentioned in the first case is often masked, and made inaudible by the high signal energy. The principal cause of degradation of the signal is the second case, which causes the *musical tones* previously mentioned, and formed by narrow bands of low level and short duration.

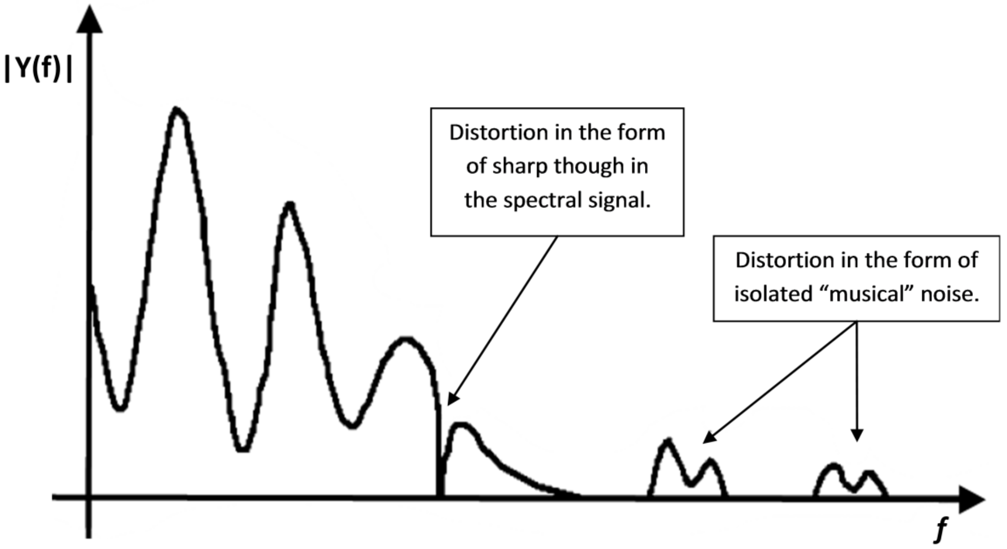

|Y(f)|
Distortion in the form of sharp though in the spectral signal.
Distortion in the form of isolated "musical" noise.
f

Figure 3.5. Example of distortions in spectral subtraction.

3.3. Algorithms to perform the spectral subtraction

The various algorithms performing the spectral subtraction are:

3.3.1. Boll Algorithm

This algorithm is the basis of all types of spectral subtraction. The main idea of this is to obtain, during the absence of speech segments, a noise spectrum estimation of the pollution noise and then subtract this estimation in the frequency domain. It performed to produce the type of noise appointed previously, musical noise. From this method, others are developed.

The main intention of the algorithm is to obtain, in the absence of voice, an estimator of noise, so as to subtract it lately from the noisy signal.

Here is when an audio signal with noise appears. The noisy signal is considered to be the sum of the signal and the noise.

We take the equation [3.1] is taken, together with the Fourier Transform:

$$Y(e^{jw}) = X\left(e^{jw}\right) + N\left(e^{jw}\right)$$

[3.9]

Where: $$y(k) \Leftrightarrow Y(e^{jw})$$

$$Y(e^{jw}) = \sum_{k=0}^{R-1} y(k)e^{-jwk}$$

$$y(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} Y\left(e^{jw}\right)e^{jwk}dw$$

[3.10]

As we cannot have *N*, we replace it with an estimate of *N*, calculated during the periods without vocal activity. Consequently, $N(e^{jw})$ would be replaced by an estimate of the noise $E\{\left|N\left(e^{jwk}\right)\right|\}$. The spectral estimator is as follows:

$$\hat{X}\left(e^{jw}\right) = \left[\left|X\left(e^{jw}\right)\right| - E\{\left|N\left(e^{jw}\right)\right|\}\right]e^{j\theta_x(e^{jw})}$$

[3.11]

As already discussed, a spectral error appears:

$$\varepsilon\left(e^{jw}\right) = \hat{X}\left(e^{jw}\right) - X\left(e^{jw}\right) = N\left(e^{jw}\right) - E\{\left|N\left(e^{jw}\right)\right|\}e^{j\theta_x}$$

[3.12]

To try to avoid the error, three processes would take place afterwards. We talk about this improvements are presented in the chapter 4, since these three processes in all the algorithms are applied there.

3.3.2. Berouti Algorithm

The aim of this algorithm is the reduction of a musical noise, as a result of the subtraction of average noise spectrum of an instantaneous spectrum, using over-subtraction and the value of the minimum spectrum.

Thus, this algorithm introduces a possible solution to musical noise. The researcher established a minimum threshold below, in which spectral subtraction cannot be applied, attenuating musical noise but increasing the mean noise level.

This spectral subtraction is carried out not in the power domain, but raised to a power γ spectrum. For the ith degraded speech frame, $Y_i(w)$, if you have an estimate of the noise power spectrum like $\left|\widehat{N}_i(w)\right|^2$, the plot is achieved as follows:

$$T_i(w) = |Y_i(w)|^{2y} - \alpha(SNR_i)\left|\widehat{N}_i(w)\right|^{2y}$$

[3.13]

Then, a full wave rectification is applied, obtaining the power spectrum estimate of the clean speech as:

$$\left|\hat{X}_i(w)\right|^2 = \begin{cases} T_i(w)^{\frac{1}{\gamma}} & ,if\ \ T_i > \beta\left|\hat{Y}_i(w)\right|^2 \\ \beta\ |\hat{y}_i(w)|^2 & \qquad otherwise \end{cases}$$

[3.14]

Where $0 < \beta \leq 1$, being β the value of the minimum power and $\alpha(SNR_i)$ a factor of pre-subtraction dependent on the signal to noise ratio of the frame I, with values between 1 and 5.

The resulting SNR will be:

$$SNR(dB) = 10\log\left(\frac{\sum_{k-0}^{N-1}|Y(k)|^2}{\sum_{k-0}^{N-1}\left|\widehat{N}(k)\right|^2}\right)$$

[3.15]

From this equation α is given by:

$$\alpha\begin{cases} 5 & SNR < -5dB \\ \alpha_0 - \dfrac{3}{20}SNR & -5\ dB \leq SNR \leq 20dB \\ 1 & SNR > 20dB \end{cases}$$

[3.16]

When performing an over-subtraction (α> 1), the spectrum is more attenuated resulting in a reduction of residual noise and an increase in audible distortion (musical noise). To avoid this issue, a suitable value for β must be selected experimentally (for example β=0.002) as this parameter limits the amount of noise removed for small values of the speech where the subtraction noise can lead to negative values.

3.3.3. Lockwood and Boudy: non-linear spectral subtraction

This algorithm is a non-linear spectral subtraction. The subtraction and thresholds established are frequency dependent. The fact that this algorithm is non-linear is an advantage, since the subtraction algorithms fixed parameters are not well adapted to the characteristics and noise level variations, differing from the Berouti's method, in which $\alpha(SNR_i)$ is a function of the frequency, i.e. the frequency-dependent SNR.

The amount of subtraction is smaller for high SNR spectral components and increases for low SNR spectral components.

To calculate the $\alpha SNR_i\ (w)$, the approach is the same as in the method of Berouti, yet taking into account all frequencies. The value of the SNR, in dB, would be:

$$SNR_i\ (w) = 10\, log_{10}\left(\frac{|Y(w)|^2}{\left|\hat{R}(w)\right|^2}\right)$$

[3.17]

From this equation [eq. 3.17], the $\alpha SNR_i\ (w)$ will be:

$$\alpha SNR_i\ (w) \ = \begin{cases} 5 & SNR < -5dB \\ \alpha_0 - \dfrac{3}{20} SNR & -5\ dB \leq SNR \leq 20dB \\ 1 & SNR > 20dB \end{cases}$$

[3.18]

The next step is the spectral subtraction:

$$|X_i(w)| = |Y_i(w)| - \alpha\_SNR_i(w)\left|\hat{N}_i(w)\right|$$

[3.19]

The estimation of noise would be:

$$\left|\widehat{N}_i(w)\right| = \sqrt{\frac{\alpha_i}{1+\gamma SNR_i}}$$

[3.20]

The parameter γ determines the influence of the SNR in the estimation of noise.

Within the next equation α_i is the maximum value of the last "M" noise spectrum:

$$\mathrm{\alpha_i(w)} = \max_{\mathrm{\tau=i-M}}\left(\left|\mathrm{\widehat{R}_\tau(w)}\right|^2\right)$$

[3.21]

Finally, the estimated signal, with a half-wave rectification, is:

$$\left|\widehat{X}_i(w)\right| = \begin{cases} \left|\overline{X_\iota(w)}\right|, & if \;\; \left|\overline{X_\iota(w)}\right| > \beta\left|\overline{Y_\iota(w)}\right| \\ \beta\left|\overline{Y_\iota(w)}\right| & \qquad otherwise \end{cases}$$

[3.22]

Where β is the minimum value of the spectrum.

3.3.4. Multiband Algorithm

All the methods that we have discussed before estimate the noise throughout all the spectrum of speech signal. However, the real noise is coloured and does not affect equally all the spectrum of the signal.

Depending on the frequency, coloured noise affects in more or less depth. And this takes it into account in the spectral subtraction to get a subtraction that fits the voice signal, the one we are working with, and get better musical noise reduction.

The target of the method is to estimate a factor able to subtract the required amount of the spectrum of noise, depending on the frequency band, so as to avoid its voice destruction.

The additive noise is assumed to be like stationary and non-correlated with the clean signal. There is an estimation of the noise $\left|\widehat{N}_i(w)\right|$ during the periods without speech, because there is no possibility obtains the noise spectrum directly from the noise signal. On the basis of the method of Berouti [eq. 3.14] we have the next equation:

$$T_i(w) = |Y_i(w)|^{2y} - \alpha(SNR_i)\left|\widehat{N}_i(w)\right|^{2y}$$

[3.23]

This supposition (eq. 3.23) assumes that the noise affects the entire signal equally and also the over-subtraction factor "α" is constant in the entire signal too. Though this does not conform to reality. The best way to be faithful to reality is divide the spectrum in *B* non-overlapping bands and apply spectral subtraction to each band separately.

The estimation of clean spectrum voice in the band "*b*" is obtained:

$$\left|\hat{X}_b(w)\right|^2 = |Y_b(w)|^2 - \alpha_b \delta_b \left|\hat{Y}_b(w)\right|^2 \qquad \mathrm{f}_b \leq \mathrm{w} \leq \mathrm{l}_b$$

[3.24]

Where $\mathrm{f_b}$ is the first frequency of the band *b* of frequency, and $\mathrm{l_b}$ is the last. And α_b if the over subtraction factor in the band *b.*

SNR is also needed. Hence:

$$SNR_b(w) = 10\log_{10}\left(\frac{\sum_{k=f_b}^{l_b}|Y_b(k)|^2}{\sum_{k=f_b}^{l_b}\left|\widehat{N}_\mathrm{b}(k)\right|^2}\right)$$

[3.25]

$$\alpha_b = \begin{cases} 5 & SNR_b < -5dB \\ 4 - \dfrac{3}{20} SNR_b & -5\ dB \leq SNR_b \leq 20dB \\ 1 & SNR_b > 20dB \end{cases}$$

[3.26]

Within the equation [3.24], δ_b is a factor that can be configured for each frequency band independently, so as to be adapted to the noise elimination properties. The valued used was obtained experimentally in other studies **[6].**

$$\delta_b = \begin{cases} 1 & f_b < 1\,kHz \\ 2.5 & 1\,kHz \leq \ f_b \leq \frac{\mathrm{f_s}}{2} - 2kHz \\ 1.5 & f_b > 20dB \end{cases}$$

[3.27]

In order to calculate the diverse frequencies by band we take into account that after FFT the spectrum is going to be in an interval of frequencies between $-\frac{\mathrm{f_s}}{2}$ and $\frac{\mathrm{f_s}}{2}$.

Therefore, the number of frequencies by band would be:

$$Number\ of\ frequencies\ by\ band = \frac{\frac{number\ of\ points\ of\ FFT}{2}}{Number\ of\ bands}$$

[3.28]

Using α_{b}, we can control the level of noise subtraction in each band can be controlled. Moreover, the use of several frequencies band and the weighting δ_{b} give an additional control in each band.

Chapter 4

4. Implementation

4.1. Stages of development

In a very general way, the process in order to clean the voice would be the one which follows:

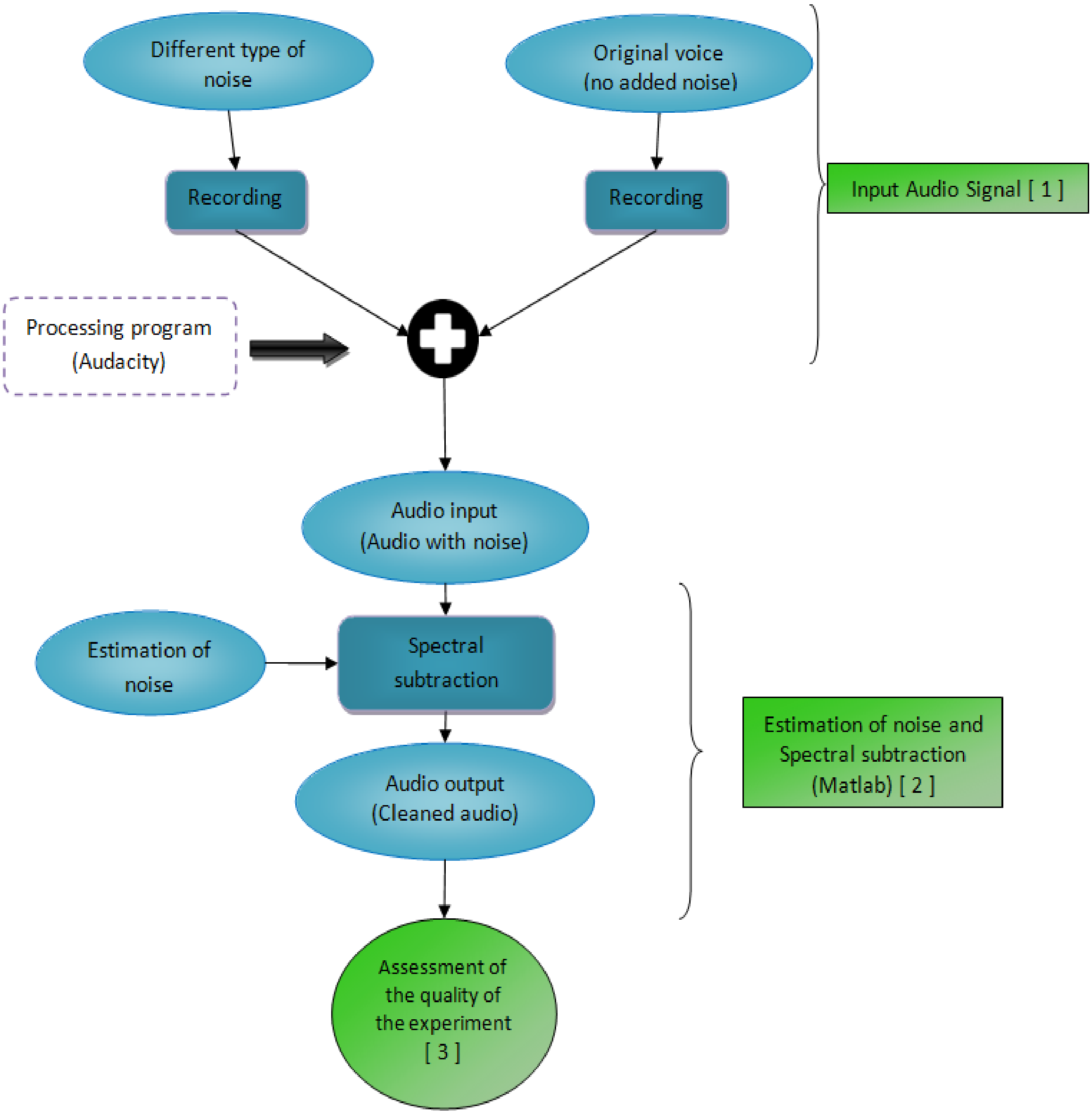

Different type of noise
Original voice (no added noise)
Recording
Recording
Input Audio Signal [1]
Processing program (Audacity)
Audio input (Audio with noise)
Estimation of noise
Spectral subtraction
Audio output (Cleaned audio)
Estimation of noise and Spectral subtraction (Matlab) [2]
Assessment of the quality of the experiment [3]

Figure 4.1. Application Process.

- *Input audio signal* [Figure 4.1(1)]*:*

  Clean voices and different types of noise have been recorded separately to create a dataset for the experimental work.

  Audio voice records include male and female speakers with speeches of diverse length, since its main characteristics are different. A variety of noises have been recorded, like white noise, pink noise, impulsive noise, etc.

  Noise records are added in the audio voice, which result is an audio signal containing noise. This is made with an audio processing program (Audacity).

- *Estimation of noise and Spectral subtraction* [Figure 4.1(2)]*:*

  The estimation of noise and the spectral subtraction was carried out in Matlab. This will be discussed in paragraph number 4.3.

- *Assessment of the quality of the experiment* [Figure 4.1(3)]*:*

  When the application is finished, and study of its effectiveness and its results must be tried.

  This will be discussed in paragraph number 4.3.5., together with its results in the paragraph number 5.6.

4.2. Matlab: Graphical Interface, estimation of noise Spectral subtraction

The section which follows now describes the software implementation of the algorithms presented in Section 3.3.

A very general outline of the program would be the Figure 4.2. And focusing more on the spectral subtraction is the Figure 4.3.

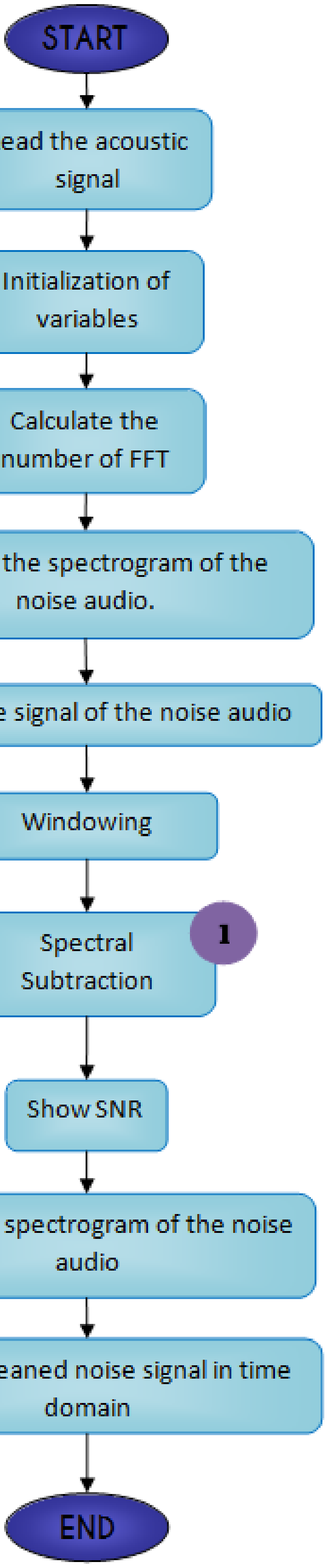

START
Read the acoustic signal
Initialization of variables
Calculate the number of FFT
Show of the spectrogram of the noise audio.
Display the signal of the noise audio
Windowing
Spectral Subtraction
1
Show SNR
Plot of the spectrogram of the noise audio
Plot the cleaned noise signal in time domain
END

Figure 4.2. Graphical Interface.

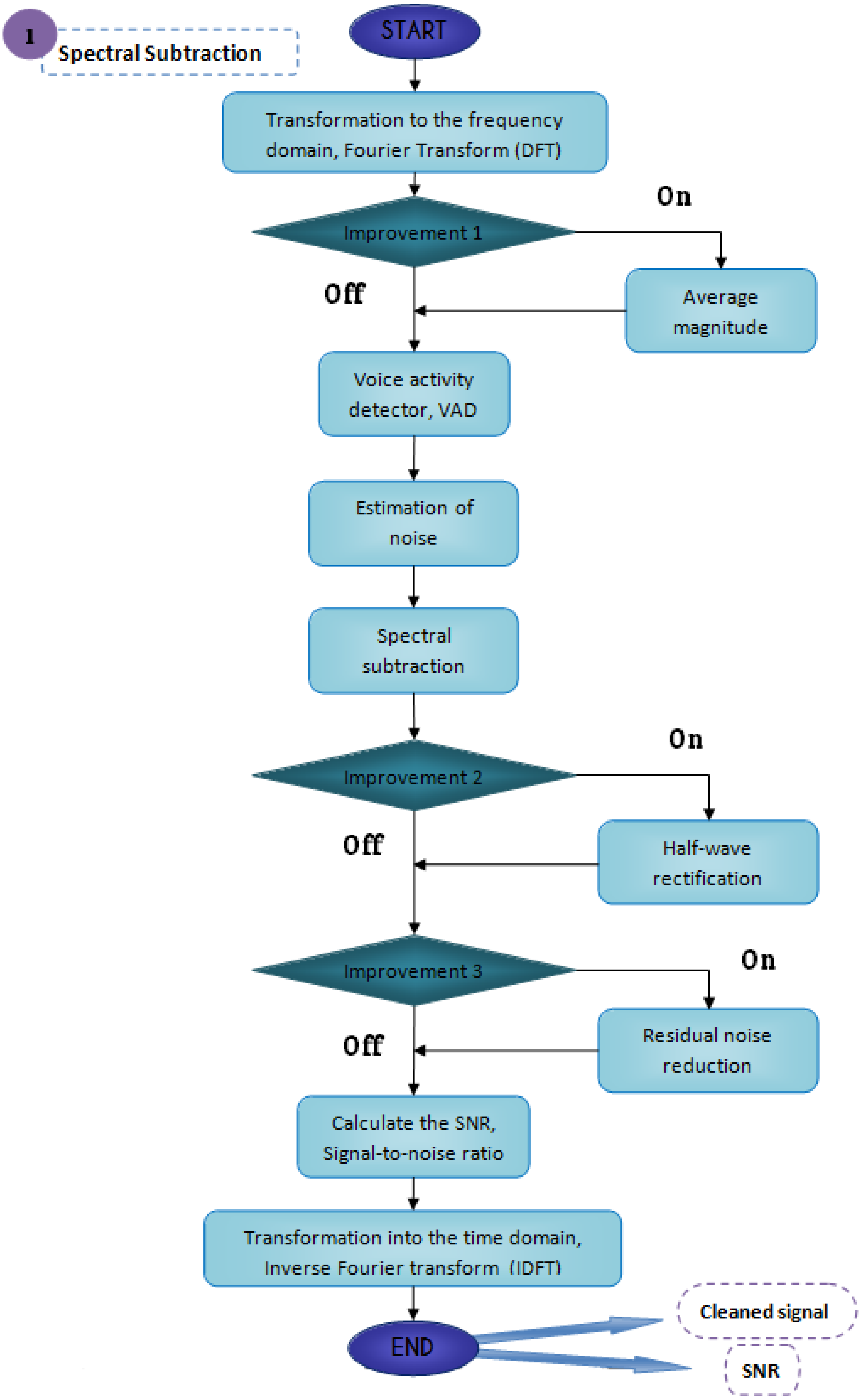


Figure 4.3. **Spectral subtraction.**

4.2.1. Graphical Interface

Figure 4.2 shows the operation of the graphical interface of the application for each spectral subtraction method.

The Interface makes easier the study of the different types of spectral subtraction, when selecting an input audio, the algorithm values are inputted, and in this interface the representation of the audio signal and its spectrum can be seen.

While clicking the button for “Clean noise” the application cleans the audio signal and shows the representation of the audio and its spectrum.

4.2.1.1. User Manual

When launched, the application shows the window in Figure 4.4 [a].

In this display one has to choose what kind of spectral subtraction is able to be use. The selection of one of the methods leads to another window that shows the processing of the spectral subtraction selected.

There are 4 displays for the spectral subtraction: Boll method (Figure 4.4 [b]), Berouti method (Figure 4.4 [c]), Lockwood and Boudy method (Figure 4.4 [d]) and Multiband method (Figure 4.4 [e]).

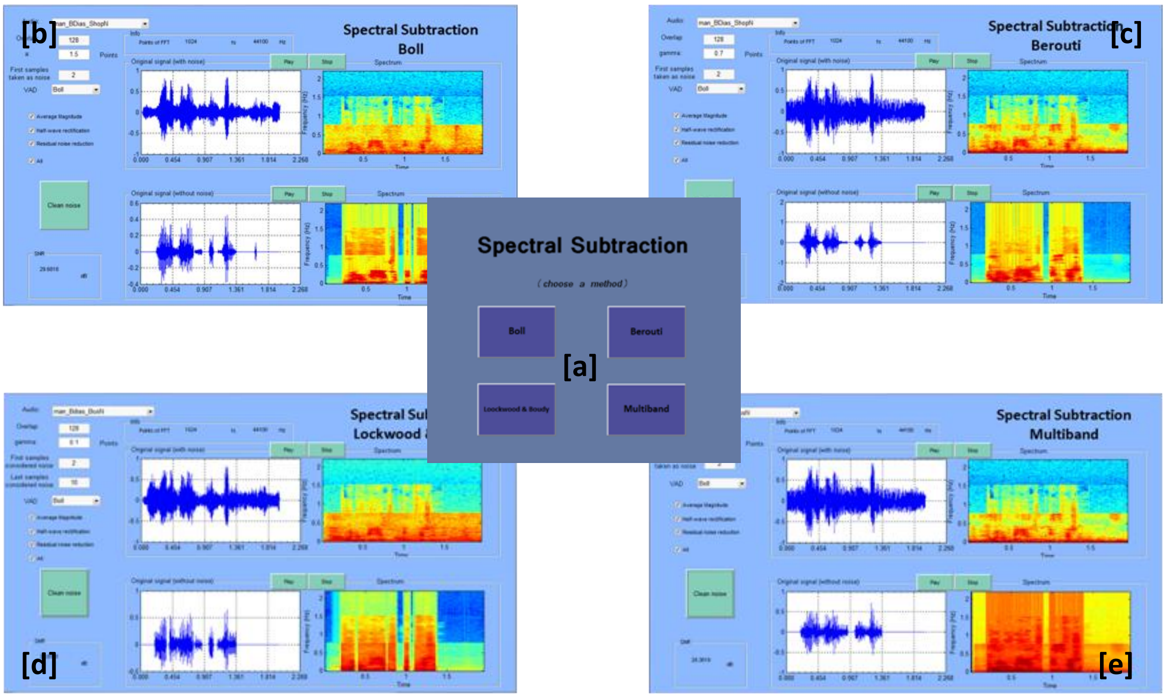


Figure 4.4. Different displays for spectral subtraction.

The different parts of the interface and its uses are shown in the Figure 4.5:

In **number 1**, the user can choose an audio signal with a specific kind of noise.

In **number 2** one can select the parameters for the spectral subtraction. This part changes according to the selected type of spectral subtraction, since each type needs a different set of parameters. In this part the user can choose the VAD.

In **number 3**, the user can select the different options to improve the algorithm (*Average Magnitude*, *Half-wave rectification* and *Reduction of residual noise*).

Once the parameters are introduced, and the audio and the improvements are selected, the audio can be cleaned by hitting “Clean noise” button [**number 4**].

When the user selects the audio, in the panel name as **number 6** in Figure 4.4 shows information about the input signal, and **number 5** (after cleaning the audio) shows SNR of the output signal.

In **number 7**, the user can play the audio and see the plot of the audio and the spectrum.

And finally, the **number 8** is like the number 7 but with the cleaned signal.

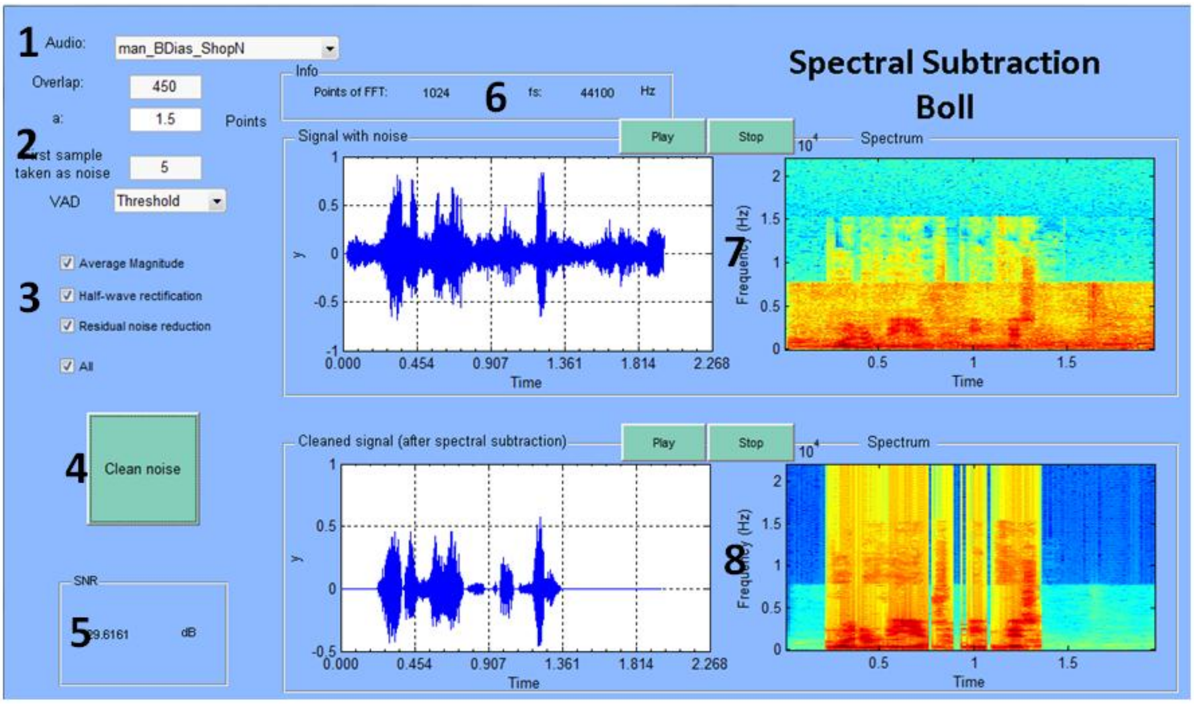

Audio: man_BDias_ShopN
Overlap: 450
a: 1.5 Points
rst sample taken as noise: 5
VAD: Threshold
☑ Average Magnitude
☑ Half-wave rectification
☑ Residual noise reduction
☑ All
Info: Points of FFT: 1024 fs: 44100 Hz
SNR: 29.6161 dB

Figure 4.5. Spectral subtraction Interface.

The parameters available for each algorithm are:

Parameters	Meaning of the parameters	Type of method			
		Boll	Berouti	L&B	Multib.
Overlap	Number of frames taken in the FFT, for example for 50% in overlap, if the number of FFT is 1024, it would be 512.	X	X	X	X
VAD	Type of Vocal Activity Detector.	X	X	X	X
Magnitude Average	Improvement 1 (explained in paragraph 4.3.4.1).	X	X	X	X
Half wave rectification	Improvement 2 (explained in paragraph 4.3.4.2).	X	X	X	X
Residual noise reduction	Improvement 2 (explained in paragraph 4.3.4.3).	X	X	X	X
A	Over subtraction factor.	X			
First sample take as noise	It is the position of the first sample taken as noise.	X	X	X	X
Gamma	It determines the influence of the SNR in the estimation of noise.			X	
Last sample considered noise	Last sample considered noise.			X	
Number of band used	Number of band used in the algorithm.				X

Table 4.1. Parameters displayed in the GUI.

4.3. Estimation of noise and spectral subtraction

Throughout this part, the different parts of the algorithm will be explained.

4.3.1. Windowing

Before starting the process of the spectral subtraction the signal is windowed. This process has been explained in paragraph 2.2.1.

4.3.2. Voice activity detector (VAD)

Once we have speech signal contaminated with noise, in order to make an estimate of the noise from this input audio, an estimate of voice is needed. Hence, so as to estimate voice, it would be necessary to recognize that part of the audio which is just noise.

To get an estimate of the noise that it would be like the real noise as much as possible, the noise along all the duration of the audio must be analysed.

For this, the noise samples taken at instants when the voice disappears must be studied like noise.

Whereupon, there is a necessity to distinguish when the audio signal is only noise, or when it is noise plus voice.

The VAD must not slow down the process of spectral subtraction, so the VAD should be simple, with low number of operations and effective.

4.3.2.1. ZCR and Energy

This method consists in the study of the signal energy and the study of the zero crossing rate, so as to discriminate which parts of the signal correspond with only noise or only voice. This study is made within the time domain.

- ***Zero crossing rate***

It measures how many times the signal cross through zero. With this measure the signal distribution can be imagined. A high ZCR means that the speech segment has a big content in high frequency, while a low rate means that the signal is in low frequency. With this we can separate the sound voice segments (containing a spectrum focused on high frequency) can be separated from the deaf ones (which has many components in high frequency). The problem which arises is that the noise has an extensive spectrum and the ZCR cannot distinguish between noise and voice.

The ZCR is applied in each windowing to obtain a vector that indicates when the audio signal passes through zero with the equation 4.1:

$$ZCR[m] = \frac{1}{ws}\sum_{n}\frac{1}{2}|sign(s[n]) - sign(s[n-1])|\, w(m-n)$$

[4.1]

Where: *s* is the voice signal.
w is the analysis window.
ws is the window size.
sign( ) is the sign function:

$$sign\,(s) = \begin{cases} +1, & s \geq 0 \\ -1, & s < 0 \end{cases}$$

[4.2]

- ***Energy***

The variation of energy over time allows us to determine if there is voice or not, and if the segment is sound or dull. In a section of sound voice the vocal cords come into vibration and the energy increase. However, in the deaf sounds the energy drops.

To calculate the localized energy:

$$E[m] = \sum_{n=0}^{ws-1} s[n]^2 w[n-m]^2$$

[4.3]

Where s is the voice signal.

w is the analysis window.

ws is the window size.

Thanks to the ZCR estimator we can know where the word starts and where it ends can be recognized, and thanks to the energy, there is also the possibility to know if this part of the signal is a word o or not.

The Figure 4.6 is the representation of the sound of the word "*burbujas*".

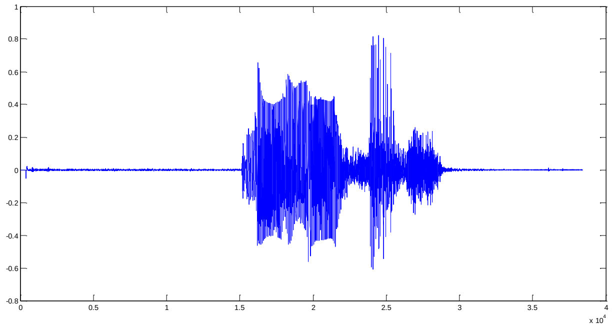


Figure 4.6. Audio signal.

The Figure 4.7 is the representation of the word "burbujas" zero crossing rate. It detects the deaf voice like the "s". Although in this graphics, we can appreciate that the noise is detected too.

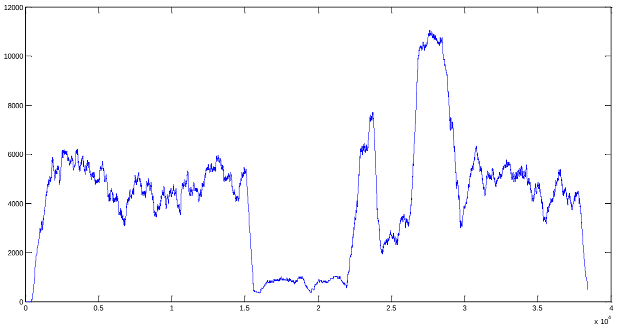


Figure 4.7. ZCR.

The Figure 4.8 is the representation of the Energy of the word "burbujas". When observing the figure, the fact that when there are sound voices, there is energy too can be appreciated.

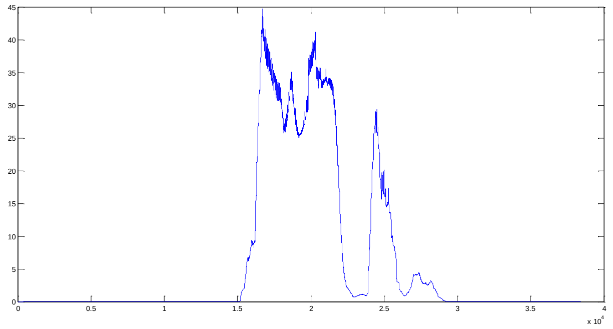


Figure 4.8. Energy.

The matlab code for this process is in appendix 1.

4.3.2.2. Use a threshold to detect voice

Boll also investigates about this kind of algorithms.

He established that when there is no voice activity, the estimate of $X(e^{jw})$ is residual noise which remains after the half-wave rectification and the minimum selection **[5]**. He determined empirically, that the value below the signal is considerate noise is at least 12 dB. If the value is below this number is considered background noise.

With this value the optimum value for the recordings was no obtained, so there was a necessity to look for empirically ways too.

The value of the threshold should depend of the audio, since for each audio; one value would work better or worse. The idea was to make average values of the audio. There was a test average with all the audio, average with the nearest neighbours (3, 100, 200 neighbours...).

Finally, the best result was obtained with a general mean with all the values of the audio. So, for this implementation, the mean as threshold have been used.

The measurement for the detection of absence of speech is:

$$\mathrm{T} = 20\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{\hat{X}(e^{jw})}{E\{|N(e^{jw})|\}}\right|dw\right]$$

[4.4]

In cases in which T is smaller than the mean, the frame is classified as if only had noise:

For T> mean, plot only with voice.

For T< mean, plot without voice, only noise.

The matlab code for this process is in appendix 2.

4.3.3. Noise estimation and spectral subtraction

This part of the process consists on implementing the different algorithms for spectral subtraction explained in the paragraph 3.3 and with its corresponding noise estimate.

Depending on the type of method, the application will do one thing or another. The methods are:

- Algorithm based on the method of Boll (Matlab code in the appendix 3).
- Algorithm based on the method of Berouti (Matlab code in the appendix 4).
- Algorithm based on the method of Lockwood and Boudy (Matlab code in the appendix 5).
- Algorithm based on the method of multiband (Matlab code in the appendix 6).

4.3.4. Methods to improve the algorithm

In order to avoid the errors that appear in the process named in the part 5, various improvements to the algorithm have been included.

The methods used to improve the algorithm are: magnitude averaging, half-wave rectification and residual noise reduction.

4.3.4.1. Magnitude Average

To reduce the spectral noise, $|Y(e^{jw})|$is replaced by an average of the noisy speech signal $|\overline{Y(e^{jw})}|$ where:

$$|\overline{Y(e^{jw})}| = \frac{1}{M}\sum_{i-0}^{M-1}|Y_i(e^{jw})|$$

[4.5]

$Y_i(e^{jw})$ is the windowed transform of $Y(\mathrm{k})$. When choosing the number of samples to average, there is a need to bear in mind that the speaker does not experiment a stationary process. Hence, for the average, short windows have been taken. If done with longer periods, losses in the intelligibility of speech will be suffered. A good approximation is to average three frames: the current, the previous and the next.

4.3.4.2. Half-wave rectification

For each frequency ω where the magnitude spectrum of the noisy signal $|Y(e^{jw})|$ is less than the estimated magnitude of the noise spectrum $E\{|N(e^{jw})|\}$, the algorithm substitutes with zero. In this case I have a 6 dB threshold, considering all the lower level signals as noise.

This technique reduces the background noise by $E\{|N(e^{jw})|\}$ and any variance of the noise tones is eliminated.

A problem can arise when the amount of noise and speech in the frequency “w” is less than$E\{|N(e^{jw})|\}$, not being possible to take any further action.

4.3.4.3. Reduction of residual noise

In the absence of voice is:

$$N_n = N - E\{|N(e^{jw})|\}e^{j\theta_n}$$

[4.6]

This difference is called *residual* noise and it takes positive values. During moments of vocal activity, residual noise will be perceived in the frequencies where it is not masked by the voice.

The residual effects of noise can be reduced if each frame is analyzed separately. If a set of frequencies are given, the residual noise varies randomly in amplitude for each frame that is analyzed. Thus, it can be removed by replacing its current value by the minimum chosen value. The minimum value will be taken only when the value of the estimate of $X(e^{jw})$ is less than the "residual" noise calculated during times without vocal activity.

The replacement of values depends on the amplitude of the estimation $X(e^{jw})$:

- If this value is below the residual noise and varies largely in the frame by frame analysis, it can mean that at this frequency the spectrum will be composed of just noise. This noise is eliminated by substituting its value for the minimum one.

- If this value is below the maximum, but it has a constant value, it would probably means that the spectrum at this frequency contained a speech with little energy. To preserve the information, it assumes a minimum value.

- If the value is greater, it means that the speech was located at that frequency.

The implementation of these three conditions over speech levels would be:

$$\begin{cases} \left|\hat{X}_i(e^{jw})\right| = \left|\hat{X}_i(e^{jw})\right|, & if \ \left|\hat{X}_i(e^{jw})\right| \geq \max\left|N_r(e^{jw})\right| \\ \left|\hat{X}_i(e^{jw})\right| = \min\left\{\left|\hat{X}_i(e^{jw})\right|_{j=i-1,i,i+1}\right\} & if \ \left|\hat{X}_i(e^{jw})\right| < \max\left|N_r(e^{jw})\right| \end{cases}$$

[4.7]

In equation 4.7, $max\left|N_r(e^{jw})\right|$ represents the maximum value of the residual noise measured in those frames without speech activity.

The matlab code for this three process are in appendix 7.

4.3.5. Assessment of the experiments, measures of quality

When the application is finished, it is necessary to evaluate its performance, in order to measure the quality of the results obtained with it. For that purpose, it is necessary a measure of quality.

In this project two methods to measure the quality of the implemented systems have been used, one objective and one subjective.

The objective one is the calculation of the *Signal to Noise Ratio* (SNR) and the other objective will be the *Mean Opinion Score* (MOS).

4.3.5.1. Signal to Noise Ratio, SNR

It is an objective measurement of the signal quality. The signal to noise ratio defines a relationship between the level of the desired signal and the background noise level (in power or in energy). In this case, it would be a relation between the final signal power and the power of the noise pollution.

The bigger the SNR value is, the result will be better, since it implies that there is more difference between the power of signal and the power of noise.

It is calculated as:

$$SNR = \frac{P_{signal}}{P_{noise}}$$

[4.8]

Where P_{signal} is the power of the signal and P_{noise} is the power of noise.

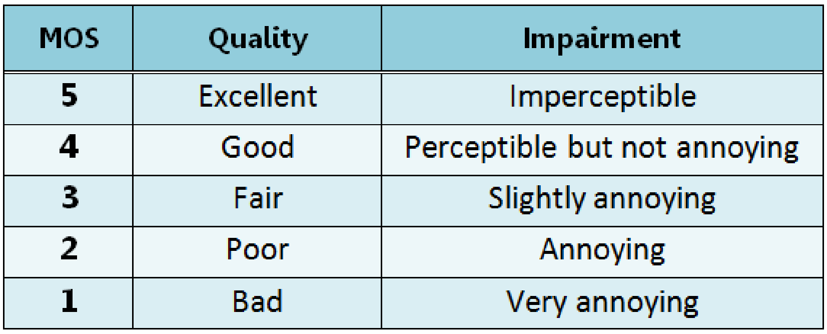
In this project the more convenient form of the SNR in dB have been used, so:

$$SNR_{dB} = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) = P_{signal,dB} - P_{noise,dB}$$

[4.9]

4.3.5.2. Mean Opinion Score, MOS

It is a subjective measure of the signal quality. The Mean Opinion Score consists of evaluating each audio file with the MOS scale. The MOS scale is the next:



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

Table 4.2 Value of MOS scale.

These measures should be made by professionals in an audio studio with professional equipment. Unable to use an equipment of that category, the measures have been carried out with ordinary headphones, and tested with 25 different students. The reported values of MOS are the sequels correspond to the average over the 25 student's evaluation.

Made in this way it cannot be considered as an official measure but it can show a general idea of the subjective quality of the audio.

Chapter 5

5. Results

In order to evaluate the performance of the proposed algorithms with real world data, the following signals of audio and noise have been recorded and then they have been merged doing all possible combinations.

On the one hand, there are different voice signals with different characteristics. On the other hand, there are diverse types of the noise.

In the voice signal, there are 2 signal characteristics studied: the length of the audio and the difference between gender of speaker (male or female voice: high pitched and low pitched). So, the recorded voices are:

- Female voice with long length (23 seconds):
- Female voice with short length (1 second):
- Male voice with long length (23 seconds):
- Male voice with short length (1 second):

In the noise signal, the different types of sound that have been recorded are:

- White noise:

  Its principal characteristic is its stationary (Section 2.3.4).

- Pink noise:

  Its principal characteristic is its stationary (Section 2.3.4.2).

- Impulsive noise:

  This noise is characterized because it concentrates much energy at the same point. (More information in section 2.3.4.3).

- Shop noise:

  This is a noise that could be separated in two components, a component of background noise which is stationary and other component consisting of voices in the background and no highlights on the other.

- Restaurant noise:

  This noise is a noise that could be separated into two components too, it is like the shop noise but in this case the voices in background predominate over the background noise.

- Bus noise:

  In this case are two components, one stationary (bus motor) that predominates over the voices in the background.

To estimate the voice, two types of voice activity detector have been used: zero crossing rate and energy and a limit as detector.

As for the spectral subtraction, 4 types of algorithms have been tested: Boll, Berouti, L&B and Multiband, explained in section 3.3.

Testing all possible combinations offered by the application is not viable because there are many combinations possible above mentioned. Consequently, the final tests with the best options have been done.

5.1. Choice of voice activity detector, VAD

Having two VAD (a threshold and ZCR & Energy), a comparison has been done between them.

In order to study the time taken to perform the same process, the ZCR & Energy method is much slower than the threshold method. So as to demonstrate this fact, both were tested in the same conditions (same audio, same variables). The results obtained were that the threshold method took about 1 second and the ZCR & Energy method about 20 seconds. If we wanted to use this application in a call in real time 20 second would be unworkable for a communication.

This VAD returns a vector with ones when the signal is voice and zero when it is not voice. Figure 5.1 shows the resulting vectors compared with the original signal.

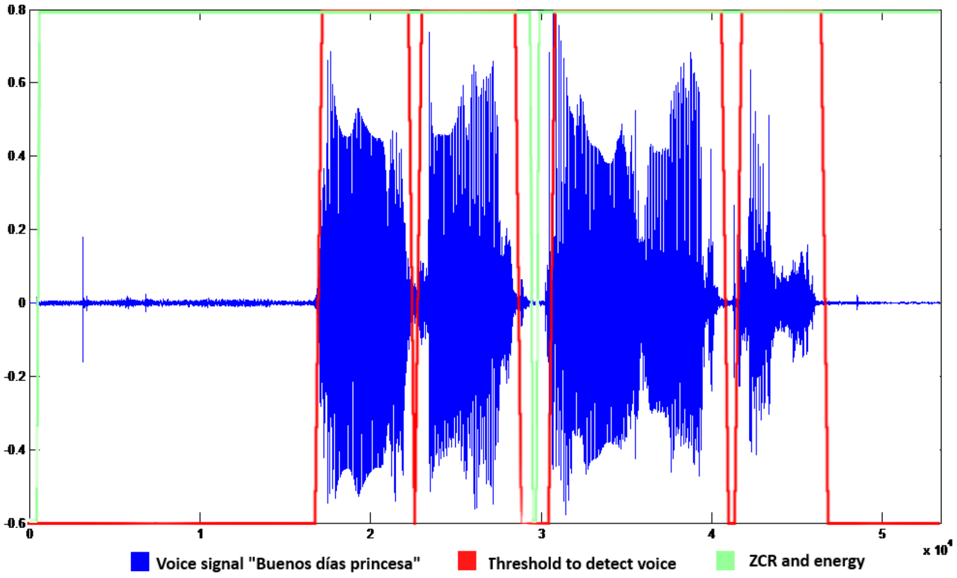


Figure 5.1. Comparison between VAD.

Figure 5.1 shows that the threshold method (red) conforms more to the shape of the original signal. However, ZCR & Energy method (green) confused voice with background noise. In addition, the first method is able to detect where the words begin and end. In this case, it divides the speech into the following segments: "*Buenos – días – prince – sa*". The second method is not able to do that, it does not detect where the word really starts and ends, only detects the separation between "*Buenosdías*" and "*princesa*" because it confused voice with background noise. We suspect that this bad performance in segmenting a clean signal would worsen with the presence of additive noise.

In conclusion, we select **the threshold to detect voice** as our method for the experiments. It would be the better of the two because it is faster and it is able to detect more precisely the presence or absence of voice in the speech signal.

5.2. Choice of the length of the recording

Initially I opted for longer recording, believing that the estimation of noise would be better because that would take more samples to take as a reference. In fact, at the beginning I took as VAD a threshold of 6 dB, considering that everything that was below would be noise **[20]**, and this happened, where the longer was the recording, the better was the estimation.

However, to the use as threshold for the VAD an average instead of a specific value improves its performance a lot. Therefore, there was no difference between long and short recordings since it estimates both equally.

Experimentally, it proved that the same results were obtained for both short and long recordings and it was decided to make experiments with short recordings because the long recordings may take more time to be processed to short, since the application need to carry out more operations to take more samples.

Thus, **the short audio** was chosen for the final experiments.

5.3. Choice of the improvement

Within the application, three additional methods have been implemented to improve the spectral subtraction. The user can choose which one to use.

The three improvements are explained in the paragraph 4.3.4 and its implementation in the appendix 7. They are:

- Magnitude average.
- Half-wave rectification.
- Reduction of residual noise.

After testing all the possible combinations, **all were used** because thanks to them, the musical noise and the background noise that was remained after the spectral subtraction was removed.

5.4. Choice of the overlap

Each voice has optimal values of overlap, if the value is too low the voice becomes metallic and if it is too high the voice is distorted making it unintelligible.

When a value of overlap is chosen, the difference between male and female speakers must be taken into account, due to the difference of frequency of each signal: the female voice is a high frequency while the man is low frequency.

Within the same period of time, the low-pitched signal varies much less in shape in comparison with a high-pitched signal, which is the main reason why a greater overlap in the male signal is required. In order to appreciate the diverse changes detected in the male signal, a larger window is necessary. Nevertheless, taking in mind that the window size is fixed, this cannot be possible. Hence, in order to fix that problem, there is a need to increase the overlap, so as to simulate the effect of a larger window, and consequently, the multiple changes of the signal would not be lost as a result of using a small window. So the voice of the woman needs lower values of overlapping than the man voice to get best results.

To study the performance of the application (and therefore of the algorithms) with voices of different frequencies, the same value for both types of voices has been used. For that reason, so as to do this, a same overlap at the half-way point between the optimum values for the two types of voices will be listen to well have been used. Empirically, a value at which is not too high to make female voices sound unintelligible and not too low to distort male voices has been chosen.

The chosen value is **450 points**.

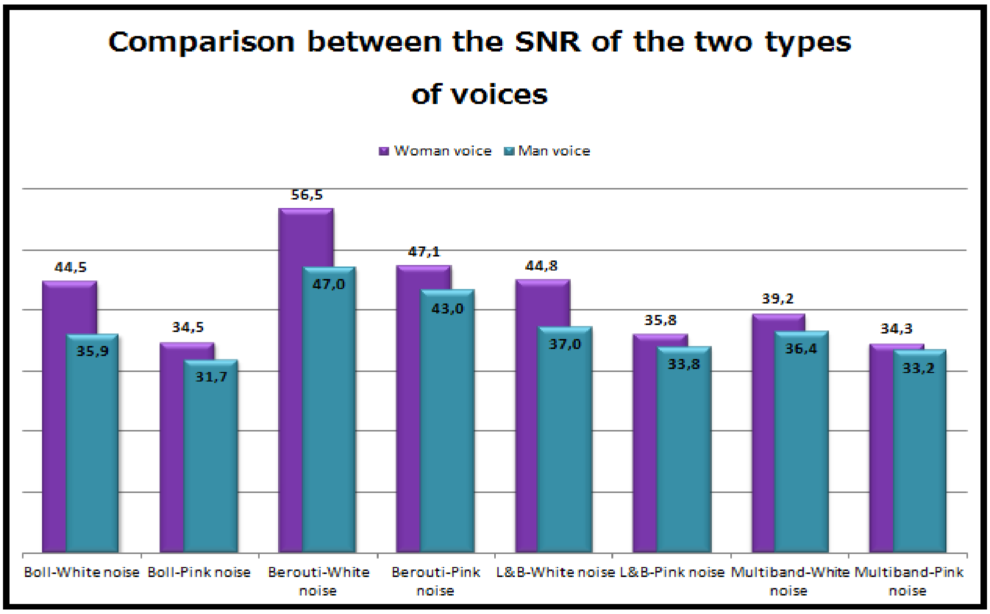


Figure 5.2 Comparison between the SNR of the two types of voices.

This Figure 5.2 shows a comparison on the SNR between male and female voice using the 4 methods and the two ideal noises (pink and white).

One can see that the SNR for each case are higher for female voice (high pitched) and for male voice (low pitched), thus demonstrating objectively what is happening and not just that a person hears the audio as quality control.

5.5. Summary of the variables and the audio used for experiments

In order to make the comparison more efficient, some values have been remain fixed through all experiments in the diverse methods, and the variables of that method have remained for the different types of noise in each process.

Two values of measurement of the signal quality were calculated to compare and study the obtained results:

- Objective: SNR.
- Subjective: MOS approximation.

The audio used in the different experiments are:

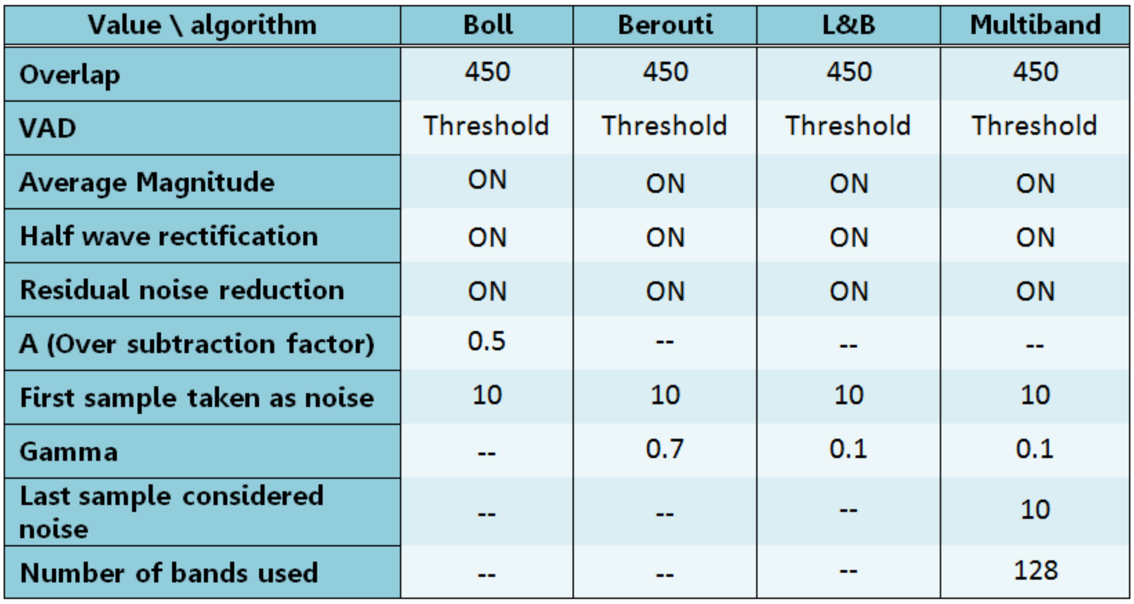
- *Female voice with short duration:*
  - Original voice:

    It is the original recorded voice without added noise, only with background noise.
  - Voice + white noise:
  - Voice + pink noise:
  - Voice + impulsive noise:
  - Voice + restaurant noise:
  - Voice + shop noise:
  - Voice + Bus noise:

- *Male voice with short duration:*
  - Original voice:

    It is the original recorded voice without added noise, only with background noise.
  - Voice + white noise:
  - Voice + pink noise:
  - Voice + impulsive noise:
  - Voice + restaurant noise:
  - Voice + shop noise:
  - Voice + Bus noise:

The parameters values used in the experiments:



Value \ algorithm	Boll	Berouti	L&B	Multiband
Overlap	450	450	450	450
VAD	Threshold	Threshold	Threshold	Threshold
Average Magnitude	ON	ON	ON	ON
Half wave rectification	ON	ON	ON	ON
Residual noise reduction	ON	ON	ON	ON
A (Over subtraction factor)	0.5	--	--	--
First sample taken as noise	10	10	10	10
Gamma	--	0.7	0.1	0.1
Last sample considered noise	--	--	--	10
Number of bands used	--	--	--	128

Table 5.1 Values used in the experiments.

The audios obtained cleaned by the application are the followings:

Female voice	Boll	Berouti	L&B	Multiband
Voice + White noise	⬇	⬇	⬇	⬇
Voice + Pink noise	⬇	⬇	⬇	⬇
Voice + Impulsive noise	⬇	⬇	⬇	⬇
Voice + Restaurant noise	⬇	⬇	⬇	⬇
Voice + Shop noise	⬇	⬇	⬇	⬇
Voice + Bus noise	⬇	⬇	⬇	⬇

Table 5.2. Cleaned female audios.

Male voice	Boll	Berouti	L&B	Multiband
Voice + White noise	⬇	⬇	⬇	⬇
Voice + Pin noise	⬇	⬇	⬇	⬇
Voice + Impulsive noise	⬇	⬇	⬇	⬇
Voice + Restaurant noise	⬇	⬇	⬇	⬇
Voice + Shop noise	⬇	⬇	⬇	⬇
Voice + Bus noise	⬇	⬇	⬇	⬇

Table 5.3. Cleaned male audios.

5.6. Results

5.6.1. Signal to Noise Ratio, SNR

The higher the value of the SNR is, the better results would be obtained, because the difference between the clean speech and the noise will be more.

These are the SNR values obtained for female voices:

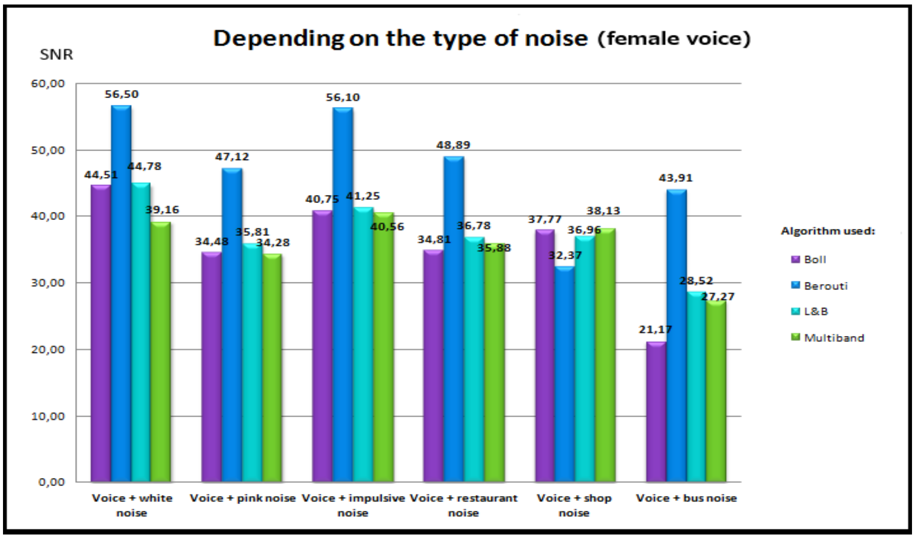


Figure 5.3. SNR depending on the type of noise (female voice).

These are the SNR values obtained for male voices:

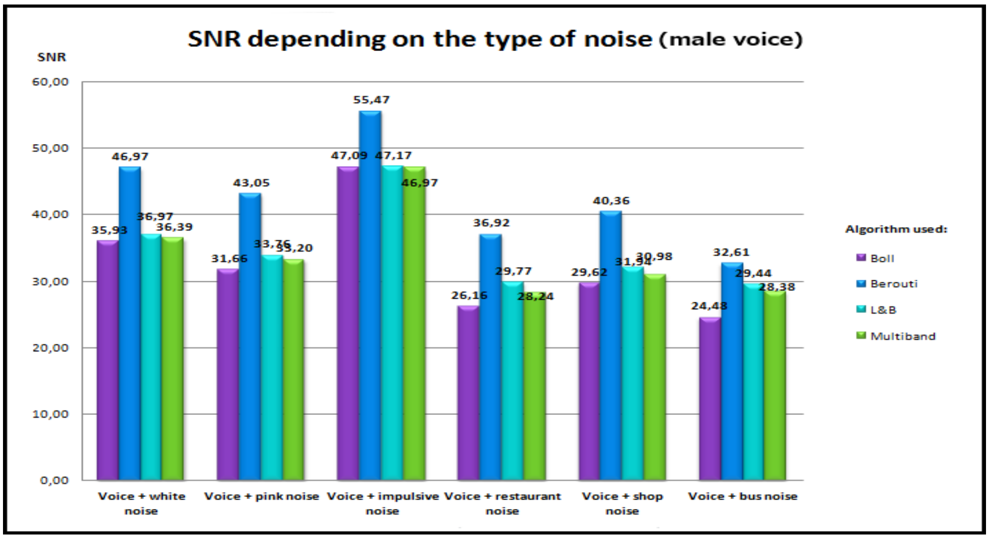


Figure 5.4. SNR depending on the type of noise (male voice).

In these two figures all SNR values are obtained for female voice and male voice in separate ways.

The difference between the results obtained with the two types of voice is due to the difference in frequencies of each signal: The female voice is of high frequency while the man is low. The same overlap has been used in order to compare both voices, obtaining the fact that the male audio gets worst results in contrast with the female one. This arises because low frequency signals requires a window with larger size so as to appreciate the changes of the signal, yet the window size is an invariable value, and it cannot be changed. In order to solve this situation, a higher overlap is used thereby achieving not lose the signal variations (It was explained in the section 5.4). Within the figures 5.3 and 5.4 in general the woman values are higher than the man values.

If the results of the voices are merged without difference between types of voices, the values of the SNR will be the following figure:

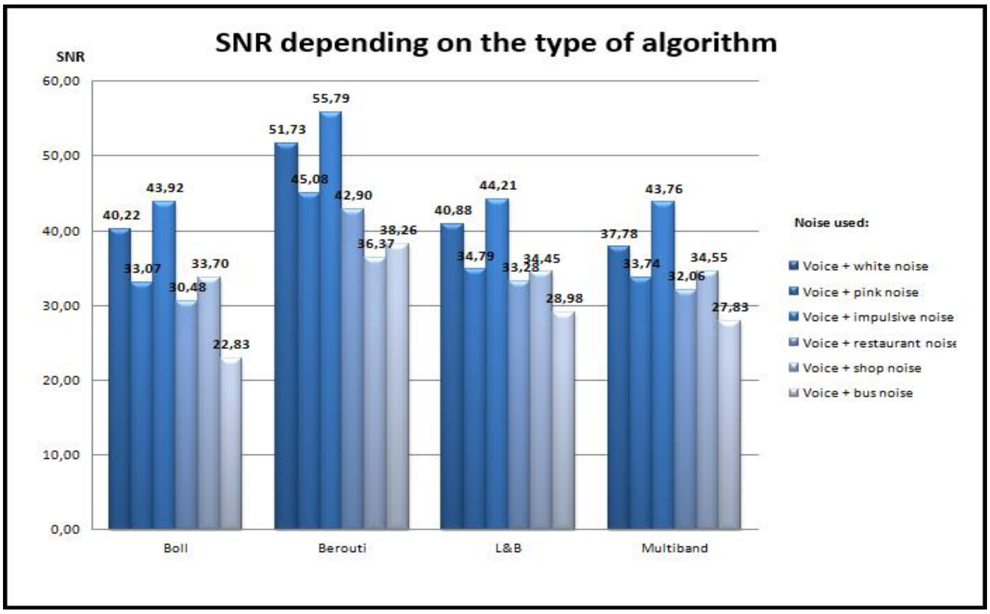


Figure 5.5. SNR value depending on the method used.

This graphic shows that for each noise the higher values are the values obtained with the method of Berouti. However, the method with the worst SNR values (the lowest of all) is Boll. **Hence, the objective method which would the best in general terms is Berouti and the worst Boll**.

The table 5.4 shows the resulting SNR values to analyze for which kind of noise the spectral subtraction works better.

Type of noise	SNR
Voice + impulsive noise	45,20
Voice + white noise	43,92
Voice + pink noise	40,30
Voice + restaurant noise	34,93
Voice + shop noise	33,71
Voice + bus noise	30,28

Table 5.4. SNR according to the type of noise.

Looking only at the SNR value it shows that you get the highest values is for voice with impulsive noise (45, 20 dB).

However, listening to the voices cleaned by the algorithms shows that these 4 methods are not able to eliminate this type of noise because it is much energy concentrated at one point and the VAD confuses it with voice and does not eliminate it. Therefore, a subjective quality meter would be necessary, in which a person listening to it will be implicated.

The following in decreasing order of performance are the voice audios with white noise (43, 92 dB) and with pink noise (40, 30 dB) are found, i.e., constant and uniform noise. The worse performances are obtained for audios with noise of restaurant, shop and bus, that they are noises mixed uniform noise with voices in background (which are considered noise).

The methods are able to eliminate much of these types of noise, which eliminates worst is the impulsive noise, despite the fact that this kind of noise gets the best SNR values. Some similar occur with the voices in background, the methods mistake with voices because despite the fact that they are voices are in background, they have less energy that the principal voice and the algorithms eliminate part.

Nevertheless, all this will be discussed further in the next section, as it is something subjective that can be only studied after listening to the audio.

The main problem of the SNR, as parameter to measure the quality of the program, is not a stake valuing a person, it does not take into account how well or bad the final result is sound, its intelligibility and other features that can only evaluate a person.

So then a subjective parameter will be studied: **MOS**.

5.6.2. Mean Opinion Score, MOS

The MOS scale is a subjective studio, which means that focuses on what the people are able to hear.

The procedure was to play to a group of person the cleaned voices obtained from the application and ask them to evaluate with the scale in the part 4.3.5.2 the quality of the audio.

The first opinion acquired from people is that any algorithm is able to cancel the impulsive noise. This fact is true, since this kind of noise does not yield good results, due to the fact that the voice is intelligible and this noise only annoys during a little fraction of a second. Although, there was people who feeling very uncomfortable with this noise, and on the contrary, some other not.

In conclusion, subjectively, the application is not able to remove the impulsive noise. This is due to the fact that the VAD is unable to detect this kind of noise, mainly because the impulsive noise means lot of energy in one instant, which can be confused with voice. In order to clean this noise, it would be necessary to study its frequency or its timbre.

As in the SNR study, in this case the MOS study shows the differences between a male voice (low pitched) and a female voice (high pitched). The listeners have given higher marks and better opinions for the female voice (Figure 5.6). That is to say, the listeners said that the quality of the female voice is sounds better than the male one when trying to make a listening test. This occurs for the same reason explained in the section on the SNR study. Thereby, to improve the low pitches voices there is a need to change the overlap with an optimum value.

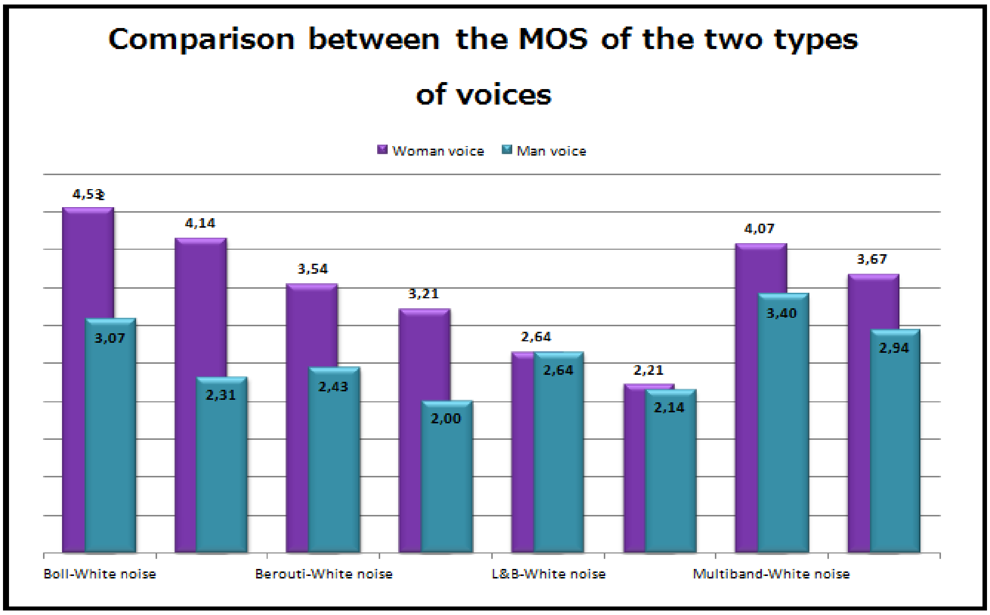


Figure 5.6. Comparison between the MOS of the two types of voices.

Studying only the audio with the optimum overlap value the reader can appreciate the best algorithm to be used is the Boll algorithm. This algorithm obtained better results than the others, since the voice is not distorted, and it also eliminates the annoying noise. The next figure shows the results depending on the type of algorithm given by the listeners.

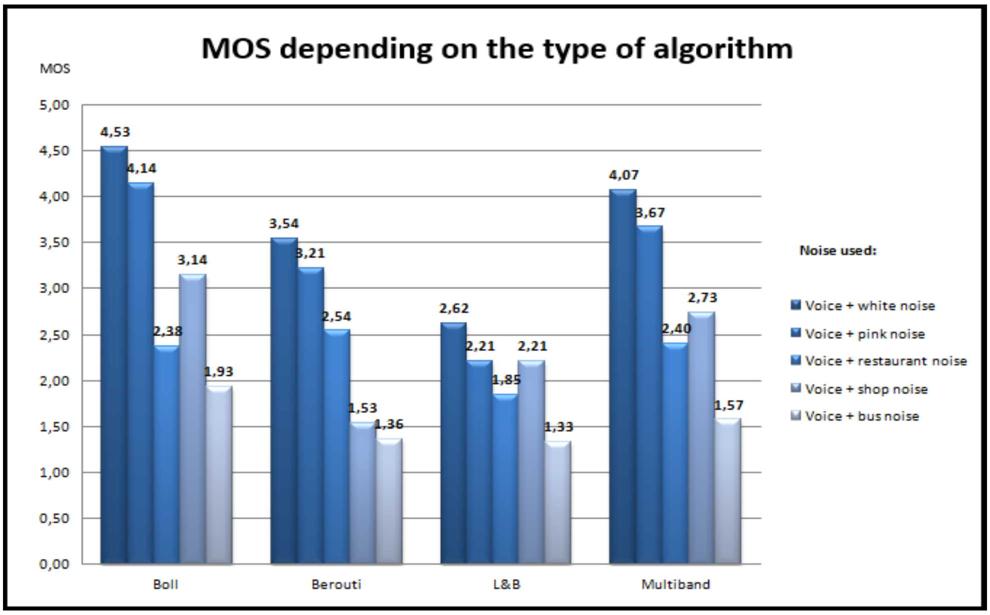


Figure 5.7. MOS depending on the type of algorithm.

With respect to the noise to be eliminated, the noise with the top results is the white noise, followed by the pink noise. If the noise is constant and invariable the algorithm is able to remove it better. Restaurant noise and shop noise are the next ones with better results, i.e. noise with background voices, and that the algorithm can eliminate mostly all the annoying noises. Here the Boll algorithm is the best for removing this type of noise with the best results. However, the worst results are obtained with bus noise, a mixture of invariable noise with background voices. This noise was the most annoying for the listeners. This information could be summarized in the Figure 5.8.

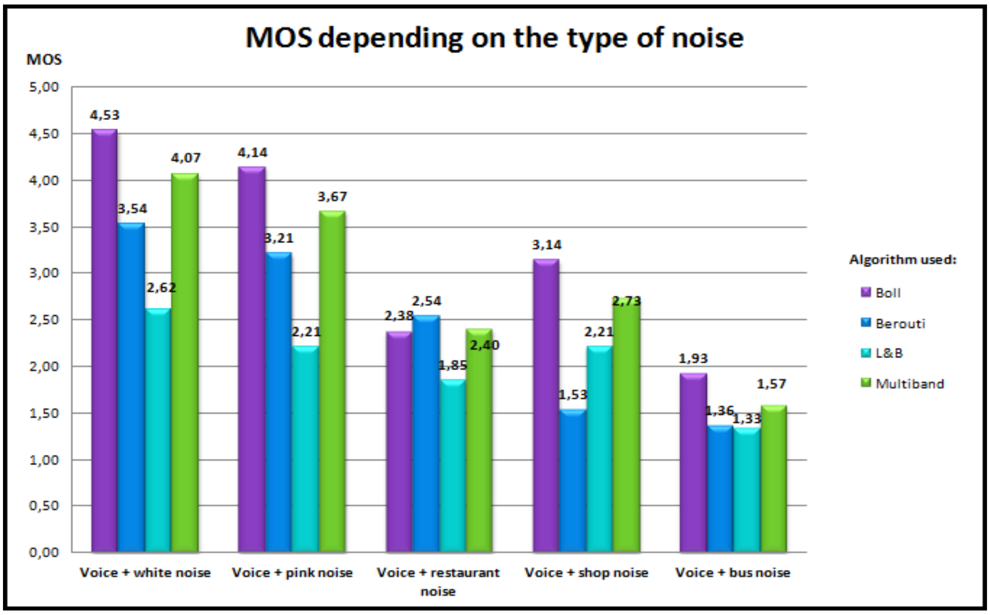
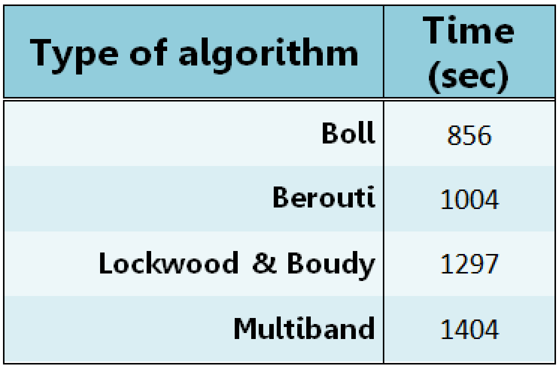


Figure 5.8. MOS depending on the type of noise.

Note that, although the Lockwood and Boudy algorithm is able to eliminate the noise between the words, it is not able to eliminate noise inside the words. In addition, the Berouti algorithm distorts the voice too much.

Comparing both evaluation methods, the results obtained are not the same. In the objective method (SNR), the best algorithm is Berouti and in the subjective method (MOS) the best is Boll. In my opinion, the subjective evaluation method should be taken into account more than the objective, since in the end people would be the expected user of the application so as to clean audio.



Type of algorithm	Time (sec)
Boll	856
Berouti	1004
Lockwood & Boudy	1297
Multiband	1404

Table 5.5. Time audio processing of female voice with white noise and long length (23 sec).

The running time of the algorithm needs to be also considered. As presented in the previous table, the Boll algorithm is the fastest, since this method is the simplest, with fewer operations to carry out, and consequently, the running time is the lowest. And as previously mentioned, this is something important to take into account. In conclusion, after researching among various algorithms to find the best one, the election has been the **Boll** one. The choice has been made considering its effectiveness and simplicity, being also the best algorithm in terms of computational cost, speed and effectiveness.

Chapter 6

6. Application uses and future lines

6.1. Future lines

The elimination of noise when working with voice is very important whereupon the main objective of a future development would be to improve the application and consequently the resulting speech after removing noise.

The first most important thing to improve would be the Vocal Activity Detector, because it is the base of all and if this part is better, the voice will suffer less losses and the clean voice will sound more natural. The ideal would be to find a voice detector difficult to deceive, i.e., it has to be able to detect with more precision the start and the end of the spoken words.

The VAD should not be based only on the power to detect the voice or in the zero crossing rate, it could be based on the pitch or even for each type of voice can specify optimal parameters for detection too.

The improvements in the VAD should also aim at alleviating the problem of removing the background voices. This is because the VAD detects the background voices (considered noise) like main voice (voice, no noise) and therefore the application does not eliminate them.

Make the application user friendly, i.e. that one user without to know the algorithms that the application uses, he can adjust the parameter knowing only the type of voice and noise with which he is working.

For example, instead of entering the value of the overlap (which influences whether the voice is low pitched, high pitched or middle pitched), it would be better if the user only has to select the type of voice and not a value. Looking at the frequency we could empirically establish some limits to determine for each interval the optimal overlap value for this kind of voice.

The next point to improve is that to introduce other audios that are not used as an example in this application is necessary to have knowledge in Matlab. For that, an user without this knowledge cannot do that. A future line will be introduce the possibility of introduce the audios with the interface.

6.2. Application uses

The main use of this application is cleaning a voice recording because this is the previous step for all the works made with voice. This is very important, since if the voice is not completely clean, the noise can interfere in the subsequent processes. Apart from this, it is necessary to clean the noise in the voice because it difficult the understanding of the speech. After recording a voice with background audio, it will be used in this application in order to get only the voice, which it is the only thing wanted, and the main goal.

This application could be used for example so as to separate the singer's voice from the music within a song.

In order to illustrate that, a test was done to check if this was possible with the next song:

It was processed with the algorithm that it had the highest mark in the subjective evaluation (Boll), and the results obtained were the following:

A further improvement of the VAD would bring better results. This application could be very useful for a person who works editing videos. Many times these people need to do promotional videos in which they have to include specific quotes from some particular actors, for that they need the cleaned quote without background noise and without other sounds from the video. With this application they will obtain the clean quote and then they will work with it in a comfortable way.

Chapter 7

7. Conclusion

An application capable of removing noise in speech using spectral subtraction method has been created successfully. Four methods have been implemented, and its performance has been examined, deciding which one is the best. As previously mentioned, always from my point of view, the subjective quality measure is the most important among them. In conclusion, for me, the best results are obtained with the Boll algorithm and voices mixture with white noise. The Boll algorithm has obtained a value around 4 in the MOS scale, which highest rating is 5. This value is two points above the worst method (Lockwood & Boudy).

I can say, the goals which were initially propose have been undoubtedly fulfilled, as the main aspiration was to study in depth how the spectral subtraction works, and how to create an application to evaluate it. In addition, a study and a research about how it would work has been presented and explained.

Although the goals have been reached, this project could be continued. The realization of a vocal activity detector could be studied in depth in order to improve the results of the later algorithms. A new goal could be imposed so as to obtain a VAD with the ability to not confuse the speech with the noise in all the cases. The VAD is a critical part of the process, as it is not able to distinguish the voice of the noise well. As a result, the later process is not carried out with the optimal values because they directly depend on this part of the process to estimate noise and to recognize whether is voice or not.

From a personal perspective, the realization of this project has supposed the entrance for me to a new working field, with various circumstances and requirements, finding out what key decisions need to be considered, and what are the main steps to follow in each situation. Furthermore, that experience has taught me how to organize myself and my schedule, since in order to combine work placement with the development of the project I had to do an effort.

Moreover, the intention to write the project entirely in English has been a difficult task for me, even though I have been using this language from many years ago. Using properly all the technical words throughout a high amount of pages is not something easy for someone whose mother tongue differs pretty much from the used one.

From my stand point, it has supposed a good experience for me, not only contributing to make me a much mature person, yet also teaching me how to face difficult and new challenges, towards my professional and personal development.

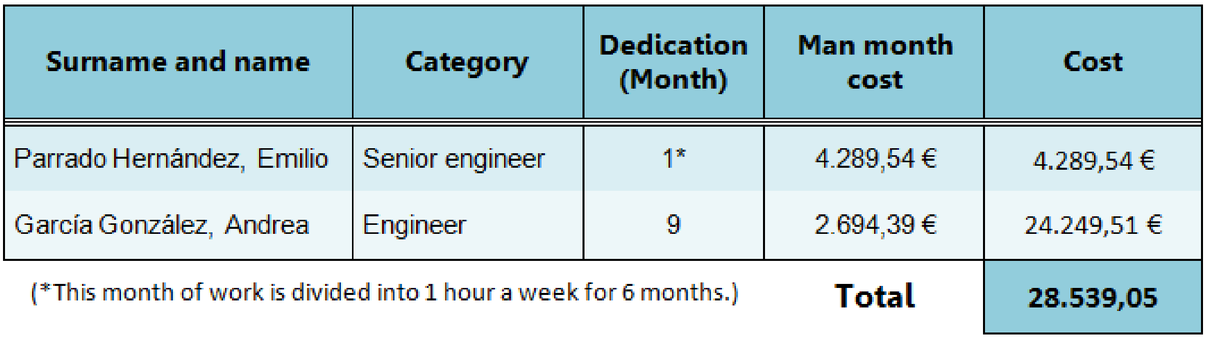
Chapter 8

8. Project budget

The project budget has been prepared according to the document provided by the Carlos III University of Madrid, within the resources section for the Bachelor thesis **[19]**.

8.1. Staff cost

The staff consists of two people whose dedication and costs are detailed below:

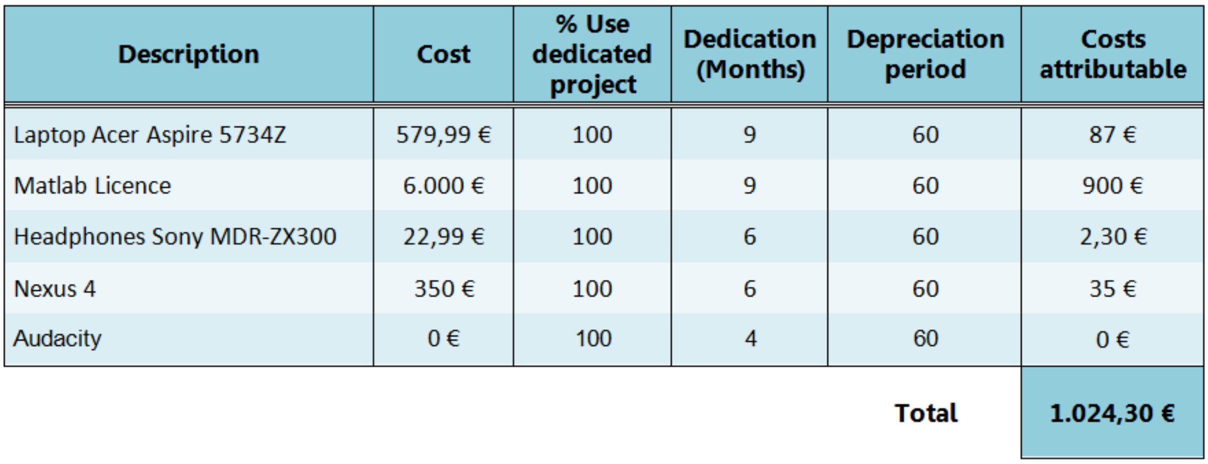


Surname and name	Category	Dedication (Month)	Man month cost	Cost
Parrado Hernández, Emilio	Senior engineer	1*	4.289,54 €	4.289,54 €
García González, Andrea	Engineer	9	2.694,39 €	24.249,51 €
(*This month of work is divided into 1 hour a week for 6 months.)			**Total**	**28.539,05**

Table 8.1. Staff cost.

8.2. Equipment costs

Hardware and software costs are budgeted considering the use of each element on a 100% and the equipment. Price excluding VAT.



Description	Cost	% Use dedicated project	Dedication (Months)	Depreciation period	Costs attributable
Laptop Acer Aspire 5734Z	579,99 €	100	9	60	87 €
Matlab Licence	6.000 €	100	9	60	900 €
Headphones Sony MDR-ZX300	22,99 €	100	6	60	2,30 €
Nexus 4	350 €	100	6	60	35 €
Audacity	0 €	100	4	60	0 €
				Total	**1.024,30 €**

Table 8.2. Equipment cost.

The following equation has been used so as to calculate the total cost within the previous table:

$$\frac{A}{B} * C \; * D \qquad [8.1]$$

Where: *A* = number of months from the date of invoice in which the equipment is used.

B = Deprecation period (60 months)

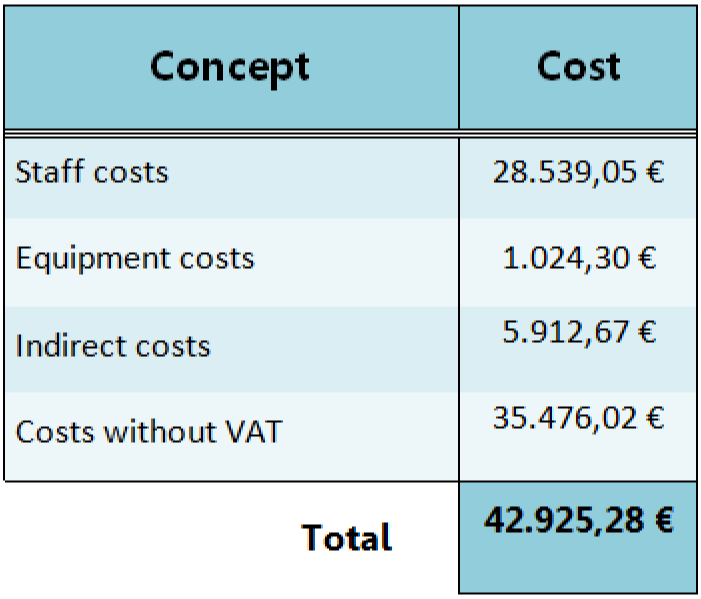
C = Cost of equipment (excluding VAT)

D = % dedicated to the project (usually 100%)

8.3. Total costs

In order to calculate the total costs, there is a need to take into account indirect costs of 20%, resulting from possible allowances, travel expenses and other costs that have not been previously warned.

Due to the fact that the project is billed in Spain a VAT of 21% has been added to the total amount.



Concept	Cost
Staff costs	28.539,05 €
Equipment costs	1.024,30 €
Indirect costs	5.912,67 €
Costs without VAT	35.476,02 €
Total	**42.925,28 €**

Table 8.3. Total cost

The costs amounted to a total of **92.753,28** €.

Chapter 9

9. References

[1] ALVARADO, D.: “Efecto del enventanado en la obtención del espectro discreto de una señal”. Universidad Técnica Particular de Loja. Ecuador. 2005.

http://www.monografias.com/trabajos20/enventanado/enventanado.shtml

[2] ARRIBAS, J.I.: “Introducción al audio digital”. Electrical Engineering. Universidad de Valladolid.

[3] BEROUTI, M.; Schwartz, R. and Makhoul, J.: “Enhancement of speech corrupted by acoustic noise”. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79. VOL 4. Bolt Beranek and Newman Inc. Cambridge, Mass. 1979.

Digital Object Identifier: 10.1109/ICASSP.1979.1170788

[4] BLEDA, S.; FRANCÉS, J.; MARINI, S.; MARTÍNEZ, J.J.: “Herramientas software para la docencia de la señal de voz en Ingeniería Técnica de Telecomunicaciones”. Universidad de Alicante. Spain. 2012.

http://web.ua.es/va/ice/jornadas-redes/documentos/posteres-exposats/246141.pdf

[5] BOLL, S.F.: “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”. IEEE Transactions on acoustics, speech, and signal processing, vol. ASSP-27, NO. 2, April. University of Utah. 1979.

Digital Object Identifier: 10.1109/TASSP.1979.1163209

[6] FRUTOS BONILLA, J.: “Detection, Analysis and correction of errors and gaps in audio signals”. Fachhochschule Braunschweig/Wolfenbüttel. Germany. End of Degree Project. 2008.

[7] GODSILL, S J. and RAYNER, P.J.W.: "Digital Audio Restoration". Springer. Cambridge, U.K. First Edition. 1998.

ISBN: 978-0-7923-8130-3

[8] GONZALEZ, G.: "Reducción de ruido en Grabaciones de Audio". Escuela Politécnica Superior. Universidad Autónoma de Madrid. Spain. Thesis Project. 2011.

[9] HÄNSLER, E. and SCHMIDT, G.: "Speech and Audio Processing in Adverse Environments". Springer. Germany. First Edition. 2008.

ISBN: 978-3-540-70602-1

[10] IFEACHOR, E.C. and JERVIS, B.W.: "Digital Signal Processing. A Practical Approach". Addison-Wesley Publishing Company. First Edition. 1993.

ISBN: 020154413X

[11] INOUE, T; Saruwatari, H.; Takahashi,Y.; Shikano, K. and Kondo, K.: "Theoretical Analysis of Musical Noise in Generalized Spectral Subtraction Based on Higher Order Statistics". IEEE Transactions on audio, speech, and language processing, VOL. 19, NO. 6, AUGUST. 2010.

Digital Object Identifier: 10.1109/TASL.2010.2098871

[12] Lockwood, P.; Boudy, J.; Blanchet, M.: "Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments." IEEE International Conference on Acoustics, Speech, and Signal Processing, VOL 1. Matra Commun., Bois d''Arcy. France. 1992.

Digital Object Identifier: 10.1109/ICASSP.1992.225921

[13] MATLAB WEBSITE:

http://www.mathworks.es/es/help/

[14] MOORER, J. A.: "DSP Restoration techniques for audio". Adobe Systems, Incorporated. 2007.

Digital Object Identifier: 10.1109/ICIP.2007.4379940

[15] OKAZAKI, M.; Kunimoto,T.; Kobayashi, T.: "Multi-stage spectral subtraction for enhancement of audio signals". IEEE. ProAudio & Digital Musical Instrum. Div., Yamaha Corp., Hamamatsu, Japan. 2004.

Digital Object Identifier: 10.1109/ICASSP.2004.1326380

[16] BORRAS, O.: "Reductor de ruido mediante resta espectral". EUIT Telecommunication. Spain. 2006.

[17] PROAKIS, J. and MANOLAKIS, D.: "Digital Signal Processing: Principles, Algorithms and Applications". McGraw-Hill Professional. 4th Edition. New Jersey, USA. 1995.

ISBN: D-13-394338-9

[18] SUNIL, D. K. and PHILIPOS, C. L.: "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise". Department of Electrical Engineering, University of Texas at Dallas. Texas. 2002.

Digital Object Identifier: 10.1109/ICASSP.2002.5745591

[19] Universidad Carlos III de Madrid: "Plantilla presupuesto TFG". Spain. 2014.

http://www.uc3m.es/portal/page/portal/administracion_campus_leganes_est_cg/proyecto_fin_carrera

[20] Universidad Carlos III de Madrid: Study materials for the course "Instrumentación acústica y control de ruido".

[21] VASEGHI, S. V.: "Advanced Digital Signal Processing and Noise Reduction". John Wiley & Sons, LTD. Second Edition. 2000.

ISBN: 978-0-470-75406-1

[22] VERA,Y. M.: “Cancelación activa de ruido”. Universidad de los Andes, Facultad de Ingeniería, Escuela de Ingeniería Eléctrica. Thesis. Mexico. 2008.

[23] WANG, G.; Wang, X. and Zhao, X.: “Speech Enhancement Based on a Combined Spectral Subtraction with Spectral Estimation in Various Noise Environment”. IEEE International Conference on Audio, Language and Image Processing. Sch. of Inf. Eng., Hebei Univ. of Technol., Tianjin. 2008.

Digital Object Identifier: 10.1109/ICALIP.2008.4590225

Appendix

APPENDIX 1. Vocal Activity Detector: ZCR and Energy

```
function [Voice_det] = vad_threshold(s,n);

frames = size(s,2);
%%% Voice Detection
noise_mean = s(:,n);
T = 20*log10(mean(s./(noise_mean*ones(1,frames))));%SNR
%%%%% Calculate of threshold
Voice_det=ones(1,frames);
threshold=mean (T);
for t = 1:frames
    if T(t) < threshold %The values that are below the limit given by
                          the user are considered noise
        Voice_det(t)=0;
    end
end
```

APPENDIX 2. Vocal Activity Detector: Threshold to detect voice

```
function [Voice_det] = vad_ZCR(s,Wind_long,bits,fs)
Signal_long=length(s);
maxS=max(abs(s));
WL_SL=Wind_long+Signal_long;
z=zeros(1,Wind_long);
aux1=round(s/maxS*2^(bits-7));
T_E=abs(s).^2;
T_zcr=1/2*abs(sign(aux1(1:Signal_long-1))-sign(aux1(2:Signal_long)));
T_zcr=[z T_zcr' z];
T_E=[z T_E' z];

%%%% Calculation of energy and ZCR:
for n=Wind_long:(WL_SL-1)
    zcr_sum=0;
    energy_sum=0;
    for r=(n-Wind_long+1):n
        energy_sum=energy_sum+T_E(r);
        zcr_sum=zcr_sum+T_zcr(r);
    end
    Energy(n-Wind_long+1)=energy_sum;
    ZCR(n-Wind_long+1)=zcr_sum;
end
ZCR=ZCR*(1/Wind_long)*fs;

ZCR_D=std(ZCR)*max(sign(ZCR-max(ZCR)/10),0);
```

```
E_D=std(Energy)*max(sign(Energy-max(Energy)/10),0);
p=1;
aux1=(E_D|ZCR_D);
for t=1:Wind_long:Signal_long
    Voice_det(p)=aux1(t);
    p=p+1;
end
Voice_det= Voice_det';

End
```

APPENDIX 3. Spectral subtraction: Boll

```
%%%% Noise estimation
i = 0;
for g = 1:frames
    if Voice_det(g) == 0 % wheter if noise.
        i = i+1; % Number of windows array voiceless sound
        N_mean= alfa*N_mean + (1-alfa)*Y(:,g);%average of noise
        n = n+1;
        Y(:,g) = c*Y(:,g); % replaces these samples by attenuated
                             noise
        Noise_array(:,i) = Y(:,g);
    end
end
N_mean = abs(N_mean);

%%%% Spectral Subtraction
X = Y - a*N_mean*ones(1,size(Y,2));
```

APPENDIX 4. Spectral subtraction: Berouti

```
%%%% Noise estimation
i = 0;
for g = 1:frames
    if Voice_det(g) == 0 % wheter if noise.
        i = i+1; % Number of windows array voiceless sound
        N_mean= alfa*N_mean + (1-alfa)*Y(:,g);%average of noise
        n = n+1;
        Y(:,g) = c*Y(:,g); % replaces these samples by attenuated
                              noise
        Noise_array(:,i) = Y(:,g);
    end
end
N_mean = abs(N_mean);
r = N_mean*ones(1,i);% Maximun deviation of noise.
```

```
%%%% Spectral Subtraction
noise_Power=sum(N_mean.^2);
Y_Power=sum((Y.^2),1);

%SNR
for f=1:frames
    SNR(f)=10*log10(Y_Power(f)/noise_Power);
    if (SNR(f)<-5)
        SNR_a(f)=4.75;
    elseif (SNR(f)>20)
        SNR_a(f)=1;
    else
        SNR_a(f)=4-((3/20)*SNR(f));
    end
    X(:,f)=Y(:,f).^(2*gamma)-(SNR_a(f)*((N_mean.^(2*gamma))));
end
```

APPENDIX 5. Spectral subtraction: Lockwood and Boudy

```
%%%% Noise estimation
i = 0;
for fra = 1:frames
    if Voice_det(fra) == 0 % wheter if noise.
        i = i+1; % Number of windows array voiceless sound
        n = n+1;
        Y(:,fra) = c*Y(:,fra); % replaces these samples by attenuated
                                   noise
        Noise_array(:,i) = Y(:,fra);
    end
    SNR(:,fra)=10*log10(Y(:,fra)./Noise_array(:,i));
    for fre=1:points
        if(SNR(fre,fra)<-5)
            a_SNR(fre,fra)=5;
        elseif (SNR(fre,fra)>20)
            a_SNR(fre,fra)=1;
        else
            a_SNR(fre,fra)=4-((3/20).*SNR(fre,fra));
        end
    end
end

%%%% Maximun values of the last frecuency
p=1;
lastM=i-M;
for fra=lastM:i;
    aux(:,p)=Noise_array(:,fra);
    p=p+1;
end
for frec=1:points
    alfai(frec)=max(aux(frec,:),[],2
end

%%%%Noise calculation
SNR=gamma.*SNR;
```

```
for fra=1:frames
    N_mean(:,fra) = alfai./(1+SNR(:,fra));
end
N_mean=abs(N_mean);

%%%% Spectral Subtraction
X = Y - a_SNR.*N_mean;
```

APPENDIX 6. Spectral subtraction: Multiband

```
%%%% Noise estimation
f=fs/points*(0:points1);
 for fra = (n+1):frames
    if Voice_det(fra) == 0 % wheter if noise.
        i = i+1; % Number of windows array voiceless sound
        N_mean= a*N_mean + (1-a)*Y(:,fra);%average of noise
        %         n = n+1;
        Y(:,fra) = c*Y(:,fra); % replaces these samples by attenuated
                                    noise
        Noise_array(:,i) = Y(:,fra);
    end

    Y2=Y.^2;
    N_mean2 = abs(N_mean).^2;
    Inicial=1;
    Final=bandwidth;

    for h=1:B
        Ysum=0;
        Nsum=0;
        for p=Inicial:Final
            Ysum=Y2(p,fra)+Ysum;
            Nsum=N_mean2(p)+Nsum;
        end
        SNR=10*log(Ysum/Nsum);
%%%% Calculate alfa
        if SNR<-5
            alfa=5;
        elseif SNR>20
            alfa=1;
        else
            alfa=4-(3/20)*SNR;
        end

%%%% Calculate delta
        if f(bandwidth)<=(1000||(fs-1000))
            delta=1;
        elseif f(bandwidth)>(((fs/2)-2000)||((fs/2)+2000))
            delta=1.5;
        else
            delta= 2.5;
        end
```

```
%%%% Spectral Subtraction

        for p=Inicial:Final
            X(p,fra)=Y2(p,fra)-alfa*delta*(N_mean(p
        end

        Inicial=Final+1;
        Final=Final+bandwidth;

    end
end
```

APPENDIX 7. Improvements

```
%%%% Option1:   Averaging module of the input signal.
    YMean = Y;
    for t = 2:(frames-1)
        YMean(:,t) = mean(Y(:,(t-1):(t+1)),2); % Average
    end
    Y = YMean;

%%%% Option 2:   Half-wave rectification
    noise_threshold = beta*Y;
    [I,J] = find(X < noise_threshold);
    X(sub2ind(size(X),I,J)) = noise_threshold(sub2ind(size(X),I,J));

%%%% Option 3:    Residual noise reduction
    for t = 2:(frames-1)
        I = find(X(:,t) < residual_noise + V_Ones);
        X_rn(I,t) = min (X(I,(t-1:t+1)),[],2);
    end
    X = X_rn;
```

(and*)