Pompeu Fabra University

Master's Thesis

Voice Transformations for Extreme Vocal Effects

Author:
Oriol Nieto

Supervisor:
Jordi Bonada

Master's Thesis submitted in partial fulfillment of the requirements for the degree:
Master's in Information Technologies, Communications and Audiovisuals

September 14, 2008

Abstract

Extreme Vocal Effects in music are so recent that few studies have been carried out about how they are physiologically produced and whether they are harmful or not for the human voice [1][2].

Voice Transformations in real-time are possible nowadays thanks to new technologies and voice processing algorithms [3]. This Master's Thesis pretends to define and classify these new singing techniques and to create a mapping between the physiological aspect of each EVE to its relative spectrum variations.

Voice Transformation Models based on these mappings are proposed and discussed for each one of these EVEs. We also discuss different transformation methods and strategies in order to obtain better results.

A subjective evaluation of the results of the transformations is also presented and discussed along with further work, improvements, and working lines on this field.

Resum

Els Efectes de Veu Extrems en la música són tan recents que pocs estudis han estat portats a terme sobre com es produeixen fisiològicament i si són realment dolents per a les cordes vocals o no [1][2].

Les transformacions de veu a temps real són possibles avui en dia gràcies a les noves tecnologies i als nous algoritmes de processament de veu [3]. Aquesta tesi pretén definir i classificar aquestes noves tècniques de cant extremes i crear una relació entre els aspectes fisiològics de cada EVE amb les seves respetives variacions de l'espectre.

Es proposen i es discuteixen Models de Transformació de Veu basats en aquestes relacions per a cadascun d'aquests EVEs. També es discuteixen diferents mètodes i estratègies de transformació per tal d'obtenir millors resultats.

També es presenta i es discuteix una avaluació dels resultats de les transformacions junt amb futurs treballs, millores i guies de treball en aquest camp.

Resumen

Los Efectos de Voz Extremos en la música son tan recientes que pocos estudios se han realizado sobre cómo se producen fisiológicamente y si son realmente dañinos para las cuerdas vocales o no [1][2].

Las transformaciones de voz a tiempo real son posibles hoy en dia gracias a las nuevas tecnologías y a los nuevos algoritmos de procesamiento de voz [3]. Esta tesis pretende definir y clasificar estas nuevas técnicas de canto extremas y crear una relación entre los aspectos fisiológicos de cada EVE con sus respectivas variaciones espectrales.

Se proponen y se discuten Modelos de Transformación de Voz basados en estas relaciones para cada uno de estos EVEs. También se discuten diferentes métodos y estrategias de transformación por tal de obtener mejores resultados.

También se presenta y se discute una evaluación de los resultados de las transformaciones junto con futuros trabajos, mejoras y guías de trabajo en este campo.

One thing I've learned.
You can know anything.
It's all there.
You just have to find it.

Neil Gaiman

Preface

This Master's Thesis has been carried out in the Music Technology Group as the final part of my *Master's in Information Technologies, Communications and Audiovisuals* in the Pompeu Fabra University, Barcelona.

It has been supervised by Jordi Bonada, and it would have been impossible to do it without his invaluable help. I want to thank him so much for all this huge effort and dedication he put in this project. I would also like to specially thank Xavier Serra, for tutoring me during the whole Master's Degree and for giving me advises everytime I asked for them.

Also many special thanks to Alex Misas, Toni González, Magali Pollac, Estefanía Figueroa and Ricard Benito for letting me dissect their amazing screams and voices and for collaborating with this project with so much enthusiasm and motivation. Many thanks to José Lozano for being so patient and helpful with me all the times I bothered him when I needed the studio.

Finally, infinite thanks to my family, to Sargon, and to all my colleagues in the MTG, for sharing with me their extensive knowledge and experiences, being so friendly and accessible all the time. This great journey is much more interesting and easy with you by my side.

Contents

Chapter 1

Introduction

1.1 Motivation

It has been many years from now that Music has been part of my life. As a child, there were two things that really fascinated me: Computers and Music. It was a big decision to eventually choose an Engineer career, but this did not mean to leave Music completely apart. I started my own band as a singer and guitarist, and I could combine perfectly my degree on Computer Engineer with the live performances of my band.

When I knew that a field called *Music Technology* existed and there were actually lots of good researchers (mostly engineers) working on it, it was finally an easy task to know which path I had to follow in this life. It was the combination of my two big passions: Engineer and Music.

As a singer and a fan of rock music, I have always been concerned about the possible damage that may occur to the vocal folds if singing in a not correct way. Moreover, this kind of music is quite aggressive and demands a lot from the singer in order to produce good effects with the voice. There is a reasonable big number of professional singers who have suffered from some vocal disorders due to a bad technique or abuse of the vocal folds.

That is why I came up with the idea of transforming a normal voice into a voice that has Extreme Vocal Effects (EVEs) on it. This would really help a lot to singers (included myself) in order to preserve their vocal folds and to create really good ways to express themselves.

Although these EVEs can be, theoretically, safely produced, they require a lot of time, patience and technique in order to not to harm the vocal folds. If a piece of software –or a small hardware device– could help them in this task, the number of voice disorders due to the production of these EVEs would decrease dramatically.

This, and the fact that few studies have been made about these EVEs, really encouraged me to start this research on this field that completely fascinates me.

1.2 Historical Background

Music is constantly changing and evolving, being always the reflection of our time and culture. Voice is probably one of the oldest and more complex instruments ever used by mankind, and it has so much potential that nowadays some of its possibilities are still being discovered.

Vocal Effects have been part of the singing music since its very beginning. In the traditional culture of some regions of the world, such as Mongolia with its characteristic Mongolian Throat Singing [4] or some tribes in Africa [5], one can find evidence that Vocal Effects are not something new either.

As music evolved and standarized in the occidental world, these Vocal Effects were put apart of the well-known classical music, where only a clean voice with or without falsetto could be used. By these times of classical and baroque music, no one could have conceived that voice would have been used in the way some singers are using it in the present time. Growls and Roughness in the voice started to be popular in the beginning of the twentieth century in blues and jazz music.

However, it has not been until the seventies, with the rock and heavy metal music, that these EVEs have been developed in a way that most singers of these styles decided to use them.

Starting from this point, more aggressive EVEs were introduced in the eighties by death metal bands such as *Sepultura* or *Morbid Angel*. Their voices sounded so broken that even pitch was impossible to detect. Their success was the beginning of a new era of music, where these EVEs started to develop and to get different colours and shapes.

In the nineties new aggressive EVEs were introduced in new kinds of extreme music such as *Grindcore* and *Hardcore*. These EVEs were dealing with much higher frequencies than the original ones, though they were aggressive and powerful as well.

These new ways of singing became so popular that many new bands using these techniques appeared, and many people started to wonder how to sing like this without hurting thier vocal folds. In June 2007, a report by the University Medical Center St. Radboud claimed that many young people were suffering from polyps and vocal disorders due to the increasing popularity of these EVEs [1].

However, a recent study in the Queens Medical Center from Nottingham claims that EVEs such as Distortion, Rattle, Growl and Grunt can be safely produced as long as they are done with the right technique [2].

By writing a software in order to produce these EVEs we could help all the singers that do not know how to produce them (or they simply do not have time to learn the right technique to produce them), assuring the safety of their vocal folds.

1.3 Voice Production

The voice organ can be divided into three different parts: the breathing apparatus, the vocal folds, and the vocal tract. Each one of these parts has different functions of their own that, put them all together, they produce the voice.

The breathing apparatus compresses the air in the lungs, and thereby an airstream is generated, passes through the vocal folds and finally through the vocal tract.

As the airstream passes through the vocal folds it creates a sound called the *voice source*. This voice source passes through the vocal tract so that it gets filtered out. Depending on the *articulation* of the vocal tract one will hear one sound or another.

In terms of engineering, we can say that the breathing system acts as a *compressor*, the vocal folds as an *oscilator* and the vocal tract as a *resonator*.

In Figure 1.1 we can see how this process is carried out. Starting from the bottom of the picture, the lungs are compressed and produce an airstream that passes through the trachea and then through the vocal folds, which vibrates and produces the voice source.

The *transglottal airflow* is the quantity of air that passes through the vocal folds in a given time.

In the figure we can see one and a half openings of the glottis (i.e. the vocal folds have been opened one and a half times in this graphic).

If we compute the spectrum of the voice source (also named glottal source) we can identify a plain sound where one can easily extract the fundamental frequency from. We can compare this type of sound to the one produced by a balloon when its opening has been slightly closed and the airflow is passing through it.

After that, the voice source is modified by the vocal tract. As said before, the vocal tract acts as a resonator, so that it adds the necessary formants to the voice to make it sound like a real voice. We can see in the final graphic (i.e the Radiated Spectrum) that the glottal source spectrum has been modified by this *filter* creating the formants that help us identify this sound as a human voice. These formants will be different deppending on the vowel that one wants to produce.

We have to point out that there are some consonants like \t or \sh that they do not need to use the voice source, since they are not phonated. In this cases only the airstream and the vocal tract take place in the production of the sound.

The preasure of the airstream that passes through the vocal folds is one of the key points in order to produce the singing voice in a safe way. This preasure will differe between singers and music styles (e.g. it is not the same to sing Rock than to sing Arias [6]).

One of the best references in order to fully understand how the voice is produced is, without a doubt, the book *The Science of the Singing Voice*, by Johan Sundberg [7].

1.4 State Of The Art

Here it is described the literature review (or State of the Art) of this Master's Thesis. It is a comprehensive review of the scientific literature regarding the Voice Transformation area of research.

1.4.1 General Voice Transformations

The crucial works that most of the modern implementations are based on are presented here as the very first approaches to the goal of this Master's Thesis.

Voice Source Models

The most well-known model for a glottal source is the one created by G. Fant and J. Liljencrants in 1985 [8]. They created a glottal model with four time parameters to characterize the sound of a glottal pulse (Figure 1.2). This model is referred to as the LF Model and has been used by many voice synthesizers and it has been shown to be a good model even to synthesize pathological voices too [9]. Even though the purpose of this thesis is to *transform* voices and not to synthesize them, it is very important to understand what our glottal source and voice folds are doing in order to create our voices. Thus, it is very important to understand this model, analyse it and, maybe, revise it in order to get the desired results.

The four paramters of this four-parameter model are:

- t_1: Time when the glottal flow opens
- t_2: Time of the maximum positive value of the glottal flow

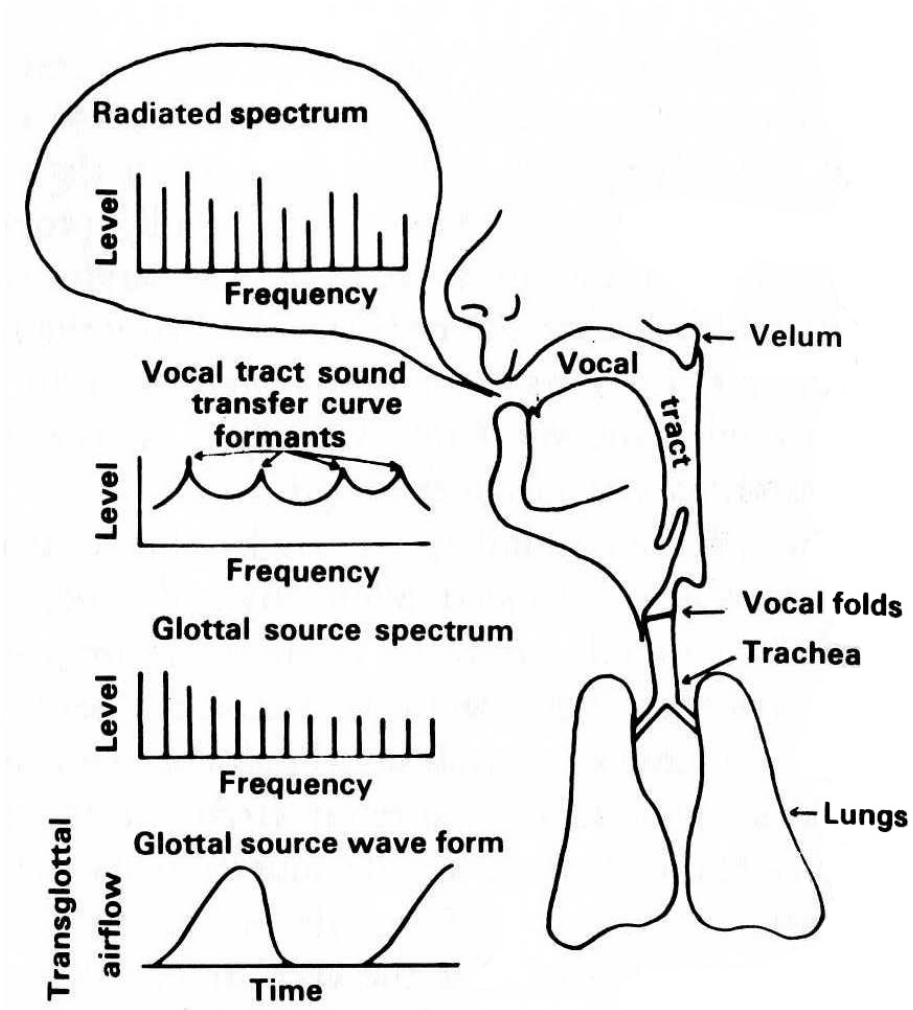

Figure 1.1: Schematic illustration of the generation of voice sounds (from [7])

- t_3: Time of the maximum negative value of the glottal flow derivative
- t_4: Time when the glottal flow closes

One of the most relevant uses of the LF Model in the topic of this Master's Thesis research is the one that A.L Lalwani and D.G. Childers made in 1991 by creating the New Unified Glottal Source [10]. This source is composed by three different types of sources: The Vocal Source, the Turbulent Noise Source and the Pitch Perturbation Source (Figure 1.3). The Vocal Source is basically a LF Model adding an IIR filter and then adding gain to the resulting sinusoid. The other two sources add Noise and/or Pitch Perturbation to the voice, thus creating a more natural voice. Not only this, but they defined five different type of voices that it could create, one of them a Rough Voice. So, this model is probably the very first approach to Roughness and Voice Transformations.

In 1995, M. Epstein, B. Gabelman, N. Antoanzas-Barroso, B. Gerratt, and J. Kreiman, analysed and revised the LF Model in order to inverse filter pathological voices. Results were very optimistic and they concluded that this model has sufficient degrees of freedom in order to reproduce pathological voices [9]. This is also relevant for this Master's Thesis since Roughness might sometimes be understood

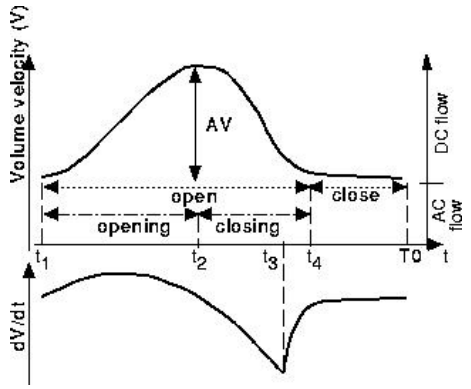

Figure 1.2: LF Model by G. Fant and J. Liljencrants. The four input parameters are t1, t2, t3 and t4 [8]

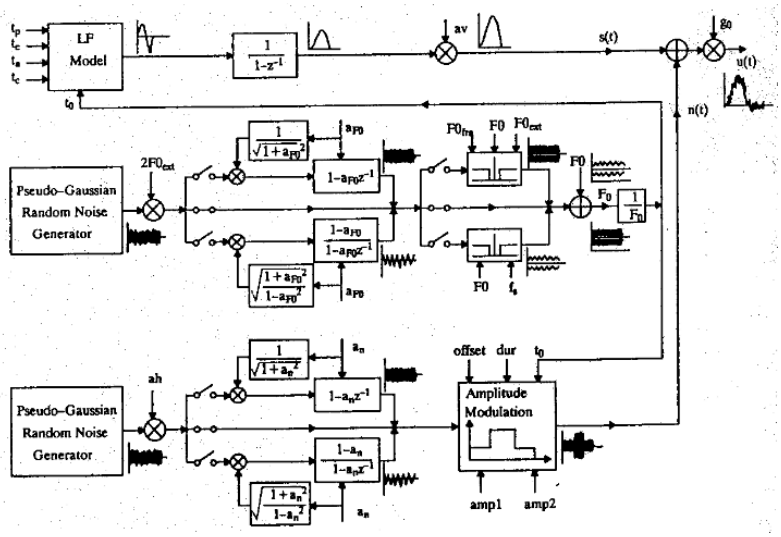

Figure 1.3: Glottal Source Model by A.L Lalwani and D.G. Childers [10]

as a pathological disease.

Voice Transformation Algorithms

Apart from having a robust voice source model, algorithms on voice transformation are also necessary in order to achieve the results of this Master's Thesis. Now we will review the most relevant ones.

The SMS (*Spectral Modeling Synthesis*)[11] is a sinusoidal model plus residual. Voice is decomposed into harmonic part and residual part. From there, features are extracted (timbre, pitch, volume, ...), then features are transformed and finally the result is synthesized. This model can also be applied to music, and not only to voice [12].

Another algorithm is the *Phase-Locked Vocoder* [13]. This algorithm takes the STFT (Short-Time Fourier Transform; see Appendix A) and decomposes it into regions associated to the spectral peaks (ideally the harmonics). These peaks are treated as sinusoids and then they are sinthesized by moving the spectral regions in frequency and amplitude.

On the other hand, the TD-PSOLA (*Time-Domain Pitch Synchronous Overlap Add Method*) [14] is an algorithm in the Time Domain. First of all, the onsets of each glottal pulse are determined. Then, the sound is segmented in overlapped frames of two periods of time. Finally, frames are overlapped in different instants.

An improvement of this last algorithm is the LP-PSOLA (*Linear Predictive Pitch Synchronous Overlap Add Method*) [15]. This algorithm filters the voice in order to obtain its formants and then have a plain spectrum. Then the voice is transformed using the TD-PSOLA algorithm and finally the formants are added again if needed.

Finally, *Wide-band Harmonic Sinusoidal Modeling* is an algorithm proposed by J. Bonada [3] and is the one that we have been using in order to do the transformations for this Master's Thesis. Each voice pulse is modeled by sinusoids. It allows to independly control each voice pulse and everyone of its harmonics. This new algorithm is the most suitable one in terms of real-time and high quality transformations, and that is why the author of this Thesis has decided to use it to create the transformations.

1.4.2 Roughness and Growling Implementations

Since these two techniques are highly related, they usually come together in the same project or even in the market products. In this part of the section the most important ones are analysed and commented.

In 1992, Perry Cook at Stanford University, developed a synthesizer based on physical models and using a wave guide articulatory vocal tract model, described in Cook [16, 17]. Even though it created a synthesized voice, it was possible to adjust the tract of the virtual voice in order to produce slightly roughness and growling (though it barely sounded as this, but as slight breathe) (Figure 1.4).

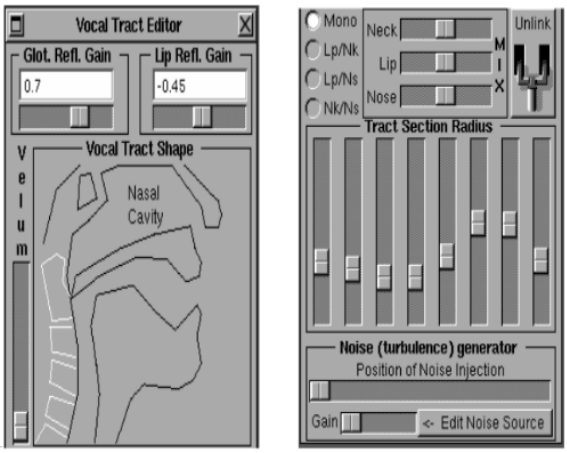

Figure 1.4: SPASM user interface [16, 17]

There's a product by Antares called THROAT which can transform voices and add them a slightly Roughness or Growling sound [18]. Although it can't achieve a lot of roughness or growling it can modulate the throat in order to try to imitate it. THROAT's breathing controls can add variable frequency noise to the model, resulting in a range of vocal effects from subtle breathing to full whisper (Figure 1.5).

One implementation for getting a Roughness Voice in real-time was included in the Vocal Processor, a VST Plugin. The Vocal Processor project was a collaboration project of the MTG with YAMAHA and some partial research results are found in [19].(Figure 1.6).

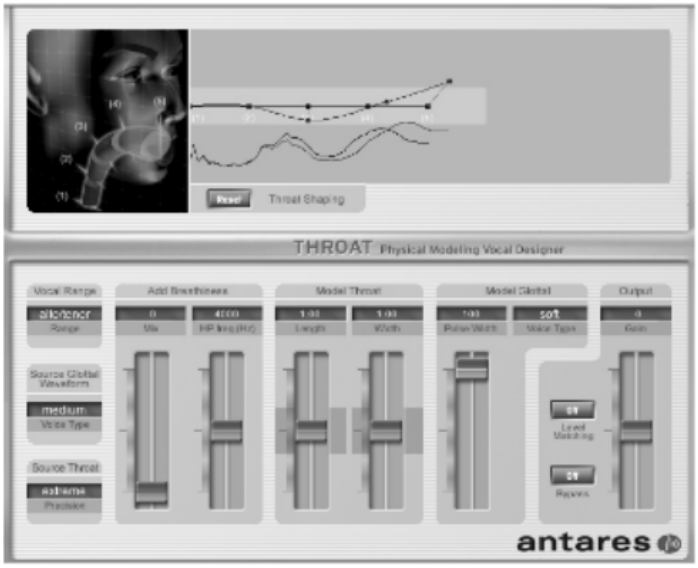

Figure 1.5: THROAT user interface [18]

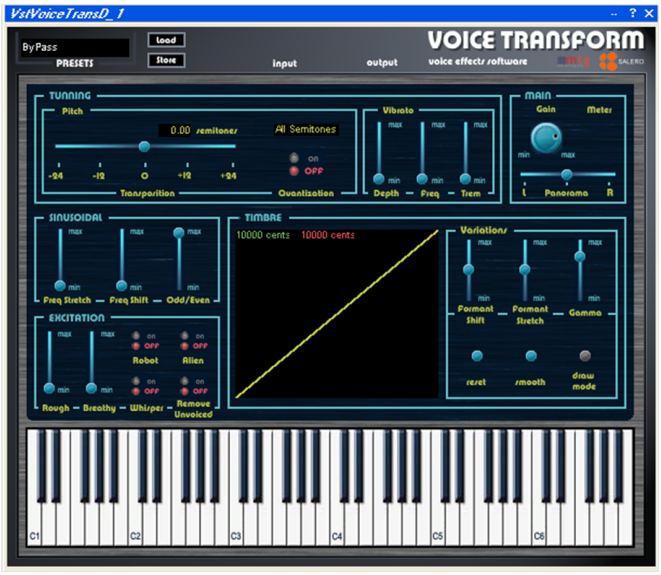

Figure 1.6: Voice Transformation VST plug-in [20, 21, 22]

This implementation for Roughness in this VST Plugin is the one proposed in the Spectral Domain by A. Loscos and J. Bonada in [23]. This implementation consists in pitch shift N octaves down the signal, add some random jitter and shimmer to each of the N octaves and overlap them all (Figure 1.7). The results were very exciting, but they could not make a satisfying approach in real-time. That is why they think that future work related to this topic is to improve their algorithm in order to be able to do a better implementation.

A. Loscos and J. Bonada proposed one Growling algorithm too. This is done by adding sub-harmonics in the frequency domain. They realized that growling is very different from one voice to another so they decided to add to their algorithm a controller to input how many sub-harmonics we

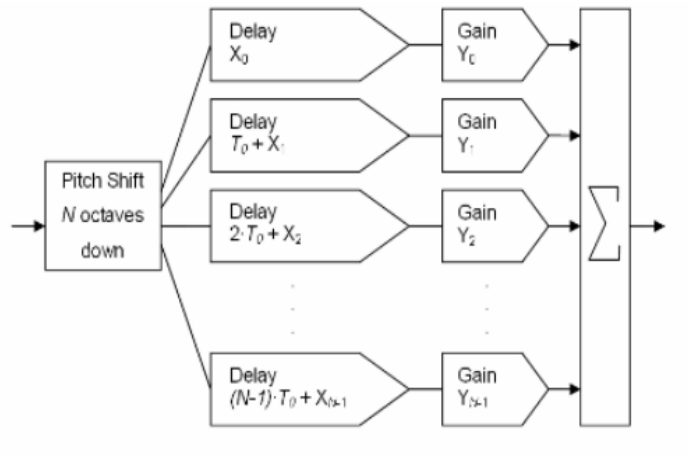

Figure 1.7: Implementation of the Roughness Algorithm by A. Loscos and J. Bonada [23]

wish to have depending on the person's voice [23](Figure 1.8).

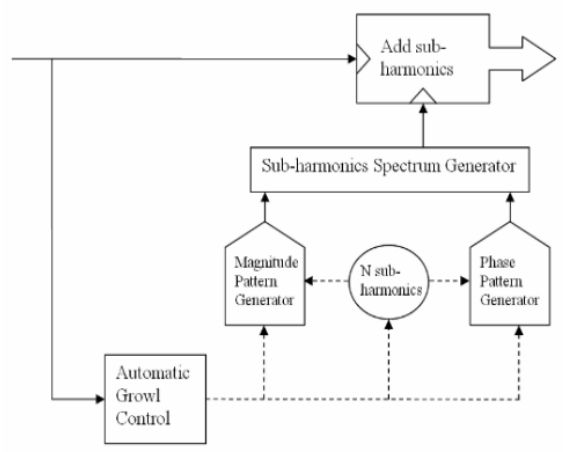

Figure 1.8: Block diagram of the Growling Algorithm by A. Loscos and J. Bonada [23]

These implementations do not sound natural in some cases depending on the voice type and characteristics. That is why the authors aim to do further work on this in order to improve them [23].

One of the most accurate implementations on Growling is the one that TC-Helicon has designed (Figure 1.9). It appears in the product called VoicePro [24]. This is one of the few professional implementations for growling but, still, it does not create extreme growl (*grunt*) or roughness like the musicians of modern rock music mentioned in section 1.2.

Research on pathological transformations has also been done. Although this is not the topic of this Master's Thesis, pathological voices are some times related to some EVEs, since some of these EVEs might sound pathological for some people. Some of this research can be found in references [25], [26], [27], [28] and [29].

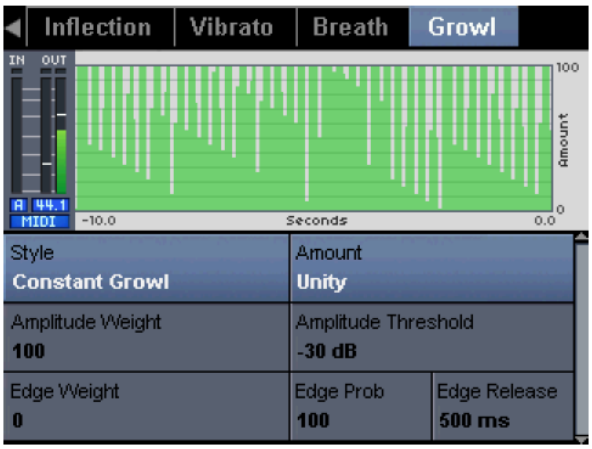

Figure 1.9: TC-Helicon VoicePro Growling interface [24]

Chapter 2

Methodology

This thesis methodology is divided into three different parts: Recordings, Analysis and Transformation Model.

One can see the block diagram of the methodology in Figure 2.1. Two different recordings were made by the same person for each EVE: the actual EVE (the *target*) and the same sound but without any vocal effect (i.e with a normal voice, or the *source*). For example, if the target EVE is a growled \o, then the source to record should be a normal \o with approximately the same pressure and volume as the source but without the EVE.

After that, these two recordings are analyzed separately and then compared with each other. Once the analysis and the comparison are done a basic Voice Transformation Model is defined in the Comparison Block and then implemented in the Transformation Block in order to proceed with the transformation. The transformation is applied to the source signal.

An iterative process begins if the transformation is not accurate enough. By subjectively evaluating the transformed signal, analysing it and comparing it again with the source, we refine our model and then we implement this new model to obtain the new transformed signal. This cycle is repeated until the result is subjectively good enough.

In the following sections we will describe and discuss each one of these parts in a comprehensive way.

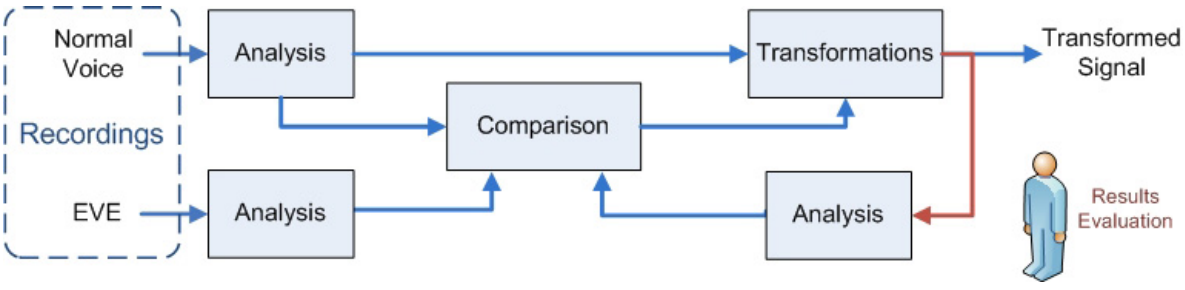

Figure 2.1: Block diagram of the Methodology

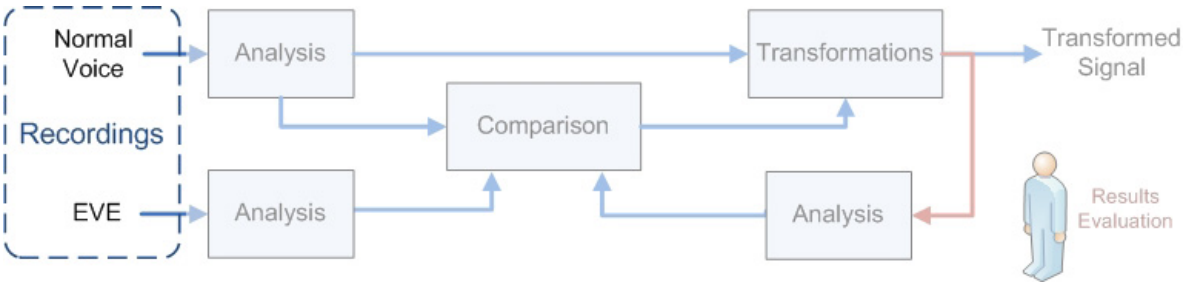

Figure 2.2: The Methodology Block Diagram, focusing on the Recordings Part.

2.1 Recordings

The Recordings are the starting point to create voice transformations. Six different people, five of them singers that use these EVEs techniques and another one a person with vocal disorders, were recorded.

Each recording was adapted to the range and capabilities of the singer. However, for each one of the EVEs recorded (the *targets*) there is also a recording produced by the same person trying to do exactly the same but without the EVE (the *source*), i.e. the same vowel and timbre but with no elements of the vocal tract vibrating, so that there is no EVE at all. Ideally, if one would apply the correspondant transformation to one source he or she would obtain the same sound as its relative target.

The recordings took place in the Audio Laboratory of the IUA (*Institute of Audio-Visual Studies*, Barcelona). The following material has been used:

- Neumann U87 Microphone
- Dynacord 16 channels mixing table
- Sony Vegas 6.0
- Cool Edit Pro 2.0

In the following subsections we will discuss each one of the recordings made, ordered chronologically. The exact number of EVEs recorded is found in Figure 2.3.

Recording 1

Name of the Singer: *Alex Misas*

In this first session the singer recorded five different types of EVEs: Distortion, Growl, Deep and High Grunt, and Scream. He also sang a whole song written by himself. This song contains all the desired EVEs except Rattle.

His voice is not very low, and yet it has enough power to perform Deep Grunt without problems. We consider his Scream EVE to be the best one among all the recordings for this work.

Recording 2

Name of the Singer: *Toni González*

This singer recorded Deep and High Grunts, Growls and a Scream. Like the first singer, he also sung a whole song which only contained Deep and High Grunts.

His voice is very hollow and dark, having the best Grunts (both Deep and High ones) recorded for this thesis.

Recording 3

Name of the Speaker: *Magali Pollac*

This recording session was made with the aim to analyze the voice of someone who have voice disorders, so that we could find some relation between a normal voice using some kind of EVE and a disordered voice.

Magali has had surgery twice on her vocal folds. By the time of this recording session she had a polyp in one of her vocal folds, so that her glottal flow could not be totally closed. This makes her voice to have a breathy or hoarse characteristic.

She recorded all the vowels in normal voice and also read a short text.

Recording 4

Name of the Singer: *Estefanía Figueroa*

This singer only focused on Grunts. It is very interesting to have female Grunts as well, since this EVE is thought to be produced only by male voices and this kind of female singers may be difficult to find. (Of course, there are important exceptions, such as the band Arch Enemy [30].)

She recorded some Deep and High Grunts. These EVEs were slightly higher pitched than the previous ones recorded, but one could hardly recognize if they come from a female or a male voice.

Recording 5

Name of the Singer: *Ricard Benito*

In this session, the singer recorded the following EVEs: Distortion, Rattle and Deep and High Grunt. He also recorded different EVEs in the same take, such as gradually going from Distortion to Deep Grunt, or from Rattle to High Grunt.

His voice is quite high pitched and clean, but when doing the EVEs it becomes dark and quite low pitched.

Recording 6

Name of the Singer: *Oriol Nieto*

In the last session the singer was the author of this thesis, that is why it was the easiest to be carried out, since he did not need anybody to tell him what to sing. Despite he is not a professional

singer, he could sing all of the EVEs for this thesis, i.e: Distortion, Rattle, Growl, Deep and High Grunt and Scream.

He has a rather clean voice, maybe that is why the Grunt effect could be improved. However the recordings are satisfactory.

As we can see in Figure 2.3, the Grunt effect is the one that has the most number of recordings, whereas the Distortion and the Scream effects are the ones that have the less. Unfortunately we could not find any other singer who could really produce these EVEs in the right way.

Another relevant aspect of the graph is that the Rattle effect was only produced by two people, which is also a drawback in order to create a general model that adapts to anybody's voice.

	Distortion	Rattle	Growl	Deep Grunt	High Grunt	Scream
Alex	2	-	2	1	2	3
Toni	-	-	2	2	2	1
Estefanía	-	-	-	4	1	-
Ricard	3	5	-	3	1	-
Oriol	3	3	3	11	2	2
TOTAL	6	8	7	21	8	6

Figure 2.3: Number of recordings made for each type of EVE

2.2 Analysis

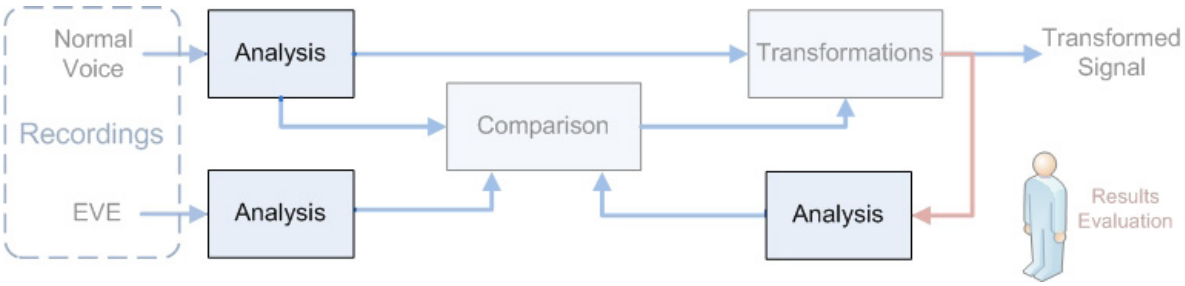

Figure 2.4: The Methodology Block Diagram, focusing on the Analysis Part.

The analysis of the recordings has been carried out by using the software SMSTools2 published by the *Music Technologpy Group* [31]. However, the specific version used for this thesis is still not published and it is exclusively used by the researchers within the group.

It is very easy to perform a good analysis with this tool, since it allows us to perform STFT with different windows types and sizes. It also provides pitch and glottal pulse onset detection among other features.

As shown in Figure 2.4, three different signals are analysed for each EVE:

- The *source* signal (upper left of Figure 2.4)
- The *target* signal (bottom left of Figure 2.4)
- The transformed signal (bottom right of Figure 2.4)

All three different signals are analysed using the same technique. However, we first analyse the *source* and *target* signals, since the transformed signal is not ready in the first iteration of the process. We perform this last analysis in the last part of our methodology (the Transformation Model part).

For each EVE we want to build a model by observing the energy variations in time of their relative spectrum. A first approach to this model will be built in the next step, in the Comparison block. When analysing the signals, the methodology used is to perform a STFT using a *Blackman Harris* window of 92dB and move it across time. The size of this FFT depends on the frequency of the glottal pulses and also on the macropulses of the signal.

Now we will explain what a macropulse is and how to identify it.

2.2.1 Finding Macropulses

As we will see in Chapter 3, some EVEs have macropulses that cover some specific number of glottal pulses. These macropulses are the key in order to transform a voice to obtain some of these EVEs.

To find these macropulses one can observe the signal in the time domain and analyze whether there are series of small pulses (the glottal pulse) that are covered by a larger pulse (the macropulse).

In Figure 2.5 there is a signal in time domain with the glottal pulses and the macropulses identified.

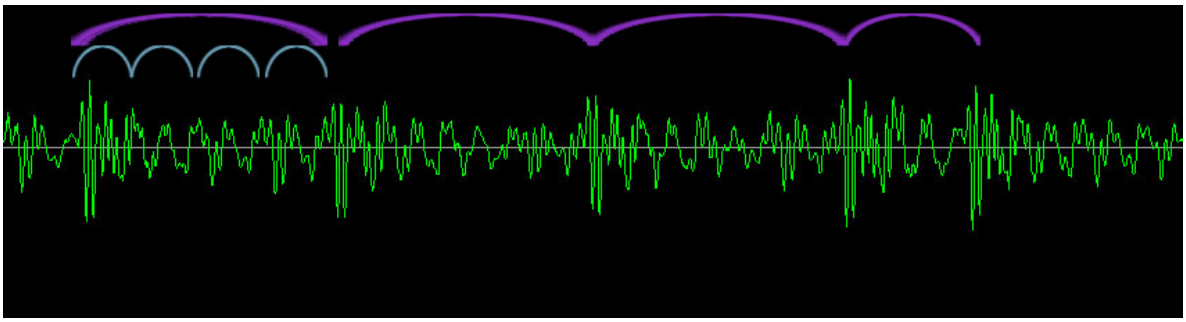

Figure 2.5: Four macropulses covering several glottal pulses. The glottal pulses of the second macropulse are also marked. This signal is a Distortion EVE.

It is important to identify these macropulses in order to build the model. For some EVEs the number of glottal pulses within a macropulse are a random number between an interval (in the previous Figure 2.5, we can see that the number of glottal pulses within a macropulse changes depending on the macropulse). For others EVEs the number of glottal pulses in a macropulse is constant (see Figure 2.6). And for others there is no macropulse at all (see Figure 2.7).

2.3 Transformation Model

The Transformation Model is the largest and most important part of the methodology. Here we compare the *source* and the *target* signals, define our EVE model, perform the transformations, perform an evaluation of the results, analyse the transformed signal and compare the transformed signal with the *target* signal. How this process is done is explained in the algorithm shown in Table 2.1.

As we can see from this algorithm, this part of the methodology is an iterative process that will not finish until the resulting transformed signal is perceptively good enough. The process starts by comparing the *source* and *target* signals to obtain the first approximation to our Transformation Model.

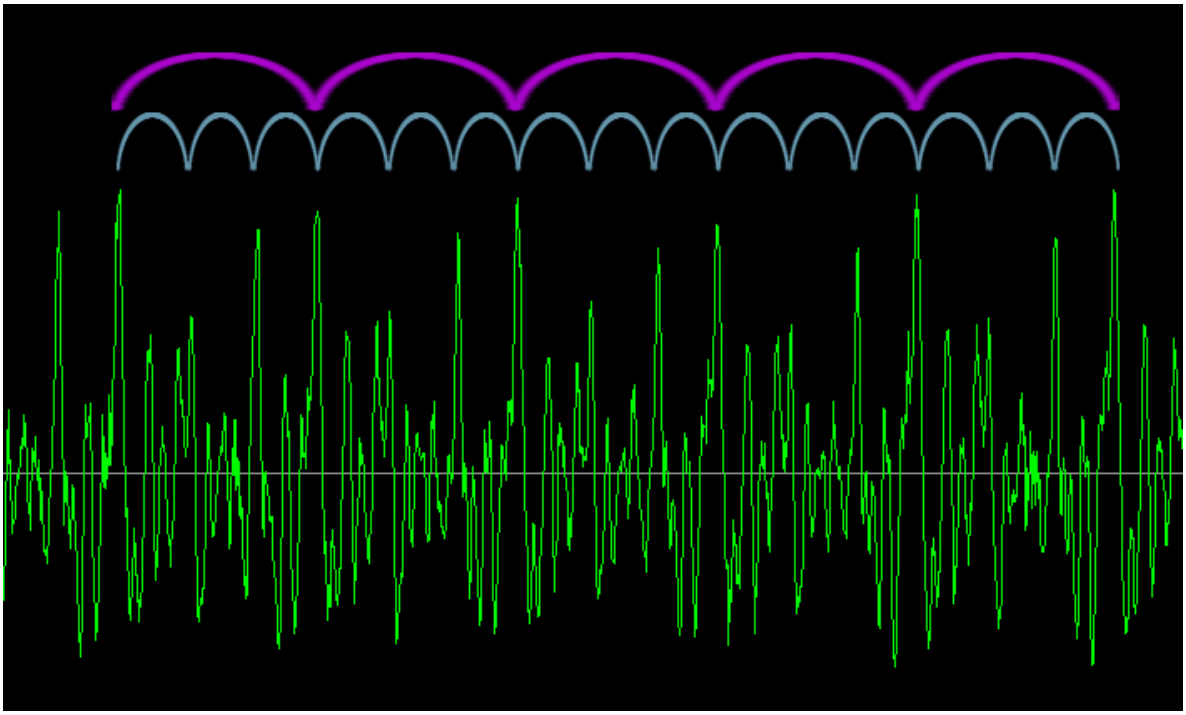

Figure 2.6: Here we see five macropulses covering three glottal pulses each. This EVE is a Growl.

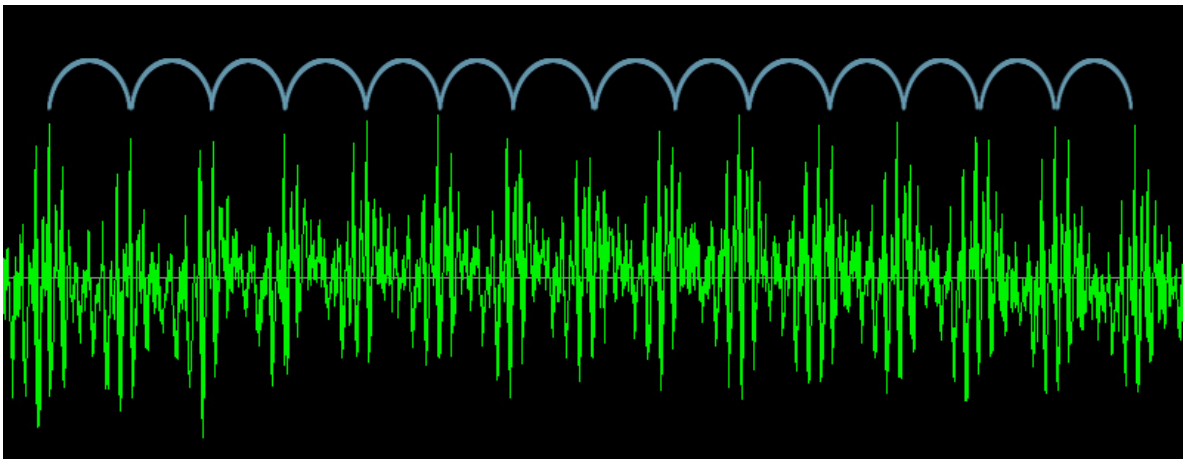

Figure 2.7: In this case, a Rattle EVE, there are no macropulses at all. The Glottal pulses are marked.

Then we define this model and implement it. We apply this transformation to the *source* signal and check if the transformed signal (the output of the transformation) is perceptively good enough (point 5 of the algorithm). If it is, we already have our model for this EVE. If not, we may start an iterative process (point 6). We analyse the transformed signal and then compare this results of the analysis with the results of the analysis of the *target* signal. In this point we try to refine the Transformation Model, implement it and apply it to the *source* again. Finally, we decide if the resulting transformed signal is perceptively good enough (i.e. listening to the sound of the transformed signal and deciding if it sounds real enough). If it is we will already have our model for this EVE, otherwise we will start another iteration in point 6 again.

```
1.   for each EVE:
2.      Compare the Source with the Target;
3.      Define the EVE Transformation Model;
4.      Implement and apply Transformation to Source;
5.      Perceptively evaluate the Result;
6.      while (Result perceptively not good enough) do:
7.         Analyse the Transformed Signal;
8.         Compare the Transformed Signal with the Target;
9.         Refine the EVE Transformation Model;
10.        Implement and perform the Transformation;
11.        Perceptively analyse the Results;
12.     end while
13.  end for
```

Table 2.1: Algorithm for the Transformation Model part of the methodology

This whole process can be divided into two separate parts: the *non-iterative process* (parts from 1–5 of the Algorithm of table 2.1, or the *source* and *target* process) and the *iterative process* (parts from 6–12 of the same Algorithm, or the transformed signal and *target* process). Now we will discuss these two different parts.

2.3.1 Non-Iterative Process

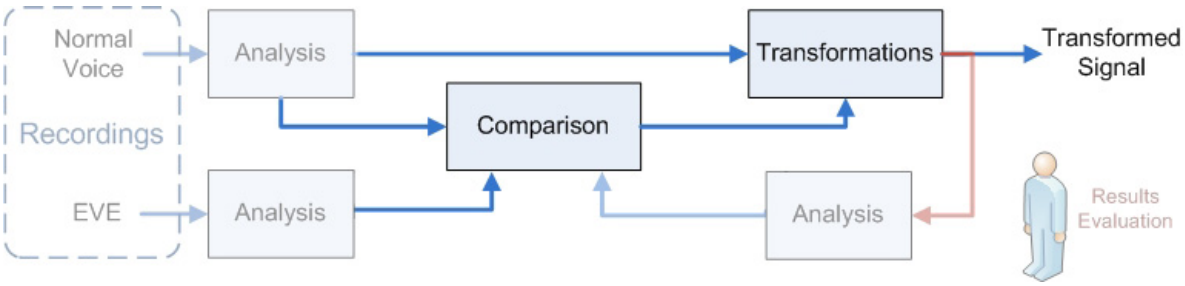

Figure 2.8: The Methodology Block Diagram, focusing on the Non-Iterative Process of the Transformed Model Part.

In this part of the methodology we compare the *source* and the *target*, define our Transformation Model, implement it and then subjectly evaluate the output transformed signal.

Source — Target Comparison

When comparing the source and the target signals we have to obtain the **first approach to the model** of the EVE we want to get. Depending on the EVE, this task is going to be more or less difficult.

For softer EVEs such as Rattle or Distortion, where pitch is audible, we have to see the spectral energy variations of the target and compare them to the source. By doing this we will be able to find in which frequency range the main part of the EVE is produced, and then build our model along with the number of glottal pulses within a macropulse (if the macropulse actually exists). We extracted this last information in the Analisys block.

In Figure 2.9 we can see the spectrum of a plain \a in the upper part and a spectrum of the same vowel but with a Distortion EVE in the bottom. We can appreciate that the timbre is roughly the same but there are changes in the energy of some parts of the spectrum. The region of the spectrum marked in the bottom part of the figure changes its energy across time, and that is the key point in order to obtain our model for the EVE.

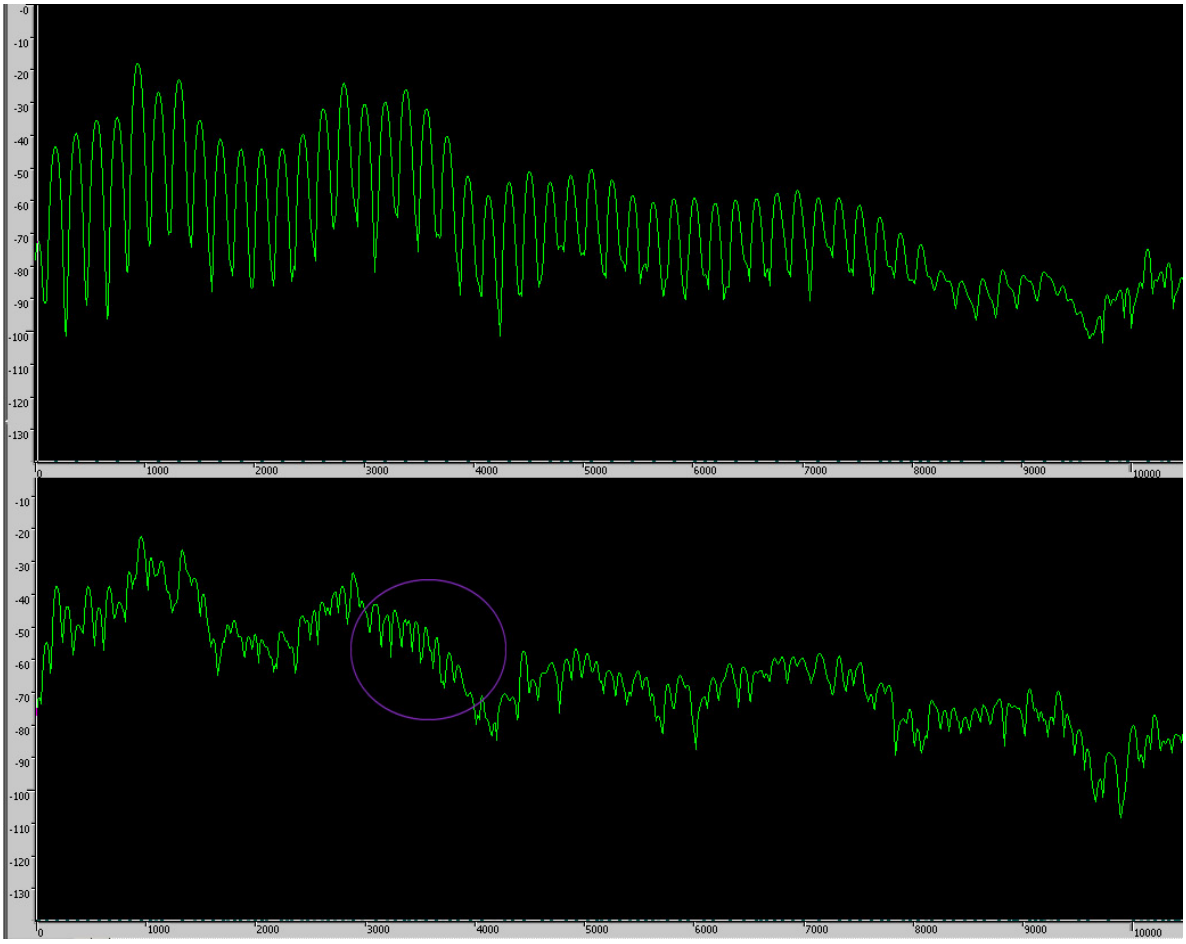

Figure 2.9: Two spectrums of the vowel \a. The upper one is a normal \a and the bottom one is one with a Distortion EVE. Marked is the region of the spectrum (from 3000–4000KHz) that varies the most across time.

Sometimes, when comparing an EVE that has pitch with its source, and depending on the EVE, we can see small variations of the pitch. This information is also very relevant in order to build our model and will have to be checked in every comparison.

For more aggressive EVEs such as Grunt, where the pitch is almost completely gone, it is going to be a more difficult task to build our first approach to the model. The model will have to be given shape on the next steps of the cycle, when comparing the *transformed signal* with the target.

In Figure 2.10 two spectograms of the vowel \u are presented. The upper one is a normal \u whereas the one in the bottom is the same \u with a Grunt EVE. This EVE is a very aggressive one and it makes the pitch disappear. Adding to this the fact that the spectrum is very unstable across time, makes a difficult task to build a model out of it.

However, when comparing these aggressive EVEs, one thing they have in common is that the

spectrum flattens in one point (in the case of Figure 2.10 it flattens from 4500KHz and on) and also that the pitch information is gone. We will take this as our first approach to the model and then we will give it shape in the next steps of the process.

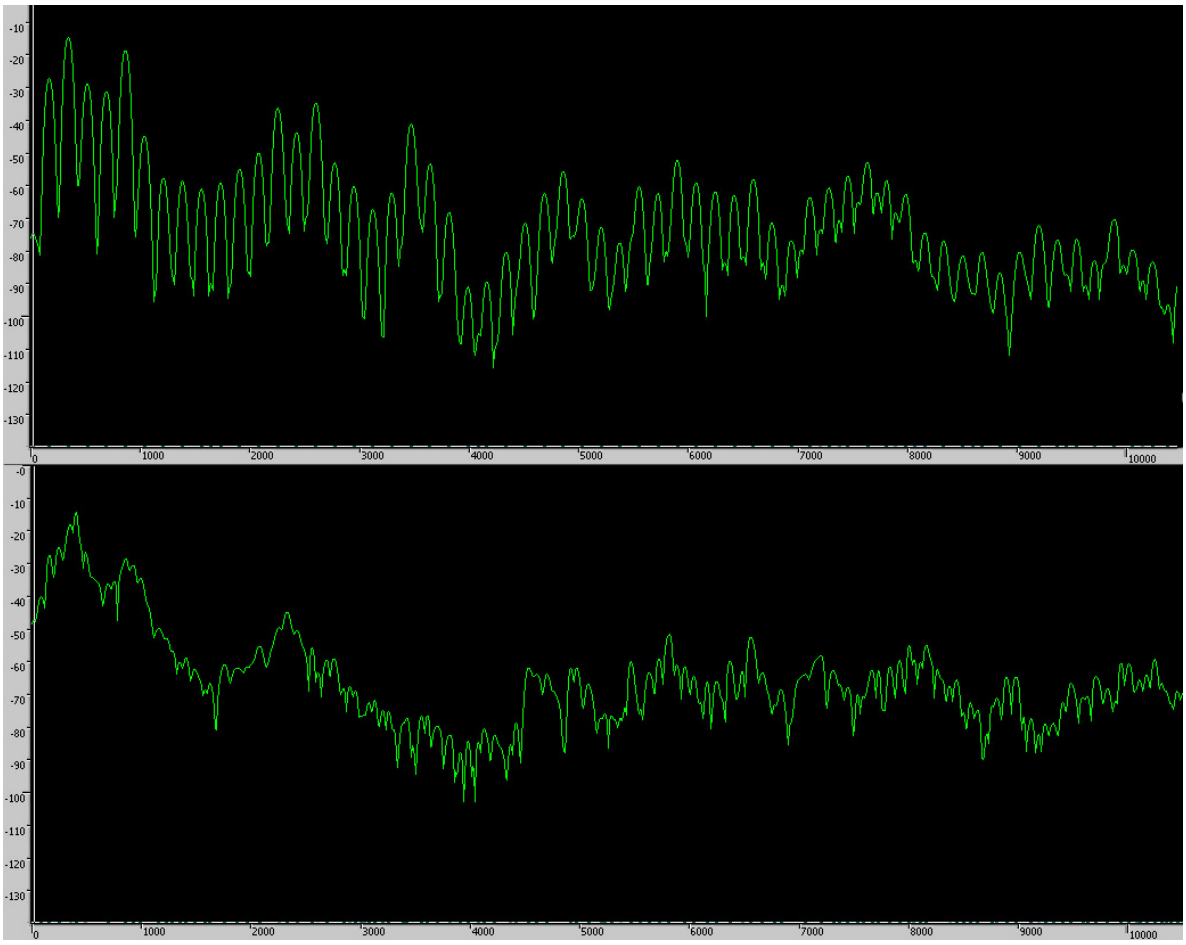

Figure 2.10: Two spectrums of the vowel \u. The upper one is a normal \u and the bottom one is one with a Grunt EVE. The spectrum of the bottom is not stable in time, and also the pitch is not audible. That makes a difficult task to extract a basic model for the transformation out of it.

For all EVEs we will also check the **timbre scaling**. Usually, the target signal has a stretched timbre compared to the source one. This is due to the vocal tract variations when producing the EVE.

We will also extract the degree of *noise* the target signal has compared to the source one. This will help us in order to change the phase of the signal components in the Transformation Block.

One of the major drawbacks when comparing sources and targets is that some of the sources are not very useful, specially these EVEs where the pitch disappears. In these cases, where the source has nothing to do with the target, we can only perceive some timbre coincidences between the source and the target. The rest, apparently, is random noise that depends on how the singer is using the vocal tract in order to create more or less *noise* for the EVE. However, we will also use this information for our model as said before.

The other problem found is that in some cases the global pitch is slightly different for the source and the target. This is because of the Recording process, where the singers could not remember the exact pitch when recording the samplers. This could lead to adding wrong information to our model

and thus, having pitch modifications in the transformed signal that should not be there. To overcome this problem, we have to analyze both *target* and *source* signals and check that the global pitches of both signals match. Having this in mind (that the pitch differences of the recordings are there due to the Recording process and not because of the EVE itself), this drawback is not a problem at all.

We have to point out that the more sources and targets we compare for each EVE, the more robust and reliable our model for this EVE will be. In this Master's Thesis it has been made a comparison for each one of the samplers recorded (see Figure 2.3), but not for the Scream EVE, since there was no time to perform the Comparison nor to build a model for this specific EVE.

Transformation

In this block we will implement the models that we have built in the previous blocks of the process (one for each EVE) and we will apply them to the *source* signal.

The tools used in order to implement and test the models are:

- Microsoft Visual Studio 2005
- SMSTools2 Source Code (in *C++*)
- CVS (*Concurrent Versions System*) [1]

The latest version of SMSTools2 has already implemented the new algorithm of glottal pulse detection by Jordi Bonada [3]. This and also because we could access the source code are the reasons why we used this software to make the transformations.

SMSTools2 is written in C++, and, with the Microsoft Visual Studio framework, the code implementation was a much easier task. With this framework it is very easy to debug your code, and you can actually *apply code changes* without having to close the debugging application (i.e. SPPTools2) and open it again every time a change is made.

We will define a new parameter to control our EVEs, which can take different values depending on what EVE we want to apply to the code. SMSTools2 reads a file with all the parameters needed for the Synthesis, and it is in this file where we will specify what type of EVE we want to transform the actual signal.

Our basic models to implement should contain, at least, the following information:

- Number of glottal pulses for each macropulse (it might be a random interval)
- Timbre modifications
- Energy variations of the spectrum (the frequency range and the amplitude)
- Pitch variations (the frequency range and the time variations)

Depending on the EVE, we could have another parameter to implement such as the % of random phasiness, depending on how much noise we want to add to the signal.

Once we have our model implemented, we will apply it to our source signal. This outcome will be our *Transformed Signal*. This new signal is going to be subjectively evaluated. If it is good enough we

[1] CVS was only used to retrieve the SMSTools2 source code.

will have our EVE model successfully constructed. Otherwise we will start the Iterative process (see subsection 2.3.2) of the Transformation Model.

In Figure 2.11 we can see a source signal (up) and the same signal transformed with a Grunt EVE (down).

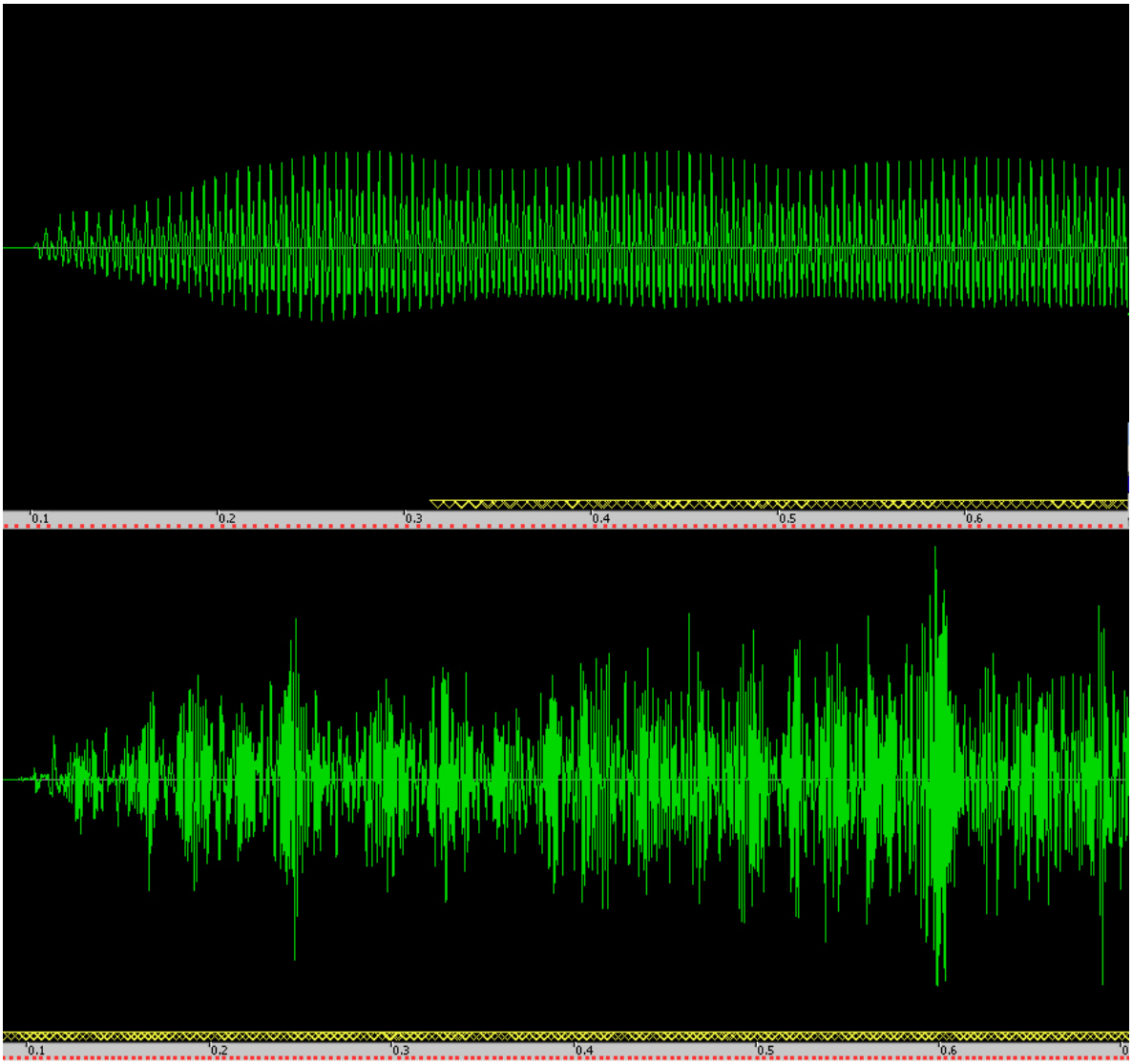

Figure 2.11: Here we see two signals in the time domain. The upper one is the vowel \u, and the bottom one is the same signal but with the Grunt EVE Transformation applied to it.

2.3.2 Iterative Process

In this process we will analyse the transformed signal as said in the Analysis block (see section 2.2), then compare the transformed signal with the *target*, refine our model and implement and apply the transformation. The transformation will be applied in the same way as in the non-iterative process (see section 2.3.1).

Transformed Signal — Target Comparison

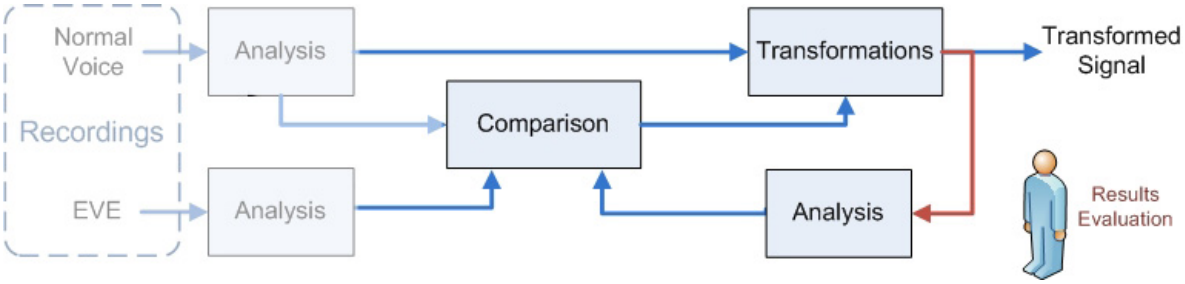

Figure 2.12: The Methodology Block Diagram, focusing on the Iterative Process of the Transformed Model Part

This time the target will be compared with the transformed signal in order to improve the model previously made for this EVE. We will have to check and review every aspect of the previous model.

Usually in this part, the range of frequencies of the energy variations are accurately changed. In some frequency regions (the lower ones specially), small changes might mean a lot in the sound, so in this comparison we will try to be as strict as possible. Fortunately SMSTools2 (the program we are using to analyze and compare the results) allows us to be accurate enough.

Another relevant aspect to have in mind at this time of the process is to check the possible pitch variations. Again, slight changes of the pitch variations (not only in frequency but also in time) might make a big difference to the sound.

In Figure 2.13 we can see two spectograms of the Grunt EVE: the transformed signal in the top and the target signal in the bottom. From only a snapshot it is not possible to visually decide if the signal is good enough or not. We should see how both signals change during time. In this case the first sinusoids are too low and there are still some frequency ranges that have to be enhanced.

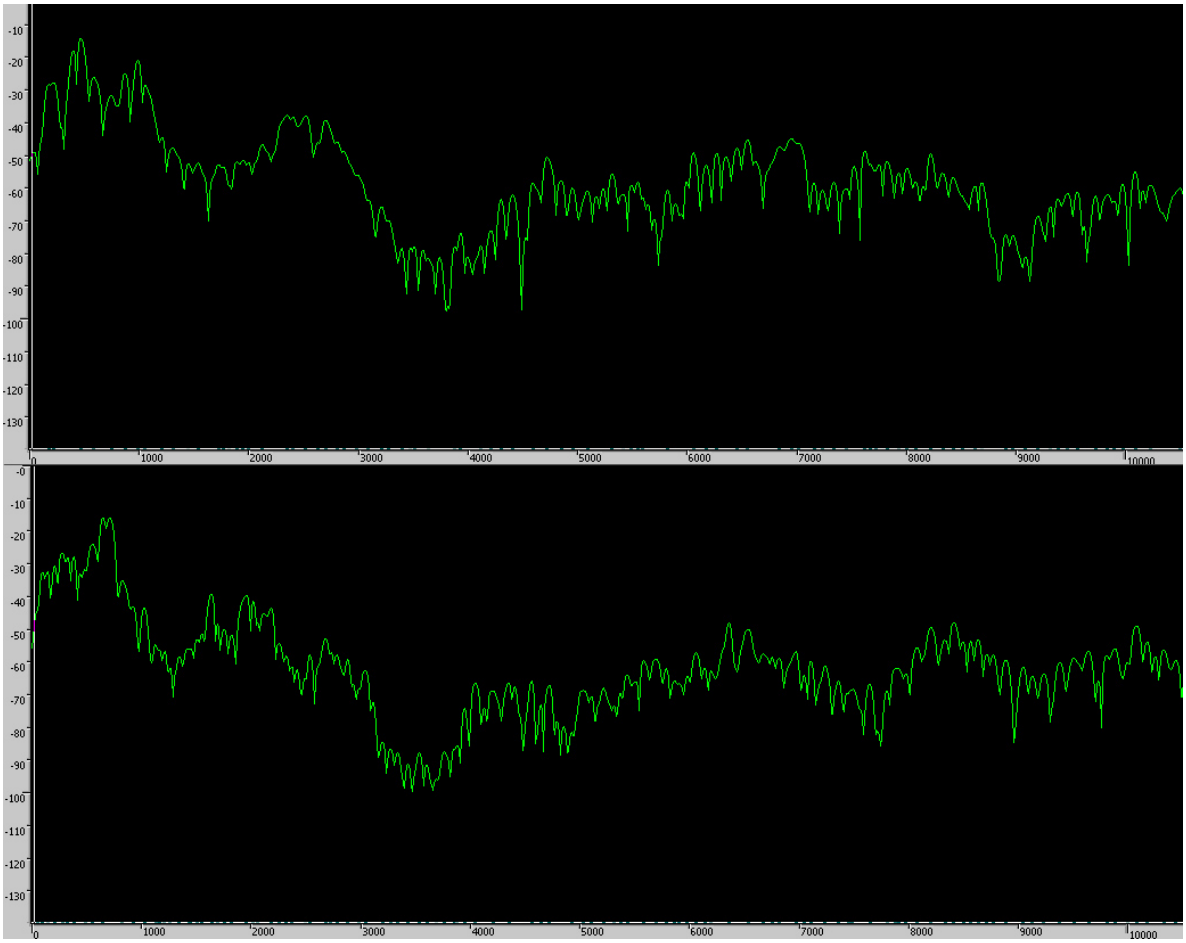

Figure 2.13: Two spectrograms of the vowel \u, both with a Grunt EVE. The upper one is a Transformed signal, the bottom one is the Target signal. Another comparison is necessary since, in this case, the Transformed Signal is not good enough.

2.3.3 Wide-Band Harmonic Sinusoidal Modeling

The *Wide-Band Harmonic Sinusoidal Modeling* is the method proposed by Jordi Bonada[3] that we have used to apply the Transformations. This method *estimates* and *transforms* harmonic components in wide-band conditions, by only taking one single period of the analyzed signal.

This method combines a good temporal resolution (typical from the time domain techniques) with the flexibility of frequency domain methods. The wide band methods only take one or two periods, so we have a good temporal resolution but it is hard to estimate the individual frequency components.

In this case, the idea is to take one single period and repeat it along the time domain, window it and extract the harmonic sinusoidal information out of it. Thus, we will have a very good temporal and frequency resolution at the same time.

In Figure 2.14 it is shown the analysis phase of this method. We have to compute the pulse onsets and then take one and overlap it along the time domain. Then a window will be applied to this overlapped signal and a FFT will be computed. We will have a high resolution spectrum for one single period. This process will be repeated for each pulse onset.

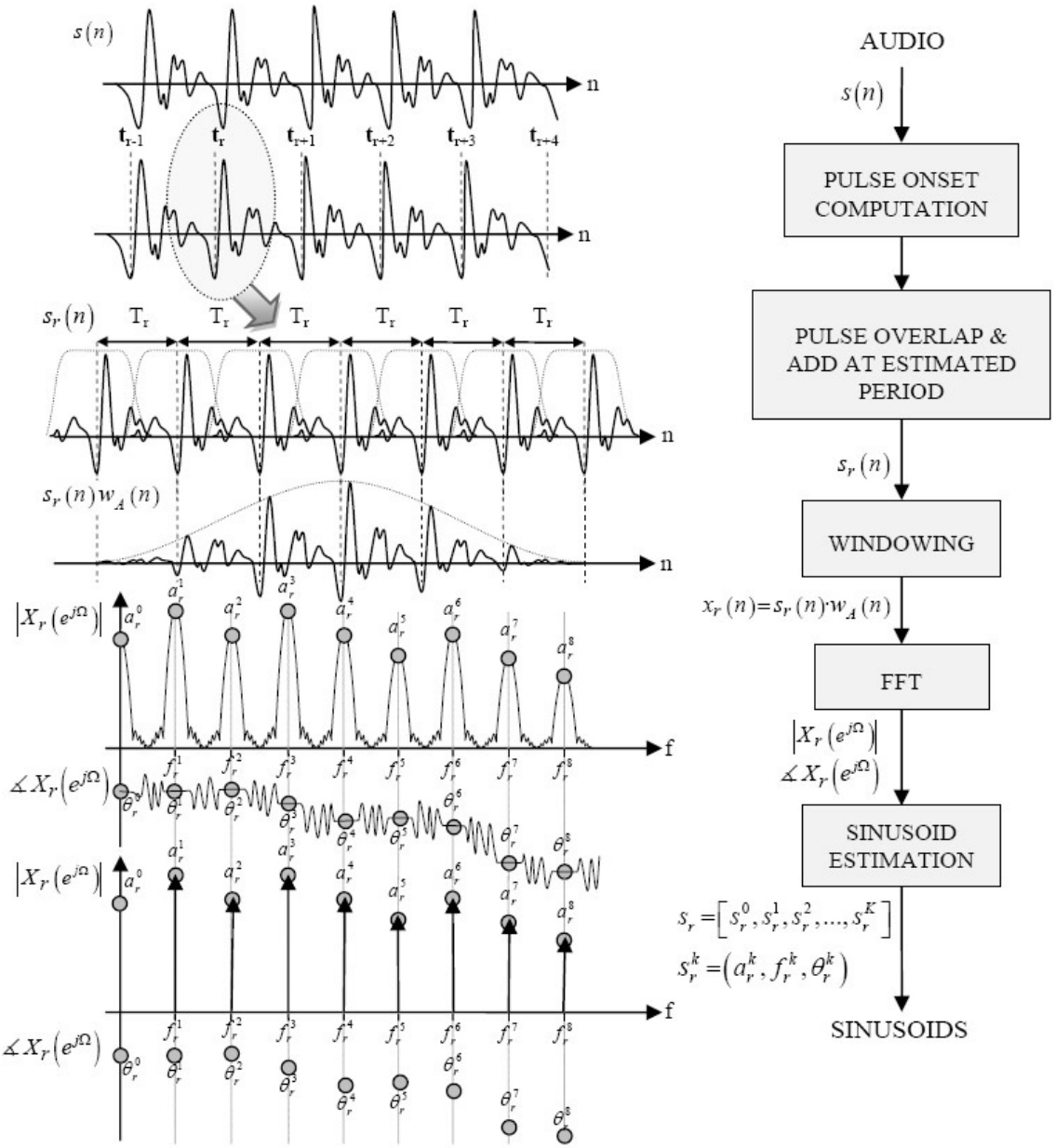

Figure 2.14: Block diagram of the analysis phase of the Wide-Band Harmonic Sinusoidal Modeling[3]

Since we will have a high resolution spectrum, we will be able to modify it in a very accurate way and also without losing temporal resolution. After applying the transformations needed to the sinusoids, we will synthesize the signal as shown in Figure 2.15.

With this algorithm we have the information of all the glottal pulses with an accurate precision of the harmonics of each one of these pulses. This makes the transformations of the spectral domain for each glottal pulse an easy task, and it is exactly this what we need in order to obtain our EVEs.

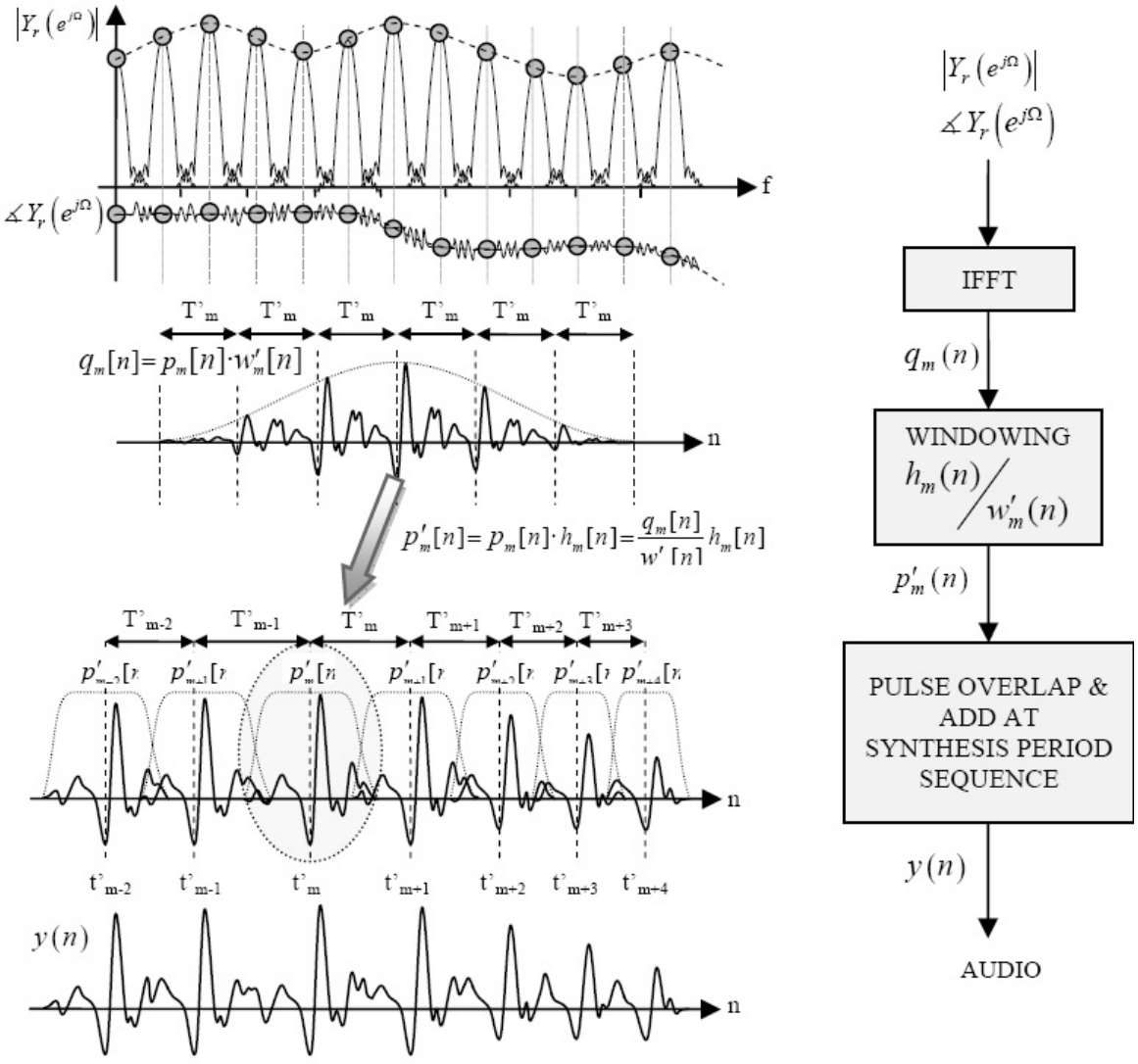

Figure 2.15: Sinusoidal Synthesis of the Wide-Band Harmonic Sinusoidal Modeling[3]

Chapter 3

Extreme Vocal Effects

In this chapter we will describe and classify the EVEs that have been recorded and analyzed in this Master's Thesis. We also propose a model for each EVE for transforming a normal voice into one of these EVEs.

The classification of the EVEs has been made based on the work by Julian McGlashan [2]. McGlashan's classification is the only one that the author of this Thesis could find. This classification divides the EVEs into four different categories: *Distortion*, *Rattle*, *Growl* and *Grunt*.

However, in none of these categories can fall the **Scream** EVE. This type of EVE is being used by a lot of new *Grindcore* and *Hardcore* bands in the past years, and that is why we think it is important to add this EVE into the classification. This final classification is:

- Distortion
- Rattle
- Growl
- Grunt
- Scream

In McGlashan's study, one of the most important conclusions is that, in order to perform this EVEs safely, the singer has to keep the vocal folds vibrating constantly, and create the EVE itself with the vocal tract, not with the vocal folds. With this in mind, it would require time and effort to develop a good technique for some singers in order not to harm their voices if they want to produce these EVEs.

All of these EVEs are effects that give more or less expression to the singing voice. They are mostly used in modern music such as blues, rock or metal, but they can also be found in other different scenarios such as in the traditional Mongolian Music. They are vocal sounds that are not connected to the melody or the lyrics and they can express different emotions depending on the music. More information about the historical background of the EVEs can be found in section 1.2.

Now we will describe each one of these EVEs physiologically and we will propose a transformation model for each one of them[1]. As said before, the key point in order not to harm the vocal folds when performing EVEs naturally is to use elements of the vocal tract (e.g. ventricular folds, cuneiform

[1]Except for the Scream EVE, which we are only going to describe.

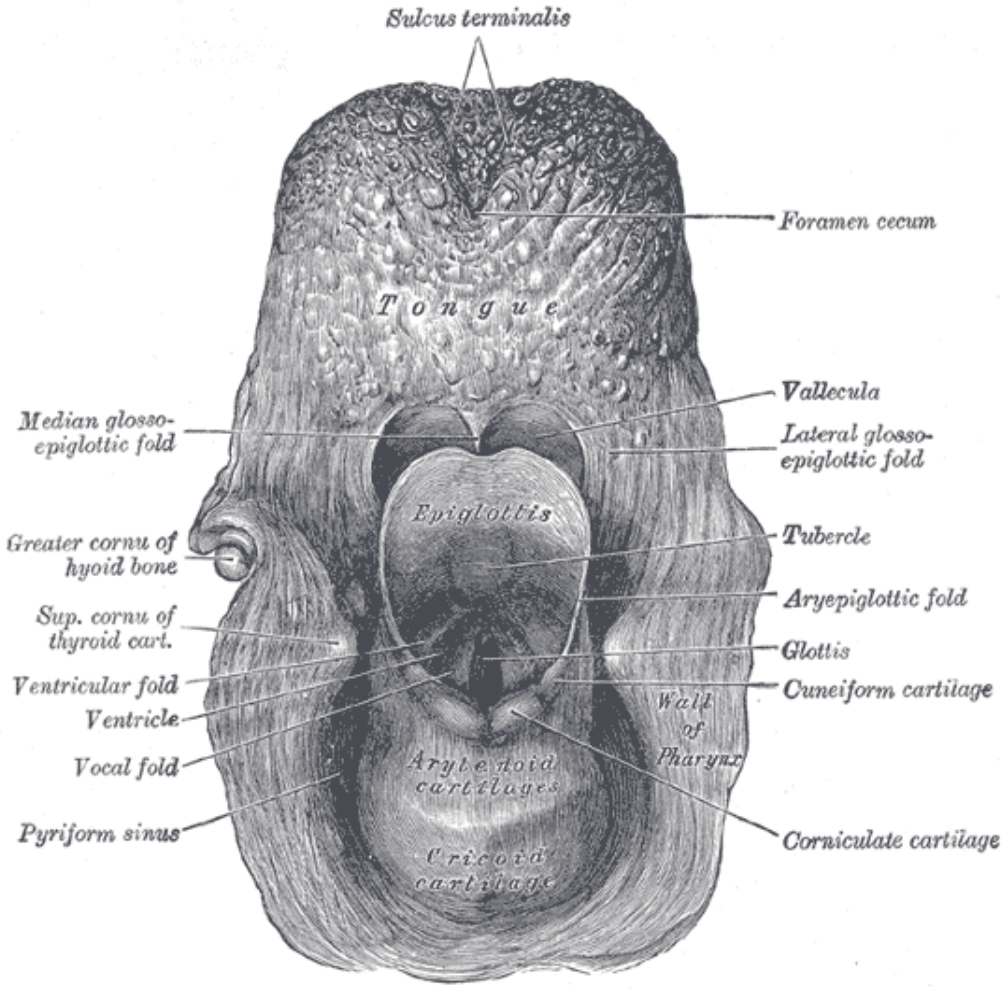

Figure 3.1: The entrance to the larynx, viewed from behind [32]

cartilage, mucosa, ...) to create these EVEs, and not the actual vocal folds. Theoretically, to produce healthy effects we should follow the three principle of singing: Open Throat, Support and No tightening of jaw and lips [7].

Major part of the physiological descriptions are based on the work by McGlashan [2]. One may refer to Figure 3.1 to fully understand the physiological description of each EVE.

3.1 Distortion

3.1.1 Description

This EVE gives a more aggressive expression to the voice. Physiologically, the *false vocal folds* (or *ventricular folds*) relax and they become part of the vibration. The vocal folds keep vibrating as usual. The sound that the ventricular folds produce is what we perceive as the Distortion.

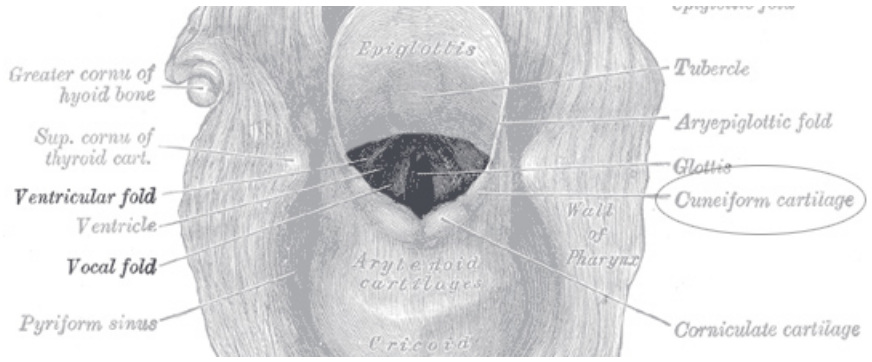

Figure 3.2: Structures of the Larynx that produce the Distortion EVE. Circled is the structure that is not necessarily needed to produce it (the *Cuneiform Cartilage*) but it might add a stronger degree of expression.

The use of the *Cuneiform Cartilages* can vary the degree of the EVE. The more tighten they are the more aggressive the EVE will be. In Figure 3.2 you can see where these structures are found in the Larynx.

This EVE is produced mostly by singers of classic rock, rock and grunge. Artist such as Kurt Cobain or Eddie Vedder often use this EVE. One recommended album to hear this EVE is *Ten* by the band *Pearl Jam* [33]. For example, in minute 0:49 of the song *Once* there is a typical Distortion EVE.

3.1.2 Transformations

In order to produce this effect, the larynx rises, stretching the vocal tract. This modifies the **timbre envelope**. It modifies it by stretching it **95%**.

According to our observations, the macropulses have a random number of **[2–5] glottal pulses**. This is due to the relaxation of the ventricular folds, which apparently vibrate between 2 and 5 times slower than the vocal folds.

The pitch is slightly moving in the analyzed samplers. Around **+3/-3 Hz**. The reason this happens is that the ventricular folds might slightly modify the pitch when relaxing too much. Thus the pitch is not completely stable for this EVE.

The Frequencies of the subharmonics change their amplitude without phase alignment. This adds some noise to the sound. We applied a **15% of noise** to the signal, so that the 15% of the phases are lost.

The sinusoid corresponding to the **F0 is lowered by 10dB**. Since the ventricular folds are also vibrating, this makes the F0 not as much audible as if they were not vibrating.

We apply different filters for the different glottal pulses within a macropulse. The filters applied are shown in Figure 3.3. It has been observed that the first and the second glottal pulses within a macropulse have different frequency enhancements than the rest of the glottal pulses.

By observing Figures 3.5, 3.6 and 3.7 we can see that the major part of the energy frequency variations are produced in the range of **[2000-3500] Hz**. In the first glottal pulse, when the ventricular folds close, there is this enhancement of these frequencies to the sound, whereas in the rest of glottal pulses we lower the energy of these frequencies.

The whole model for this transformation is found in Figure 3.4.

Frequencies (Hz)	First Glottal Pulse (dB)	Second Glottal Pulse (dB)	Rest of Glottal Pulses (dB)
0	3	3	3
400	2	-0.6	-0.33
500	2	-1.5	-1.67
600	2	-2.4	-2
700	1	-3.3	-3
800	0	-4.2	-4
1000	-6	-6	-6
1400	0	0	0
1500	0	0	0
2000	0	-10	-10
3000	2	-10	-10
3500	-15	-15	-15
5000	-15	-15	-15
5500	0	0	0
6000	0	6	0
10000	0	6	0
12000	0	0	0

Figure 3.3: Filters for the Distortion EVE

Model for the Distortion EVE	
Timbre Envelope	95%
Macropulse	[2–5] Glottal Pulses
Noise	15%
Pitch Variation	-3/+3 dBs
Filters	See Figures 3.3, 3.5, 3.6 and 3.7
F0	-10 dBs

Figure 3.4: Model for the Distortion EVE

All frequency ranges in each pulse vary randomly a number of times between [0.5–1.5]dBs. This adds a more realistic sound to the signal. There are always apparently small random variations to the sound, and these make the sound more real.

3.2 Rattle

3.2.1 Description

This EVE adds a raspy sound to the voice. It can be used in conjunction with Distortion. As in Distortion, vocal folds vibrate as usual, but now it is the Cuneiform cartilage mucosa which vibrates against each other or against the epiglottis that produces this rasp sound. In Figure 3.8 we can see these structures in the Larynx.

It is very common to find this EVE in the blues, rock and roll and hard rock music. Artists such

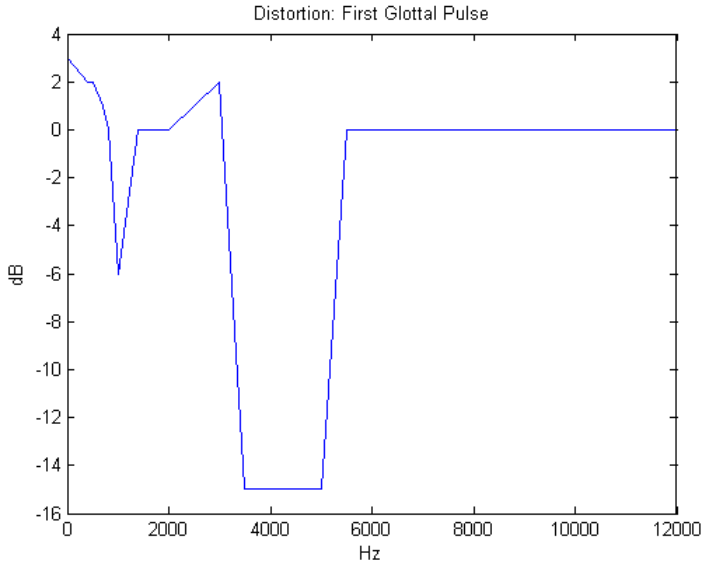

Figure 3.5: Filter for the first Glottal Pulse of a Macropulse for the Distortion EVE.

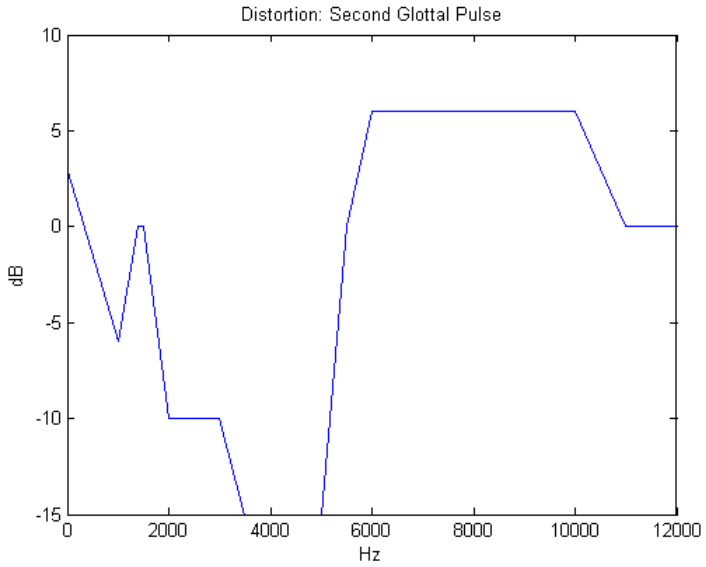

Figure 3.6: Filter for the second Glottal Pulse of a Macropulse for the Distortion EVE.

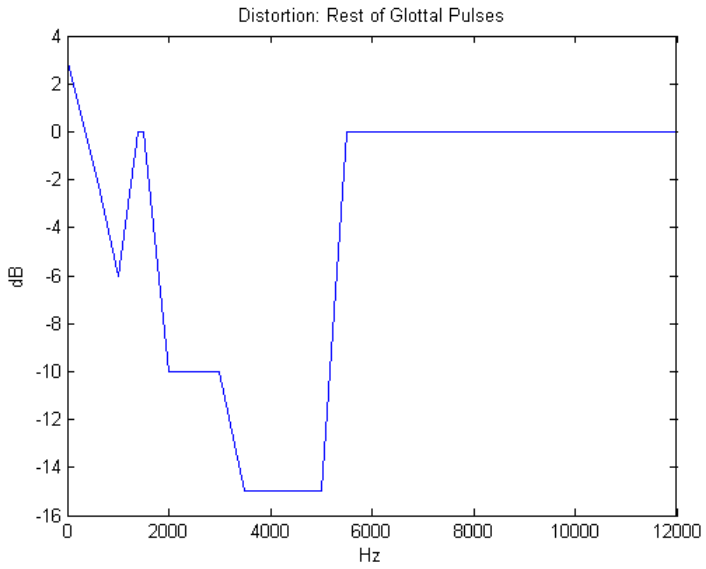

Figure 3.7: Filter for the rest of Glottal Pulse of a Macropulse for the Distortion EVE.

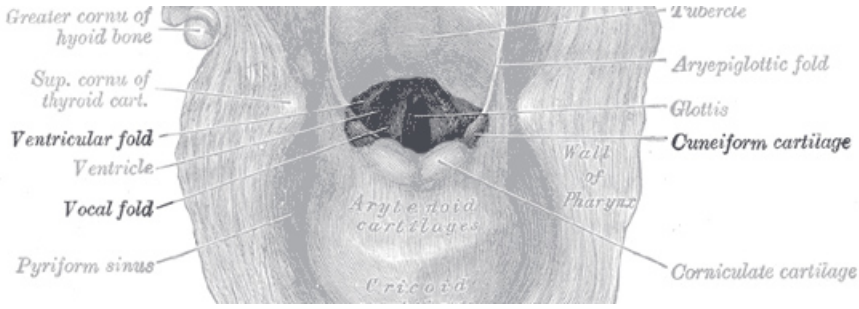

Figure 3.8: Structures of the Larynx that produce the Rattle EVE.

as Steve Tyler, Eric Martin or Joe Cocker use this technique a lot during their recordings and live performances. One of the recommended albums in order to hear this EVE is *Bump Ahead* by the band *Mr. Big* [34]. Eric Martin, the singer, is producing this EVE almost during the whole album.

3.2.2 Transformations

In this case, the **timbre envelope** does not need to be stretched, since the vocal tract stays the same when performing this EVE. So, there is no need to scale the timbre envelope (**100% of the Timbre Envelope**).

The number of Glottal Pulses within a Macropulse is a constant number. The Cuneiform cartilage mucosa vibrates the half of the vocal folds, that is why **the number of Glottal Pulses for a Macropulse is 2**.

The vibrations that create this EVE does not distort the signal at a big degree, so we can assume that **the quantity of noise is 0%**. The phases will always be aligned.

The energy of the first sinusoid is not affected because of the vibration of the structures. The

ventricular folds, the ones who have a strong influence in this sinusoid, do not vibrate as intensively as in the Distortion EVE, and the Cuneiform cartilage mucosa does not affect it at all.

Filters for second glottal pulse and the rest of glottal pulses are exactly the same as the one for Distortion due to the same vibrations of the vocal folds as in the Distortion during these glottal pulses (where the ventricular folds do not take place). On the other hand, the first glottal pulse enhances the frequencies from 1400 to 3000Hz, due to the sound that produce the Cuneiform cartilage mucosa. The Filters are shown in Figures 3.9, 3.11, 3.12 and 3.13.

The whole Model is shown in Figure 3.10

Frequencies (Hz)	First Glottal Pulse (dB)	Second Glottal Pulse (dB)	Rest of Glottal Pulses (dB)
0	3	3	3
400	2	-0.6	-0.33
500	4	-1.5	-1.67
600	3.5	-2.4	-2
700	3	-3.3	-3
800	0	-4.2	-4
1000	-4	-6	-6
1400	2	0	0
1500	2	0	0
2000	2	-10	-10
3000	4	-10	-10
3500	-15	-15	-15
5000	-15	-15	-15
5500	0	0	0
6000	0	6	0
10000	0	6	0
12000	0	0	0

Figure 3.9: Filters for the Rattle EVE

Model for the Rattle EVE	
Timbre Envelope	100%
Macropulse	2 Glottal Pulses
Noise	0%
Pitch Variation	0 dBs
Filters	See Figures 3.9, 3.11, 3.12 and 3.13
F0	0 dBs

Figure 3.10: Model for the Rattle EVE

All frequency ranges in each pulse vary randomly a number of times between [0.5–1.5]dBs. This, as in the Distortion EVE, adds a more realistic sound to the signal.

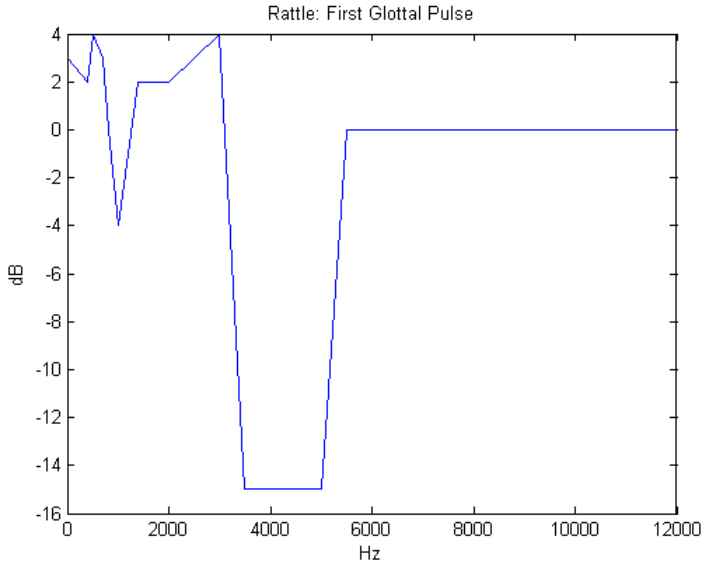

Figure 3.11: Filter for the first Glottal Pulse of a Macropulse for the Rattle EVE.

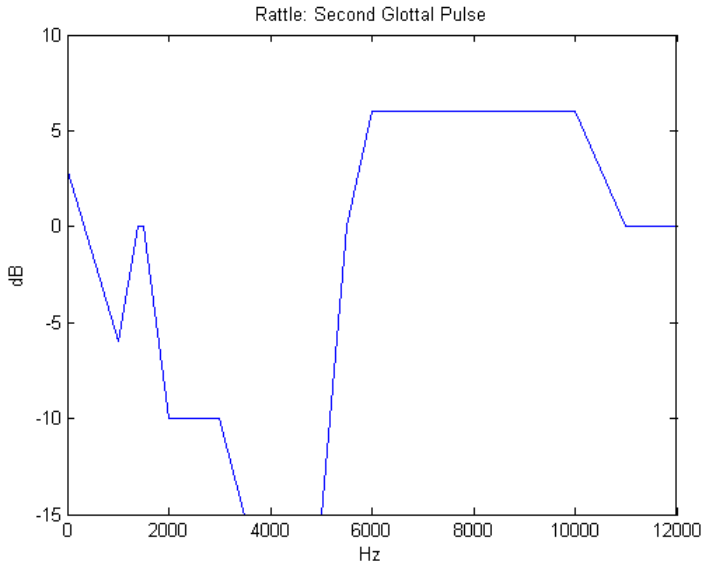

Figure 3.12: Filter for the second Glottal Pulse of a Macropulse for the Rattle EVE.

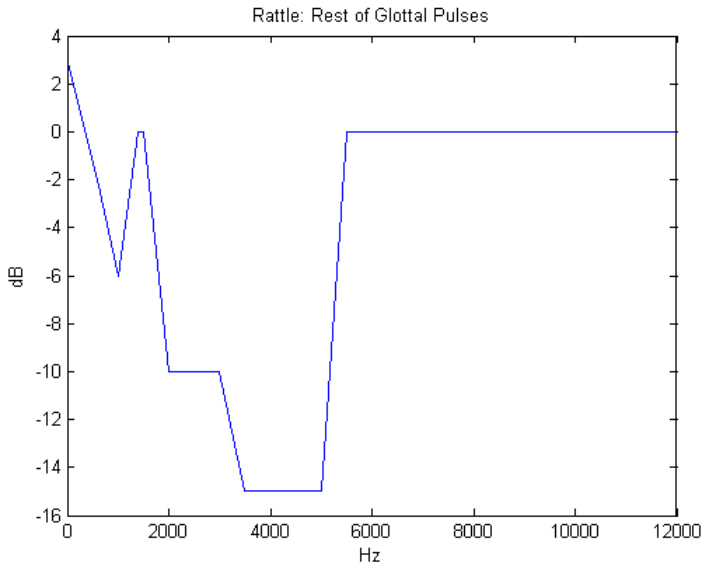

Figure 3.13: Filter for the rest of Glottal Pulse of a Macropulse for the Rattle EVE.

3.3 Growl

3.3.1 Description

This EVE is the most popular of the EVEs described in this Thesis. It is also the oldest one, since it not only appeared in the first blues songs, but it has been used by the Mongolian Throat Singers since the ancient times.

Growling adds a more aggressive expression to the voice. It is more aggressive than Distortion, but some singers such as Louis Armstrong used it even in their soft songs. It is used in a wide variety of music styles, from jazz to metal music.

Physiologically, ventricular folds vibrate with a large amplitude, and also the *Aryepiglotic folds* are vibrating. The Cuneiform Cartilages may also vibrate against the epiglotis, depending on the degree of the effect. These are the structures that create the Growl EVE. The vocal folds, as usual, keep vibrating the same. In Figure 3.14 we can see which are the structures that vibrate with this EVE inside the Larynx.

There are a lot of artists who use this type of EVE. Some of the most popular ones would be: Louis Armstrong, Freddy Mercury, Michael Jackson or Whitney Houston. The classical example of this EVE is the popular song *What a Wonderful World* recorded by Louis Armstrong in 1967.

3.3.2 Transformation

The position of the vocal tract in order to create the Growl EVE makes it stretch at about 80%. That is why **the Timbre Envelope is 80% stretched**.

Due to the amount of structures in the Larynx that are vibrating, an important quantity of noise is added to the sound. There is approximately **25% of noise**.

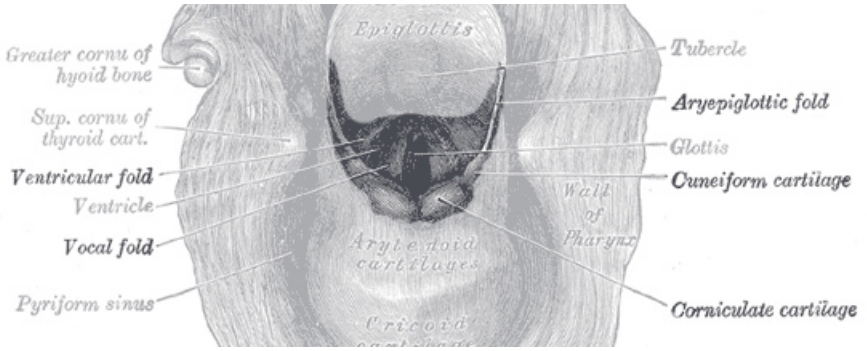

Figure 3.14: Structures of the Larynx that produce the Growl EVE.

The energy of **the first sinusoid is lowered 5dB**. The ventricular folds are vibrating in a way that is not as strong as in the Distortion but neither as soft as in Rattle.

The **number of Glottal Pulses for each Macropulse is 3**. This might differe between singers, but it has been observed that it is always a constant number, and the mean of all the analyzed samplers is 3.

The Filters for the Growl are found in Figures 3.15, 3.17, 3.18 and 3.19. There is a strong energy variation between the first and the rest of glottal pulses, specially in regions from 0 to 260Hz and 800 to 3000Hz. The responsible for these are the Aryepiglotic folds and also the Cuneiform cartilage.

The whole model for the Growl EVE is found in Figure 3.16.

Frequencies (Hz)	First Glottal Pulse (dB)	Second Glottal Pulse (dB)	Rest of Glottal Pulses (dB)
0	3	0	0
160	7	0	0
260	0	0	0
800	2.92	0	0
1000	4	-16	-6
1400	10	0	0
1500	4	0	0
2000	4	0	0
2500	10	6	1
3000	4	12	2
3500	0	0	0
5000	0	0	0
5500	4	0	0
12000	4	0	0

Figure 3.15: Filters for the Growl EVE

3.4 Grunt

3.4.1 Description

This is the darkest effect. It is the most aggressive one along with the Scream EVE. This EVE produces a very deep sound, nearly as dark and hollow as the monsters of some horror movie. This sound erases

Model for the Growl EVE	
Timbre Envelope	80%
Macropulse	3 Glottal Pulses
Noise	25%
Pitch Variation	0 dBs
Filters	See Figures 3.15, 3.17, 3.18 and 3.19
F0	-5dB

Figure 3.16: Model for the Growl EVE

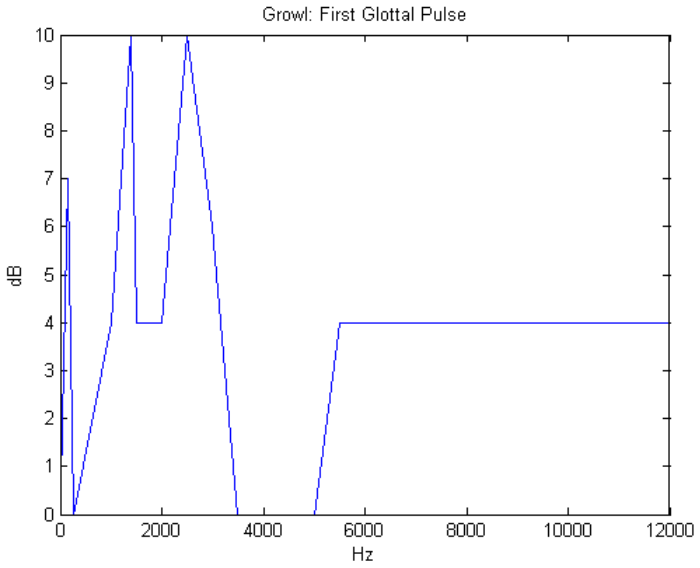

Figure 3.17: Filter for the first Glottal Pulse of a Macropulse for the Growl EVE.

the real pitch of the singer due to the enormous quantity of noise that the structures of the vocal tract add to the sound.

In this case, the false cords, the Aryepiglottic folds and the whole supraglottis structure are vibrating. The vocal folds also vibrate, but the other structures make the pitch of the vocal folds disappear. In Figure 3.20 we can see the structures of the Larynx that vibrate when producing this EVE.

This EVE is used in metal and all its extreme subgenres such as Death Metal, Thrash Metal or Black Metal. Some of the singers that often use these techniques are Max Cavalera or Mikael Åkerfeldt. One recommended album with a high quantity and quality Grunts is *Blackwater Park* by the band *Opeth* [35]. For example, in the song *The Lepper Affinity* almost the whole song is sung with the Grunt EVE.

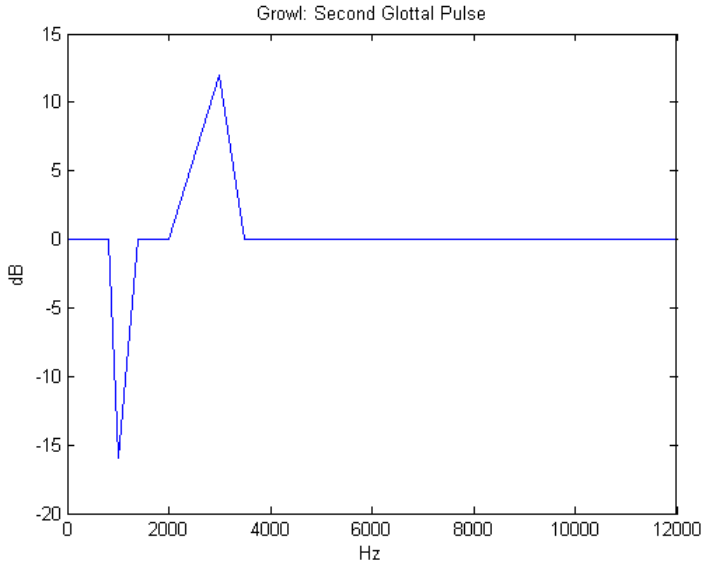

Figure 3.18: Filter for the second Glottal Pulse of a Macropulse for the Growl EVE.

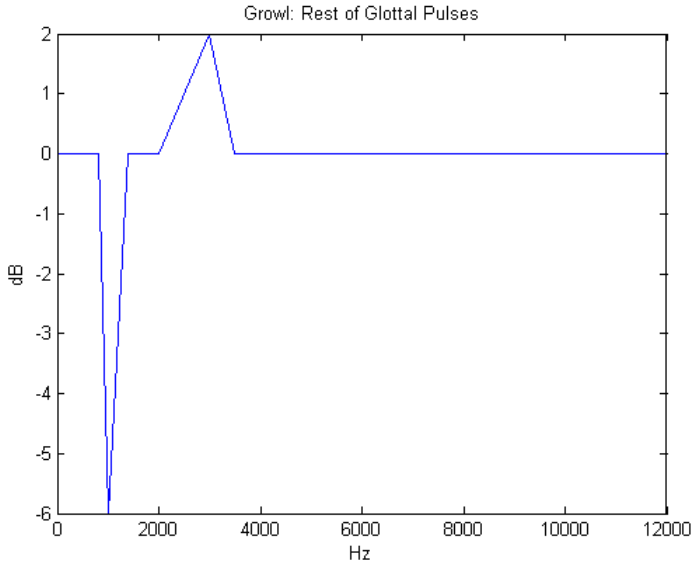

Figure 3.19: Filter for all the Glottal Pulse of a Macropulse for the Growl EVE except for the first and second ones.

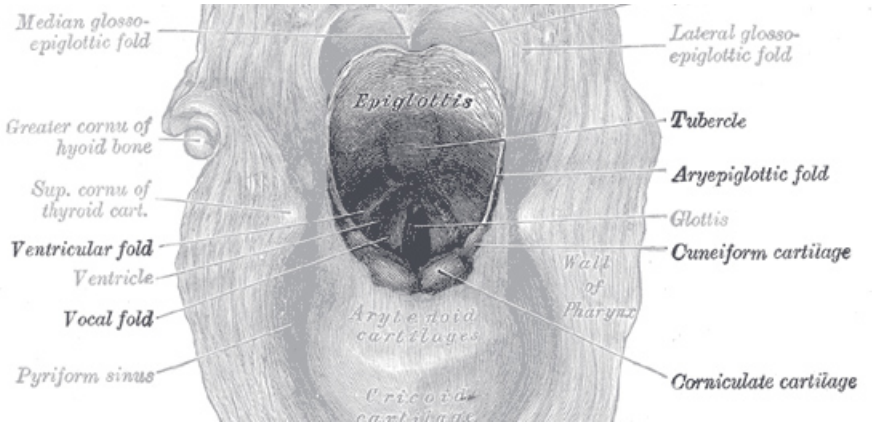

Figure 3.20: Structures of the Larynx that produce the Grunt EVE.

3.4.2 Transformations

For this EVE we can say that the vocal tract gets stretched by a 85%. That is that the **Timbre Envelope stretches 85%**. The position of the Larynx is not as low as in the Growl EVE, but still it has an important modficiation.

Since there are a lot of structures distorting the sound, the pitch is completely lost. There is a lot of noise added to the signal, and that is reflected in **the loss of 80% of the phasiness** (i.e. 80% of noise).

The **number of Glottal Pulses within a Macropulse is a random number between 3 and 6**. The number of structures that are vibrating makes it difficult to extract a model out of it. In this case, this number aparently is a random number. The energy of the first and second sinusoids are also lowered by 2 and 5 dBs respectively.

The Filters also appear to be random in time, and we eventually extracted the filter from a *target* signal and then apply it to the first glottal pulse of the macropulse. This adds a lot of randomness to the signal and makes it much better. This is the only transformation that has a **morphing component**. This filter is found in Figure 3.24. For the rest of glottal pulses, we built a filter that enhances high frequencies and lowers the lower ones. The Filters are found in Figures 3.21 and 3.23.

The whole model is found in Figure 3.22.

Frequencies (Hz)	Rest of Glottal Pulses (dB)
0	-3
250	-10
400	0
600	10
1400	0
5000	3
9000	20
22050	0

Figure 3.21: Filter for the Grunt EVE. Applied to all glottal pulses except for the 1st and the 2nd ones.

Model for the Grunt EVE	
Timbre Envelope	85%
Macropulse	[3–6]
Noise	80%
Pitch Variation	0 dBs
Filters	See Figures 3.21 and 3.23
F0	-2dB
F1	-5dB

Figure 3.22: Model for the Grunt EVE

3.5 Scream

3.5.1 Description

This EVE is also a very aggressive one. It is difficult to say whether it is more or less aggressive than the Grunt, although they are very different compared to each other. Scream is not dark or deep as Grunt is, but it neither has any pitch due to the amount of *noise* by the vocal tract added to the sound.

This kind of sound is mainly produced by Hardcore and Grindcore bands, and it is definitively the newest EVE to be described on this Thesis. It should be necessary to be analyzed with a laryngectomy in order to rigorosuly explain how it is phisiologically produced.

However, from the knowledge acquired by how the other EVEs are produced, we can say that probably this effect is produced by high pitching the vocal folds (i.e. making them vibrate really fast) along with the ventricular folds. The *Arytedoid Cartilage* might contract and it might allow little air to pass through. In Figure 3.25 these structures are marked inside the Larynx.

Some of the most popular singers in this genre that use this EVE are Greg Puciato and Mike Patton. One recommended album is *Miss Machine* by the band *Dillinger Escape Plan* [36]. This album is full of Screams and Grunts, but, for example, the song *Panasonic Youth* starts with a Scream EVE that identifies its sound clearly.

3.5.2 Transformations

There was no time to perform a transformation for this Master's Thesis. In the future work it would be a good start to build a new model for the Scream EVE.

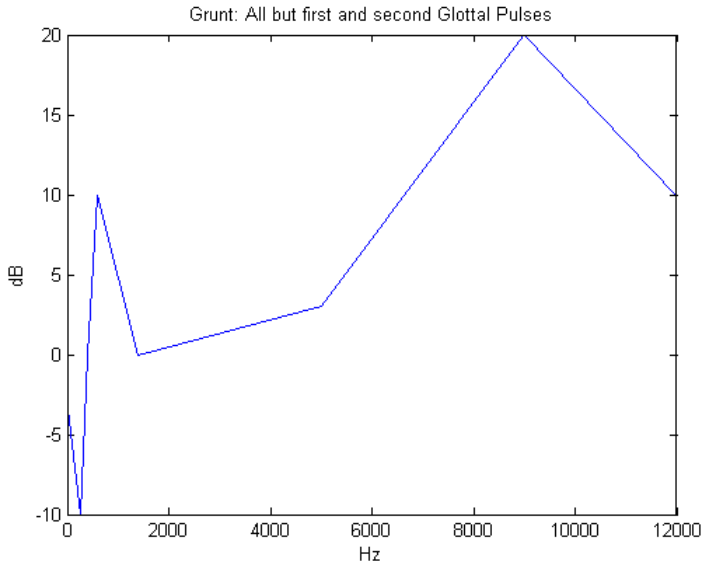

Figure 3.23: Filter for all the Glottal Pulse of a Macropulse for the Grunt EVE except for the first and second ones.

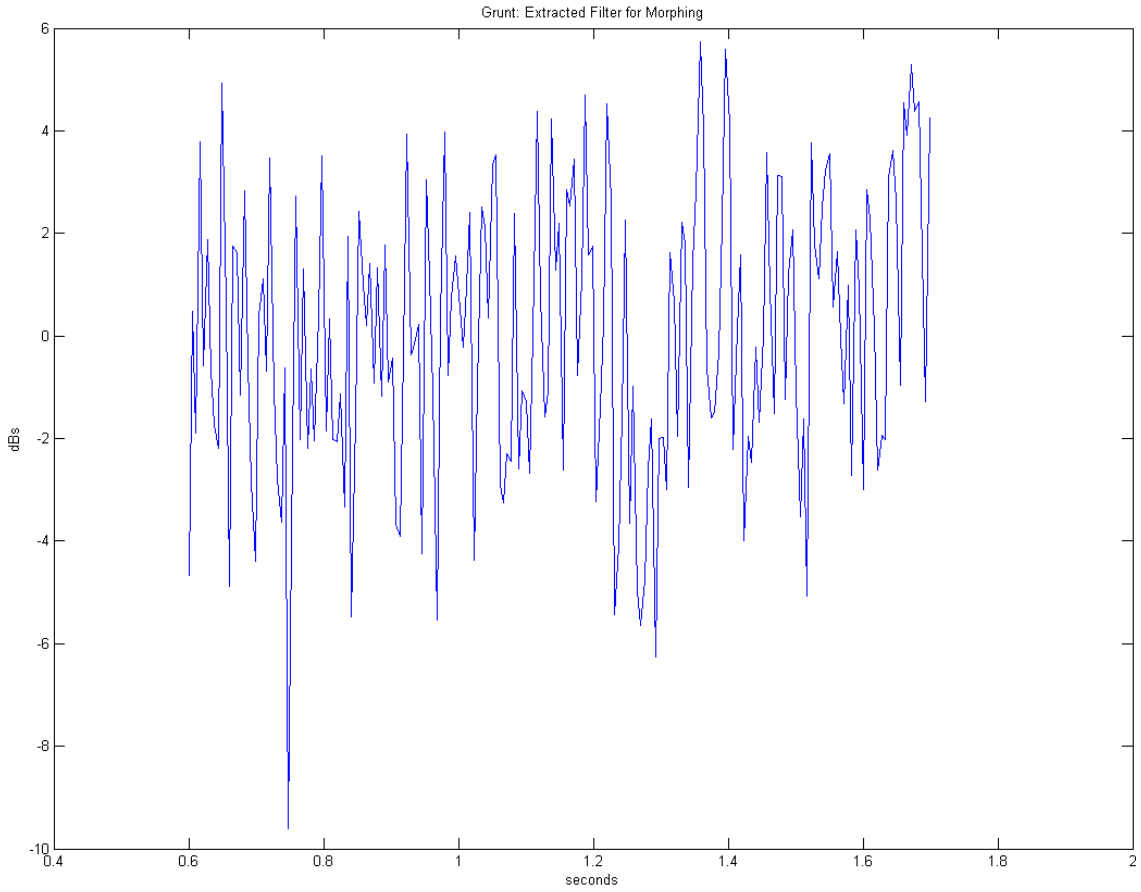

Figure 3.24: Filter in Time Domain for the Grunt EVE. It has been extracted from a real Grunt EVE recording.

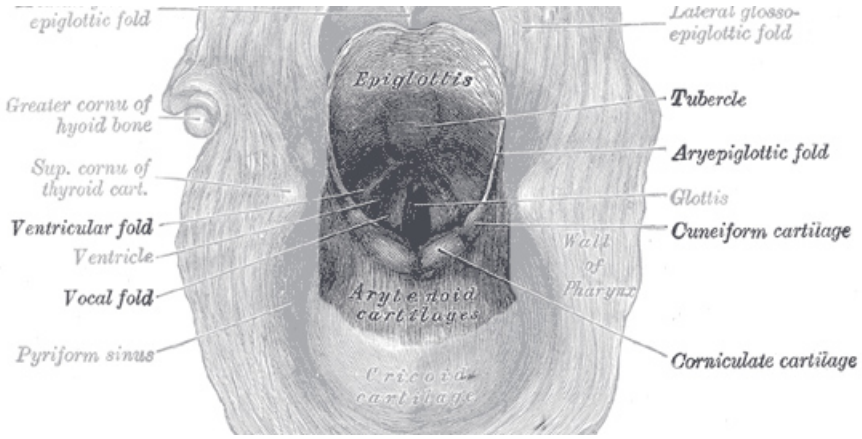

Figure 3.25: Structures of the Larynx that produce the Scream EVE.

Chapter 4

Results

In order to evaluate the results a test has been carried out. It is very difficult to evaluate this Master's Thesis results in an objective way (it is not trivial to write a piece of code that tells you whether the *transformed* signal sounds real or not). That is why we decided that a ***subjective* test** would be the best option in order to evaluate the results.

In this chapter we explain how this subjective test has been carried out and we show and discuss the results in a comprehensive way.

4.1 Subjective Test

The test is a collection of sounds and the evaluator has to decide how realistic these sounds sound to him or her. The evaluator can use 1 to 5 to evaluate each sound, 1 for the most artificial sound, 5 for the most realistic sound. In the test it was also asked the name of the evaluator, his or her familiarity with the rock/metal singing and optional comments.

The sounds that have to be evaluated are not only the transformed ones but also some **real ones**. By doing this, the evaluator does not really know if he or she is evaluating a transformed or a real sound, so that we can also evaluate how *real* a real EVE sounds for the evaluators, and then compare it with our transformed sounds. In total there are **19 sounds to evaluate** (5 for Distortion, 3 for Rattle, 4 for Growl and 5 for Grunt).

This test was uploaded in the Internet[1] and **40 people** took it. The familiarity with the subject of the people who took the test is as follows:

- Not familiar at all: 1 person
- A little bit familiar: 12 people
- Fairly familiar: 7 people
- Quite familiar: 10 people
- Very familiar: 9 people

[1] http://sargonmetal.com/uri/thesis/evaluation.php

4.2 Results of the Test

Here they are presented the results of the test. The results are divided into the four different EVEs' results and the overall results. The results are presented in boxplots. This is the easiest way to visualize the results, since the shortest observation, the lower quartile (Q1), the median (Q2), the upper quartile (Q3), and the largest observation are shown in the boxplots.

4.2.1 Distortion

In Figure 4.1 we can see the results for the Distortion EVE. In this figure there are two boxplots, the one for the Transformed sounds (left) and the one for the Real sounds (right). In both of them, the evaluators always chose values between 1 and 5, and in both of them, the majority of them voted between 2 and 4.

However, in the Real Distortion boxplot, the median is 3.5 instead of the 3 of the Transformed Distortion boxplot because there were more evaluators who evaluated these results with 4.

Although the median is half a point higher in the real sounds, this is a very positive result for the Distortion EVE. The transformed sounds sounded nearly as real as the real sounds for the evaluators.

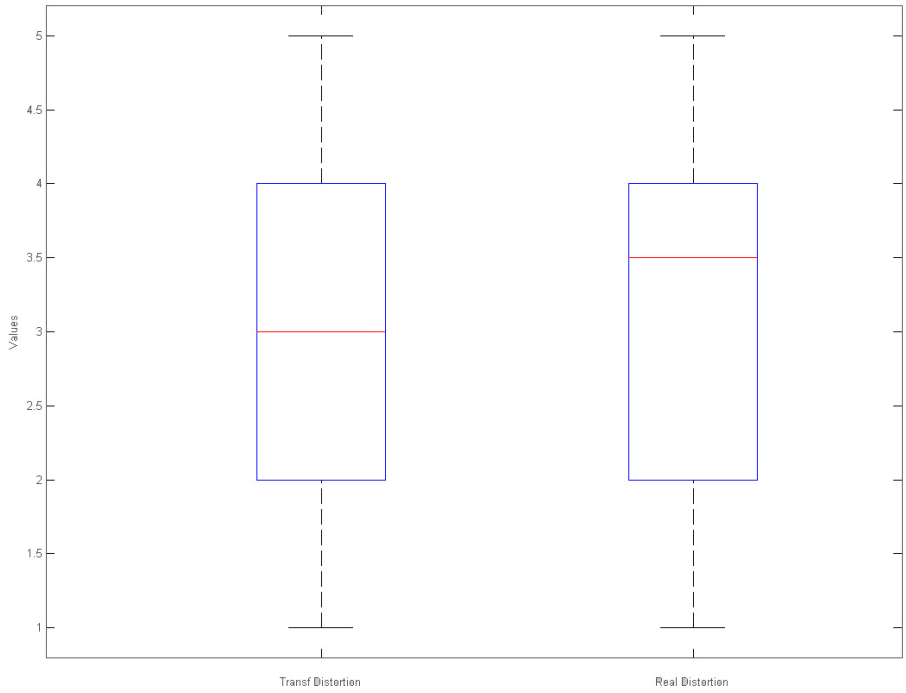

Figure 4.1: Results of the Test: Transformed Distortion (left) and Real Distortion (right)

4.2.2 Rattle

The results for the Rattle EVE are shown in Figure 4.2. We can see the Transformed Rattle results in the left boxplot and the Real Rattle results in the right boxplot. In both cases there were evaluators

voting these EVE from 1 to 5, however, the majority of them voted from 1 to 3, having the majority of people voting it as 1 (the most artificial one).

This might seem a very bad result (the mean in both cases is 1), but in fact it is totally the opposite. The real sounds of this EVE sound, in general, false to the evaluators. This means that this EVE in general sounds false, but it is not false at all. The transformed sounds were evaluated exactly the same as the real sounds.

As pointed out by some comments in the test, maybe these sounds might sound more realistic if they were played with background music.

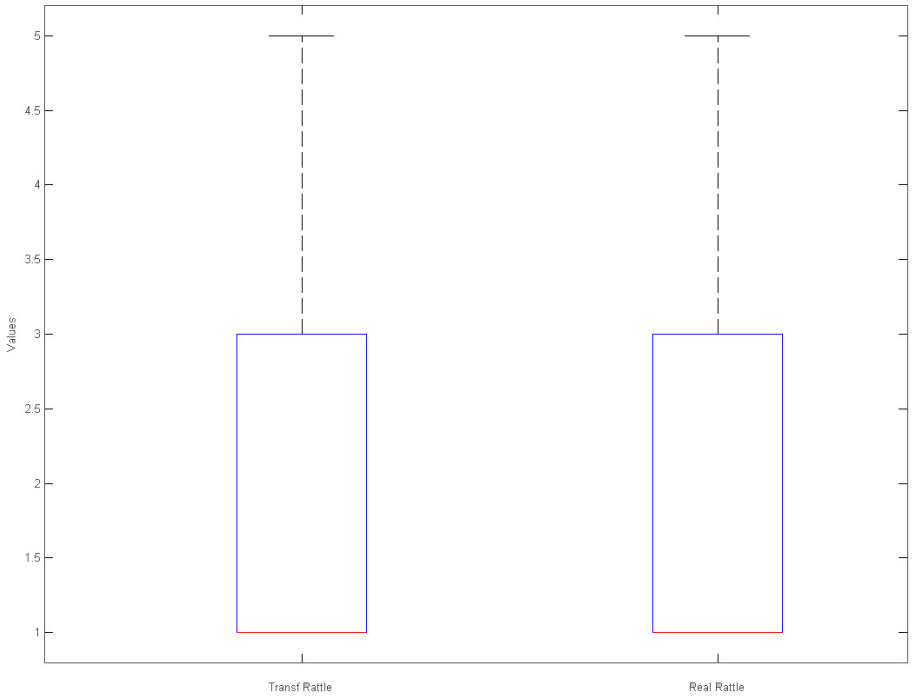

Figure 4.2: Results of the Test: Transformed Rattle (left) and Real Rattle (right)

4.2.3 Growl

The Growl EVE results can be be found in Figure 4.3. Again, there are two boxplots in the figure: the Transformed Growls in the left and the Real Growls in the right.

In both cases the evaluators chose between 1 and 5 to evaluate the results, but in the Transformed Growls the majority of evaluators chose between 2 and 4 whereas in the Real Growls the majority of evaluators chose between 3 and 5.

This is the worst result, since the median of the Transformed sounds is 3 and the median of the Real sounds is 5. There is a big difference between these two sounds, and although the Transformed sounds do not sound very artificial for the evaluators (median of 3), the Real EVE sound very realistic (5).

This Transformation should be improved in further work. This will be discussed in Chapter 5.

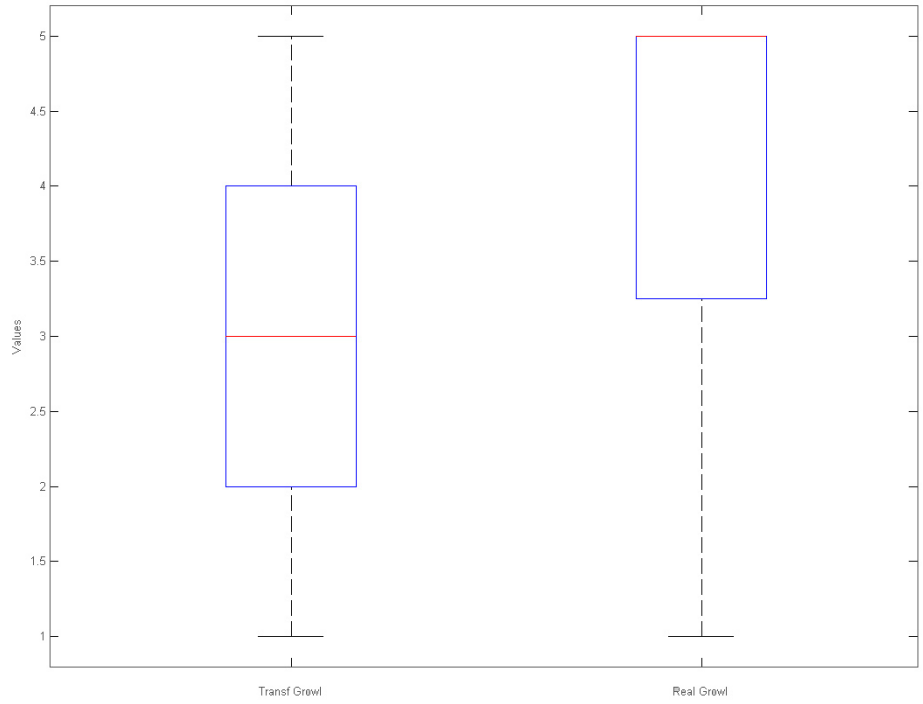

Figure 4.3: Results of the Test: Transformed Growl (left) and Real Growl (right)

4.2.4 Grunt

The results of the Grunt EVE evaluation are shown in Figure 4.4. In the left there is the boxplot of the Transformed Grunts and in the right there is the boxplot of the Real Grunts.

All evaluators chose bewteen 1 and 5 in both the Transformed and Real Grunts. However, in the Transformed Grunts, the majority of people vote from 1 to 3, whereas in the Real Grunts the majority of people voted from 2 to 4.

This is not a very bad result, because the means are only one point away (the Transformed Grunts mean is 2 and the Real Grunts mean is 3). This means that, although the *real* EVE did not sound very real to the majority of evaluators (3), the Transformed EVE sounded a little bit more artificial but not a lot more (2). In any case, the results should be improved in the future.

4.2.5 All EVEs

Finally, in Figure 4.5 we can see the overall evaluation of the Test. The two boxplots correspond to all the Transformed EVEs evaluations (left) and all the Real EVEs evaluations (right).

In both of them, the majority of people have voted between 2 and 4. The big difference is that in the Transformed EVEs the median is 2, whereas in the Real EVEs the median is 3. Still, there is an important number of people who voted 4 in the Transformed EVEs.

This results are satisfying enough, since, although the mean for the Transformed EVEs is one point lower, the majority of people voted from 2 to 4 in both Transformed and Real EVEs. However, this results can still be improved

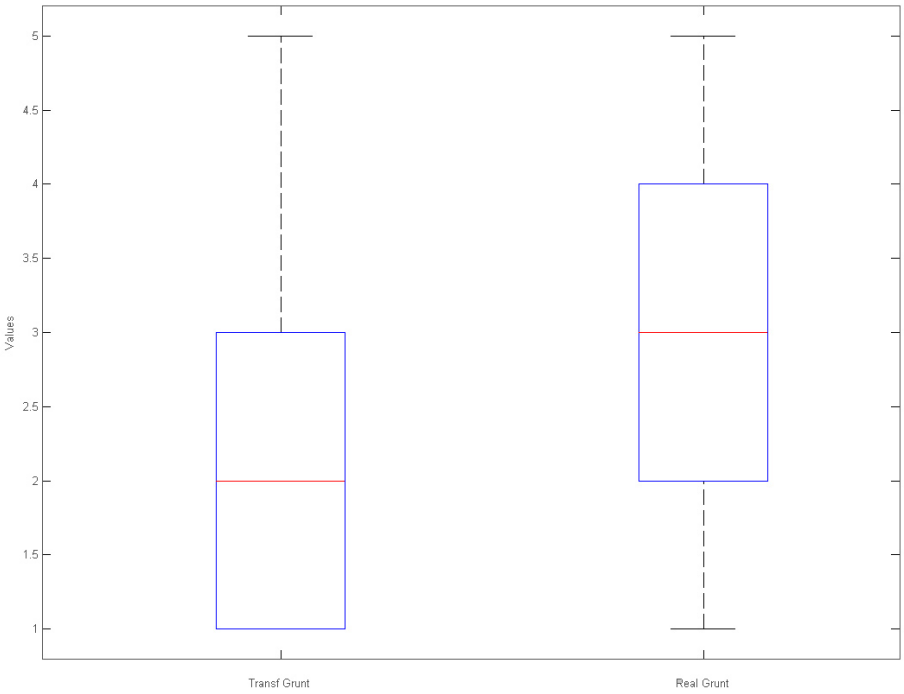

Figure 4.4: Results of the Test: Transformed Grunt (left) and Real Grunt (right)

Another important thing to point out about this test is that, in general, the evaluators do not find very *real* the real EVEs. On the other hand they do not find very real our transformations either, but this is a good point, since they apparently sound to them as nearly as good as the real ones.

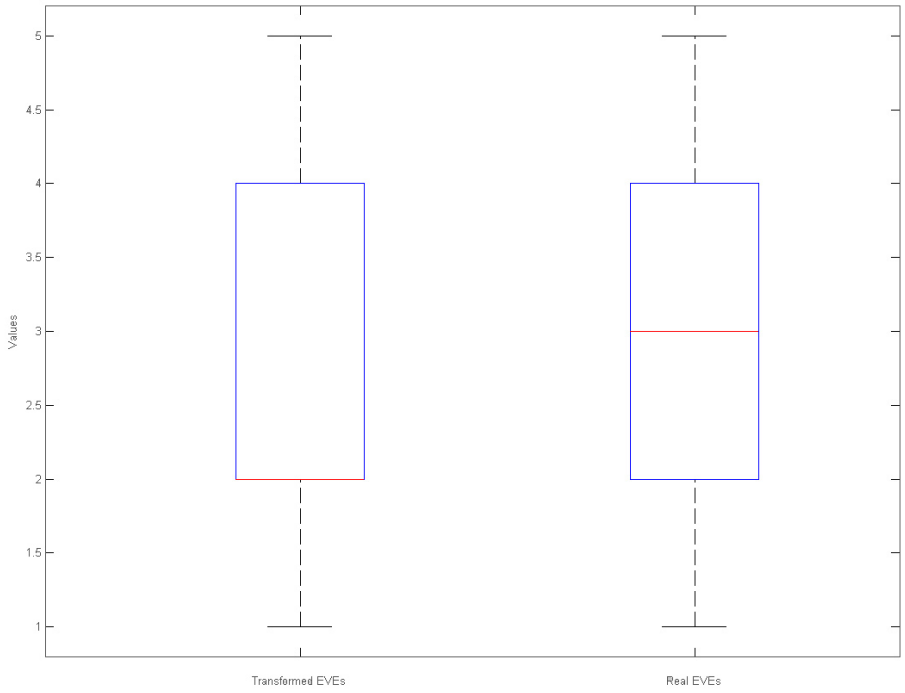

Figure 4.5: Results of the Test: All Transformed EVEs (left) and All Real EVEs (right)

Chapter 5

Conclusions and Further Work

5.1 Conclusions

In this Master's Thesis it has been proposed a classification for the EVEs into 5 different categories:

- Distortion
- Rattle
- Growl
- Grunt
- Scream

It has also been described all of these EVEs and commented and discussed how they are physiologically produced. Some examples of the singers that use these techniques have been also presented.

A Voice Transformation Model for each one of these EVEs —except for the *Scream* EVE— has been presented and discussed. The methodology about how to obtain this model has also been shown, including all the process of recording, analysing, comparing and transforming. These models reflect what physiologically produces the EVEs into the frequency-domain modifications needed in order to obtain the desired EVEs.

The algorithm used for the transformations is the *Wide-Band Harmonic Sinusoid Modeling*[3] (WBHSM), and it has been shown that this algorithm is accurate enough in order to produce realistic transformed EVEs.

Implementations of the Voice Transformation Model using the WBHSM algorithm have been carried out, and applied to several recorded *sources*. The results were evaluated in a survey done by 40 people. The results of the test were positive and encouraging, since the transformed EVEs were nearly as good as the real EVEs from the evaluators' point of view. However, improvements should be done in future work, since there are some transformed EVEs that are more realistic than others. The evaluations also revealed that some of our models are as realistic as real EVEs to some of the people who evaluated them, which is a very good result for this Master's Thesis work.

The study of the EVEs is a recent topic, and few people have carried out scientific work on this field. We could not find any specific work related to audio processing about the EVEs such as *Rattle*,

Grunt or *Scream*, nor find any Voice Transformation Model that maps the physiological aspects of the EVE to its frequency domain modifications in order to transform a normal voice into these EVEs.

This research has not only been a successful enhancement of my knowledge in Signal and Voice Processing as an Engineer, but also as a Musician and Singer. The understanding of these EVEs and how they are physiologically produced made a strong impact on my atempts to produce these EVEs in my own band. As a personal opinion, I would find really useful to have a device to transform my voice to create more ways to express myself, and I believe that this work is the first step in order to create this.

Finally, we have implemented all these EVEs transformations in a framework that works off-line, with streaming that has a high latency (about 200ms[1]). The same implementation has been made in a VST Pluggin, working real-time with this latency. We believe that real-time transformations with lower latency time are possible, and that a regular desktop computer could manage them well. However, our implementation already works for off-line projects and it can be useful to many singers (included myself) in the studio. In a near future some hardware devices may appear for singers to transform their voices live or in the studio off-line.

5.2 Further Work

Some of the transformed EVEs are still not realistic enough, and this is the most important aspect to improve in order to carry on with this working field. In possible future work, it would really help to have much more recordings for each one of the EVEs. This would require to be in contact with more singers that use these techniques, which sometimes are hard to find.

Another idea in order to improve the EVEs is to transform the very end of the EVE in a different way than the rest of the EVE. It has been observed that, in the very end of some EVEs, some formants change place, so that the vowel and the degree of the EVE vary. This would be also a relevant aspect to work on, and it could improve the results dramatically.

Once all the transformed EVEs are realistic enough, it would be a good feature to add a degree of the EVE as a parameter in order to obtain different colours and shapes for a single EVE. One idea is to control this degree of the EVE with some extracted features of the own voice (research on using the voice as a controller has been carried out [37]), so that different EVE with different degrees would be applied deppending on some features extracted from the voice itself. Then, this implementation could be done in a hardware device in order to produce these EVEs live or in the studio, without the need to require a whole computer to transform the signal.

The classification of the EVEs might grow in a near future, since it has only been in the last two decades that some of these EVEs such as *Grunt* or *Scream* have become popular in music. Thus, in a near future this classification should be reviewed and maybe new categories should be added.

The singing voice is fascinating, and we are still discovering how to use it in order to have new ways to express ourselves. Therefore, future work should evolve as fast as these new singers come up with new different EVEs to add more expression and colours to their voices.

[1] SMSTools2 latency

Appendix A

Fourier Analysis

The *Fourier Transform* is one of the basic methods in order to analyze an audio signal. It transforms one function into another, and transforms it into what is called the *frequency domain representation* of the original function. The *Continuous Fourier Transform* equation is as follows:

$$F(v) = \int_{-\infty}^{\infty} f(t)e^{-i2\pi vt}\,dt \tag{A.1}$$

where t is time, f the original function and F the new transformed function, for any real number v.

This transformation is *continuous*, but since in this Thesis we are dealing with digital audio, we will use the *Discrete Fourier Transform* (DFT) instead. This discrete transformation takes a sequence of N complex numbers $x_0, \ldots, x_{N-1}$ and transforms them into a sequence N complex numbers $X_0, \ldots, X_{N-1}$ as follows:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn} \qquad k = 0, \ldots, N-1 \tag{A.2}$$

where $e^{\frac{2\pi i}{N}}$ is a primitive N'th root of unity.

Once a function is transformed, one can inverse the function in order to obtain the *time domain representation* of an already transformed function. This is done by applying the *Inverse Discrete Fourier Transform* (IDFT), which is as follows:

$$x_n = \frac{1}{N}\sum_{n=0}^{N-1} X_k e^{\frac{2\pi i}{N}kn} \qquad n = 0, \ldots, N-1 \tag{A.3}$$

In this Master's Thesis, the IDFT is done after having modified the frequency domain in order to obtain the desired EVEs.

Since we are dealing with signals that change over time, in order to perform an analysis we will have to use the discrete-time *Short Time Fourier Transform* (SFTFT). This related Fourier transform uses a window w in order to analyse a part of the signal. It follows this equation:

$$\mathbf{STFT}\{x[n]\} \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{A.4}$$

where $x[n]$ is the signal we want to transform using the window $w[n]$. $X(m, \omega)$ is essentially the DFT of $x[n]w[n - m]$, where m is the discrete-time index and ω is the continuous frequency axis. However, most typical applications of STFT are performed using the *Fast Fourier Transform* (FFT), an efficient algorithm to compute these transformations.

The discrete time STFT is the one used in Bonada's *Wide-Band Harmonic Sinusoidal Modeling* algorithm[3]. This algorithm is the one we use in this Master Thesis in order to perform the transformations for obtaining the EVEs.

Bibliography

[1] Nederlands Dagblad. "Grunten" sloopt de stem. University Medical Center St Radboud, June 2007. (Dutch).

[2] Julian McGlashan. Can vocal effects such distortion, growling, rattle and grunting be produced without traumatising the vocal folds? Queens Medical Center, Nottingham, 2008.

[3] Jordi Bonada. Wide-band harmonic sinusoidal modeling. *Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08)Espoo, Finland, September 1-4*, 2008.

[4] I. Sakakibara, L. Fuks, H. Imagawa, and N. Tayama. Growl voice in ethnic and pop styles. *International Symposium on Musical Acoustics, Nara, Japan*, 2004.

[5] H. Zemp. *Les Voix du Monde. Une anthology des expressions vocales.* CMX374 1010.12, CNRS/Musée de l'Homme, 1996. 3 vol. CDs with book.

[6] M. Thomasson. *From Air to Aria. Relevance of Respiratory Behaviour to Voice Function in Classical Western Vocal Art.* PhD thesis, Royal Institute of Technology, Department of Speech, Music Hearing, Stockholm, Sweden, 2003.

[7] J. Sundberg. *The Science of the Singing Voice.* Northern Illinois University Press, Dekalb, IL, USA, 1987.

[8] G. Fant, J. Liljenctrants, and Qi guang Lin. A four-parameter model of glottal flow. *STL-QPSR 4/1985, pp 1-13*, 1985.

[9] M. Epstein, B. Gabelman, N. Antoñanzas-Barroso, B. Gerratt, and J. Kreiman. *Source Model Adequacy for Pathological Voice Synthesis.* Voice Lab, Division of Head and Neck Surgery, 1995.

[10] A.L. Lalwani and D.G. Childers. Modeling vocal disorders via formant synthesis. *Acoustics, Speech, and Signal Processing, ICASSP-91.*, 1991.

[11] X. Serra and J. Bonada. Sound transformations based on the sms high level attributes. *Proceedings of COST G6 Conference on Digital Audio Effects, Barcelona, Spain*, 1998.

[12] X. Serra. *Musical Sound Modeling with Sinusoids plus Noise*, pages 91–122. Swets Zeitlinger, 1997.

[13] J. Laroche and M. Dolson. New phase-vocoder techniques for real-time pitch-shifting, chorusing, harmonizing, and other exotic audio effects. *Journal of the Audio Engineer-ing Society*, 47:928–936, November 1999.

[14] C. Hamon, E. Moulines, and F. Charpentier. A diphone synthesis system based on time-domain prosodic modifications of speech. *Acoustics, Speech, and Signal Processing ICASSP, Glasgow, UK*, pages 238–241, May 1989.

[15] E. Moulines, C. Hamon, and F. Charpentier. High-quality prosodic modifications of speech using time-domain over-lap-add synthesis. *Twelfth GRETSI Colloquium. Juan-les-Pins, France*, 1989.

[16] P. Cook. *Identification of control parameters in an articulatory vocal tract model with applications to the synthesis of singing*. PhD thesis, Stanford University, CCRMA, USA, 1990.

[17] P. Cook. Spasm: a real-time vocal tract physical model editor/controller and singer: the companion software synthesis system. *Computer Music Journal, pages 30-44.*, 17, 1993.

[18] Antares Web Site. Avox antares vocal toolkit, Nature. last time checked March 26th 2008.

[19] A. Loscos. *Spectral Processing of the Singing Voice*. PhD thesis, Pompeu Fabra University, 2007.

[20] I. Sakakibara, L. Fuks, H. Imagawa, and N. Tayama. Spectral modeling for higher-level sound transformation. *MOSART Workshop on Current Research Directions in Computer Music, Barcelona*, 2001.

[21] I. Sakakibara, L. Fuks, H. Imagawa, and N. Tayama. Spectral processing. udo zölzer. *International Conference on Digital Audio Effects (DAFx'02)*, 2002.

[22] I. Sakakibara, L. Fuks, H. Imagawa, and N. Tayama. Content-based transformations. *Journal of New Music Research*, 32, 2003.

[23] A. Loscos and J. Bonada. Emulating rough and growl voice in spectral domain. *International Conference on Digital Audio Effects (DAFx'04), Naples, Italy*, October 2004.

[24] TC-Helicon Web Site. Voice pro manual http://www.tc-helicon.com/media/voicepro_manual_101us.pdf, Nature. last time checked March 31st 2008.

[25] H. Sawada, N. Takeuchi, and A. Hisada. Real-time clarification filter of a dysphonic speech and its evaluation by listening experiments. *Conference on Disability, Virtual Reality and Associated Technologies (ICDVRAT2004)*, 2004.

[26] A. Loscos and J. Bonada. Esophagela voice enhancement by modeling radiated pulses in frequency domain. *Audio Engineering Society, San Francisco, CA, USA*, October 2006.

[27] J. Lu K. Shikano T. Doi, S. Nakamura. Improvement in oesophageal speech by replacing excitation components in cepstrum domain. *The Acoustical Society of Japan*, pages 253–254, 1996. Autumn Meeting 2-4-17.

[28] J. Gonzlez T. Cervera, J.L. Miralles. Acoustical analysis of spanish vowels produced by laringectomized subjects. *Journal of Speech, Language, and Hearing Research*, 4:988–996, 2001.

[29] E. Blom M. Singer J. Robbins, H. Fisher. A comparative acoustic study of normal, oesophageal and tracheoesophageal speech production. *Journal of Speech and Hearing Disorders*, 4:202–210, 1984.

[30] Angela Gossow, Michael Amott, Christopher Amott, Sharlee D'Angelo, Daniel Erlandsson. Arch Enemy – Doomsday Machine. Century Media, July 2005. Music Album.

[31] Smstools2, spectral modeling synthesis tools software. http://mtg.upf.edu/technologies/sms/. Website last time checked: July 21, 2008.

[32] Henry Gray. Henry Gray's Anatomy of the Human Body, 1918.

[33] Dave Krusen, Jeff Ament, Eddie Vedder, Mike McCready, Stone Gossard, Rick Parashar. Pearl Jam – Ten. Epic, August 1991. Music Album.

[34] Eric Martin, Paul Gilbert, Billy Sheehan, Pat Torpey. Mr. Big – Bump Ahead. Atlantic Records, July 1993. Music Album.

[35] Mikael Åkerfeldt, Peter Lindgren, Martin Mendez, Martin Lopez, Steven Wilson. Opeth – Blackwater Park. Music for Nations, February 2001. Music Album.

[36] Benjamin Weinman, Brian Benoit, Chris Pennie, Greg Puciato, Liam Wilson. Opeth – Blackwater Park. Music for Nations, February 2001. Music Album.

[37] J. Janer. *Singing-driven Interfaces for Sound Synthesizers.* PhD thesis, Universitat Pompeu Fabra, Barcelona, 2008.