UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE ELECTRÓNICA

Real-Time Voice Quality Modification using WORLD Vocoder

Thesis presented by

Benjamín Opazo Campusano

as a partial requirement for the degree of

Electronics Engineer

and the degree of

Master of Science in Electronics Engineering

Supervisor
Dr. Matías Zañartu

Valparaíso, 2021.

TITLE OF THE THESIS:

Real-Time Voice Quality Modification using WORLD Vocoder

AUTHOR:

Benjamín Antonio Opazo Campusano

THESIS WORK, presented in partial compliance with the requirements for the title of Electronics Engineer and the Master's degree in Sciences of Electronics Engineering from Universidad Técnica Federico Santa María.

Dr. Matías Zañartu _____

Dr. William M. Buttlicker _____

Dr. Michael Scarn _____

Dr. Lloyd Gross _____

Valparaíso, 2021.

All we have to decide is what to do with the time that is given us.

-Gandalf

ACKNOWLEDGMENTS

TABLE OF CONTENTS

ABSTRACT

The role of the self-perception of voice quality has not been explored in the literature and it is believed to play a critical role in the development of hyperfunctional voice disorders. By modifying the auditory feedback, researchers can further investigate the underlying laryngeal motor control mechanisms. Thus, this thesis introduces a Vocoder capable of real-time resynthesis with voice quality changes. The proposed Vocoder is based on the known WORLD synthesizer, and can simultaneously modify fundamental frequency, spectral envelope, and voice aperiodicity. An excitation signal is generated using the Rosenberg++ glottal pulse and a wave shape parameter, along with parameters that allow the fine-tuning of the Rosenberg++ pulse. Frequency and amplitude modulations are also utilized using Brownian noise and sinusoidal signals, as needed. It is also possible to modify the fundamental frequency of the input signal with different parameters such as a multiplier, filtering, and added vibrato. Results illustrate that the resulting voice quality is natural and that it is possible to synthesize Breathy, Rough, Disphonic, Vocal Fry and Modal voice. Current implementation is performed offline, but the Vocoder implementation in an embedded system is underway.

ABSTRACT

El rol de la auto-percepción de la calidad vocal no ha sido explorado en la literatura y se cree que juega un rol critico en el desarrollo de desórdenes vocales hiperfuncionales. Al modificar la retroalimentación auditiva, investigadores pueden explorar los mecanismos de control laríngeo subyacentes. Por lo tanto, esta tesis introduce un Vocoder capaz de resintetizar voz en tiempo real aplicando cambios de calidad vocal. El Vocoder propuesto está basado en el conocido sintetizador WORLD, y puede modificar al mismo tiempo la frecuencia fundamental, la envolvente espectral y la aperiodicidad de la voz. Una señal de excitación es generada usando el pulso glotal de Rosenberg++ y un parámetro de forma de onda, al mismo tiempo que se utilizan parámetros que permiten el ajuste fino del pulso de Rosenberg++. También se aplica modulación de frecuencia y amplitud usando ruido Browniano y señales sinusoidales a medida que se necesiten. Es también posible modificar la frecuencia fundamental de la señal de entrada con distintos parámetros como un multiplicador, filtrado y *vibrato*. Los resultados muestran que la calidad vocal de la voz sintetizada es natural y que es posible sintetizar voz Aspirada, Áspera, Disfónica, *Vocal Fry*, y voz Modal. La implementación actual funciona *offline*, pero la implementación del Vocoder en un sistema embebido está en camino.

Chapter 1

INTRODUCTION

1.1 Problem description

Auditory feedback plays an important role in speech production [1]. For example, the lack of auditory feedback in adults is associated with deterioration in speech production over time [2], a Delayed Auditory Feedback (DAF) of 200 ms may induce increased speech errors and decreased speech rate [3], while in stuttering patients DAF can help with speech fluency [4]. Another well known example of auditory feedback altering speech production is the Lombard Effect[5], in which, speakers have the tendency to raise their voice level when the ambient noise level increases, and when the auditory feedback level decreases[1].

By modifying the auditory feedback, researchers can further understand the underlying mechanisms of speech production, particularly those related to feedback. The role of the self-perception of voice quality has not been explored in the literature and it is believed to play a critical role in the development of hyperfunctional voice disorders [6]. Voice quality can also affect the transmitted emotion of a speaker [7], and as such, the real-time modification of voice quality can help further explore this relationship. Another aspect of real-time voice modification is in the entertainment business, for example, voice quality modification can be used in social media apps to generate different voices of the speaker.

In order to study the effects of auditory feedback on speakers, and in the particular case of this thesis, the feedback on running speech, it is a necessary condition to determine what is the maximum amount of latency that is detectable by a subject, and if that delay disrupts continuous speech or not. Subjects without a DAF have a proportion of disfluences per 300 sillables on the order of 0.005, with 25 ms of DAF this proportion rises to 0.01, and with 200 ms it is of the order of 0.1 disfluences, regardless of gender [8]. Based on the previous results, the criteria for real-time speech synthesis for auditory feedback that will be used in this thesis is a system that has a delay of at most 25 ms.

The objective of this thesis is to disturb the auditory feedback of a subject with real-time modification of voice quality by introducing pathological labels in it. This proposes an important challenge in regards to the definition of voice quality. Voice quality is fundamentally perceptual in nature, both patient and doctors assess the

effectiveness of a treatment based on how the voice *sounds* better [9]. This means that to modify voice quality, it is important to define to some extent the different kinds of voice quality, and in particular, the pathological labels of it. To define and objectively assess these labels, two consensus are usually used, the CAPE-V scale [10], and the GRBAS scale [11]. The assessment still relies on the subjectivity of the evaluator, but it is based on objective descriptions of the pathological labels of voice quality. These issues are discussed in [12], and an important conclusion is that voice quality is not only a function of the voice signals alone, but an interaction between the listeners and the signals.

To modify voice quality in real-time, the proposed method is based on the WORLD voice Vocoder [13]. The World Vocoder is a solution to the problem of high-quality, real-time voice synthesizing. Previous approaches include STRAIGHT [14], which is presented as a real-time Vocoder that performs an order of magnitude worse than the WORLD Vocoder and is difficult to implement in real-time [15], and simplified implementations of the same Vocoder such as TANDEM-STRAIGHT [16]. The main advantage of the WORLD Vocoder over other real-time Vocoders is the high quality of the synthesized voice, which is superior to the previously discussed implementations. The code is also publicly available on Github and the author's website, and is constantly being updated, with implementations on MATLAB, and C, and contributions by other authors, with wrappers on Python, C++, JavaScript, C# and Swift[17]. The version of WORLD used on this thesis is the WORLD V0.2.3 (D4C edition[18]).

1.2 Thesis contribution

The main contributions of this thesis are the following:

1. Propose a framework that allows to modify voice quality in real-time with a high quality Vocoder with sustained vowels and running speech. The modification of voice quality includes, but is not limited to, breathy, rough, vocal fry, disphonic and modal voice.

2. Design and implement a set of algorithms that allows for the modification of voice quality.

The designed system and algorithms will be implemented in a high-level language (MATLAB), with a future work intended to implement the code in a low-level language such as C.

1.3 Outline

Chapter 1. This chapter introduces the problem to solve and what this thesis contributes.

Chapter 2. This chapter explores the background of this thesis. It starts with a quick overview of voice production, then it shows the problem with evaluating voice quality along with some solutions, then different glottal modals are explored, and ends with different methods, systems and algorithms that synthesize voice.

Chapter 3. This chapter explores the solutions to modify voice quality. First, it shows that it is possible to synthesize voice with the WORLD Vocoder in real-time, then it explores the implementation of the Rosenberg++ pulse that allows for part of the voice quality modification. Then a modification to a module of the WORLD Vocoder is presented. Afterwards, two experiments are proposed, Experiment 1 explores the perceptual results of the proposed Vocoder, and Experiment 2 explores the results of objective evaluation of some of the parameters of the proposed Vocoder.

Chapter 4. This chapter shows the future work and conclusions of the thesis.

Chapter 2

BACKGROUND

2.1 Voice Production

This section will explore voice production models that can be used to resynthesize voice so that voice quality can be modified, narrowing down the models that are useful.

Voice production can be modeled as a mechanic process [19] that involves parts of the articulatory system such as the larynx, the mouth, the tongue, and how the airflow interacts with each part. This way of approaching voice production, along with tests and analysis on real subjects, can help precisely model the mechanisms that are present on the production of pathological voices for example. This approach was used to develop the Voice Profile Analysis (VPA) [20], and its biomechanical description of different settings of voice production can be useful for research on voice quality. For the scope of this thesis, the biomechanical description of voice production is not useful so it will not be explored, although, the results of descriptions such as VPA can be used to draw parallels on, for example, Glottal Flow Models.

Another approach of voice production is the Source Filter Theory. This is a mathematical approximation of voice production, and different models can be found in the literature with different degrees of freedom and tunable parameters [13][21][22], which can have an impact on the specific set of algorithms that can be used to synthesize voice along with the quality and perceived naturalness of the synthesized voice. For the scope of this thesis this consequences are fundamental.

2.1.1 Source Filter Theory

Voice can be approximated as the convolution of an excitation signal given by the glottal folds, and the impulse response of the vocal tract. This process can be described as follows

$$y(t) = h(t) * \sum_{n=-\infty}^{\infty} \delta(t - nT_0) \tag{2.1.1}$$

Where $y(t)$ is the voice signal, $h(t)$ is the impulse response of the vocal tract, $\delta(t)$ is the Dirac delta, and $\sum_{n=-\infty}^{\infty} \delta(t - nT_0)$ is the impulse train with fundamental

frequency T_0, representing the glottal source. The symbol * represents the convolution. By applying the Fourier Transform, the previous equation can be written as [13]

$$Y(\omega) = \frac{2\pi}{T_0} H(\omega) \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0) \tag{2.1.2}$$

Note that $\omega_0 = 2\pi/T_0$. $H(\omega)$ is the Fourier Transform of the impulse response of the vocal tract and is known as the spectral envelope, including in this case, the effects of the glottal folds. Estimating the spectral envelope is a difficult challenge by itself [23][24]. An important remark about this model is that in a real-world setting the glottal source is not a perfect impulse train but rather a sine-like waveform, nevertheless, this approximation can be ignored, because it is included in $H(\omega)$

It is useful to represent $H(\omega)$ on the Z-domain with the following equation [22]

$$H(z) = \frac{b_0 \Pi_{k=1}^{M} \left(1 - d_k z^{-1}\right)}{\Pi_{k=1}^{N} \left(1 - c_k z^{-1}\right)} \tag{2.1.3}$$

Where c_k and d_k represent the poles and zeroes of the spectral envelope respectively. For example, in [19], different criteria are applied to the design of zeroes and poles depending on the specific kind of voice that is being synthesized: poles can be used to represent the speech formants, and in unvoiced consonants, nasals and nasalized vowels, the filter is designed with an additional pole-zero pair on top of the usual vowel formants.

The equation (2.1.2) is a compact representation of voice production. In a more general form, the voice signal can be rewritten as [21]

$$Y(\omega) = \Big[H^{f_0}(\omega) G(\omega) + N(\omega) \Big] C(\omega) L(\omega) \tag{2.1.4}$$

Where $H^{f_0}(\omega)$ is the Fourier Transform of an impulse train of fundamental frequency f_0, $G(\omega)$ is the shape of the Fourier Transform of the Glottal Source. It depends on the model of glottal source, or on the estimation of the glottal source with inverse filtering if that is the case. The shape of $G(\omega)$ resembles a lowpass filter, and its slope determines the spectral slope of the voice signal, a normal spectral slope is approximately $12[dB/octave]$ [25]. A simple model of G for voiced speech in the Z-domain can be generated with a third order filter with a complex conjugate pole representing the glottal formant, and one real pole representing the spectral tilt [26]:

$$G(z) = \frac{1}{(1 - az^{-1})(1 - a^* z^{-1})(1 - bz^{-1})} \tag{2.1.5}$$

Where a^* is the complex conjugate of a. In the case of unvoiced speech, G can be modeled as a Gaussian noise. $N(\omega)$ is the aperiodic component of the glottal source signal [27], these aperiodicities appear due to the erratic oscilation of the glottal folds. $C(\omega)$ is the Vocal Tract Filter, representing its resonances and antiresonances

(poles and zeroes), and it includes the effects of the Vocal Tract from the glottis up to the lips. $L(\omega)$ corresponds to the radiation at the lips, approximated as a first order time derivative with the following FIR filter on the Z-domain [28]

$$L(z) = 1 - \alpha z^{-1} \tag{2.1.6}$$

Where $\alpha \lesssim 1$. The equation (2.1.4) can be adapted so that its components can be parametrized by different variables. For example, the glottal model, and its spectral component $G(\omega)$, can be parametrized by the R_d parameter [29][30][21], adding a degree of freedom to the Source Filter Model. If the objective is to resynthesize a given speech signal, then Glottal Inverse Filtering [31] is a useful process that extracts the glottal source $G(\omega)$ and the Vocal Tract Filter $C(\omega)$. The aperiodicity $N(\omega)$ can be estimated, some examples of aperiodicity estimation can be found in the STRAIGHT Vocoder [32], the WORLD Vocoder with D4C [13] or in the SVLN method [21].

It is common to combine the Glottal Source G and the lips radiation L to obtain the glottal flow derivative instead of the glottal flow [26], this is particularly useful because the vocal tract can be modeled as an inductive impedance, where the glottal flow velocity is transformed to pressure by taking the derivative of the signal, which is similar to what the microphone captures.

2.2 Voice Quality

Defining voice quality is a challenging task, mainly because there is no general consensus about what is voice quality, or how to define different aspects of it. According to the US National Library of Medicine, voice quality refers to the different properties of speech that 'gives the primary distinction to a given speaker's voice when pitch and loudness are excluded', this definition, even when broad, is useful as a first approach to voice quality, it defines voice quality as a compound characteristic of voice, rather than the sum of it parts (breathy, hoarse, etc), i.e., a listener is not only perceiving one aspect of voice quality, but it is perceiving the whole signal and evaluating the combination of its components, this definition also excludes pitch and loudness, but to the purposes of this thesis, pitch will be considered when modifying voice quality, because there is some correlation between some pathological voices and its pitch [7]. The role of voice quality in speech production is undeniably important, a teacher may be unable to teach if they present problems in their speech, a person with Parkinson Disease is likely to develop early dysphonia and subtle speech impairment [33], making even more difficult their daily life, or a singer may lose their ability to perform if their voice degrades. Vocal hyperfunction plays a critical role in the development of voice disorders, and compensatory mechanisms can counteract this effect [6].

Voice quality is perceptual in its origin, its a multidimensional property of speech that is not only dependant on the speaker but also on the listener [12], this brings

different challenges, in the first place, any algorithm that modifies voice quality, has to take into account both the multidimensional aspect of voice quality, and the listener's perception, for example, as stated by [7], modifying voice to a tense state involves modifying different parameters of the glottal flow, as well as varying the f_0 parameter and the first formant bandwidth. These modifications are backed by research, but to be meaningful, they have to take into account how it is perceived.

The objective of this thesis is to modify voice quality in real-time, and to achieve that, the multidimensional aspect of voice quality is taken into account. An important issue that arises on the perceptual dimension of voice quality, is that listeners agree on the extremes of the scale (normal phonation, or severe pathology), but not necessarily agree on the milder, more intermediate sections of the scale [9].

There are different approaches for voice quality analysis: articulatory, acoustic and perceptual [34]. Only the perceptual approach will be discussed in this thesis, as that is the approach that makes sense in the scope of this work: the voice feedback is perceptual, and the modifications of voice quality will be used in a purely perceptual setting. The results of this thesis will be objectively evaluated as well by using algorithms that are known to be related to voice quality.

The perceptual evaluation of voice quality is a task that has been proven difficult, given the subjective and multidimensional nature of voice. The methods of evaluation that will be discussed, are going to be useful in two ways for this thesis. First, they provide an objective description of different dimensions of voice quality, and in some cases with a description of the underlying physiopathological processes, this is useful for an objective model of voice quality in order to modify it via resynthesis, and second, these methods can be used to evaluate the resulting synthesized voice, at least in an informal manner, but using state of the art methods.

The modeling of voice quality for this thesis will loosely use the definitions that will be discussed in the following section, taking advantage of the fact that many of them overlap. The objective definitions or VPA will be useful to objectively model voice qualities for synthesis. The multidimensional aspect of voice quality presents a challenge for the modeling of voice quality, but the definitions of compound settings in VPA can help to overcome this problem.

2.2.1 Voice Quality Evaluation

As discussed before, voice quality is perceptual in its origin which means that the listeners perception of the signal is and important factor in the assessment of voice quality. Several consensus have been designed evaluate voice quality in a perceptual manner as objective as possible. These consensuses label different pathological aspects of voice quality, and give a precise definition and method for the evaluator to assess them. Furthermore, these consensuses were created to standardize the labeling of voice quality [35] and to also improve, if possible, the inter and intra listeners variability [36]. Classical definitions of voice quality labels are usually based on physiological terms which can lead to confusion between evaluators when trying

to judge a vague term like bright, or thin [34].

A caveat in these methods, is that in order for the evaluation to be reliable, the evaluator has to be trained in the assessment of voice pathologies, otherwise, inter and intra listeners variability is higher [37].

The following section discusses the perceptual evaluation scales that can be found in the literature

Buffalo Voice Profile (BVP)

Proposed by Wilson in 1972 [38] as a method to assess voice disorders in children. This voice profile was used to establish a baseline of a child under therapy, which can then be used as a reference for the improvement during therapy. The rating scale of the BVP is composed of equal-appearing intervals from 1 to 7, where the lower numbers indicates a slight deviation, and the higher numbers indicate a severe deviation. The author also discusses a disadvantage of using an equal-appearing interval scale, which is the "end effect", where the evaluator is biased to not to use the extreme ends of the scale. This voice profile evaluates, Pitch, Vocal Inflections, Laryngeal Tone, Laryngeal Tension, Vocal Abuse, Resonance, Nasal Emission, Loudness, Rate and Overall Voice Efficiency. To evaluate the voices of the children, the evaluator has to obtain different types of voice samples, connected speech, oral reading, isolated speech sounds and counting.

Vocal Profile Analysis (VPA)

VPA is one of the most used methods to assess voice quality [34]. VPA was defined in [20], first defining basic concepts as a theoretical framework, like articulatory setting and segments.

An articulatory setting, or just setting, is the common sequences of individual segments i.e., the long-term articulatory, phonatory, or muscular tendency of the speaker[39]. The setting gathers the shared, or even the average segments of a speech, and includes different aspects of speech such as positions of the lips, the jaw, vibration of vocal folds or overall muscular tension on the vocal apparatus, or more generally, the supralaryngeal and phonatory characteristics. Some characteristics of a setting can be shared between multiple individual as a shared characteristic of their culture or language. The setting is described with its relationship to a neutral setting, which is a setting with specific characteristics, and not necessarily means a normal speech of an individual. Some of the neutral settings characteristics are [20]

- Lips not protruded
- Jaw neither closed nor too open
- Root of the tongue is neither advanced nor retracted

The reasoning behind this neutral setting is to compare with a vocal tract with known acoustic characteristics, including but not limited the formants frequency or

the bandwidths of the formants for a typical male or female subject. The relationship of these characteristics of the neutral setting, and non neutral settings, where a characteristic is modified, for example, lowering or raising the larynx, can be studied and described. The length of a setting is not exactly defined, it has to be longer than a segment, although no upper bound is defined. Depending on the length of a setting, and considering the context, settings can be used to identify an individual speaker, or even as a way to identify a language.

In speech, segments correspond to the short intervals that are approximately the length of a consonant or vowels. The long term characteristics of a segment's continuum can be considered as an articulatory setting. As stated by Laver, some segments can be unrelated to the underlying settings, and even reverse the value of an articulatory parameter, thats why Laver proposed a distinction between segments that are susceptible and non-susceptible to the underlying setting, and subsequently updates the setting's defintion including only the susceptible segments.

To describe the settings, Laver classifies them as supralaryngeal and phonatory settings, the former being further classified as longitudinal, latitudinal and velopharyngeal settings. For this thesis, it is of interest the phonatory settings, from which the following settings can be found

- **Modal Voice**: Includes the range of fundamental frequencies that are normally used by people when talking or singing. It is characterized by having moderate adductive tension, moderate medial compression and moderate longitudinal tension. The specific description of this and subsequent voices can be seen in [20].

- **Falsetto**: It opposes with respect of the modal voice by having a high adductive tension of the interarytenoid muscles, a large medial compression of the glottis and a high longitudinal passive tension of the vocal ligaments. The fundamental frequency tends to be higher than modal voice, slope of the spectrum is steeper than the modal voice and the harmonics of the voice are separated.

- **Whisper**: Physiologically, the glottis is opened in a triangular manner. It has a low adductive tension, and a moderate to high medial compression.

- **Creak**: (vocal fry, glottal fry) It is characterized by having a low fundamental frequency, below that of the modal voice. The vocal folds are thick and compressed when adducted, in a similar manner, the ventricular folds are also adducted. The glottal waveform is irregular, and the glottal spectrum is less steep than modal voice.

- **Harshness**: Harshness is characterized by an irregular glottal wave-form and spectral noise, it has a high jitter, which is the aperiocidity of the fundamental frequency, it has a big laryngeal tension given by an extreme adductive tension and extreme medial compression.

- **Breathiness**: The vibration of the vocal folds is inefficient, and a slight audible friction can be heard. The vocal folds are relaxed, and as a consequence, they do not close completely, leading to a higher rate of air flow. It has minimal adductive tension and weak medial compression, longitudinal tension is low, but can be controlled, which means that fundamental frequency can be controlled.

Some of the previously described characteristics can be combined, creating a *compound phonation*. This combination of phonatory settings can create a voice where both settings can be identified, and the interaction between them does not modify each setting individually substantially.

The way these phonatory settings can combine is a follows:

- Modal and Falsetto can occur individually but not mixed with each other
- Whisper and Creak can occur individually, combined with each other (whisper and creak), combined with either Modal or Falsetto voices, or even in a triple combination, like whisper-creak-modal voice or whisper-creak-falsetto voice.
- Harshness and Breathiness can only occur in combination with other voices: Harshness can be combined with Modal and Falsetto voice, and Breathiness can only be combined with Modal voice. These voices can also be combined with Whipser and Creak voices if they are also combined with Modal or Falsetto Voices.

An issue with VPA is that, while it models voice production, there is no reference for a listener of the different settings [12], it also assumes that the listener can discern between different features of voice quality. VPA uses between 30 and 40 settings to evaluate, being an exhaustive method of evaluation [34], and as a consequence, evaluation using VPA is more complex, and inter-rater agreement may decrease.

CAPE-V

The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) is a clinical evaluation tool for the assessment of voice quality. It was developed in 2002 by the American Speech-Language-Hearing Association (ASHA) on the Consensus Conference on Auditory-Perceptual Evaluation of Voice [10]. The consensus was designed following considerations such as having a minimal set of clinically meaningful perceptual voice parameters, the consensus can be applied to a broad range of vocal pathologies and it has a demonstrable low inter-rater reliability.

To assess the voice of a patient, three tasks have to be performed, the patient has to sustain the vowels /a/ and /i/, read six sentences, and converse naturally. The scale used to assess the voice quality of a patient consists of a visual analog scale, that is, a 100 mm scale with unlabeled anchors, representing normal voice on

the left side, and the right side of the scale is the most extreme example of deviance for the listener.

The vocal attributes to be evaluated are the following [40]

- **Overall Severity**: Global impression of voice deviance.
- **Roughness**: Irregularity in the voicing source.
- **Breathiness**: Audible air escape in the voice.
- **Strain**: Perception of vocal hyperfunction or vocal effort.
- **Pitch**: Perceptual correlate of fundamental frequency. It rates the deviation from a normal person at that age, gender and referent culture.
- **Loudness**: Perceptual correlate of sound intensity. It also rates the deviation from a normal person given the gender, age and referent culture.

There are also blank scales that the evaluator may use to rate additional attributes.

GRBAS

The GRBAS scale (Grade, Rough, Breathy, Asthenic, Strained) is a four-point scale used to asses voice quality, where the scale is divided in regular intervals, ranging from no abnormality (0) and severe abnormality (3). GRBAS was proposed by the Japan Society of Logopedics and Phoniatrics and it is widely used to evaluate voice disorders [41]. The five scales are defined as following [42]:

- **Grade**: Degree of hoarseness or voice abnormality.
- **Rough**: Psychoacoustic impression of the irregularity of vocal folds vibrations. It describes the irregularities in fundamental frequency and amplitude of the glottal source sound.
- **Breathy**: Psychoacoustic impression of the air leakage trhough the glottis. It describes turbulence.
- **Asthenic**: It is the weakness or lack of power in the voice. It represents a weak intensity of the glottal source sound, or lack of higher harmonics.
- **Strained**: Psychoacoustic impression of a hyperfunctional state of phonation. It represents abnormally high fundamental frequency, noise in the high frequency range and rhichness in high frequency harmnoics.

Like other scales, it needs a trained professional in order to be reliable [43]

2.3 Glottal Models

An important aspect of the voice quality modification that will be applied in this thesis relies on modifying the source signal of a Vocoder. The source signal of a Vocoder will be modeled as the glottal source, and for that, a glottal model is required. For the scope of this thesis, the glottal model has to be:

- Parametrized, allowing a wide range of glottal pulses
- Simple, so that pulses can be synthesized in real-time
- Precise, in order to synthesize a realistic voice

The first condition allows the Vocoder to synthesize a wide range of voice qualities, for example, when comparing the Rosenberg Model with the Rosenberg++ Model, the former model restricts the parameter t_p, and lacks a factor added in the Rosenberg++ Model. The second condition is necessary for the real-time aspect of this thesis, as the glottal pulse has to be calculated several times per second, and the third condition is related to the perceptual nature of voice and in particular of voice quality: it is a desired feature for the synthesis of voice but not completely needed.

The glottal model that will be used is the Rosenberg++ (R++) Model as it fulfills the conditions imposed by this work. The parameters are not as relaxed as in the LF Model, but with a set of algorithms that can be implemented in real-time, it is possible to synthesize voice as natural as the LF Model [44].

2.3.1 LF Model

The Liljencrants-Fant Model [45] [29] is a four parameter model of the glottal flow. The parameters of the model are t_p, t_e, t_a and E_e and they uniquely determine the shape of the pulse. Another parameter, T_0, determines the length of the pulse but not its shape. The shape parameter are defined as follows [46]

- t_p: Instant of the Glottal Flow maximum
- t_e: Instant of maximum excitation
- t_a: Return phase constant
- E_e: Amplitude of the derivative of the Glottal Flow

Note that E_e corresponds to the amplitude (or the maximum value) of the *derivative* of the glottal flow, whereas U_0 will be defined as the maximum amplitude of the glottal flow. The Glottal Flow is defined by the following equation

$$E(t) = \begin{cases} E_0 e^{\alpha t} \sin(\pi t/t_p) & 0 < t < t_e \\ \frac{-E_0}{\varepsilon t_a} \left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)} \right] & t_e < t < t_c \end{cases} \tag{2.3.1}$$

The parameter ε is defined by the following implicit equation

$$\varepsilon t_a \begin{cases} 1 & \text{for a small } t_a \\ 1 - e^{-\varepsilon(t_c - t_e)} & \text{otherwise} \end{cases} \tag{2.3.2}$$

The Glottal Flow is defined as having zero net gain of flow during a fundamental period, that means, that the following integral is zero:

$$\int_0^{T_0} E(t) = 0 \tag{2.3.3}$$

On the other hand, the residual flow is defined as

$$\int_{t_e}^{t_c} E(t) = U_e = \frac{E_e t_a}{2} K_a \tag{2.3.4}$$

Where K_a is calculated as follows

$$R_a = \frac{t_a}{t_c - t_e} \tag{2.3.5}$$

$$Ka = \begin{cases} 2 - 2.34R_a^2 + 1.34R_a^4 & R_a < 0.5 \\ 2.16 - 1.32R_a + 0.64(R_a - 0.5)^2 & R_a > 0.5 \\ 2 & R_a < 0.1 \end{cases} \tag{2.3.6}$$

The parameter α is calculated by solving α or $R_d = 2\alpha t_p/\pi$ of equations (2.3.3) and (2.3.4). This leads to the following implicit equation of α [46]

$$\frac{1}{\alpha^2 + \left(\frac{\pi}{t_p}\right)^2} + \left(e^{-\alpha t_e} \frac{\pi/t_p}{\sin(\pi t_e/t_p)} + \alpha - \frac{\pi}{t_p} \text{cotg}(\pi t_e/t_p) \right) = \frac{t_c - t_e}{e^{\varepsilon(t_c - t_e)} - 1} - \frac{1}{\varepsilon} \tag{2.3.7}$$

The previous equation can be solved, but it is computationally expensive to solve, creating a problem for real-time voice quality modification. On the other hand, it has been reported that the LF-model is able to synthesize voice with high quality[47].

The R_d parameter, defined as [29]

$$R_d = \frac{2\alpha t_p}{\pi} \tag{2.3.8}$$

has been used to modify voice quality of synthesized voices [21] [30] [48]. The R_d parameter, is used as a shape parameter that is related to the quantification of the covariance of the parameters of the LF-Model. Low values of the R_d parameter represents relaxed glottal pulses, generating breathy voices, whereas higher values are related to a tighter phonation, with low Open Quotient and high F_a, with

$F_a = 1/(2\pi t_a)$, which is related to the spectral tilt. The R_d parameter can also be defined as follows

$$R_d = (U_o/E_e)(F_0/110) \tag{2.3.9}$$

It is also useful to define the following parameters

$$R_a = t_a/t_0 \tag{2.3.10}$$
$$R_g = t_0/(2t_p) \tag{2.3.11}$$
$$R_k = (t_e - t_p)/t_p \tag{2.3.12}$$

Where R_a is used as an alternative to F_a, R_k is related to the duration of the falling branch, and R_g is inversely related to the rise time. The values of R_a and R_k can be predicted given a value of the R_d parameter with the following equations

$$R_a^* = (-1 + 4.8R_d)/100 \tag{2.3.13}$$
$$R_k^* = (22.4 + 11.8R_d)/100 \tag{2.3.14}$$

Conversely, the R_d parameter can be estimated given the measured values of R_a, R_k and R_g as follows

$$R_d = (1/0.11)(0.5 + 1.2R_k)(R_k/4R_g + R_a) \tag{2.3.15}$$

For the scope of this thesis, the R_d parameter will be used as an input parameter, and with it, calculate R_a, R_k, and R_g, afterwards, t_a, t_0 and t_e can be calculated.

2.3.2 Rosenberg Model

In [49], Rosenberg proposed six different models of glottal pulses and tested them to determine which model works better. The two best models are model *b* and model *c*. These are the proposed models:

$$f_b(t) = \begin{cases} a\left(3\left(\frac{t}{t_p}^2\right) - 2\left(\frac{t}{t_p}\right)^3\right) & 0 \leqslant t \leqslant t_p \\ a\left(1 - \left(\frac{t-t_p}{t_n}\right)^2\right) & t_p \leqslant t \leqslant t_p + t_n \end{cases} \tag{2.3.16}$$

$$f_c(t) = \begin{cases} \frac{a}{2}\left(1 - \cos\left(\frac{t}{t_p}\pi\right)\right) & 0 \leqslant t \leqslant t_p \\ a\cos\left(\left(\frac{t-t_p}{t_n}\right)\frac{pi}{2}\right) & t_p \leqslant t \leqslant t_p + t_n \end{cases} \tag{2.3.17}$$

In the literature, a variant of $f_b(t)$ where $t_p = 2t_n$ or equivalently $t_p = 2t_e/3$ with $t_e = t_p + t_n$ is cited as the rosenberg model [44][50]:

$$g(t) = \begin{cases} t^2(t_e - t) & 0 \leqslant t \leqslant t_p + t_n \\ 0 & t_p + t_n \leqslant t \end{cases} \tag{2.3.18}$$

The previous expression generates a glottal shape similar of that of $f_b(t)$, but not precisely the same. The maximum of the previous equation is at $t = \frac{2}{3}t_e$ and it is equivalent to $g(t = \frac{2}{3}t_e) = \frac{4}{27}t_e^3$. With the purpose of maintaining the equivalency between different models, it is convenient to scale equation 2.3.18 by the reciprocal value of the maximum, and subsequently multiply by an a parameter. The glottal pulse proposed in [49] does not necessarily restricts $t_p = 2t_n$, which gives more range to the shapes that the pulse can attain.

The parameters of the Rosenberg Model are defined as follows:

- a: Maximum amplitude of the pulse
- t_p: Opening time, the slope is positive
- t_n: Closing time, the slope is negative

A fourth parameter can be found in the description of the model, t_e which corresponds to the length of the full pulse, that is $t_p + t_n$. A shortcoming of this model is the absence of a return phase, which is related to the rate at which the pulse returns to zero, and it can affect the perception of the pulse.

2.3.3 Rosenberg++ Model

The Rosenberg++ Model (R++ Model) is an extension of the Rosenberg model, proposed by [44]. Each plus sign represents an improvement over the original Rosenberg Model. It is based on the Rosenberg-B Model discussed in equation (2.3.18). The first problem that the R++ model fixes is adding an exponential return phase, and the second addition is an extra factor that allows the specification of t_p. An important remark is that on [49], the Rosenberg-B model does have a t_p parameter, but subsequent literature assumes that this parameter is fixed with $t_p = 2t_n = 2t_e/3$.

The R++ Model is defined by the following equation.

$$f(t) = \begin{cases} Kt^2(t^2 - \frac{4}{3}t(t_p + t_x) + 2t_p t_x) & 0 \leqslant t \leqslant t_e \\ f(t_e) + t_a f'(t_e)\frac{1-\exp(-(t-t_e)/t_a)-((t-t_e)/t_a)\exp(-(t_0-t_e)/t_a)}{1-\exp(-(t_0-t_e)/t_a)} & t_e \leqslant t \leqslant t_0 \end{cases} \tag{2.3.19}$$

The net gain of flow during a fundamental period is zero, resulting in the following expression

$$\int_0^{\tau} f'(\tau)d\tau + t_a f(t_e) D(t_0, t_e, t_a)$$
$$= Kt^2 \left(t^2 - \frac{4}{3}t(t_p - t_x) + 2t_p t_x \right) + t_a f(t_e) D(t_0, t_e, t_a) = 0 \tag{2.3.20}$$

Solving for t_x, the following expression is obtained

$$t_x = t_e \left(1 - \frac{\frac{1}{2}t_e^2 - t_e t_p}{2t_e^2 - 3t_e t_p + 6t_a(t_e - t_p) D(t_0, t_e, t_a)} \right) \tag{2.3.21}$$

$$D(t_0, t_e, t_a) = 1 - \frac{(t_0 - t_e)/t_a}{\exp((t_0 - t_e)/t_a) - 1} \tag{2.3.22}$$

The factor $D(t_0, t_e, t_a)$ appears as a consequence of the exponential return phase, as defined in the LF-model. With the previous equations, the pulse can be defined with its derivative

$$f'(t) = \begin{cases} 4Kt(t_p - t)(Tx - t) & 0 \leqslant t \leqslant t_e \\ f'(t_e) \frac{\exp(-(t-t_e)/t_a) - \exp(-(t_0-t_e)/t_a)}{1 - \exp(-(t_0-t_e)/t_a)} \end{cases} \tag{2.3.23}$$

To ensure that the pulse is non-negative $t_x \geqslant t_e$ and $tx \leqslant 0$, and these limitations can be enforced with the following restrictions

$$\frac{1}{2}t_e \leqslant t_p \leqslant \frac{3}{4}t_e \left(\frac{t_e + 4t_a D(t_0, t_e, t_a)}{t_e + 3t_a D(t_0, t_e, t_a)} \right) \tag{2.3.24}$$

The factor K does not necessarily means that the amplitude of the glottal pulse is K, to solve this, a factor of $-\frac{1}{3}t_p^3(t_p - 2t_x)$ is added to the pulse, so that K is, in fact, the amplitude of the glottal pulse.

The parameters of the model are as follows

- K: Maximum amplitude of the pulse
- t_p: Opening time, the slope is positive
- t_e: Minimum of the glottal flow derivative waveform (excitation instant). Also known as Glottal Closure Instant (GCI)
- t_a: Time constant for the return phase

This model has the advantage that it does not require solving non linear equations to obtain the pulse shape, which can be beneficial to the objective of synthesizing voice in real-time, it is also similar to the LF-model, the parameters have a rather large range of operation.

2.3.4 KLGLOTT88 Model

The KLGLOTT88 Model [51] is a model based on the Rosenberg B Model discussed previously. It is a four parameter model, where the extra parameter with respect to the Rosenberg model controls the spectral tilt of the pulse, adding a return phase with a filter. The KLGLOTT88 Model is defined as [52]

$$U_k(t) = at^2 - bt^3 \tag{2.3.25}$$

$$a = \frac{27AV}{4t_0O_q^2} \tag{2.3.26}$$

$$b = \frac{27AV}{4t_0^2O_q^3} \tag{2.3.27}$$

Where AV is the Amplitude of Voicing and O_q is the open quotient. To add a return phase, the glottal pulse is filtered with a first order low-pass filter with cutoff frequency F_t. This method has the shortcoming that the pulse is modified during the filtering, obtaining a pulse that is not exactly the one defined with A_v and O_q. It is important to note that the maximum amplitude of the filter is given by A_vT_o, where T_0 is the fundamental period of the pulse.

2.4 Voice Synthesis

This section will discuss different tools for analysis-synthesis of voice. For the objective of this thesis, the tools have to:

- Analyze and synthesize in real-time
- Synthesize high quality signals, that is, signals that are sufficiently similar to the original, so that a listener wont be able to notice
- Be parameterizable, meaning that with the use of parameters the tool can change the vocal quality and other characteristics of voice

Many systems can achieve the first condition, to synthesize in real-time, especially with modern computers, although in many occasions minimizing the latency is related to a loss of quality due to the filters using less samples or reasons alike. Usually, to solve this problem, the analysis-synthesis tool has to be designed with real-time in mind. Another cause for delay is the Operative System the synthesizer is working on. Usually it is more convenient to implement the synthesizers on an embedded device that can assure a given latency rather than to implement a synthesizer over an Operative System like Linux or Windows. Without limiting the foregoing, the system can be implemented and measured offline, so that the analysis can emulate a real world scenario.

The second condition imposed by this work, which is to synthesize high quality signals, is usually the objective that researchers and developers have prioritized, however, synthesizing high quality real-time signals is usually more problematic and the results are not necessarily good.

The third condition, a paremeterizable Vocoder, is useful to modify voice quality with ease, specially for a user of the Vocoder. If these conditions are met, along with a sufficiently good model of voice quality, the objective is then to match the model of voice quality to the parameters of the synthesizer and its effects on voice.

The following subsections explore different voice synthesizers found in the literature. The synthesizer that will be used in this thesis is the WORLD Vocoder [13] because it fulfils the previously stated conditions. The WORLD Vocoder is a high-quality Vocoder designed with real-time capabilities. An important feature of the WORLD Vocoder that will be useful in this thesis is that the latency between analyzing and synthesizing is also parameterizable, although quality degrades with lower latencies.

The STRAIGHT Vocoder fulfills the previous conditions, but it performs worse than the WORLD Vocoder in terms of latency by an order of magnitude and performs slightly worse in subjective evaluations of the quality of the synthesized voice.

2.4.1 Linear Predictive Coding

Linear Predictive Coding (LPC) is a widely used method (e.g.: [53] [54]) for representing the spectral envelope of speech. It is based on the notion that over short time intervals, the linear system that represents the speech signal can be described by an all-pole system function of the form [22]

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.4.1}$$

Where $U(z)$ is the z-transform of the excitation signal, and $S(z)$ is the speech signal. Following the previous equation, the speech signal $s(n)$ is given by the following equation

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Gu[n] \tag{2.4.2}$$

The linear predictor, is then

$$\hat{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k] \tag{2.4.3}$$

And the prediction error is given by the difference of the predicted signal and the real signal

$$e[n] = s[n] - \hat{s}[n] = s[n] - \sum_{k=1}^{p} \alpha_k s[n-k] \tag{2.4.4}$$

Reordering the previous equation in the z domain, it can be expressed as

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} = \frac{E(z)}{S(z)} \tag{2.4.5}$$

Namely, the transfer function of a FIR linear system. If $a_k = \alpha_k$ for all $k = 1, ..., p$, then the error is $e[n] = Gu[n]$, and $H(z) = G/A(z)$. So the problem is to predict $\alpha = (\alpha_1, ..., \alpha_p)$ from $s = (s[n-1], ..., s[n-p])$. An optimal estimation of $\hat{s}[n]$ is by minimizing the minimum mean square error, i.e. minimizing $E[e^2[n]]$ over a short segment, where $E[\cdot]$ is the expected value. The specifics of this method can be found in the literature, and the resulting equation to estimate α_k is given by

$$R_s \alpha = r_s \tag{2.4.6}$$

Where R_s is the autocorrelation matrix of s and r_s is constructed with the following expression

$$r_s = \begin{bmatrix} E[s[n]s[n-1]] \\ E[s[n]s[n-2]] \\ \vdots \\ E[s[n]s[n-p]] \end{bmatrix} \tag{2.4.7}$$

Note that this last expression is different from any column in the autocorrelation matrix. Equation (2.4.6) can be efficiently solved using the Levison-Durbin Method.

Even though this Method is computationally efficient, and it is widely used in real-time applications, the synthesized signals are not necessarily of high quality, depending on the noise conditions and the length of the windows of analysis. The analyzed signal is not necessarily modifiable with high reliability, which is a problem for the objective of this thesis.

2.4.2 Audapter

Audapter [55] [56] is a real-time Vocoder that allow for the real-time manipulation of speech. It allows the perturbation of the following parameters:

1. Formant Frequencies F1 and F2
2. Fundamental Frequency
3. Local Timing
4. Local intensity

5. Global time delay (DAF)

6. Global intensity

For the scope of this thesis, these modifications are not enough to modify voice quality with the desired results.

2.4.3 VocalTractLab

VocalTractLab (Vocal Tract Laboratory) [57] is a software designed to synthesize voice using articulatory synthesis, which is the synthesis of voice by exciting a physical model of the vocal tract. The Vocoder has the potential of generating natural sounding voice, along with more flexibility of the parameter compared to the other Vocoders discussed in this section. An important difficulty on simulating voice by its physical model is that it requires vast amounts of computing power to be able to simulate in real-time the voice. Another aspect is the modeling of the parameters of the speaker so that the Vocoder sounds like its target.

2.4.4 Machine Learning

Voice synthesis using machine learning has been applied in different solutions and in general the quality of the synthesized voice scores better than many other Vocoders [58] in scales like the Mean Opinion Score (MOS) test. Machine learning based Vocoders can synthesize voice in real-time or faster [59], and encode high quality, low bitrate, lossy audio in real-time [60].

A difficulty associated with machine learning based Vocoders, and machine learning in general is the requirement of high volumes of high quality datasets for the training of the model, for example, Google develops state of the art Vocoders, but does not distributes the trained models. The Mozilla Foundation has released an open source voice dataset [61] that can be used for the training of machine learning based Vocoders. A shortcoming of this dataset is that is not as big as the datasets that Google has access to, which means that the trained models will not be as good as Google's trained models. Another important shortcoming is the language specific nature of these trained models, which means that models have to be trained for different language, and even different accents of the same language.

The biggest shortcoming of the Machine learning based Vocoders for this thesis is the black box nature of Machine Learning models, which means that modifying the voice quality using these Vocoders is not necessarily possible. Machine Learning are not deterministic but stochastic by nature, which means that the result of the synthesis is not assured to be known before hand, which is not optimal for the use cases of the result of this thesis.

Machine Learning could be useful to modify specific parameters of a Vocoder, for example, to recover or synthesize lost information due to problems in the input signal, which could help to repair (for example) the spectral envelope of a voice prior to synthesizing it.

2.4.5 STRAIGHT

The STRAIGHT Vocoder [62], is a Vocoder that allows for a flexible speech modification framework, by performing real-time analysis-synthesis of speech, synthesizing high quality audio. The STRAIGHT Vocoder implements a method for F0 estimation based on a new parameter called fundamentalness, it also implements a method to reduce the buzziness of the Vocoder, a problem that was prevalent in previous Vocoders. The STRAIGHT Vocoder eliminates periodicity interference, which is a periodic variation in the time domain of the power spectrum.

An advantage of the STRAIGHT Vocoder over other Vocoders is that the extracted spectrogram has no trace of the source periodicity, which means that the synthesized voice can be modified with flexibility.

It is important to highlight that the STRAIGHT Vocoder allows for the real-time analysis-synthesis of voice, and the resulting signal is of high quality, and it also opens the doors for modification of voice. This Vocoder fulfills all the imposed conditions by this thesis. The reason that it was not chosen is due to the WORLD Vocoder, which is based on similar ideas of the STRAIGHT Vocoder but performs better in synthesizing high quality speech, and it requires less processing than the STRAIGHT Vocoder.

2.4.6 WORLD

The WORLD [13] Vocoder, and in particular the WORLD (D4C edition[18]) Vocoder is a real-time, high quality analysis-synthesis Vocoder, that allows real-time modification of voice, with a wide range of possible modifications of voice due to the flexibility of the Vocoder's parameters. Its main advantage over other real-time Vocoders is that it takes an order of magnitude less processing to achieve superior results. An important caveat of the WORLD Vocoder is that it requires high Signal to Noise Ratio (SNR) to work properly, which can limit the applications in the real world, but in the particular use case of this thesis this is not a problem, because experiments will be performed in controlled setups with low ambient noise.

There are different implementations of the WORLD Vocoder that differ in some aspects with respect to the original specification of the Vocoder in [13]. These implementations represent upgrades in different parts of the Vocoder and the overall result is a higher quality analysis-synthesis. Some modifications of the original specification are well documented in follow up papers, but some modifications are not discussed, although, this is an expected outcome of a software that is under development. Some of the new functions introduced are slower than the originals, but perform better on their respective objectives. This trade off has to be taken into account when designing the real-time voice quality modification. In a followup chapter this topic will be further discussed.

The WORLD Vocoder consists of four main steps

- Fundamental Frequency (F_0) estimation: DIO or Harvest

- Spectral Envelope estimation: CheapTrick
- Aperiodic Parameter estimation: PLATINUM or D4C (lovetrain)
- Synthesis

Each step is done sequentially because they depend on the previous steps.

To estimate F_0, the algorithm used is called DIO [63], which then uses StoneMask to make the estimation more reliable, whereas in later revisions of WORLD, the used algorithm is Harvest [64], which estimates F_0 with more precision, but requires more computational power.

The main idea behind DIO is to calculate F_0 candidates based on a criterion called *fundamentalness*, which is defined as the "variance of negative and positive going zero-crossing intervals and the intervals between successive peaks and dips". The input signal is separated in n frequency bands and the fundamentalness is calculated for each band, where a lower value means that its respective band is more likely to be on the fundamental frequency.

An improved DIO Method is also found in the literature [65], which claims that it performs better than DIO and is faster, but no code was found online as open source.

The Harvest algorithm [64] is not suitable for real-time speech analysis/synthesis applications according to the author, but offline tests done in a computer suggests that Harvest can be used in real-time applications. The algorithm first filters the input signal in different bands along the spectrum, then F_0 candidates are chosen. The estimate is refined by calculating the instantaneous frequency, defined as the derivative of the phase of the signal. This refined F_0 is then used to generate a contour of the F_0. The resulting algorithm performs better than other state of the art algorithms, in expense of computational efficiency.

The Spectral Envelope is estimated using CheapTrick [66]. The problem to be solved is to separate the effect of the source signal (i.e.: the glottal impulse) with the spectrum of the rest of the vocal tract. The general idea behind CheapTrick is to lifter in the quefrency domain the power spectrum. An issue of this method is that the power spectrum of a windowed waveform depends on the temporal position of the window, and to avoid this, a Hanning window of length $3T_0$ is used. This means that the F_0 has to be known or estimated. The power spectrum is smoothed with a rectangular windows with width of $2\omega_0/3$.

The Aperiodic Parameter estimation is done by using Platinum or D4C (lovetrain). The Platinum method is used in the legacy WORLD Vocoder, whereas the D4C (lovetrain) method is present in the latest versions of the WORLD Vocoder. The aperiodic parameter corresponds to the component of the source signal that is given by the erratic movement of the glottal pulses that deviates from a perfect pulse [27]. For the purpose of the D4C (lovetrain) Method, the aperiodicity is defined as the power ratio between the speech signal and the aperiodic component of the signal. The algorithm first generates a temporally static parameter based on

the concept of group, this temporally static parameter has the shape of a periodic signal with period ω_0 and does not depends on time, hence, the temporal stability. The next step is to take this temporally stable parameter and strip it of the noise, so that the resulting signal has a frequency of ω_0, which is given by the input signal. The next step is to estimate the aperiodicity of different frequency bands, given the temporally stable signal that was previosuly generated.

The last step is to synthesize the signal given the previously estimated parameters, corresponding to the fundamental frequency, spectral envelope and aperiodic parameter. The WORLD Vocoder uses impulses as an excitation signal instead of glottal pulses (based on glottal flow models). These impulses are then arranged so that they emulate the instantaneous fundamental frequency of the signal. The aperiodic component is synthesized separately from the periodic component, by adding a velvet noise [67] [68], which is used to synthesize a more natural sounding, less *buzzy* signal. The periodic component is synthesized using the spectral envelope and then both synthesized signals are added together. The aperiodic signal is weighted with the aperiodic component estimated with D4C.

The generated signal is a voice that was analyzed and then synthesized in real-time.

Chapter 3

VOICE QUALITY MODIFICATION

This chapter describes the methods and algorithms used to modify voice quality in real-time using the WORLD Vocoder. The first section explores the performance of the WORLD Vocoder with different settings and the feasibility of real-time voice modification. The following section explores the implementation of the Rosenberg++ pulse. Afterwards, Cheaptrick from the WORLD Vocoder is discussed, its limitations and the modifications done to the algorithm in order to adapt it to the Real-Time modification of Voice Quality. The following section discusses the methods used for the Voice Quality Modification, and the last section summarises the parameters that can be used to control the Vocoder.

3.1 WORLD Vocoder Performance

The authors of WORLD [13], measure the performance of the Vocoder with a parameter named Real-Time Factor (RTF) which corresponds to the ratio of the processing time with the length of the input signal, that means that an RTF of 1 corresponds to a signal of length n [s] processed in the same n [s].

The tests were done on a Dell XPS 13 Model 9343 with a 5th generation Intel core i5-5200 @2.20 GHz, with 8 GB of DDR3 Ram @1600 MHz Running Ubuntu 20.04, the input signal is sampled at 48$[KHz]$ with a depth of 16 bit in a *.wav* file, with a duration of 2.5 [s] or 120000 samples. The signal corresponds to a middle-aged male saying "Esta es una grabación de prueba" (This is a test recording).

Analysis was done using Cheaptrick for the spectral estimation, D4C for the aperiodic parameter estimation, Stonemask for the F_0 smoothing (DIO only) and either DIO or Harvest for F_0 estimation. Synthesis was done using three different available methods, standard synthesis, a synthesis where all the frames are added at the same time, and a synthesis with a more efficient ring buffer. Because the shortest synthesis was the standard method, and no audible differences were produced between different methods, it will only be presented the results of the standard synthesis. For both DIO and Harvest three different analysis windows were used: 5

[ms], 3 [ms] and 1 [ms]. A longer window generates a better synthesized signal and requires less processing overall, but the delay given by the buffer is longer. The results are shown in tables 3.1 and 3.2.

Results suggests that Harvest is not suitable for real-time voice quality modification, its RTF was less than 1 only with a 5 [ms] window. In the case of DIO, the RTF is less than 1 when the window is of length 3 and 5 [ms]. The problem that arises when the window is longer, is that the delay due to the filling of the window buffer is not negligible, considering that for this application, a real-time scenario will be considered when the delay is 25 [ms] or less [8]. The benefit of the highly parameterized nature of the WORLD Vocoder is that the window length can be modified without disrupting the rest of the algorithm, this means that in a real world scenario, the window length will be shortened as much as possible, while respecting the 25 [ms] criterion. A latency that is not considered in this analysis is the audio latency on the input, which corresponds to the microphone-Vocoder latency, and latency on the output, which corresponds to the Vocoder-headphone delay. This latency is not considered in this analysis and its subject to the hardware/software specific implementation, a desktop PC or a laptop can have a fast processor that allows for a quick analysis-synthesis, but the input and output latencies can be longer than this thesis supports, and in this case an embedded hardware with DSP features will probably be able to implement this algorithm in real-time. These problems will not be further discussed nor taken into account in the following section, because these are implementation-specific issues that will be solved when the algorithms are implemented in a real-world scenario, rather, this thesis will discuss the technical aspects of the voice quality modification, assuming that it can be implemented in real-time given the results of tables 3.1 and 3.2.

Window [ms]	**DIO [ms]**	**StoneMask [ms]**	**Cheaptrick [ms]**	**D4C [ms]**	**Synthesis [ms]**	**Total [ms]**	**RTF**
1	69	437	541	3387	89	4523	1.81
3	65	145	179	1108	92	1589	**0.64**
5	70	56	104	445	124	799	**0.32**

Table 3.1: RTF of WORLD Vocoder analysis-synthesis using DIO

Window [ms]	**Harvest [ms]**	**Cheaptrick [ms]**	**D4C [ms]**	**Synthesis [ms]**	**Total [ms]**	**RTF**
1	1426	501	2807	111	4845	1.94
3	1422	169	942	109	2642	1.06
5	1477	106	580	113	2276	**0.91**

Table 3.2: RTF of WORLD Vocoder analysis-synthesis using Harvest

3.2 The Rosenberg++ Pulse

The implementation of the Rosenberg++ Pulse developed in this thesis follows the equations (2.3.19) to (2.3.24). The pulse is defined based on t_e, t_a and t_p, which are defined as follows

- t_p: Opening time, the slope is positive
- t_e: Minimum of the glottal flow derivative waveform (excitation instant). Also known as Glottal Closure Instant (GCI)
- t_a: Time constant for the return phase

The pulse also includes as an input parameter the fundamental period $t_0 = 1/f_0$, the sampling frequency fs, and the amplitude k of the pulse, which is defined as the maximum value of the pulse at $t = t_e$. This definition of the amplitude of the pulse is used for convenience, that way, the WORLD Vocoder can reliably synthesize pulses of the desired amplitude. The original specification of the amplitude of the Rosenberg++ pulse is related to the derivative of the pulse, which is inconvenient when synthesizing the voice.

The parameters are restricted based on the specifications of [44] and [46] and to avoid numerical instabilities of the pulse. The restrictions of t_e, t_a and t_p are the following

$$\frac{t_e}{2} \leqslant t_p \tag{3.2.1}$$

$$t_e < t_o \tag{3.2.2}$$

Depending on the closing phase t_a, the restriction and parameters differ. If $t_a = 0$, then

- The $D(t_0, t_e, t_a)$ parameter disappears
- The pulse is not synthesized if $t_p = \frac{2t_e}{3}$
- t_x is defined as

$$t_x = \frac{t_e}{2}\frac{(3t_e - 4t_p)}{(2t_e - 3t_p)} \tag{3.2.3}$$

Conversely, if $t_a > 0$, then $D(t_0, t_e, t_a)$ is defined according to equation (2.3.22), and

- t_p is restricted by

$$t_p > \frac{3}{4}\frac{t_e(t_e + 4t_a D(t_0, t_e, t_a))}{(t_e + 3t_a D(t_0, t_e, t_a))} \tag{3.2.4}$$

- t_x is defined as

$$t_x = t_e \left(1 - \frac{(\frac{1}{2}t_e^2 - t_e t_p)}{2t_e^2 - 3t_e t_p + 6t_a(t_e - t_p)D(t_0, t_e, t_a)} \right) \tag{3.2.5}$$

The amplitude of the pulse K, defined as the maximum value of the pulse, or $f(t_p)$, is used to calculate an internal amplitude parameter K_{int}. The pulse is then multiplied by K_{int}, and the resulting pulse has the desired amplitude. The following expression is used to calculate K_{int}.

If $t_a > 0$

$$K_{int} = \begin{cases} K/(t_p^3(2t_x - t_p)/3) & t_p < \frac{(4D(t_0,t_e,t_a)t_a t_e + t_e^2)}{(2D(t_0,t_e,t_a)t_a + t_e)} \\ -K/(t_p^3(2t_x - t_p)/3) & \text{otherwise} \end{cases} \tag{3.2.6}$$

If $t_a = 0$, then K_{int} is defined as

$$K_{int} = \begin{cases} 3K/(t_p^3(2t_x - t_p)) & t_p < t_e \\ -3K/(t_p^3(2t_x - t_p)) & \text{otherwise} \end{cases} \tag{3.2.7}$$

It is convenient that $|t_p - t_e| > 0.001$, otherwise numerical instabilities appear as a consequence of dividing by a number close to zero, and compensating that division with K_{int}, where it grows to exponentially larger number as t_p approaches t_e.

3.3 Cheaptrick modification

Cheaptrick is the module of the WORLD Vocoder that extracts the spectral envelope of the voice. To achieve the spectral envelope extraction, Cheaptrick is done in three steps [66]:

1. Using a Hanning Window with a length of $3T_0$, a temporally stable power spectrum is generated

2. The power spectrum is smoothed by filtering with a rectangular window of size $2\omega_0/3$

3. A liftering on the quefrency domain is done, where two steps are done simultaneously

   - Remove frequency fluctuations caused by discretization
   - Spectral Recovery

The general idea behind the spectral recovery and the *discretization* described on the third step relies on the premise that voice can be modeled as follows

$$y(t) = h(t) * \sum_{-\infty}^{\infty} \delta(t - nT_0) \tag{3.3.1}$$

$$Y(\omega) = \frac{2\pi}{T_0} H(\omega) \sum_{-\infty}^{\infty} \delta(\omega - n\omega_0) \tag{3.3.2}$$

where $y(t)$ is the voice signal, $h(t)$ is the composite effect of the vocal tract, vocal folds and, in general, all the factors that affect voice production, excluding the fundamental frequency, and $\sum_{n=-\infty}^{\infty} \delta(t - nT_0)$ is the impulse train of period T_0 that represents the excitation signal of the vocal folds. This model of voice production is equivalent to the equations that describe the discretization of a continuous signal, and by using this idea, Cheaptrick extracts the spectral envelope.

The WORLD Vocoder is consistent with the previously mentioned model of voice production, using an impulse train as an excitation signal for synthesis. This approach to extract the spectral envelope assuming an impulse train and subsequent synthesis with the same impulse train is not compatible with the proposed method of using a glottal flow model as an excitation signal. To solve this, the Cheaptrick algorithm has to be modified.

If the Cheaptrick algorithm is not modified, then, using the Rosenberg++ pulse as an excitation signal results in a muffled audio with the lower frequencies exacerbated. The reason why this happens is that the spectral envelope extracted with Cheaptrick includes the effect of the glottal folds (similar to that of a low-pass filter), and synthesizing with a custom glottal fold excitation signal is equivalent as cascading two sets of glottal folds as excitation signal, the first set used as an excitation signal along with an spectral modulation, and the second set further modulates the spectral envelope of the voice, resulting in exacerbated lower frequencies. Three solutions are proposed for this problem.

3.3.1 Solution 1: High-pass filtering

A quick solution that was first proposed was to use a high-pass filter, either on the input signal (as a pre-emphasis), on the Rosenberg++ pulse, or on the resulting signal. The main benefit of this solution is that a high-pass filter is trivial to apply in real-time.

High-pass filtering the input signal was not an optimal solution in terms of perceptual quality of the synthesized signal. The resulting signal does not sound good probably because of the loss of information prior to analyzing the signal.

The result of high-pass filtering the Rosenberg++ excitation signal, synthesized an output signal that was perceptually similar to the input signal, although lacking some of the lower frequencies. To solve this issue, the output signal was linearly combined with the input signal using the following equation

$$y[n] = y_{filtered}[n] * Bias + x[n] * (1 - Bias) \tag{3.3.3}$$

By adjusting the $Bias$ parameter, it was possible to generate an output signal that sounded similar to the input signal. This solution allows for the modification of the Rosenberg++ pulse, which generates a similar, but different, output signal.

The problem with this solution, is that when using Rosenberg++ pulses synthesized with high values of the R_d parameter, the energy distribution in the frequency domain is predominantly in the lower frequencies. The resulting filtered Rosenberg++ pulse is a signal with an amplitude that is negligible with respect to the noise amplitude used as part of the excitation signal. This problem does not allow for voice quality modification as intended.

Another problem with filtering the excitation signal is the need of combining the synthesized signal with the input signal, because if a modification requires modifying the fundamental frequency of the voice, then the resulting combined signal sounds like two voices speaking at the same time, instead of a mix of both that sound like one modified voice.

Filtering the output signal generates a synthesized voice that does not sound natural, that is not similar to the input signal, and that is dependant of the modification done during synthesis and the characteristics of the input voice.

3.3.2 Solution 2: Inverting the glottal folds

The idea behind this solution is to first estimate the behaviour of the vocal folds of the input signal using a Glottal Inverse Filtering algorithm, and with this information invert in frequency the estimated signal, and use this inverted signal to filter either the spectral envelope, the excitation signal, or add an additional filter to the Cheaptrick filters.

This solution, if implemented correctly, is probably the solution that allows the widest range of modification for the synthesized voice, because the Vocoder then turns into a machine that emulates the vocal tract, and can use an arbitrary (not only the Rosenberg++ pulse) signal as an excitation signal.

The first problem of this approach is the feasibility of implementing a reliably Glottal Inverse Filtering (GIF) algorithm that works in real-time. GIF is an ongoing challenge in the literature, and adding the real-time constraint adds a complexity to the solution that is out of the scope of this thesis.

The second problem of this solution is inverting the estimated glottal folds signal. The low energy of higher frequencies means that a naive implementation of the inverse of the glottal folds can add numerical instabilities on higher frequencies. This means that an optimal solution has to take into consideration these instabilities.

3.3.3 Solution 3: Modifying the spectral envelope

An alternate solution, and the one that was used in the implementation of this thesis is to modify the extracted spectral envelope. By analyzing the recovered spectral envelopes, it is noted that the lower frequencies have a higher energy density. After a threshold that is roughly correlated with the fundamental frequency of the voice at that time, the spectrum has a lower energy density.

The proposed solution is to lower the values of the spectral envelope where the energy is higher, and leave intact the rest of the spectral envelope. Figure 3.1 shows the recovered spectrum of a frame, figure 3.2 shows the resulting spectrum after the modification and figure 3.3 shows the modifier used. Note that the modifier only affects the first 75 samples in this example, which corresponds to a section of the spectrum just before the second peak of the spectrum. In this example, the modified spectrum keeps the second peak intact, but its amplitude is one order of magnitude less than the original peak. The modifier is defined by

$$f[n] = \exp(n\frac{p}{m} - p) \tag{3.3.4}$$

Where $n \leqslant m$ and p is related to the slope of the exponential.

The examples shown in figures 3.1 to 3.3 have been synthesized with an input signal of a middle aged male subject, with a maximum F_0 of 211 $[Hz]$, a minimum F_0 of 76 $[Hz]$ and a mean F_0 of 128 $[Hz]$.

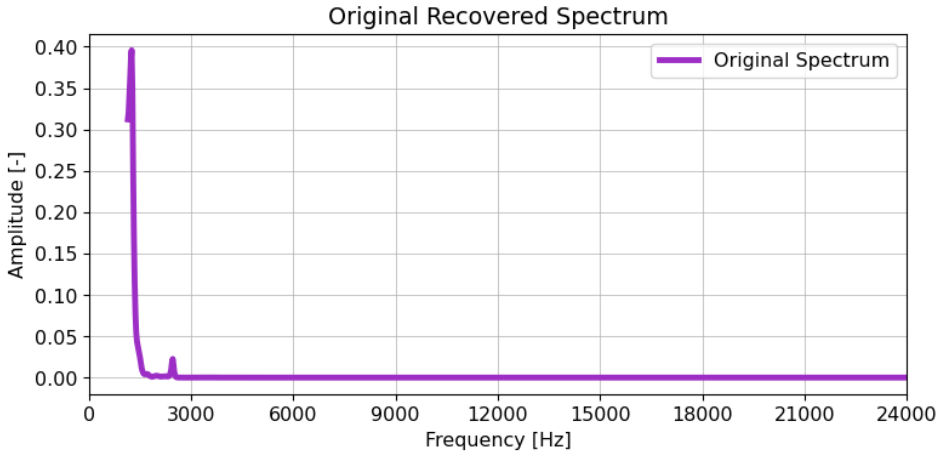

Figure 3.1: Original Spectrum

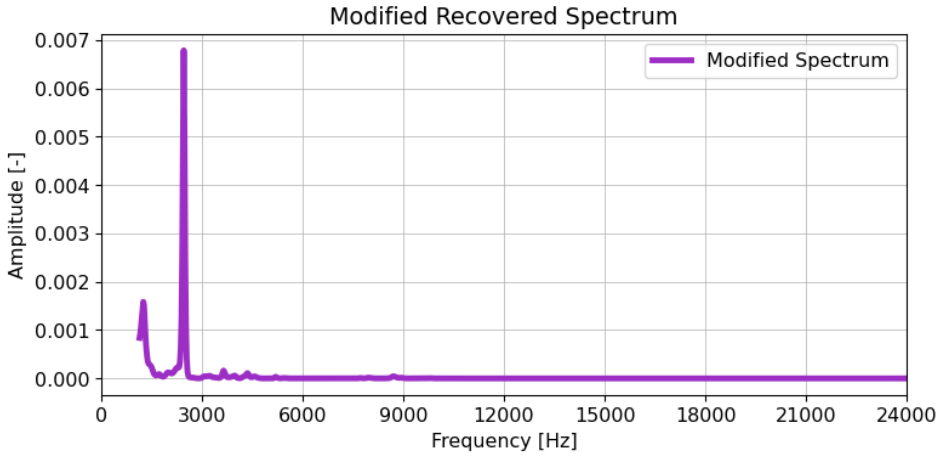

Figure 3.2: Modified Spectrum

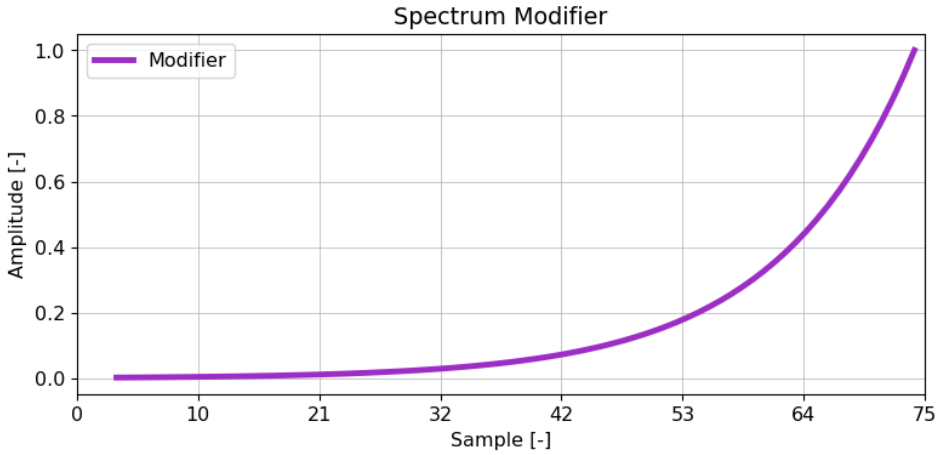

Figure 3.3: Spectrum Modifier

3.4 Voice Quality Modification

This section describes the methods and algorithms used for the voice quality modification. To accomplish the intended modifications, it is necessary to add parameters to the Vocoder. This section consists of a subsection that introduces and explains the parameters of the Vocoder, then the experiments carried out.

3.4.1 Vocoder Parameters

The following subsections are named according to the names of the parameter on the MATLAB code that was developed for this thesis.

rd_param

This parameter is used to control the R_d parameter of the pulse. The value of the parameter is used to obtain the t_e, t_p, and t_a parameters from the Rosenberg++ pulse by using the Cooperative Voice Analysis Repository (COVAREP) toolbox [69](v1.4.2).

f0_filter_frequency, f0_filter_order

These parameters are used to filter the variable that corresponds to the fundamental frequency. The WORLD Vocoder implementation uses an array which corresponds to the estimated fundamental frequency at each frame. This array is low-pass filtered with a Butterworth filter of order *f0_filter_order* and with cutoff frequency *f0_filter_frequency*. These parameters may be used to reduce jitter.

f0_multiplier

This parameter multiplies the estimated fundamental frequency array with a factor corresponding to *f0_multiplier*. This parameter, then, controls the fundamental frequency of the synthesized signal.

jitter_amplitude, jitter_frequency

These parameters are used to add jitter to the synthesized voice. *jitter_amplitude* controls the amplitude of the jitter and *jitter_frequency* controls the frequency of the added jitter. To generate jitter, first a Brownian noise is generated, and this Brownian noise is added to the estimated fundamental frequency array.

To generate the Brownian noise, first a white Gaussian noise is generated which is then filtered by a low-pass Butterworth filter. The idea behind this process is to generate a Brownian noise that does not drift. A naive implementation of a classical Brownian noise (by integrating Gaussian noise for example), generates a signal where its amplitude is proportional to the square root of the power of the Gaussian noise. This property of Brownian noise is undesired, hence, the implementation with a low-pass Butterworth filter that avoids this problem.

The *jitter_amplitude* parameter controls the power of the Gaussian noise, and *jitter_frequency* the cutoff frequency of the Butterworth filter, effectively controling the amplitude and frequency of the Brownian noise used to add jitter.

An important aspect to take into account when choosing the value of *jitter_frequency* is that the vector that carries the information of the fundamental frequency is not sampled at the same sample frequency of the input signal, but rather it is composed of the frames that the WORLD Vocoder is using. For example, if the frames are 0.2 [ms], then the sample frequency of the vector that has the fundamental frequency is 50 times higher.

shimmer_amplitude, shimmer_frequency

These parameters are used to add shimmer to the synthesized voice, where *shimmer_amplitude* controls the amplitude of the shimmer and *shimmer_frequency* controls the frequency of the shimmer. The implementation is similar to the jitter implementation, but differs on how the Brownian noise is added. The amplitude of the Rosenberg++ pulses is multiplied by $(1 + BN)$, where BN is the Brownian noise.

Note that the *shimmer_frequency* is acting over the amplitude of the excitation signal which is sampled at the same rate that the input signal.

vibrato_amplitude, vibrato_frequency

These parameters modulate in frequency the array of the estimated fundamental frequency. The array is modified with the following equation

$$F_0^{-*}[n] = F_0[n]VA * \sin(n * VF) \tag{3.4.1}$$

Where VA corresponds to the *vibrato_amplitude* parameter and VF to the *vibrato_frequency* parameter.

spectrum_filtering_exponential, spectrum_filtering_samples

These parameters implement the solution described in subsection 3.3.3, where *spectrum_filtering_exponential* corresponds to p in equation (3.3.4), and the parameter *spectrum_filtering_samples* corresponds to m of the same equation.

rpp_multiplier_te, rpp_multiplier_tp, rpp_multiplier_ta, rpp_k

These parameters are used to fine tune the Rosenberg++ pulse used for the excitation. While the R_d parameter controls the shape of the Rosenberg++ pulse, by defining the values of t_a, t_e and t_p, these multipliers allow controlling each individual parameter independently. Finally, the *rpp_k* parameter controls the amplitude of the pulse.

band_aperiodicity_multiplier

One of the internal parameters of the WORLD Vocoder is the *band_aperiodicity*, which is obtained after estimating the aperiodicity of the input signal with the D4C algorithm. This parameter corresponds to the estimated aperiodicity in bands of 3 [KHz]. The *band_aperiodicity_multiplier* parameter is an array that has the same number of elements as the *band_aperiodicity* parameter (of the WORLD Vocoder), that multiplies the amplitude of each band to its corresponding value.

3.4.2 Experiment 1: Perceptual evaluation (informal)

The objective of this experiment is to synthesize voice using the proposed Vocoder and do a perceptual evaluation of the synthesized signals.

Setup

Four input signals are synthesized and modified, two of a male subject and two of a female subject. The signals were taken from the Perceptual Voice Quality Database [70], the male signal corresponds to audio file *LA_9011_ENSS* and the female signal to audio file *LA9023_ENSS*. Two sections were taken from each audio file, one where the subject sustains the vowel /ɑː/, and one where the subject says "*Peter will keep at the peak*". Table 3.3 shows the maximum, minimum and mean value of the fundamental frequency of the input signals. *Female Running Speech* and *Male Running Speech* corresponds the sustained utterances, and *Female Sustained Vowel* and *Male Sustained Vowel* corresponds to the "*Peter...*" section. Note that the difference in fundamental frequency between the male and female subject is considerable, and that the sustained vowel has a more stable fundamental frequency, but maintains the same mean fundamental frequency with respect to the running speech section.

The voices were synthesized so that they sound like five kind of voices: modal voice, breathy voice, vocal fry, disphonia and rough voice. All voices are based on the modal voice, and to synthesize each particular voice parameters are added or modified.

Subject	**Max F0 [Hz]**	**Min F0 [Hz]**	**Mean F0 [Hz]**
Male Running Speech	181	74	114
Male Sustained Vowel	180	84	106
Female Running Speech	317	71	225
Female Sustained Vowel	268	149	227

Table 3.3: Fundamental frequency of input voices

Modal Voice

The modal voice is used as a *ground truth* for the following voice quality modifications. The objective is to synthesize a voice that sounds as similar as possible with respect to the original signal, while using the Rosenberg++ pulse as an excitation signal.

Parameter	Male Voice	Female Voice
rd_param	0.35	1
spectrum_filtering_exponential	8.5	12
spectrum_filtering_samples	45	75
rpp_multiplier_te	0.95	1
rpp_multiplier_tp	0.94	1
rpp_multiplier_ta	1	1
rpp_k	0.9	2.5

Table 3.4: Parameter values for modal voice

Breathy Voice

Breathy voice is related to relaxed glottal folds and with air passing through the folds even when they are the closest. This effect is obtained by raising the value of the R_d parameter. Depending on the voice, breathiness will no be perceived only by raising the value of the R_d parameter, and to exacerbate the effect, it was also necessary to lower the value of the pulse, so that the effect of the noise (which corresponds to the turbulent air passing on the glottal folds) is more noticeably, but by doing this the noise effect was too strong, so it was necessary to lower the *band_aperiodicity* parameter of the WORLD Vocoder.

Parameter	Male Voice	Female Voice
rd_param	2.8	4
spectrum_filtering_exponential	8.5	12
spectrum_filtering_samples	45	75
rpp_multiplier_te	0.95	0.8
rpp_multiplier_tp	0.94	1
rpp_multiplier_ta	1	1
rpp_k	0.5	0.8
band_aperiodicity_multiplier	0.5*[1 1 1 1 1 1 1]	0.5.*[1 1 1 1 1 1 1]

Table 3.5: Parameter values for breathy voice

Vocal Fry

To synthesize vocal fry, the fundamental frequency was lowered to the range of $50 - 75$ [Hz] [71], and 1% of jitter and shimmer was added to the synthesized voice.

Parameter	**Male Voice**	**Female Voice**
rd_param	0.35	1
f0_multiplier	0.5	0.35
spectrum_filtering_exponential	8.5	12
spectrum_filtering_samples	45	75
rpp_multiplier_te	0.95	-
rpp_multiplier_tp	0.94	-
rpp_multiplier_ta	1	-
rpp_k	0.5	-
jitter_amplitude	12	18
jitter_frequency	50*0.7	50*0.7
shimmer_amplitude	12	15
shimmer_frequency	50*0.7	50*0.7

Table 3.6: Parameter values for vocal fry

Disphonia

To synthesize disphonia, the amplitude of the excitation pulse is lowered, so that the sound is generated mainly by the noise generated with the aperiodicity estimation. To lower the background noise given by the low values of *rpp_k*, the first component of the *band_aperiodicity_multiplier* is lowered 10 times more than the other components. The *spectrum_filtering_exponential* of the female voice was lowered from 12 to 10.

Parameter	**Male Voice**	**Female Voice**
rd_param	1.35	1.35
spectrum_filtering_exponential	8.5	10
spectrum_filtering_samples	45	75
rpp_multiplier_te	0.95	0.45
rpp_multiplier_tp	0.94	1
rpp_multiplier_ta	1	0.1
rpp_k	0.5	0.5
band_aperiodicity_multiplier	0.05*[0.1 1 1 1 1 1 1]	0.05*[0.5 1 1 1 1 1 1]

Table 3.7: Parameter values for disphonia

Rough Voice

Rough voice is synthesized by adding jitter and shimmer along with lowering the fundamental frequency.

Parameter	**Male Voice**	**Female Voice**
rd_param	0.35	1.35
f0_multiplier	0.95	0.95
spectrum_filtering_exponential	8.5	12
spectrum_filtering_samples	45	75
rpp_multiplier_te	0.95	0.45
rpp_multiplier_tp	0.94	1
rpp_multiplier_ta	1	0.1
rpp_k	0.9	0.8
jitter_amplitude	14	17
jitter_frequency	50*0.5	0.8
shimmer_amplitude	15	17
shimmer_frequency	50*0.8	0.8

Table 3.8: Parameter values for rough voice

3.4.3 Experiment 2: Objective evaluation

The objective of this experiment is to generate objective measures of the most relevant parameters, so that they can be compared and evaluated with measures of voice quality.

Setup

The synthesized voices are the same of Experiment 1. The objective parameters that were applied to the Vocoder parameters for each voice were the Cepstral Peak Prominence (CPP), Harmonic-to-Noise Ratio (HNR), Mean Jitter, Mean Shimmer, Perceptual Evaluation of Speech Quality (PESQ), Peak Slope (PS) and Spectral Envelope H1-H2 (SE). The CPP, PS and SE were calculated using the COVAREP[69] toolbox, the HNR, mean jitter and mean shimmer were calculated using openSMILE[72] and PESQ was calculated using python-pesq[73].

The parameters that were modified were the *rd_param*, *f0_multiplier*, *rpp_k*, *jitter_amplitude*, *jitter_frequency*, *shimmer_amplitude* and *shimmer_frequency*. The parameters were evaluated with only the minimum parameters necessary to synthesize voice, and using the same values to synthesize modal voice in table 3.4. The *rd_param* was synthesized with the *spectrum_filtering_exponential* and *spectrum_filtering_samples* parameters, the *rpp_k* parameter was synthesized with the *rd_param*, *spectrum_filtering_exponential* and *spectrum_filtering_samples* parameters, *jitter_amplitude* was synthesized with the *jitter_frequency* parameter and viceversa, and *shimmer_amplitude* was synthesized with the *shimmer_frequency* parameter and viceversa. The objective measure SE was only calculated in the *rd_param*, *f0_multiplier* and *rpp_k* because it is the measure that takes the most to be generated, and it is expected to vary with the aforementioned parameters, otherwise, the parameter generation can take long times. The Vocoder parameters were then varied in different ranges as shown in table 3.9 and a measure was generated for each value.

Parameter	**Min Value**	**Step Size**	**Max Value**
rd_param	0.35	0.05	4
rpp_k	0	0.2	5
f0_multiplier	0.5	0.1	4
jitter_amplitude	0	2	50
jitter_frequency	0	100	22000
shimmer_amplitude	0	2	50
shimmer_frequency	0	100	22000

Table 3.9: Minimum, maximum and step values of the parameters

3.5 Results

3.5.1 Experiment 1

When synthesizing the different voices, the running speech had consistently better results than the sustained vowel speech. This probably due to the fact that the small imperfections of synthesis pass unnoticed, whereas with sustained vowels, the imperfections of the synthesis are repeating.

Modal Voice

The discrepancies with the parameter on both voices can be attributed to individual differences, and more importantly, differences in fundamental frequency. This is noted in the differences of the *spectrum_filtering_samples* parameter and with the R_d parameter. The resulting voice sounds similar to the original voice, although listening them side by side some differences can be notices, specially generating sounds that are more *brilliant*. These differences could probably be minimized by adding the *band_aperiodicity_multiplier* and modifying the different bands individually, but this fine tunning is easier to do on a setup where the parameters can be modified online rather than offline.

Breathy Voice

The generated voices sound breathy, it is noticeably the turbulence of the glottal folds not completely closing. In the female voice the effect was stronger, although, a noise in the higher frequencies was perceived. The use of the R_d parameter is useful to synthesize breathy voices, but other parameters that modify the pulse are needed to create a stronger effect.

Vocal Fry

The effect of lowering the fundamental frequency near the range $50 - 75[Hz]$ is strong enough for both male and female subject that the synthesized signal sounds with an important vocal fry. To exacerbate the effect jitter and shimmer was added, so that the voice sounds rougher. The synthesized voice sounds with vocal fry, but the characteristics of the input signals are still present, so that it is possible to identify the speaker.

Disphonia

To generate a disphonic voice, the amplitude of the excitation pulse was lowered to half the original amplitude, while maintaining low values of the R_d parameter. The synthesized voice sounds *brilliant*, which is undesired. Another problem is that it can be faintly perceived a modal voice that sounds at the same time with the disphonic voice, this can be a problem if the user is looking to synthesize a slightly disphonic voice, where the pulse amplitude with the *rpp_k* parameter is higher. The

main difference between the synthesis of breathy and disphonic voice is that breathy voice is based on using higher values of the R_d parameter, whereas the disphonic voice lowers the amplitude of the excitation pulse.

Rough Voice

To generate rough voice, the main component of it was the added jitter and shimmer, with a slight lowering of the fundamental frequency. The synthesized signal does not sound good. When only the fundamental frequency is modified, specially with jitter and shimmer, the resulting synthesized signal sounds robotic and noisy, which is an undesired result. The jitter and shimmer implementation is based on how it occurs in a real-world situation, but does not perform good in the WORLD Vocoder.

3.5.2 Experiment 2

In each figure, the *Female Speech* and *Male Speech* legends corresponds to the "*Peter...*" section of the female and male voice respectively, whereas the *Female Sust* and *Male Sust* legends corresponds to the corresponds to the sustained vowel a. In general, the results in the sustained vowel signals are more reliable, and the running speech signals have more noise, and in general are less reliable. This difference is mainly due to how the objective measurements are calculating the values and not with the quality of the synthesized voice of the running speech.

An important aspect to notice, is that Voice Quality is multidimensional [74] [75]. Whereas experiment 1 explores the synthesized voices in a multidimensional aspect, experiment 2 isolates the different parameters and evaluates the effect of the parameter with objective measures, and in this aspect, these experiments may show incompatible results.

rd_param

In accordance with the perceptual results and the literature [76] [77] [78][79], figure 3.4 shows that CPP is 6$[dB]$ lower when the R_d parameter has high values, figure 3.10 shows that the spectral envelope varies from 10 to 40$[dB]$ and in figure 3.9 the peak slope varies from -0.5 to -0.4, all of these values are indicative that the voice is changing from tense to breathy. The value of PESQ in figure 3.8 shows that in general, higher values of the R_d parameter synthesizes voice that are more different with respect to the original signal. It can be noted that for lower values of the R_d parameter, the female sustained signal has values of over 4 in the PESQ measurements. The other figures do not show correlations to the R_d parameter.

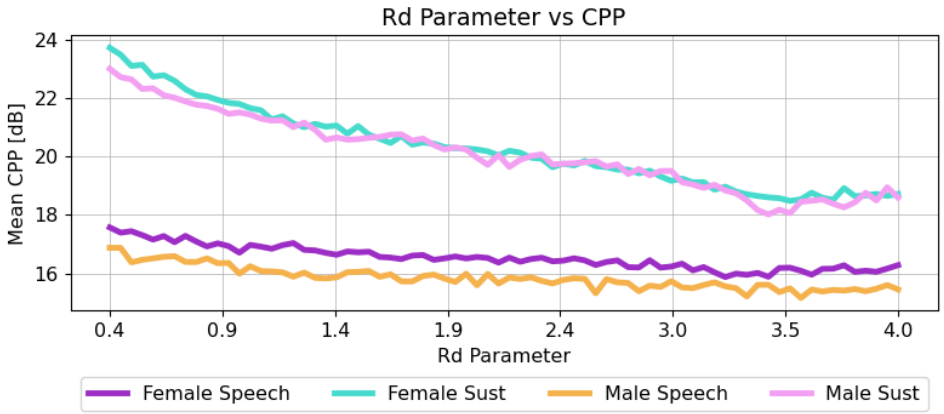

Figure 3.4: *rd_param* vs Mean CPP

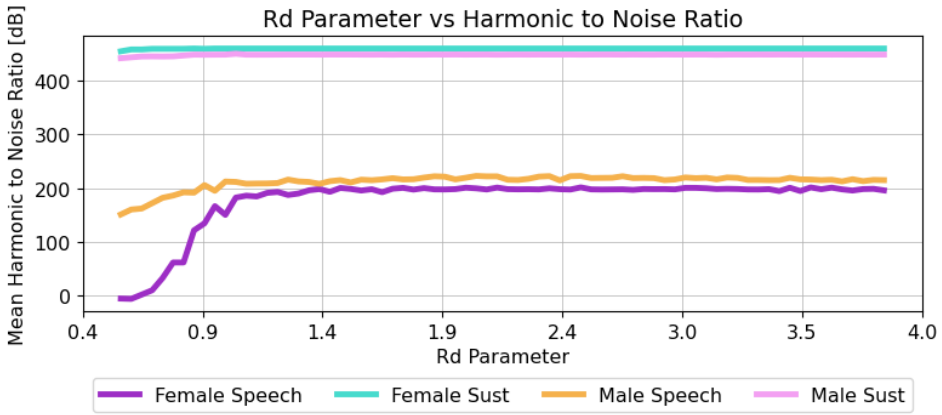

Figure 3.5: *rd_param* vs HNR

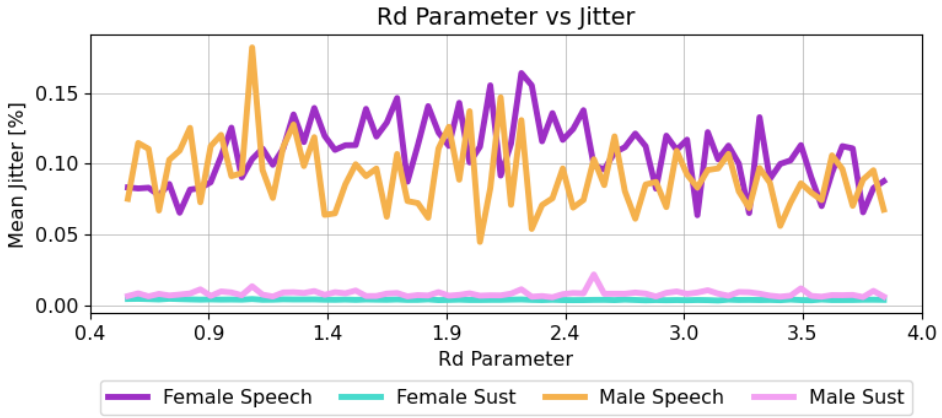

Figure 3.6: *rd_param* vs Mean Jitter

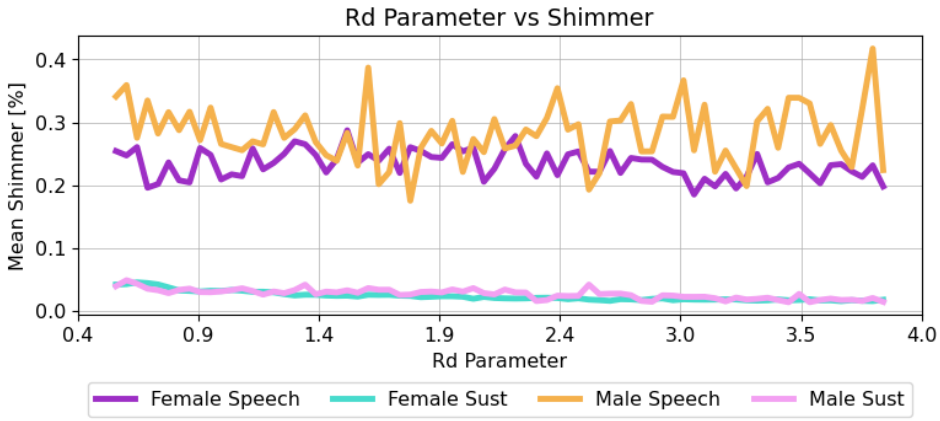

Figure 3.7: *rd_param* vs Mean Shimmer

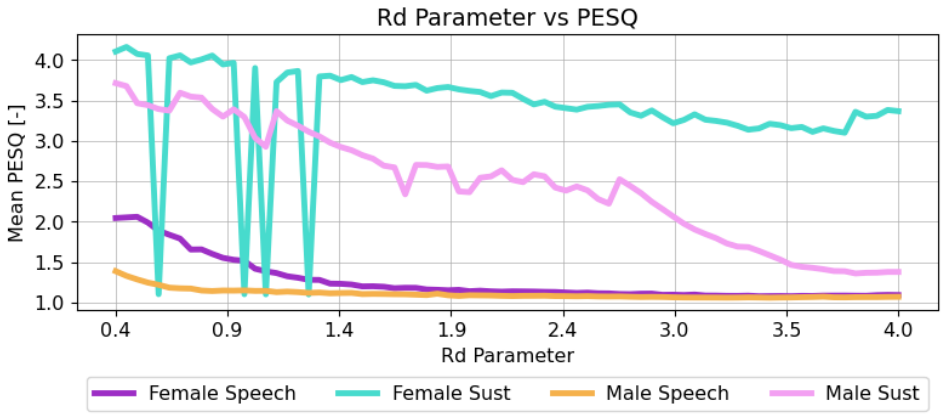

Figure 3.8: *rd_param* vs PESQ

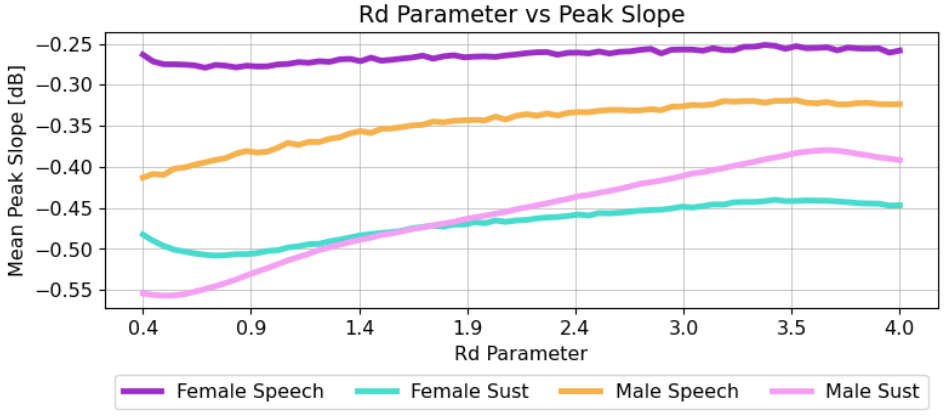

Figure 3.9: *rd_param* vs PS

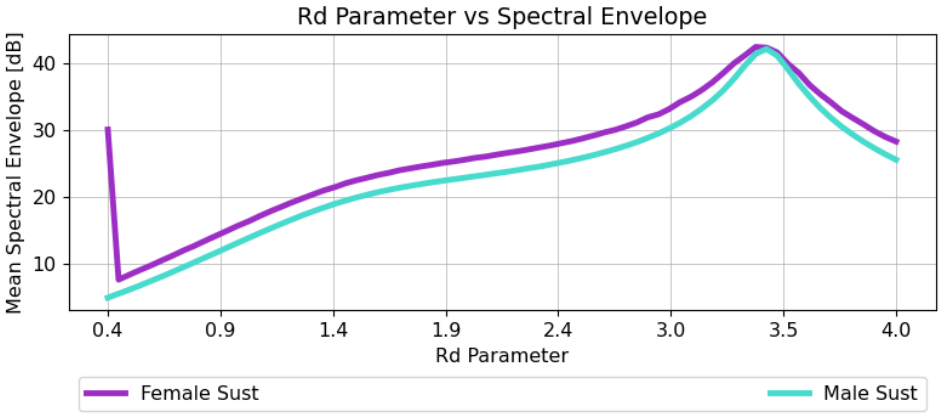

Figure 3.10: *rd_param* vs SE

f0_multiplier

It is clear from figure 3.15 that the peak value of the PESQ is when the *F0_multiplier* is 1, which is consistent to the idea that not modifying the fundamental frequency allows for a voice that is perceptually similar to the original voice. It is important to notice that this happens to all four signals, and that the peak values are similar to that of the peak values of 3.8, this is an important result because when the *F0_multiplier* is 1, the Vocoder is not modifying in any aspect the voice, which means that the synthesized voice with R_d parameter can synthesize modal voice like the input signal. The other measurements are not necessarily related to the parameter, although the results of figure 3.16 are attention drawing, in the sense that there is a strong correlation with the PS, but it can be a spurious correlation.

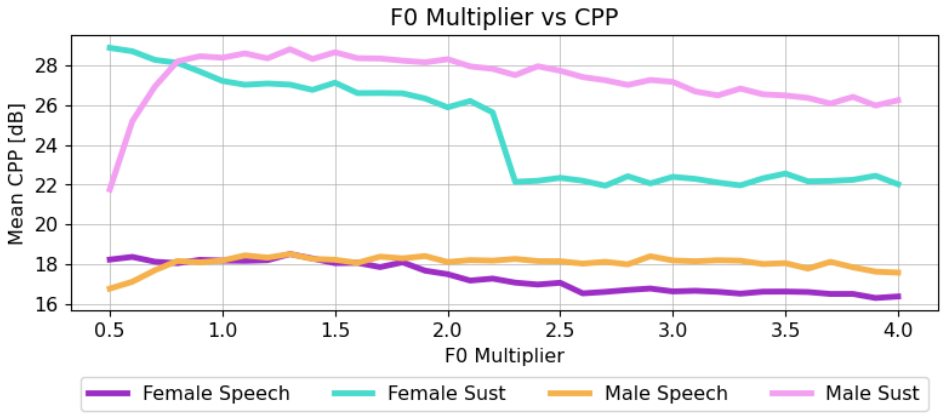

Figure 3.11: *f0_multiplier* vs Mean CPP

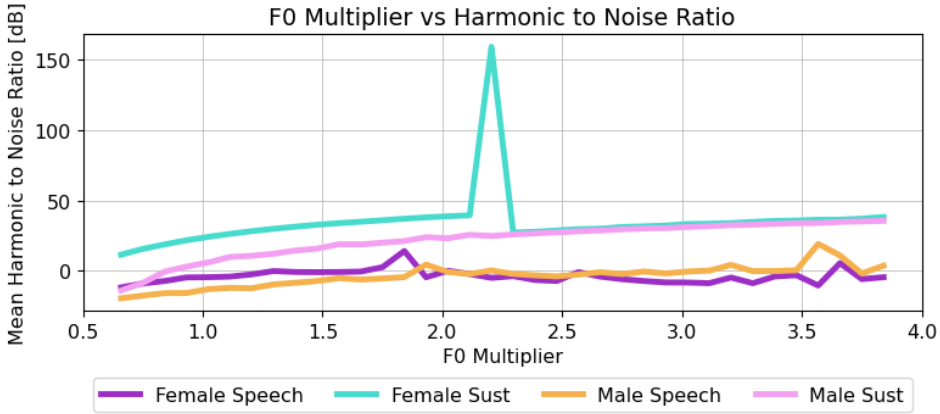

Figure 3.12: *f0_multiplier* vs Mean HNR

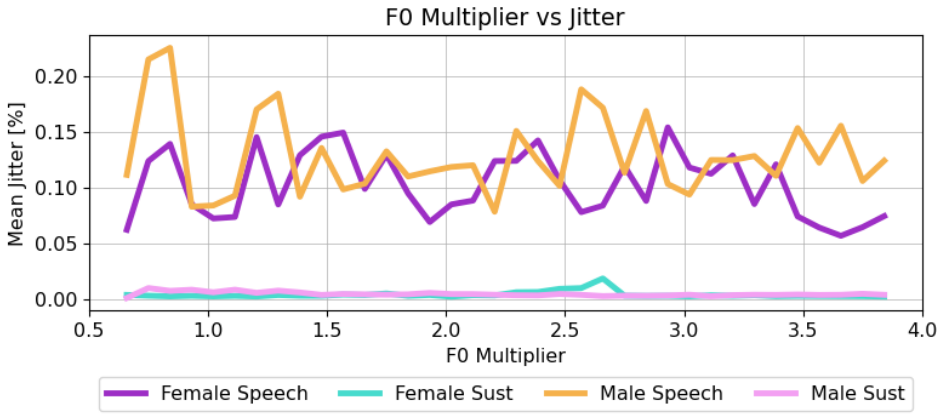

Figure 3.13: *f0_multiplier* vs Mean Jitter

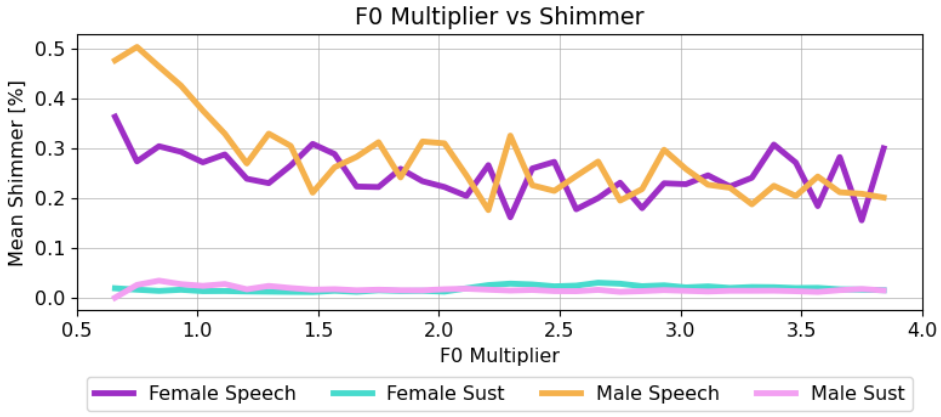

Figure 3.14: *f0_multiplier* vs Mean Shimmer

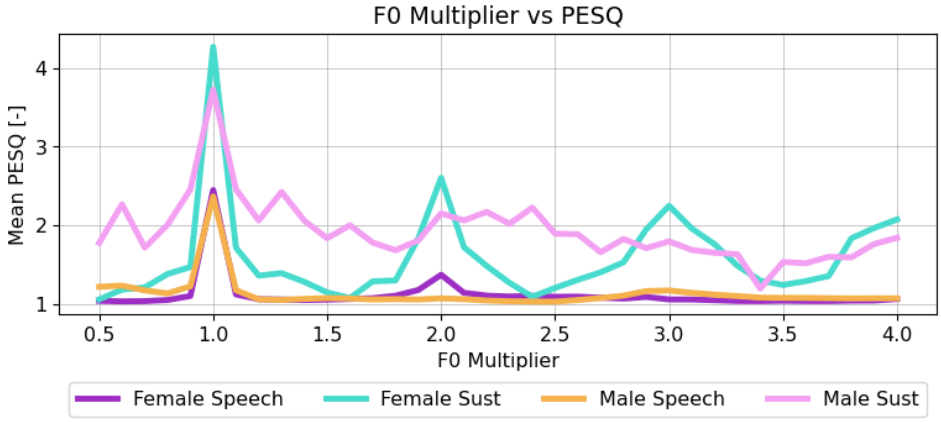

Figure 3.15: *f0_multiplier* vs Mean PESQ

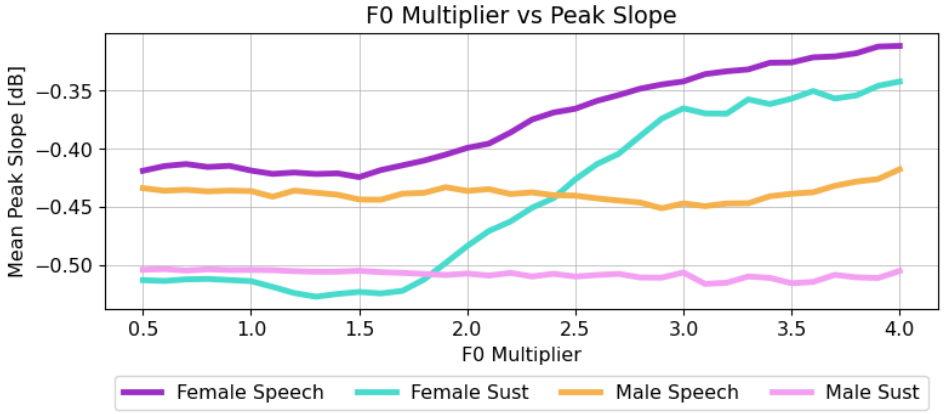

Figure 3.16: *f0_multiplier* vs Mean PS

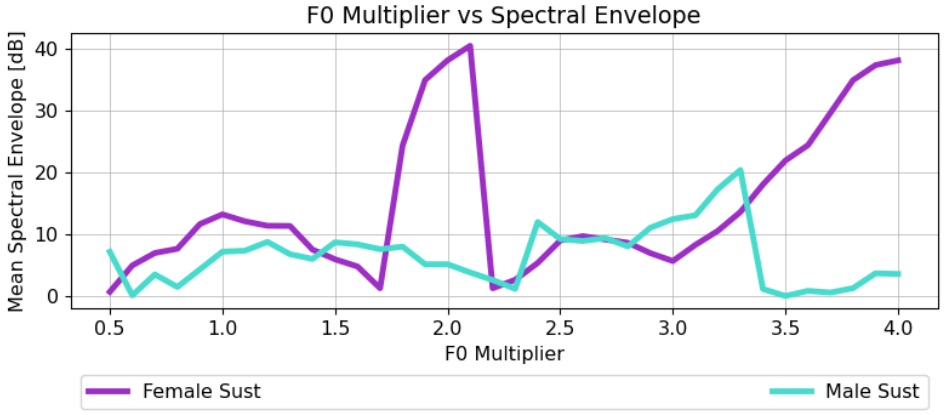

Figure 3.17: *f0_multiplier* vs Mean SE

rpp_k

In figure 3.18 it can be noted the drastic change of the CPP of 6[dB] in the case of the sustained vowels when the pulse amplitude changes from 0 to a higher value. This change is due to the fact that when this parameter is 0, there is no excitation pulse, and the synthesized signal is excited only by noise, emulating a breathy or disphonic voice. It can be seen that this drastic change is repeated in figures 3.19, 3.20, 3.22, 3.23 and 3.24. An unexpected result is in the case of the HNR, because it was expected that the HNR would be correlated to the pulse amplitude, in the sense that the noise component of the synthesized signal is higher with respect to the noise. The PESQ value converges quickly to its maximum value when the amplitude is non-zero, the PS and the SE have the same behaviour. Jitter and shimmer are not correlated in any way as expected.

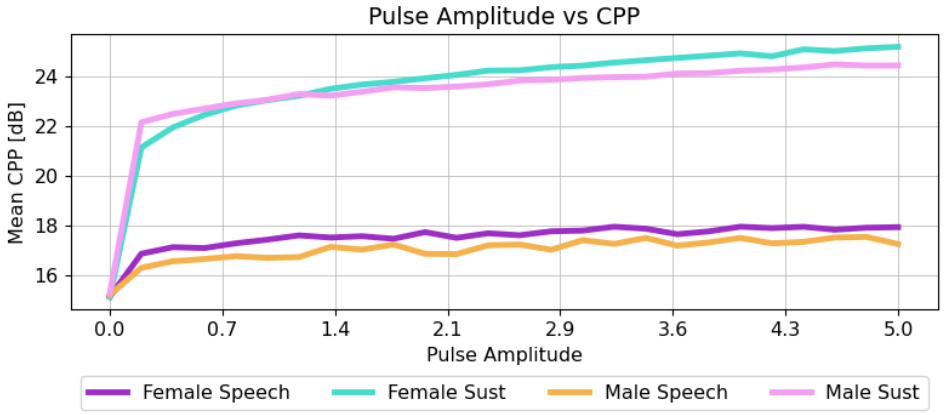

Figure 3.18: *rpp_k* vs Mean CPP

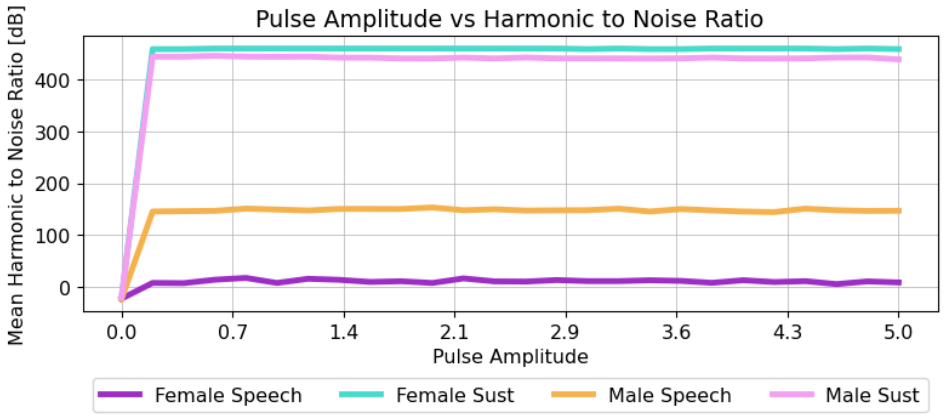

Figure 3.19: *rpp_k* vs Mean HNR

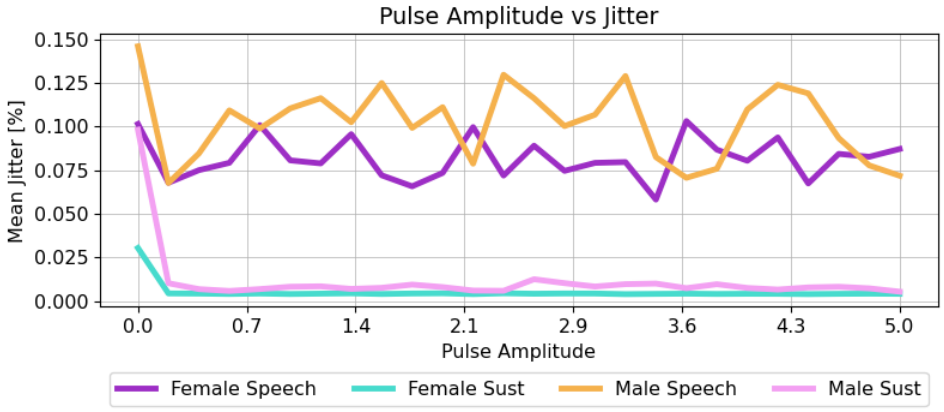

Figure 3.20: *rpp_k* vs Mean Jitter

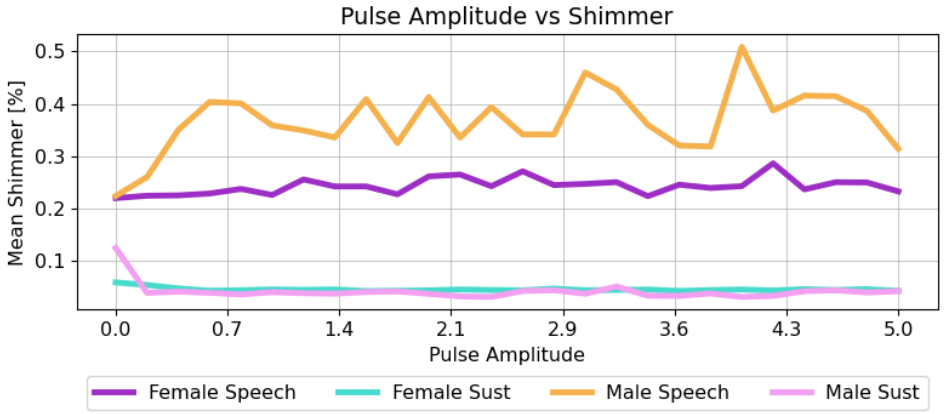

Figure 3.21: *rpp_k* vs Mean Shimmer

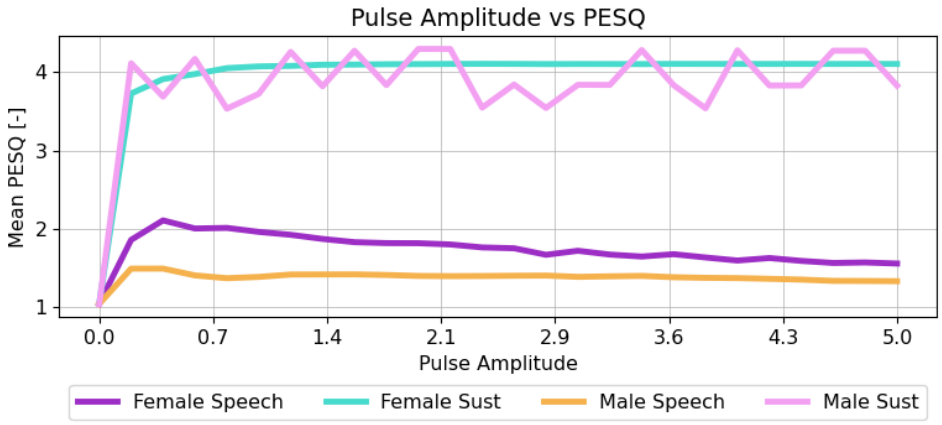

Figure 3.22: *rpp_k* vs Mean PESQ

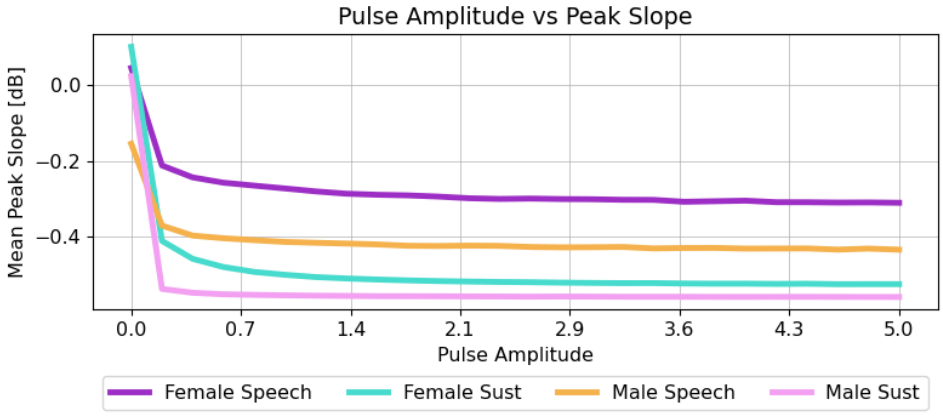

Figure 3.23: *rpp_k* vs Mean PS

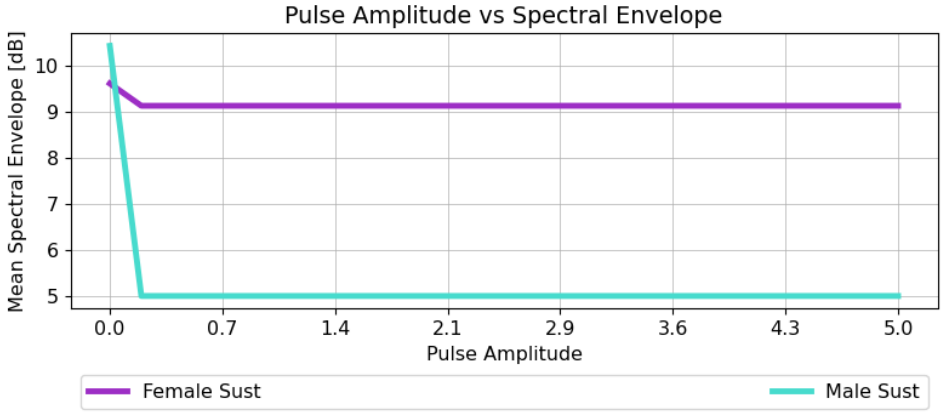

Figure 3.24: *rpp_k* vs Mean SE

jitter_amplitude and jitter_frequency

This section will analyze the results of both the *jitter_amplitude* parameter and the *jitter_frequency* parameter. As expected, PESQ lowers when either the amplitude or the frequency of the jitter increases, being more susceptible to changes of the frequency, as it can be seen in figures 3.29 and 3.35. The measured value of the jitter increases from 0.35% to 31% from changes of the frequency (figure 3.33) and from 1.2% to 25% from changes in amplitude (figure 3.27), which is consistent with the values of the parameters. The measured shimmer is also affected in a similar manner which is unexpected. The amplitude and frequency of jitter are inversely correlated with the HNR, which is what is expected, as higher values of jitter are related with a lower harmonic component of the signal.

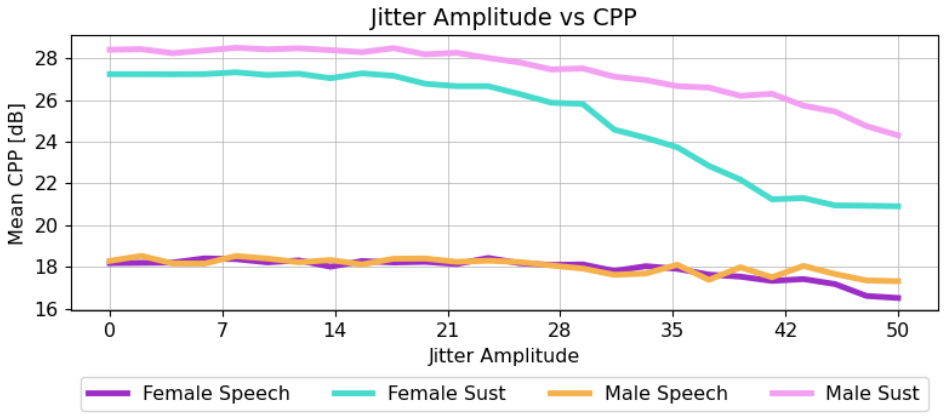

Figure 3.25: *jitter_amplitude* vs Mean CPP

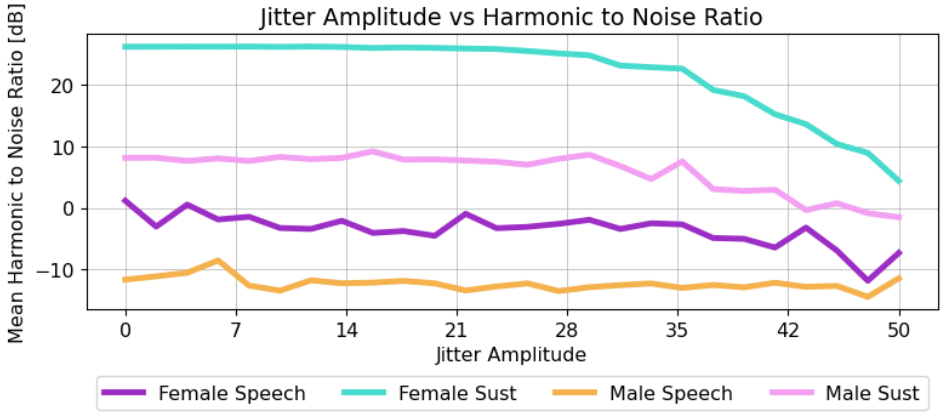

Figure 3.26: *jitter_amplitude* vs Mean HNR

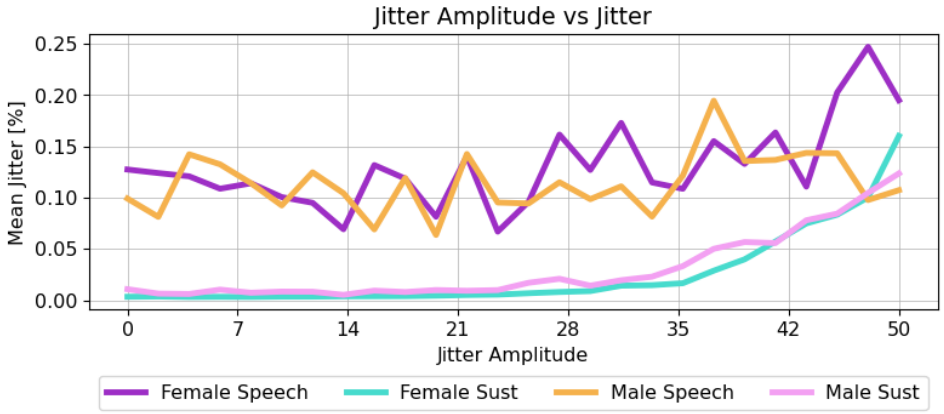

Figure 3.27: *jitter_amplitude* vs Mean Jitter

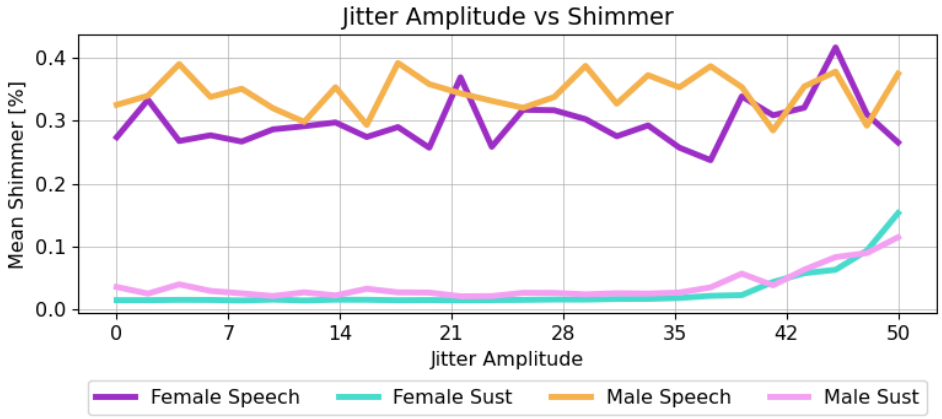

Figure 3.28: *jitter_amplitude* vs Mean Shimmer

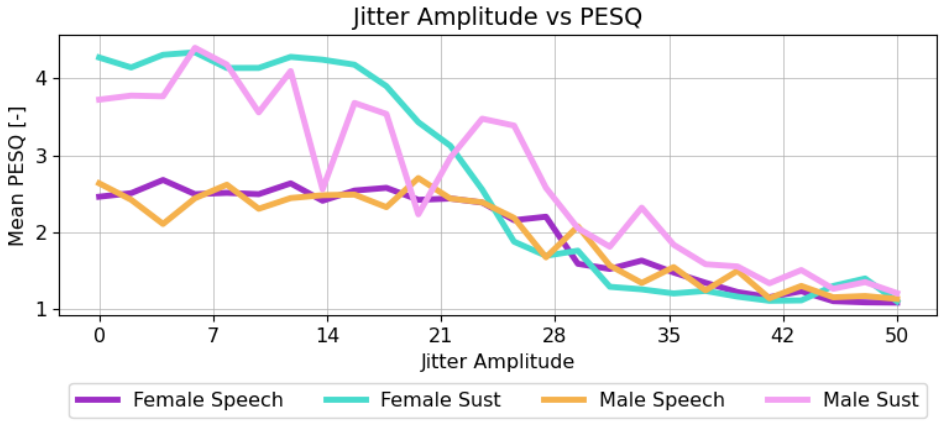

Figure 3.29: *jitter_amplitude* vs Mean PESQ

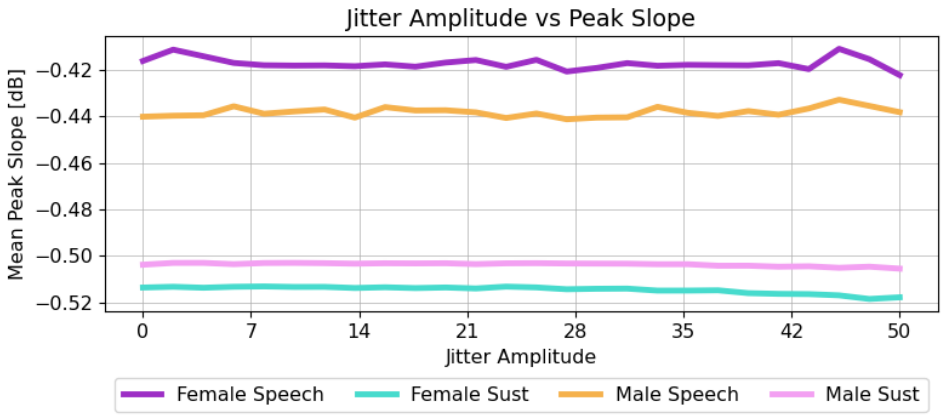

Figure 3.30: *jitter_amplitude* vs Mean PS

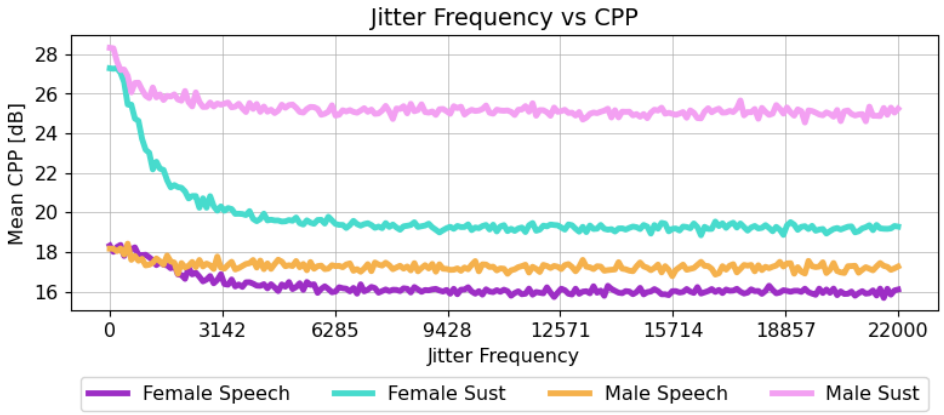

Figure 3.31: *jitter_frequency* vs Mean CPP

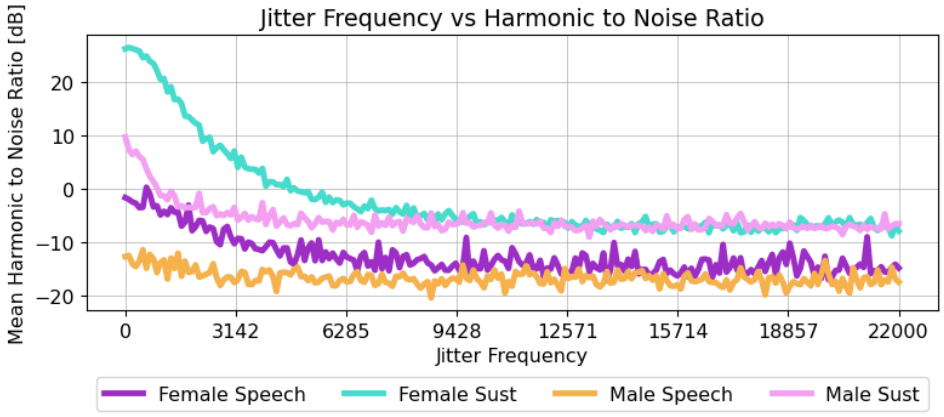

Figure 3.32: *jitter_frequency* vs Mean HNR

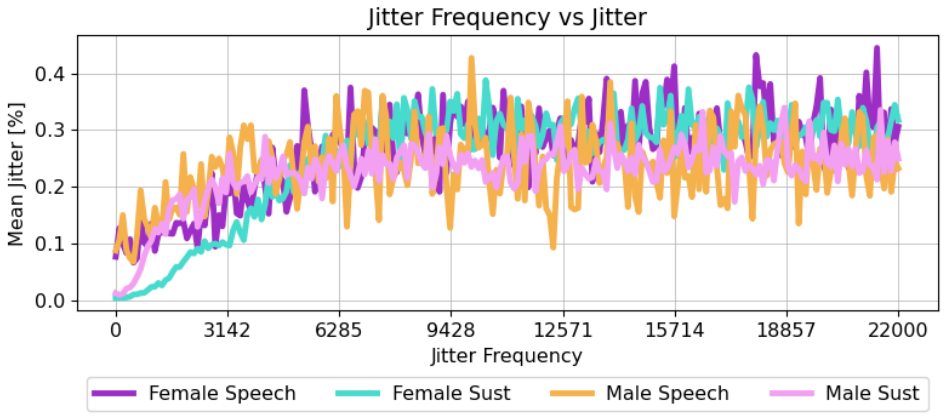

Figure 3.33: *jitter_frequency* vs Mean Jitter

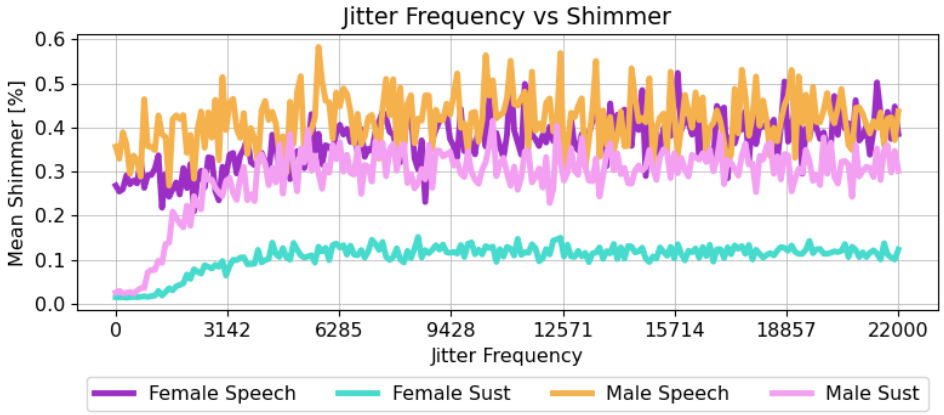

Figure 3.34: *jitter_frequency* vs Mean Shimmer

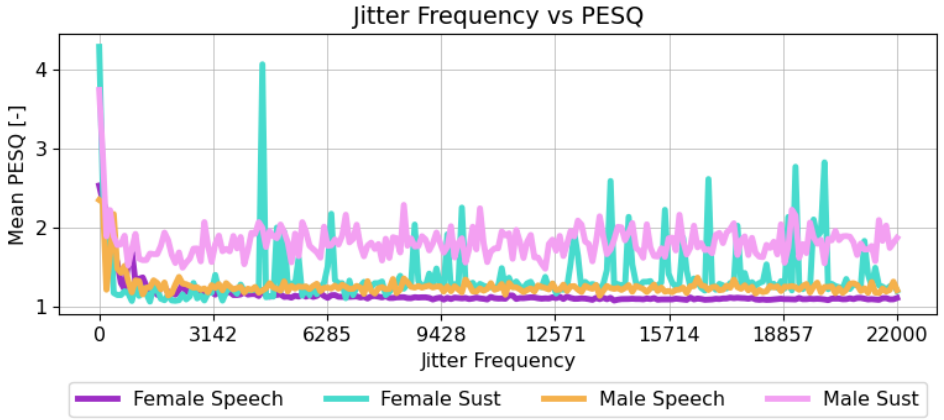

Figure 3.35: *jitter_frequency* vs Mean PESQ

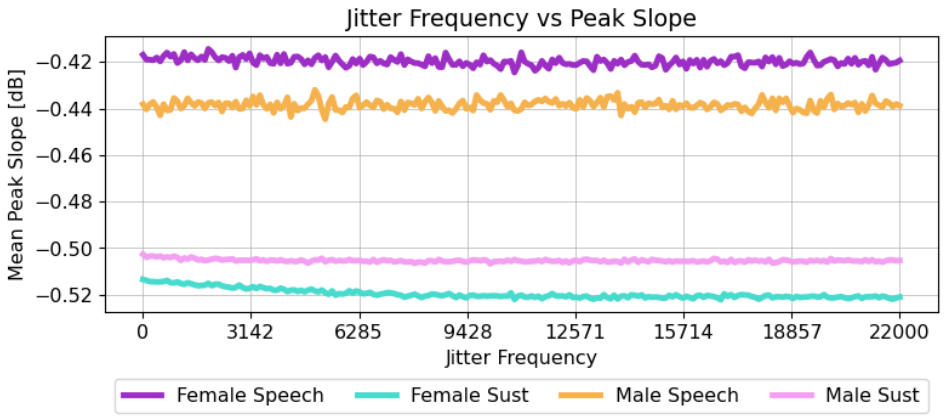

Figure 3.36: *jitter_frequency* vs Mean PS

shimmer_amplitude and shimmer_frequency

This section will analyze the results of both the *shimmer_amplitude* parameter and the *shimmer_frequency* parameter. The *shimmer_amplitude* and *shimmer_frequency* affect the measured shimmer as it is shown in figures 3.40 and 3.46, changing from 2.5% to 30% for changes in frequency and from 1.4% to 12% for changes in amplitude, which is consistent with the values of the parameters. Following a similar trend with respect to the *jitter_frequency* and *jitter_amplitude* parameters, the measured jitter for both amplitude and frequency of shimmer is affected. As expetected, the PESQ values drops from near 4 to 1 when the *jitter_amplitude* and *jitter_frequency* parameters are higher. CPP and HNR also drops with higher values of both parameters, which for the case of the HNR is expected.

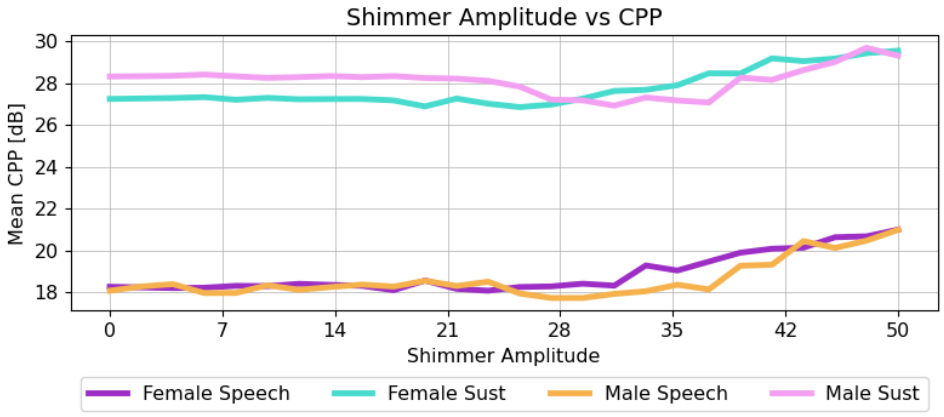

Figure 3.37: *shimmer_amplitude* vs Mean CPP

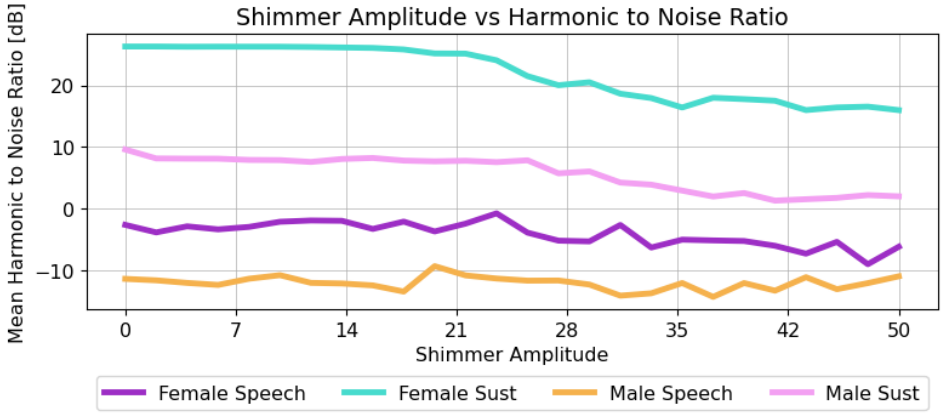

Figure 3.38: *shimmer_amplitude* vs Mean HNR

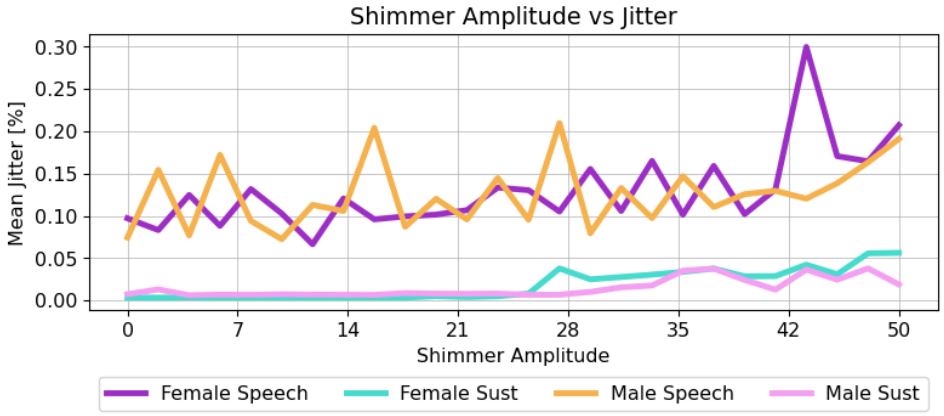

Figure 3.39: *shimmer_amplitude* vs Mean Jitter

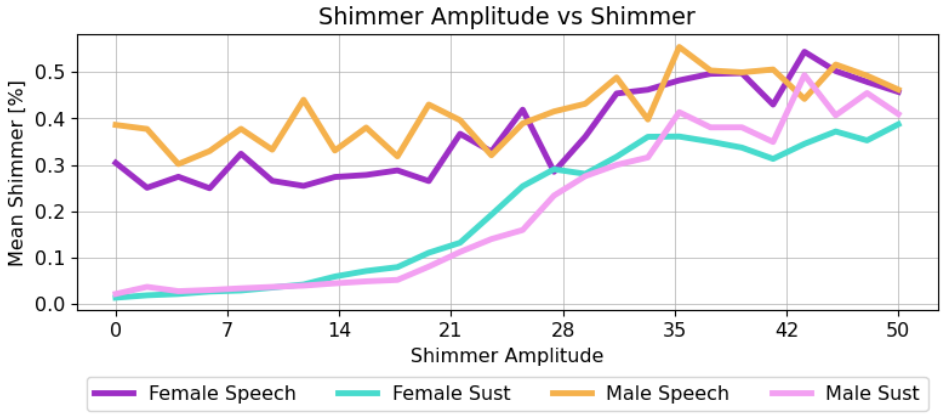

Figure 3.40: *shimmer_amplitude* vs Mean Shimmer

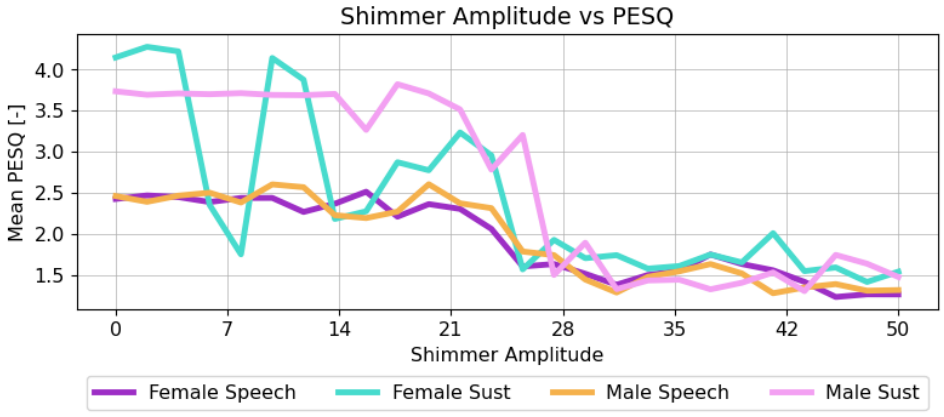

Figure 3.41: *shimmer_amplitude* vs Mean PESQ

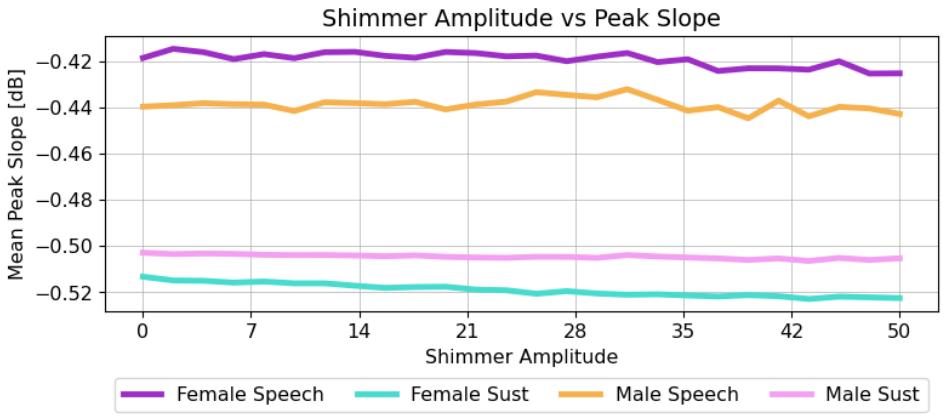

Figure 3.42: *shimmer_amplitude* vs Mean PS

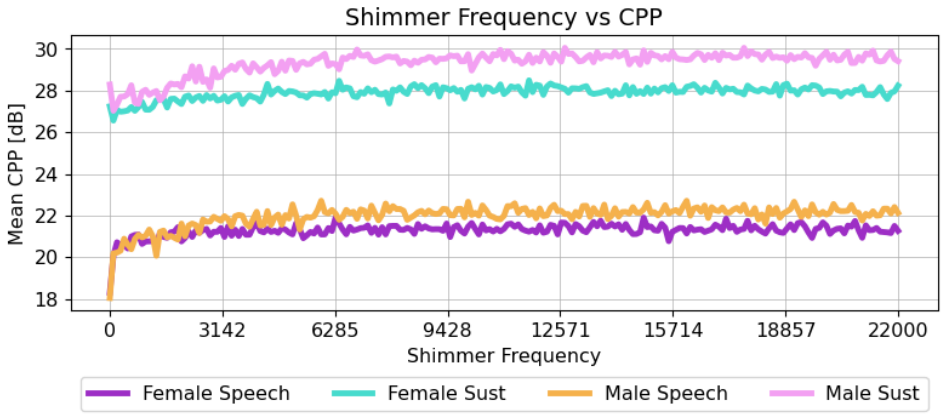

Figure 3.43: *shimmer_frequency* vs Mean CPP

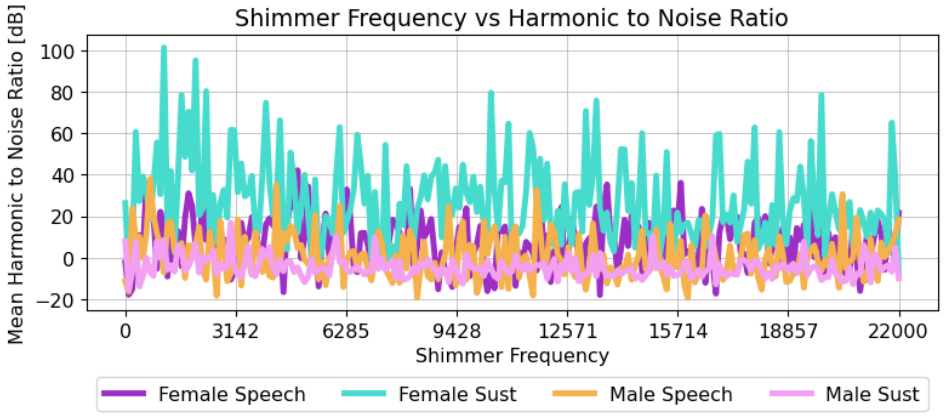

Figure 3.44: *shimmer_frequency* vs Mean HNR

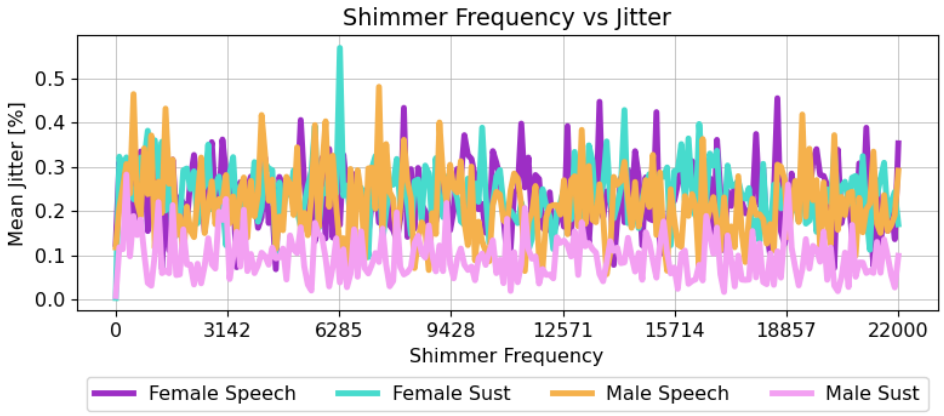

Figure 3.45: *shimmer_frequency* vs Mean Jitter

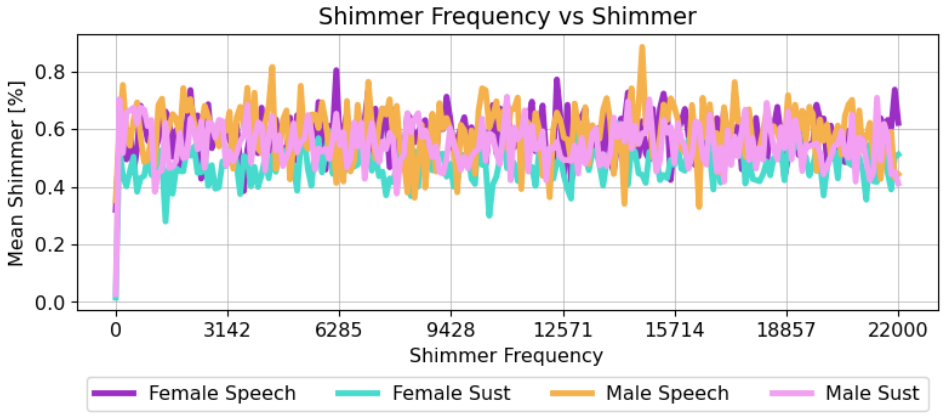

Figure 3.46: *shimmer_frequency* vs Mean Shimmer

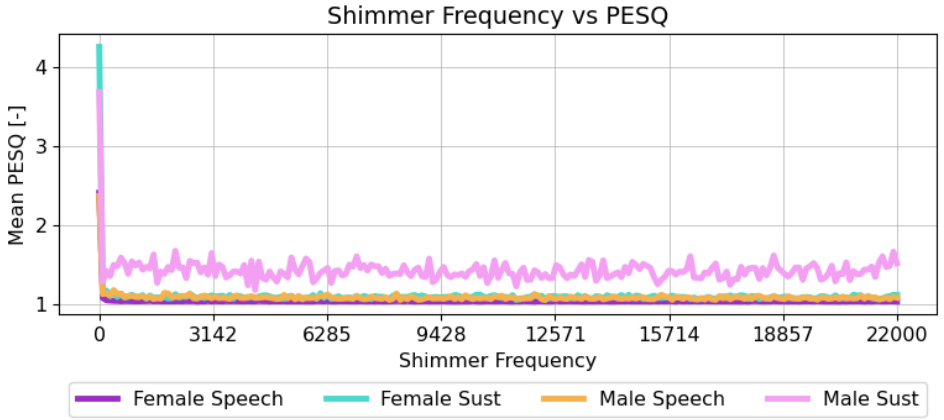

Figure 3.47: *shimmer_frequency* vs Mean PESQ

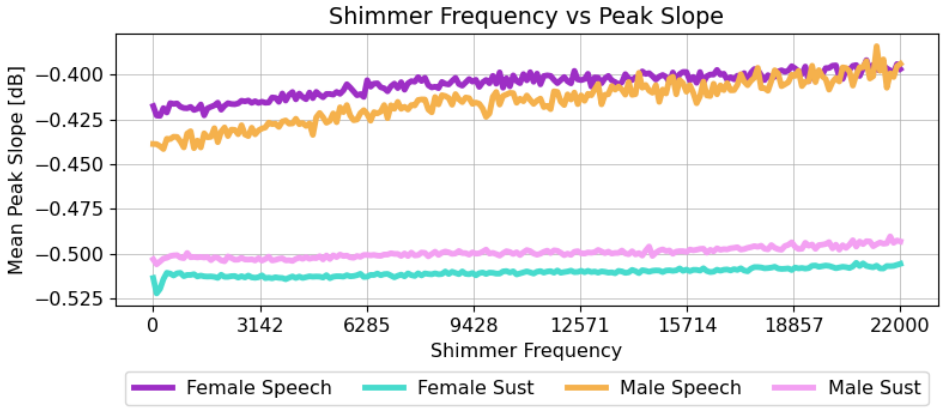

Figure 3.48: *shimmer_frequency* vs Mean PS

Chapter 4

CONCLUSIONS AND FUTURE WORK

4.1 Future Work

This work has shown that it is possible, at least in theory, to modify Vocal Quality in real-time, but no meaningful implementations has been done in real-time. The next step is to implement this work in real-time, and for this objective there are several options. As it has been explored in section 3.1, it is feasible to implement the WORLD Vocoder in real-time, with a real-time Factor (RTF) that is well below 1 in a modern processor. The potential problems that appear is that in a processor based on, for example, x86 architecture, which uses an Operating System like Windows or Linux, the delay between an input signal and that signal (even without any modifications) leaving the system to the headphones of a user can be well over the 25 [ms], which corresponds to the real-time threshold in this work. Another solution is to implement this system in an embedded system, where the bottleneck can be the processing power of such a system. Then, the following work should address that, and work around this idea on how to properly implement this system in real-time. Probably the best way to proceed is using highly specialized DSPs to do this work.

The modification of Cheaptrick in this work plays a fundamental role to synthesize a voice that sounds natural by using the Rosenberg++ pulse as an excitation signal. The proposed solution of this work is not ideal, in the sense that no physical model is used to fix the Spectral Envelope, just a perceptual *tinkering* of the parameters and proposed solutions. This is probably where Machine-Learning can play an important role for a related work. The proposed solution in this thesis only modifies the first samples of the spectral envelope, but it is not only the samples of the lower parts of the spectrum that plays an important perceptual role in voice production. By using Machine Learning, it is hypothesized that it will be possible to modify the spectral envelope, not only to filter the effect of the glottal pulse, but even to change some aspects of Vocal Quality. This work has shown that it is possible to worsen the voice quality of a subject, but fixing it has not been done satisfactorily. By training a deep learning machine to change the spectral envelope to fix vocal quality will probably solve a problem to the millions of persons that need to work

with their voices and are affected by different illnesses related to their voices. There are Text-To-Speech synthesizers[80], and voice separation machines[81] that use the WORLD Vocoder, using the idea to synthesize an spectral envelope to generate voice.

The proposed parameters are probably insufficient for the modification of all kinds of voice quality, but the developed framework of this thesis opens the doors for further development in this area. Fine tuning the algorithms associated to the parameters, and proposing new parameters can lead to better results of voice quality modification. Another issue related to the proposed parameters, is that there are probably too many parameters for an inexperienced user, which can lead to frustration or inability to use the Vocoder. An automated calibration step is probably a good way to solve part of this problem. This calibration step can be done offline, as it is dependant of the vocal characteristics of the subject under test, and probably these characteristics do not change drastically during a session using the Vocoder.

With the idea of calibration in mind, it would be convenient to estimate the glottal pulse in real-time of the subject under test, which could lead to better, more focalized modification of voice quality.

A shortcoming of using only the Rosenberg++ pulse is that it does not let the user to generate an arbitrary glottal pulse, and there is space to research alternative excitation pulses. For example, to generate vocal fry, the fundamental frequency is approximately divided by 2. Probably a better solution is to emulate with more precision the process of vocal fry, where there is a second pulse that is lower than the first pulse and near it.

4.2 Conclusions

The results of this thesis can be put into three categories, first, the technical challenge to develop and further modify parameters that are related to voice quality, second, the perceptual dimension of the synthesized signals, which is probably the most important aspect of this thesis, because it directly solves the motivation of this work, which is to do neuroscience experiments based on the self perception of voice quality. The third aspect of this thesis are the objective measures that can be done to the modifications of voice quality.

The technical challenge was, in its most part, solved. Using the WORLD Vocoder, and the idea that modifying the excitation signal of the Vocoder, it was shown that it is possible to modify voice quality in real-time, although, it is important to notice that this work has not been implemented in an embedded system at the moment, and there could be problems with this real-time *promise*. The Rosenberg++ pulse as an excitation signal has shown to be useful, but with that, modifications to the WORLD Vocoder other than the Synthesis module, has been needed. In this thesis, the main work has been done on modifying the Cheaptrick

module, and modifying the Synthesis module. To give a researcher or a potential user more flexibility, several parameters have been proposed, which in conjunction are able to modify voice quality. The result is a framework in which future work can be based that allows for the real-time modification of voice quality.

The perceptual dimension was not fully explored. Sanitary restrictions due to COVID-19 made it difficult to perform serious and controlled experiments to rigorously evaluate the results, rather, perceptual work was done informally. Probably, an important challenge for the near work is to perceptually evaluate the results of this work in a controlled environment, although, even when informally, the results look promising. The synthesized signals that emulate modal, breathy, disphonic voice, and vocal fry shows good results, but the rough voice needs more work, probably using a new implementation of jitter and shimmer.

The objective evaluation of the synthesized voice resulted gave some insight on the performance of the proposed parameters. The R_d parameter showed strong results that are concordant with the literature along with perceptual results that look promising. The F_0 multiplier showed results that were in line with what was expected, being the main factor when synthesizing vocal fry and rough voice. The pulse amplitude was showed to be a parameter that was correlated as expected with the objective measures, standing out its effect when the excitation pulse is non existent to synthesize disphonic voices. The jitter and shimmer parameters shows strong results that shows that the implementation works as expected with the objective parameters, although, in the perceptual dimension, this is not the case.

In general, the sustained vowel shows stronger results in the objective measures and the running speech shows stronger results in the perceptual setup. In the case of the objective measures, this is expected because these measurements are typically designed for sustained vowels, and for the running speech the parameters are averaged for (sometimes drastically) different frames. In the case of the perceptual setup, the opposite effect is in play: when listening running speech, the small imperfections of synthesis pass unnoticed, whereas with sustained vowels, the imperfections of the synthesis repeats again and again.

REFERENCES

[1] J. L. Elman, "Effects of frequency-shifted feedback on the pitch of vocal productions," *The Journal of the Acoustical Society of America*, vol. 70, pp. 45–50, jul 1981.

[2] R. Taitelbaum-Swead, M. Avivi, B. Gueta, and L. Fostick, "The effect of delayed auditory feedback (DAF) and frequency altered feedback (FAF) on speech production: cochlear implanted versus normal hearing individuals," *Clinical Linguistics & Phonetics*, vol. 33, pp. 628–640, jan 2019.

[3] J. Chesters, L. Baghai-Ravary, and R. Möttönen, "The effects of delayed auditory and visual feedback on speech production," *The Journal of the Acoustical Society of America*, vol. 137, pp. 873–883, feb 2015.

[4] G. A. Soderberg, "Delayed auditory feedback and stuttering," *Journal of Speech and Hearing Disorders*, vol. 33, pp. 260–267, aug 1968.

[5] H. Lane and B. Tranel, "The lombard sign and the role of hearing in speech," *Journal of Speech and Hearing Research*, vol. 14, pp. 677–709, Dec. 1971.

[6] G. E. Galindo, S. D. Peterson, B. D. Erath, C. Castro, R. E. Hillman, and M. Zañartu, "Modeling the pathophysiology of phonotraumatic vocal hyperfunction with a triangular glottal model of the vocal folds," *Journal of Speech, Language, and Hearing Research*, vol. 60, pp. 2452–2471, sep 2017.

[7] C. Gobl, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, pp. 189–212, apr 2003.

[8] A. Stuart and J. Kalinowski, "Effect of delayed auditory feedback, speech rate, and sex on speech production," *Percept Mot Skills*, vol. 120, pp. 747–765, jun 2015.

[9] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality," *Journal of Speech, Language, and Hearing Research*, vol. 36, pp. 21–40, Feb. 1993.

[10] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," *American Journal of Speech-Language Pathology*, vol. 18, pp. 124–132, May 2009.

[11] M. Hirano, *Clinical examination of voice.* Wien New York: Springer-Verlag, 1981.

[12] J. Kreiman, D. Vanlancker-sidtis, and B. Gerratt, "Kreiman et al. defining and measuring voice quality defining and measuring voice quality."

[13] M. MORISE, F. YOKOMORI, and K. OZAWA, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. & Syst. Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, apr 1999.

[15] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, "Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation," *Acoust. Sci. & Tech.*, vol. 28, no. 3, pp. 140–146, 2007.

[16] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, mar 2008.

[17] M. Morise, "World - a high-quality speech analysis, manipulation and synthesis system." `https://github.com/mmorise/World`, 2021.

[18] M. Morise, "D4c, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, nov 2016.

[19] J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer, and N. Umeda, "Synthetic voices for computers," *IEEE Spectr. Spectrum*, vol. 7, pp. 22–45, oct 1970.

[20] J. Laver, *The Phonetic Description of Voice Quality.* Cambridge Studies in Linguistics, Cambridge University Press, 1980.

[21] G. Degottex, A. Roebel, and X. Rodet, "Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter," in

2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, may 2011.

[22] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *FNT in Signal Processing in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.

[23] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1–7, mar 2015.

[24] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *International Conference on Digital Audio Effects*, pp. 30–35, 2005.

[25] I. Titze, *Principles of voice production*. Englewood Cliffs, N.J: Prentice Hall, 1994.

[26] O. Perrotin and I. V. McLoughlin, "On the use of a spectral glottal model for the source-filter separation of speech," *arXiv preprint arXiv:1712.08034*, 2017.

[27] A. McCree and T. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Process. Transactions on Speech and Audio Processing*, vol. 3, pp. 242–250, jul 1995.

[28] M. Airaksinen, T. Bäckström, and P. Alku, "Automatic estimation of the lip radiation effect in glottal inverse filtering," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[29] G. Fant, "The lf-model revisited. transformations and frequency domain analysis," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, vol. 2, no. 3, p. 40, 1995.

[30] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, pp. 278–294, feb 2013.

[31] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, pp. 109–118, jun 1992.

[32] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, mar 2008.

[33] J. R. Lechien, S. Blecic, Y. Ghosez, K. Huet, B. Harmegnies, and S. Saussez, "Voice quality and orofacial strength as outcome of levodopa effectiveness in patients with early idiopathic parkinson disease: A preliminary report," *Journal of Voice*, vol. 33, pp. 716–720, sep 2019.

[34] E. S. Segundo, P. Foulkes, P. French, P. Harrison, V. Hughes, and C. Kavanagh, "The use of the vocal profile analysis for speaker characterization: Methodological proposals," *Journal of the International Phonetic Association*, vol. 49, pp. 353–380, jun 2018.

[35] W. H. Perkins, "Vocal function: Assessment and therapy," *Handbook of speech pathology and audiology*, vol. 1, pp. 481–503, 1971.

[36] C. J. Bassich and C. L. Ludlow, "The use of perceptual methods by new clinicians for assessing voice quality," *J Speech Hear Disord*, vol. 51, pp. 125–133, may 1986.

[37] J. Kreiman and B. R. Gerratt, "Perceptual assessment of voice quality: Past, present, and future," *Perspectives on Voice and Voice Disorders*, vol. 20, pp. 62–67, July 2010.

[38] D. Wilson, *Voice Problems of Children*. Williams & Wilkins, 1972.

[39] E. S. Segundo and J. A. Mompean, "A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity," *Journal of Voice*, vol. 31, pp. 644.e11–644.e27, sep 2017.

[40] A. S.-L.-H. Association *et al.*, "Consensus auditory–perceptual evaluation of voice (cape-v): Asha special interest division 3, voice and voice disorders," *Rockville, MD: Author. Retrieved from https://www.asha.org/siteassets/uploadedfiles/ASHA/SIG/03/CAPE-V-Procedures-and-Form.pdf*, 2002.

[41] K. Nemr, M. Simões-Zenari, G. F. Cordeiro, D. Tsuji, A. I. Ogawa, M. T. Ubrig, and M. H. M. Menezes, "GRBAS and cape-v scales: High reliability and consensus when applied at different times," *Journal of Voice*, vol. 26, pp. 812.e17–812.e22, nov 2012.

[42] M. Hirano, *Clinical Examination of Voice*. Disorders of human communication, Springer-Verlag, 1981.

[43] T. L. Eadie and C. R. Baylor, "The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice," *Journal of Voice*, vol. 20, pp. 527–544, dec 2006.

[44] R. Veldhuis, "A computationally efficient alternative for the liljencrants–fant model and its perceptual evaluation," *The Journal of the Acoustical Society of America*, vol. 103, pp. 566–571, jan 1998.

[45] G. Fant, J. Liljencrants, and Q.-g. Lin, "A four-parameter model of glottal flow," 1985.

[46] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta acustica united with acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[47] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Institute of Electrical and Electronics Engineers.

[48] A. Murphy, I. Yanushevskaya, A. N. Chasaide, and C. Gobl, "The role of voice quality in the perception of prominence in synthetic speech," in *Interspeech 2019*, ISCA, sep 2019.

[49] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *The Journal of the Acoustical Society of America*, vol. 49, pp. 583–590, feb 1971.

[50] G. Degottex, *Glottal source and vocal-tract separation*. Theses, Université Pierre et Marie Curie - Paris VI, Nov. 2010.

[51] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America*, vol. 87, pp. 820–857, feb 1990.

[52] B. Doval and C. d'Alessandro, "Spectral correlates of glottal waveform models: an analytic study," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE Comput. Soc. Press.

[53] M. Schroeder and B. Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates," in *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Institute of Electrical and Electronics Engineers.

[54] and K. Vos and T. Terriberry, "Definition of the opus audio codec," tech. rep., sep 2012.

[55] S. Cai, M. Boucek, S. S. Ghosh, F. H. Guenther, and J. S. Perkell, "A system for online dynamic perturbation of formant trajectories & results from perturbations of the mandarin triphthong iau," in *In Proceedings of the 8th ISSP*, pp. 65–68, 2008.

[56] J. A. Tourville, S. Cai, and F. Guenther, "Exploring auditory-motor interactions in normal and disordered speech," ASA, 2013.

[57] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS ONE ONE*, vol. 8, p. e60603, apr 2013.

[58] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *ICASSP*, 2015.

[59] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016.

[60] W. B. Kleijn, F. S. C. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "Wavenet based low rate speech coding," 2017.

[61] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4211–4215, 2020.

[62] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, apr 1999.

[63] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*, Feb 2009.

[64] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Interspeech 2017*, ISCA, aug 2017.

[65] R. Daido and Y. Hisaminato, "A fast and accurate fundamental frequency estimator using recursive moving average filters," in *Interspeech 2016*, pp. 2160–2164, 2016.

[66] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1–7, mar 2015.

[67] H. Järveläinen and M. Karjalainen, "Reverberation modeling using velvet noise," in *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*, Audio Engineering Society, 2007.

[68] H. Kawahara, K.-I. Sakakibara, M. Morise, H. Banno, T. Toda, and T. Irino, "Frequency domain variants of velvet noise and their application to speech processing and synthesis," in *Proc. Interspeech 2018*, pp. 2027–2031, 2018.

[69] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP: A collaborative voice analysis repository for speech technologies," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2014.

[70] P. R. Walden, "Perceptual voice qualities database (pvqd)," 2020.

[71] S. A. Memon, "Acoustic correlates of the voice qualifiers: A survey," 2020.

[72] X. Na, "opensmile." `https://github.com/naxingyu/opensmile`, 2019.

[73] M. Wang, "python-pesq." `https://github.com/ludlows/python-pesq`, 2021.

[74] J. Kreiman, B. R. Gerratt, and G. S. Berke, "The multidimensional nature of pathologic vocal quality," *The Journal of the Acoustical Society of America*, vol. 96, pp. 1291–1302, Sept. 1994.

[75] J. Kreiman, D. Vanlancker-sidtis, and B. Gerratt, "Defining and measuring voice quality," 2003.

[76] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control*, vol. 14, pp. 42–54, Nov. 2014.

[77] Y. D. Heman-Ackah, D. D. Michael, and G. S. Goding, "The relationship between cepstral peak prominence and selected parameters of dysphonia," *Journal of Voice*, vol. 16, pp. 20–27, Mar. 2002.

[78] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform."

[79] J. Kreiman, B. R. Gerratt, M. Garellek, R. Samlan, and Z. Zhang, "Toward a unified theory of voice production and perception," *Loquens*, vol. 1, p. e009, June 2014.

[80] Mozilla, "Tts: Text-to-speech for all." `https://github.com/mozilla/TTS`, 2021.

[81] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer modeling timbre and expression from natural songs," *Applied Sciences*, vol. 7, p. 1313, Dec. 2017.