UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
Departamento de Electrónica

# Thesis Proposal

*Thesis Proposals presented by*
***Benjamín Antonio Opazo Campusano***,
*as a partial requirement to qualify for the*
*Master of Sciences in Electrical and Electronics Engineering,*
*Telecommunications and Signal Processing Degree.*

**Thesis Director: Matías Zañartu**

**ROL USM**
201321022-0

**RUT**
18.782.330-7

**e-mail**
benjamin.opazo.13@sansano.usm.cl

July 25, 2019

# Contents

# 1 Definition of thesis topic

## 1.1 Title

Real-Time voice quality Modification

## 1.2 Abstract

Auditory feedback is an important aspect of speech production, affecting the way a person speaks. Modifying auditory feedback can help researchers to better understand how speech production works. Another aspect associated with speech production is voice quality. 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. The main focus of this research, is to develop a voice synthesizer that focuses on modifying in real-time different aspects of voice quality, thus, generating a modified auditory feedback of a subject. In order to accomplish this task, a set of algorithms will be designed and implemented on both high and low level language, modifying auditory feedback with an expected latency of less than 20 ms.

# 2 Introduction

Auditory feedback plays an important role in speech production [6]. For example, the lack of auditory feedback is associated with deterioration in speech production over time [26], and a Delayed Auditory Feedback (DAF) of 200 ms may induce increased speech errors and decreased speech rate [4], while in stuttering patients DAF can help with speech fluency [25]. By modifying the auditory feedback, researchers can further understand the underlying mechanisms of speech production and feedback effects. On the same subject, voice resynthesis can be used as a tool to modify and measure different parameters of auditory feedback, like voice quality.

Voice quality is an important aspect of voice perception, which refers to the different properties of speech that 'gives the primary distinction to a given speaker's voice when pitch and loudness are excluded' according to the US National Library of Medicine. Patients seek treatment for voice disorders because they do not sound normal. Furthermore, physicians classify the voice quality of patients in order to assess voice disorders [15].

On the other hand, pathological labels of voice quality like "*breathiness*", "*roughness*", "*hoarseness*" and related qualities have never received widely accepted definitions in the clinical literature [14]. There have been efforts to standardize the different voice quality labels and definitions, like CAPE-V [12], although this standard focuses on the perceptual evaluation of voice, and lacks and objective measurement and definition of the voice quality. In [3], three of the six pathological voice quality labels defined in CAPE-V (*breathy* phonation, *strained* phonation and *rough* phonation) were given a more precise and objective definition. *Pitch*, as defined by the American National Standards Institute is "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from high to low", and according to [9], it cannot be expressed in physical units or measured by physical means. *Loudness* is defined of term of its spectral parameters in [5]. 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 [31][7]. There is a need to develop engineering tools for this purpose, particularly those associated with transforming voice quality in real-time for perceptual studies.

# 3 Formulation of the problem and evaluation of the theoretical and technical framework

## 3.1 Definition of the problem

Defining voice quality, and its associated pathological labels, is a task that has not been fully solved. There are some efforts in the literature to solve this problem [16][14][3][12], but almost all of them fail to give a quantitative definition for the different voice quality labels. Furthermore, it is not currently possible to modify in real-time different aspects of the voice quality of a person in order to measure the way the subject responds to different alterations of its own voice. voice quality is also a key issue in speech synthesis research [5].

A machine capable of modifying different voice quality labels in real-time can solve the previously exposed problems. Such machine should be able to choose which voice labels to modify and to what extent, using previously chosen parameters. The method the machine uses to modify the voice quality labels could be either by resynthesizing the voice or by applying a Glottal Inverse Filtered algorithms.

## 3.2 Solutions and approaches made by other authors

At the moment of writing this proposal, no software could be found that can modify multiple aspects of voice quality in real-time, however, there are different software and papers that synthesize or modify voice in real-time or with delays.

### 3.2.1 Voice Synthesis

As a simplified model, speech is generated in different ways depending on the sound that it produces. For example, voiced sounds, like vocals or nasal sounds (like $n$) is produced by the excitation of the tract with the vocal folds, which is making quasi-periodic pulses of air pressure [23]. Plosives (like $p$ or $t$) are generated by closing the vocal tract at some point above the focal folds [22]. There are many different sounds that can be generated and each sound has an associated movement on the vocal tract, mouth, tongue or a combination of them.

**Source-filter theory** As a way to generalize voice synthesis [24], source-filter theory models the voice in terms of the z-transform as

$$S(z) = E(z)G(z)V(z)R(z) \tag{1}$$

Where $S(z)$ represents the speech segment, $E(z)$ represents the impulse train on the vocal folds, $G(z)$ represents the glottal waveform, $V(z)$ represents the vocal tract impulse response, and $R(z)$ is the acoustic impedance at the lips. Different zeros and poles can generate different sounds and emulate the voice of a person. Using techniques like Linear Predictive Coding (LPC), it is possible to obtain the spectral envelope of the voice, and thus, some, but not all, parameters of the voice.

**Klatt Synthesizer** The Klatt Synthesizer is a speech synthesis software designed by Dennis Klatt in 1980 [1]. It is a software for cascade/parallel formant synthesizer [13] consisting in 60 parameters that can be modified to synthesize voice. It is currently widely used in the academia [19][21][27].

**Audapter** Audapter is a software package developed by Guenther Lab at Boston University [17]. It can modify in real-time acoustic parameters of voice and it runs on desktop and laptop computers. The acoustic parameters that it can modify are formant frequencies, fundamental frequency (pitch), local timing, local intensity, global time delay (DAF) and global intensity.

Although some of the previously mentioned parameters can be considered as voice quality labels, it does not modify all of the commonly known pathological labels of voice quality like *breathiness*. The method used to modify voice quality labels is through voice resynthesis. The code is free to access which means that it is a good starting point for real-time voice modification.

**VocalTractLab** VocalTractLab is an articulatory speech synthesizer packaged in an interactive multimedia software tool [2]. It focuses on articulation, acoustics and control, and it is meant to be used by students of phonetics and related disciplines. The software simulates the human vocal tract with a three-dimensional model.

VocalTractLab uses a technique called Articulatory Speech Synthesis, where the voice is synthesized as a direct simulation of the principles of speech production, that is, 'by creating a synthetic model of human physiology and making it speak' [22]. An articulatory synthesizer must have at least a geometric description of the vocal tract, a mechanism to control the parameters during an utterance and a model for acoustic generation.

This software lacks the real-time voice synthesis feature, and it focuses on generating voice rather than modifying a preexisting voice, as it can be a difficult task to correctly obtain the necessary parameters of a subject's voice in order to generate voice that sounds like the subject under test.

**Machine Learning Assisted Synthesis** Machine Learning assisted synthesis is a promising topic of research for realistic voice synthesis, for example, in [8] Generative Adversarial Networks (GAN) are used in order to train a machine to impersonate a target speaker, that is, mimicking the pitch and other qualities like the *style* of the target.

The problem of this approach for the objective of this research, is that the machine needs to be trained with a big enough dataset in order to improve, and the internal parameters are not accessible to the final user, that means that the modification of voice quality becomes a non achievable task.

### 3.2.2 Voice Quality Modification

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 [31][7]. There is a need to develop engineering tools for this purpose, particularly those associated with transforming voice quality in real-time for perceptual studies. In order to modify voice quality it is necessary to define what is voice quality, and which aspects of voice quality will be modified. there are two consensus of voice quality that will be used in this research, CAPE-V and GRBAS, the former being a North-American consensus, and the latter a Japanese-European consensus.

#### Voice Quality Consensus

**CAPE-V** Is a tool made for perceptual evaluation of voice, used for clinical assessment of voice. It was developed in 2002 and defines 5 pathological labels of voice quality [12]: *breathy* phonation, *strained* phonation, *rough* phonation, *pitch* and *loudness*. It lacks an objective definitions of the pathological labels of voice quality, so further research will be needed to model these labels.

**GRBAS** Is a scale that measures *grade*, *roughness*, *breathiness*, *asthenia* and *strain* (thus making the acronym GRBAS). Like CAPE-V, the scale relies on highly trained personnel due to its subjectivity [10].

Both tools define a subjective measurement of voice quality that needs highly trained personnel. *This brings two problems*, the first problem, is that it is needed an *objective definition* of

the different pathological labels of voice quality in order to design algorithms that can modify this labels. The second problem that arises is *the need for an objective measurement* of the voice modification.

It is important to note that both scales share similar labels for the different pathological labels of voice quality, and literature suggest that both are scales are reliable and there is some agreement between scales if they are used by trained personnel [20][11].

**Algorithm design** There has been some effort to objectively standardize pathological labels of voice quality, for example, in [3], there are equations defining *breathy* phonation, *strained* phonation and *rough* phonation. In order to design the necessary algorithms to modify voice quality, more research in this topic is needed. This is the first tasks that needs to be solved after choosing which pathological labels of voice quality will be modified.

**Objective measurements** Although CAPE-V and GRBAS are subjective tools, some efforts to objectively measure voice quality has been done. Recently, the objective measures that have been developed, make use of Machine Learning technologies. In the scope of this work, this is not a problem, because the objective measurement will be used to validate the output signal, and does not need to be in real-time.

An example of an objective measurement consist in a Multidimensional Acoustic Analysis [28], that is, using different measurements of the voice to assess the voice quality, and then using different classification algorithms like Support Vector Machine (SVM) or Extreme Learning Machine (ELM), to determine the voice quality. The problem of this method, is that it can determine with high accuracy if a voice is pathological or not, but cannot reveal more subtle aspects of pathological voices, meaning that for the scope of this research, with this method it will not be possible to assess to what extent the voice quality is being modified.

Another machine learning based method is used by [30], where they carefully curate their dataset in order to be consistent. Their results, that is, the objective measurements made by the machine, are reasonably consistent.

In this research an objective measurement is needed, thus, the previously discussed method will be used. The issue that arises by this decision is that the code is not publicly available, but an attempt will be made to contact the researchers and ask for the code. The second problem is the lack of data to compare the results, and that fine-tuning will be made heuristically.

# 4 Design and research method

## 4.1 Hypothesis

It is possible to develop a tool to modify in real-time the voice quality of a person, using a dedicated audio DSP board with an audio latency of less than 20 ms, and using Digital Signal Processing algorithms to modify the auditory feedback of a subject by using the proposed approach. The modified voice quality output is consistent with the respective pathological label from the CAPE-V or GRBAS standards.

## 4.2 Objectives

- General Objective
  - Disturb the auditory feedback of a subject with real-time modification of voice quality by introducing pathological labels in it.

- Specific Objectives
  - Design a set of algorithms that can be implemented in a high-level language to modify different pathological labels of voice quality of a prerecorded voice.
  - Adapt the selected algorithms to a low-level language to modify different pathological labels of voice quality in real-time, using a dedicated audio DSP board.

### 4.3 Methodology

In order to design a set of algorithms that modify different pathological labels of voice quality, it is necessary to:

- *Establish which pathological labels of voice quality will be modified* by reading the current literature on the subject of voice quality and assess which pathological labels are more important. Some pathological labels of voice quality may have more incidence in speech production than others, or may be more prevalent than others. It is important to also *determine which criteria will be used to select the pathological labels that will be modified.*
- *Model the chosen pathological labels of voice quality* either by using preexisting models found in the literature or in specialized open source software.
- *Establish whether it is convenient to resynthesize the voice or to apply Glottal Inverse Filtered (GIF) algorithms.* At the moment of writing this proposal and without knowing the models that will be used, it is unclear which method will prove to be better, more reliable and with less time delay. This choice must be made once the models are chosen, and in parallel to the algorithm design.
- *Design the algorithms that will be used to modify the pathological label of voice quality.* The algorithm design will be focused on translating the previously modeled pathological labels to a mathematical language, taking into account the chosen method to modify the voice, that is, voice resynthesis or GIF.

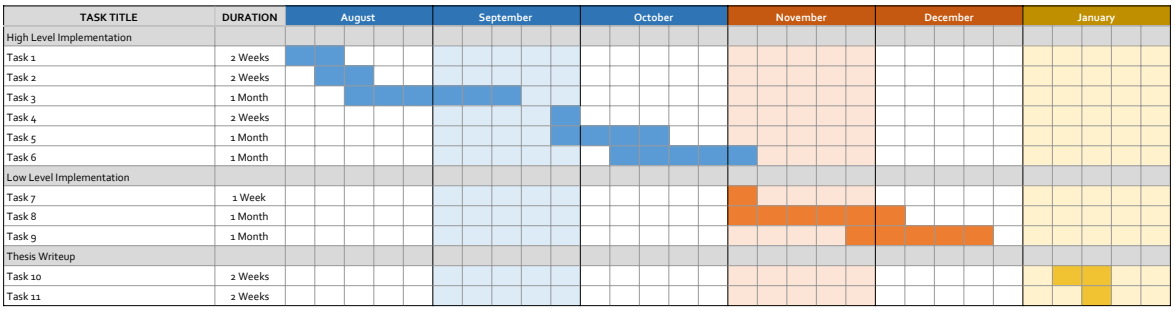
After selecting the proper framework and designing the algorithms that will be used, *the algorithms will be implemented in a high-level language such as MatLab or Python* as a proof of concept, and to fine tune the parameters of the algorithm. This step is recommended when implementing low level design [29].

Once the algorithms are implemented and tested in a high-level language, the work will be focused on implementing the algorithms in a low-level language, to do that, it is necessary to:

- *Choose the proper hardware that fits the needs given the designed algorithms and the restrictions that it imposes*, that is, sufficiently low audio latency in both input delay and processing delay, good ADC and DAC, and other needs that will appear during the models and algorithm research, and the human perception of voice. A good candidate is the TI 66AK2G evaluation module.
- *Adapt and implement the algorithms in a low-level language on the chosen hardware.* On a first approach, the only restriction is to implement the designed algorithms on the hardware, focusing on the viability of the implementations and not on the restrictions given by timing.

Once the algorithms are implemented in a low-level language, then the task is to *improve the latency up to the required levels given by human perception* in order to modify voice quality in real-time.

## 4.4 Work plan




Task 1: Establish pathological labels that will be modified
Task 2: Choose the criteria to select pathological labels
Task 3: Model the pathological labels of voice quality
Task 4: Choose between GIF or voice resynthesis
Task 5: Design the algorithms
Task 6: High level implementations of the algorithms
Task 7: Choose the hardware that fits the needs
Task 8: Adapt and implement the algorithms in a low-level language
Task 9: Improve the latency
Task 10: Draft writing
Task 11: Final write up

## 4.5 Expected results

It is expected to design a high level implementation of a set of algorithm that modify different pathological voice quality labels. The voice modification is parametric. It is also expected to design a real-time solution to reliably modify different voice quality labels, using specialized hardware, such as a dedicated audio DSP, or a more suited hardware. Finally, it is expected that this work will be published in a conference paper, and that it will be useful in future work in this field, for example, in modeling auditory feedback.

# References


[1] UC Berkeley Berkeley Linguistics. Klatt Synthesizer, 2017.

[2] Peter Birkholz. VocalTractLab towards high-quality articulatory speech synthesis, 2017.

[3] Michal Borsky, Daryush D. Mehta, Jarrad H. Van Stan, and Jon Gudnason. Modal and nonmodal voice quality classification using acoustic and electroglottographic features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2281–2291, dec 2017.

[4] Jennifer Chesters, Ladan Baghai-Ravary, and Riikka Möttönen. The effects of delayed auditory and visual feedback on speech production. *The Journal of the Acoustical Society of America*, 137(2):873–883, feb 2015.

[5] Cristophe d'Alessandro and Boris Doval. Experiments in voice quality modification of natural speech signals: The spectral approach. *SSW3-1998*, pages 277–282, nov 1998.

[6] Jeffrey L. Elman. Effects of frequency-shifted feedback on the pitch of vocal productions. *The Journal of the Acoustical Society of America*, 70(1):45–50, jul 1981.

[7] Gabriel E. Galindo, Sean D. Peterson, Byron D. Erath, Christian Castro, Robert E. Hillman, and Matías Zañartu. Modeling the pathophysiology of phonotraumatic vocal hyperfunction with a triangular glottal model of the vocal folds. *Journal of Speech, Language, and Hearing Research*, 60(9):2452–2471, sep 2017.

[8] Yang Gao, Rita Singh, and Bhiksha Raj. Voice impersonation using generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2018.

[9] Adrianus J.M. Houtsma. Pitch perception. In *Hearing*, pages 267–295. Elsevier, 1995.

[10] Farideh Jalalinajafabadi. Automatic assessment of voice signals according to the grbas scale, 03 2014.

[11] Michael P. Karnell, Sarah D. Melton, Jana M. Childes, Todd C. Coleman, Scott A. Dailey, and Henry T. Hoffman. Reliability of clinician-based (GRBAS and CAPE-v) and patient-based (v-RQOL and IPVI) documentation of voice disorders. *Journal of Voice*, 21(5):576–590, sep 2007.

[12] Gail B. Kempster, Bruce R. Gerratt, Katherine Verdolini Abbott, Julie Barkmeier-Kraemer, and Robert E. Hillman. Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18(2):124–132, May 2009.

[13] Dennis H. Klatt. Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, 67(3):971–995, mar 1980.

[14] Jody Kreiman, Bruce R. Gerratt, and Gerald S. Berke. The multidimensional nature of pathologic vocal quality. *The Journal of the Acoustical Society of America*, 96(3):1291–1302, September 1994.

[15] Jody Kreiman, Bruce R. Gerratt, Gail B. Kempster, Andrew Erman, and Gerald S. Berke. Perceptual evaluation of voice quality. *Journal of Speech, Language, and Hearing Research*, 36(1):21–40, February 1993.

[16] Jody Kreiman, Diana Vanlancker-sidtis, and Bruce Gerratt. Defining and measuring voice quality, 2003.

[17] Guenther Lab. Audapter, 2016.

[18] Harlan Lane and Bernard Tranel. The lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, 14(4):677–709, dec 1971.

[19] Gabriella Musacchia, Silvia Ortiz-Mantilla, Cynthia P. Roesler, Sree Rajendran, Julie Morgan-Byrne, and April A. Benasich. Effects of noise and age on the infant brainstem response to speech. *Clinical Neurophysiology*, 129(12):2623–2634, dec 2018.

[20] Katia Nemr, Marcia Simões-Zenari, Gislaine Ferro Cordeiro, Domingos Tsuji, Allex Itar Ogawa, Maysa Tibério Ubrig, and Márcia Helena Moreira Menezes. GRBAS and cape-v scales: High reliability and consensus when applied at different times. *Journal of Voice*, 26(6):812.e17–812.e22, nov 2012.

[21] Alessandro Presacco, Jonathan Z. Simon, and Samira Anderson. Evidence of degraded representation of speech in noise, in the aging midbrain and cortex. *Journal of Neurophysiology*, 116(5):2346–2355, nov 2016.

[22] Du Qinsheng, Zhao Jian, Wang Lirong, and Shi Lijuan. Articulatory speech synthesis: A survey. In *2011 14th IEEE International Conference on Computational Science and Engineering*. IEEE, aug 2011.

[23] Lawrence R. Rabiner. *An Introduction to Digital Speech Processing (Foundations and Trends in Signal Processing,)*. Now Publishers Inc, nov 2007.

[24] J.C. Rutledge, K.E. Cummings, D.A. Lambert, and M.A. Clements. Synthesizing styled speech using the klatt synthesizer. In *1995 International Conference on Acoustics, Speech, and Signal Processing*. IEEE.

[25] George A. Soderberg. Delayed auditory feedback and stuttering. *Journal of Speech and Hearing Disorders*, 33(3):260–267, aug 1968.

[26] Riki Taitelbaum-Swead, Michal Avivi, Batel Gueta, and Leah 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*, 33(7):628–640, jan 2019.

[27] Shunsuke Tamura, Miduki Mori, Kazuhito Ito, Nobuyuki Hirose, and Shuji Mori. Study on interactions between voicing production and perception using auditory feedback paradigm. Acoustical Society of America, 2017.

[28] Zhijian Wang, Ping Yu, Nan Yan, Lan Wang, and Manwa L. Ng. Automatic assessment of pathological voice quality using multidimensional acoustic analysis based on the GRBAS scale. *Journal of Signal Processing Systems*, 82(2):241–251, jun 2015.

[29] Greg Wilson, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, Kathryn D. Huff, Ian M. Mitchell, Mark D. Plumbley, Ben Waugh, Ethan P. White, and Paul Wilson. Best practices for scientific computing. *PLoS Biology*, 12(1):e1001745, jan 2014.

[30] Zheng Xie, Chaitanya Gadepalli, Farideh Jalalinajafabadi, Barry M G Cheetham, and Jarrod J Homer. Machine Learning Applied to GRBAS Voice Quality Assessment. *Advances in Science, Technology and Engineering Systems Journal*, 3(6):329–338, 2018.

[31] Matías Zañartu, Gabriel E. Galindo, Byron D. Erath, Sean D. Peterson, George R. Wodicka, and Robert E. Hillman. Modeling the effects of a posterior glottal opening on vocal fold dynamics with implications for vocal hyperfunction. *The Journal of the Acoustical Society of America*, 136(6):3262–3271, dec 2014.

# 5 Signatures of the Student and the Thesis Director

Benjamín Opazo Campusano Matías Zañartu