

Singing Voice Synthesis for Real-Time Applications

Master Thesis

Submitted by:

William Rodewald



Primary Supervisor:

Prof. Dr. Stefan Weinzierl

Secondary Supervisor:

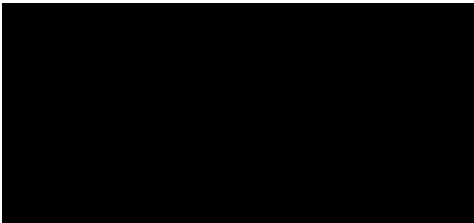
Dr. Henrik Hahn
Ableton AG

A thesis submitted to Technische Universität Berlin in partial fulfilment of the requirements for the degree Master of Science in *Audiocommunication and -technology*.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, den 28. Februar 2021



Abstract

Recently, many contributions in the field of singing voice analysis and synthesis utilized the advancements in machine learning to propose powerful models capable of synthesizing fully phonetic singing voice samples. However, while these advancements in machine learning can be used to solve various complex problems, they have yet to reach the robustness and computational efficiency of traditional signal processing methods. This thesis tries to utilize the advantages of both domains and proposes a singing voice analysis and synthesis method combining traditional signal processing with machine learning and optimization.

The proposed method is based on a simple but robust voice synthesis method which models the glottal source using an LF- R_d wavetable oscillator and the vocal tract with a pole-zero filter bank. To control said synthesis method, a neural network is trained to predict the required synthesis parameters from a freely controllable pitch trajectory. Thus, central to the analysis is the estimation of said parameters from a source sample so that the network can be trained in a supervised context. This is achieved with a novel synthesis model parameter estimation method which implements the synthesis model as part of a gradient descent optimizer, tasked to minimize the reconstruction error between the source sample and its resynthesis.

The proposed singing voice analysis and synthesis method was evaluated in an empirical study consisting of two experiments. In the first experiment, participants were asked to rate the perceived audio quality between a reference stimuli and multiple alterations thereof reconstructed using the different subcomponents of the proposed method. Both a Friedman test with post hoc Wilcoxon signed rank test and a linear mixed model confirmed that the synthesis model parameter estimation \mathcal{E} produces the largest drop in audio quality, indicating that the method still leaves room for improvements. Additionally, the mixed model suggests that the perceived audio quality of the synthesized sample depends on the pitch range present in the source audio sample, the singers gender and the vowel. In a second experiment, participants were asked to rate the relative presence of four timbre qualities, namely naturalness, breathiness, brightness and roughness, between a reference stimuli and its reconstruction. A Wilcoxon signed rank test confirmed that participants perceived the reference stimuli to sound significantly more natural, breathier and brighter while a weaker but still statistically significant effect was found for roughness.

Zusammenfassung

In den letzten Jahren haben viele wissenschaftliche Veröffentlichungen in dem Gebiet der Sprachsynthese die Fortschritte im Bereich des maschinellen Lernens genutzt, um leistungsstarke Modelle zu entwickeln, die in der Lage sind, vollständig phonetischen Gesang zu synthetisieren. Während Methoden des maschinellen Lernens zur Lösung verschiedener komplexer Probleme verwendet werden können, erreichen sie häufig noch nicht die Robustheit und Effizienz der traditionellen Signalverarbeitung. Aus diesem Grund untersucht diese Arbeit eine Methode zur Analyse und Synthese gesungener Vokale, die die traditionelle Signalverarbeitung mit maschinellem Lernen kombiniert.

Die vorgeschlagene Methode basiert auf einer einfachen, aber robusten Methode für die Synthese von Vokalen, die die in den Stimmenbändern entstehende Grundschiwingung mit einem Wavetable-Oszillator und den Vokaltrakt mit einer seriellen Filterbank modelliert. Zur Steuerung dieser Synthesemethode wird ein neuronales Netz trainiert, das aus einer frei steuerbaren Tonhöhenkurve die erforderlichen Syntheseparameter bestimmt. Im Mittelpunkt der Analyse steht die Schätzung dieser Parameter aus einem Audiosample. Dazu wird das Synthesemodell mittels automatischer Differenzierung als Teil einer gradientenbasierten Optimierung implementiert, dessen Ziel es ist, den Rekonstruktionsfehler zwischen dem Audiosample und seiner Resynthese zu minimieren.

Die vorgeschlagene Methode zur Analyse und Synthese gesungener Vokale wurde in einer empirischen Studie evaluiert. In einem ersten Experiment wurden die Teilnehmer gebeten, die wahrgenommene Audioqualität zwischen einem Referenzstimulus und mehreren mit der Methode rekonstruierten Varianten zu bewerten. Die Auswertung bestätigte, dass die Parameterschätzung des Synthesemodells für den größten Verlust an Audioqualität verantwortlich ist. Zudem legt die Auswertung nahe, dass die wahrgenommene Audioqualität des synthetisierten Samples vom Tonhöhenbereich, dem Geschlecht des Sängers und dem Vokal abhängt. In einem zweiten Experiment wurden die Teilnehmer gebeten, die Präsenz von vier Klangfarbenqualitäten, nämlich *Naturalness*, *Breathiness*, *Brightness* und *Roughness*, zwischen einem Referenzstimulus und seiner Rekonstruktion zu bewerten. Die Auswertung bestätigte, dass die Teilnehmer die Referenzstimuli als signifikant natürlicher, atmungsreicher, heller und, in einem geringeren Maße, auch rauer empfanden.

Acknowledgements

This thesis would not have been possible without the substantial support of my colleagues, friends and family members. First and foremost, I would like to thank Dr. Henrik Hahn for his guidance and unending patience. I would like to thank Dr. phil. Steffen Lepa for his advice on designing and evaluating an empirical study for this thesis.

I am grateful to my colleagues at Ableton for giving me the opportunity to work on this thesis for as long as it took. Furthermore, I would like to thank my former colleagues at u-he for giving me a place to gather valuable experience throughout my years as a Master's student.

Finally, I would like to thank my friends and family for their support. Particular thanks go to Maximilian Weber and Gabriel Dernbach for their help and for the numerous inspirational discussions we had about this work.

Table of Contents

List of Figures	VIII
List of Tables	X
List of Symbols	XI
List of Abbreviations	XII
1 Introduction and Motivation	1
1.1 Research Goal and Questions	2
1.1.1 Thesis Scope	3
1.1.2 Application Cases	3
1.2 Structure of the Thesis	4
2 Fundamentals	5
2.1 Physiology and Source-Filter Theory	5
2.1.1 Glottal Flow	6
2.1.2 Vocal Tract	7
2.1.3 Lip and Nostril Radiation	8
2.1.4 Source-Filter Theory	8
2.2 Synthesis Models	9
2.2.1 Source Models	10
2.2.2 Filter Models	12
2.3 Source-Filter Analysis	15
2.3.1 Spectral Envelope Estimation	16
2.3.2 Iterative Adaptive Inverse Filtering	17
2.3.3 Analysis-by-Synthesis	19
2.3.4 Minimum Phase Assumption	20
2.4 Machine Learning in Voice Synthesis	20
2.4.1 WaveNet	20
2.4.2 Differentiable Digital Signal Processing (DDSP)	21
2.5 Automatic Differentiation	22
2.6 B-Spline Modeling	24
3 Method	25
3.1 Overview	25

3.2 Synthesis Model \mathcal{S}	28
3.2.1 Source Model $S(z)$	29
3.2.2 Vocal Tract Model $V(z)$	31
3.2.3 Summary	32
3.3 Harmonic Analysis \mathcal{A}	34
3.3.1 Partial Extraction and Harmonic Reconstruction	35
3.3.2 Joint Optimization	36
3.3.3 Summary	37
3.4 Synthesis Model Parameter Estimation \mathcal{E}	39
3.4.1 Method Overview	40
3.4.2 Parameter Preprocessing	41
3.4.3 Synthesis	43
3.4.4 Optimization Loss	44
3.4.5 Optimization and Implementation	47
3.4.6 Summary	47
3.5 Parameter Predictor Network \mathcal{P}	49
3.5.1 Network	49
3.5.2 Loss	50
3.5.3 Implementation and Training	51
3.5.4 Summary	51
4 Evaluation	53
4.1 Dataset and Stimuli Generation	54
4.2 Experiment 1	57
4.2.1 Method	57
4.2.2 Results	59
4.2.3 Discussion	64
4.3 Experiment 2	66
4.3.1 Method	67
4.3.2 Results	69
4.3.3 Discussion	70
5 Conclusion	71
Literature	75
Appendix A Mixed Model Post Hoc Test Results	83

List of Figures

1.1	A simplified method overview.	2
1.2	The relation between vocal score and performance.	4
2.1	The human voice production system.	5
2.2	The glottal oscillation	6
2.3	The vocal tract tube model	7
2.4	Vocal tract tubes and spectra for three vowels	7
2.5	Parameters of the LF-model	10
2.6	The LF glottal pulse and derivative for three parameter sets	11
2.7	The LF- R_d glottal pulse derivative and spectrum for three R_d values . . .	12
2.8	The vocal tract Kelly-Lochbaum model	14
2.9	Two spectral envelopes estimated with LPC and True-Envelope	16
2.10	An overview of IAIF methods	18
2.11	The structure of the WaveNet neural network	20
2.12	The structure of the DDSP neural network	21
2.13	Two 3rd order B-spline curves	24
3.2	A detailed overview of the proposed method	26
3.3	The synthesis model	28
3.4	The source synthesis	29
3.5	The wavetable used during synthesis.	30
3.6	The decibel magnitude response of an exemplary vocal tract filter	31
3.7	An overview of the harmonic analysis method	35
3.8	Magnitude spectra of harmonic signals with and without pitch change . .	36
3.9	Pitch and pitch slope estimates	37
3.10	Frequency and magnitude of estimated harmonic partials	38
3.11	The structure of the synthesis model parameter estimation method	40
3.12	Parameter upsampling and scaling	41
3.13	The parameter preprocessing used in the proposed synthesis model param- eter estimation method	41
3.14	Filter coefficients for the temporal regularization loss filter	45
3.15	Comparison between low and high fluctuation of parameter trajectories .	46
3.16	Spectrum of a source sample and reconstruction after parameter estimation	47
3.17	The structure of the predictor network \mathcal{P}	49
4.1	Stimuli generation for the evaluation.	55

4.2	Frequency response of the anchor condition filter	56
4.3	MUSHRA trial page	58
4.4	MUSHRA experiment boxplot	59
4.5	Mixed model residual Q-Q plot and histogram.	62
4.6	Mixed model estimates	63
4.7	A trial page for experiment 2	68
4.8	Histogram plots for the four ratings in experiment 2	69
5.2	A combined synthesis parameter estimation and prediction network. . . .	74

List of Tables

1.1 Research fields within linguistics	2
4.1 A breakdown of the VocalSet dataset	54
4.2 A breakdown of the stimuli used in the evaluation.	55
4.3 Experiment 1: Wilcoxon signed-rank post hoc tests	60
4.4 Experiment 1: General information on the mixed model.	61
4.5 Experiment 1: Mixed model omnibus test results.	61
4.6 Experiment 1: Mixed model fixed effects parameter estimates	61
4.7 Experiment 1: Mixed model post hoc tests for main effects. (extract) . . .	63
4.8 Experiment 2: Wilcoxon signed-rank tests	69
A.1 Experiment 1: Mixed model post hoc tests for main effects	83
A.2 Experiment 1: Mixed model post hoc tests for Vowel * Gender	83
A.3 Experiment 1: Mixed model post hoc tests for Condition * Vowe	84
A.4 Experiment 1: Mixed model post hoc tests for Condition * Gender . . .	86
A.5 Experiment 1: Mixed model post hoc tests for Gender * Vowel * Cond. .	86

List of Symbols and Notations

General notations.

\hat{x}	Approximation or estimation of x .
\bar{x}	The complex conjugate of x .
$x(t)$	A continuous signal with time t .
$x[n]$	A discrete signal with sample n .
$\mathbf{A} = (a_{i,j})$	A matrix \mathbf{A} with element $a_{i,j}$ at position i,j .
$\mathbf{b} = (b_i)$	An array \mathbf{b} with element b_i at position i .

Reoccurring variables.

f_0	The fundamental frequency.
f_s	The sampling rate.
ω	Phase in the Z-domain, with $\omega = 2\pi f/f_s$.
z	The complex Z-domain variable if not otherwise stated.
s	The complex Laplace domain variable if not otherwise stated.
$\mathbf{Y} = (y_{n,m})$	Audio frames with $0 \leq n < N_B$ samples and $1 \leq m \leq M$ frames.
$\mathbf{O} = (o_{k,m})$	Complex harmonic partials (magnitude and phase) for $1 \leq k \leq K$ partials and $1 \leq m \leq M$ frames.
$\mathbf{X} = (\chi_m)$	Synthesis parameters for $1 \leq m \leq M$ frames.
θ	State of synthesis method \mathcal{S} .
R_d	Glottal flow parameter of the synthesis model.
g_{dB}	Gain parameter for the synthesis model in decibel.
$\mathbf{p} = (p_j)$	Complex numbers representing $1 < j < J$ poles in the Z-domain.
$\mathbf{q} = (q_j)$	Complex numbers representing $1 < j < J$ zeros in the Z-domain.
C	Defines stimuli condition with $C \in \{C_R, C_H, C_F, C_P, C_A\}$
G	Defines singers gender with $G \in \{G_F, G_M, G_P\}$
V	Defines vowel with $V \in \{V_A, V_I, V_O\}$

List of Abbreviations

- ANOVA** analysis of variance 59, 60
- API** application programming interface 22, 44, 47, 51, 52
- AR** autoregressive 13
- ARMA** autoregressive-moving-average 13, 14, 19
- ARX** autoregressive with exogenous input 13, 14, 19
- CALM** causal-anticausal linear model 12, 20
- DDSP** differentiable digital signal processing 20–23, 25
- DSP** digital signal processing 3, 5, 21–23, 25, 26, 32, 71, 72, 74
- FIR** finite impulse response 15, 22
- GCI** glottal closure instance 6
- GFM-IAIF** glottal flow model-IAIF 17, 18
- GIF** glottal inverse filtering 15
- GOI** glottal opening instance 6
- IAIF** Iterative Adaptive Inverse Filtering 17–19
- IIR** infinite impulse response 12, 13, 22, 32, 43, 74
- IOP-IAIF** iterative optimal preemphasis-IAIF 17, 18
- LPC** linear predictive coding 16–18
- MSE** mean squared error 50
- MUSHRA** multi stimulus with hidden reference and anchor 53, 54, 56–60, 64, 65, 72
- OQ** open quotient 6
- SIMD** single instruction, multiple data 13, 33
- SMS** spectral modeling synthesis 9, 10
- TE** true envelope 17, 18

1 Introduction and Motivation

The human voice is omnipresent in our day-to-day life and the focus of research efforts across multiple scientific fields, including linguistics and literature, music, performing arts, biology, medicine, signal processing, telecommunication and many more. This thesis focuses on one specific role of our voice; its role as a musical instrument. With the growing popularity of machine learning in the field of sound synthesis, we have seen an influx of realistic singing voice and text-to-speech models. However, while virtual real-time capable emulations of natural instruments such as pianos, violins or guitars are already widely used in music making, emulations of the human voice are rare and mostly still researched within the scientific community. That is even though playable choirs exist since the 60s in form of the sample-based Mellotron. Arguably, the multilayered complexity of the human voice is at fault for why modern emulations are still a rather rare occurrence. Unlike most acoustic instruments, the voice isn't solely conveying auditory impressions which are characterized mainly by temporal and spectral properties. Instead, it also relays semantics through the spoken language and, in addition, may convey speaker identity, regional and socioeconomic background and the speakers emotional and biological state [1]. Because of that, it seems reasonable that speech and voice perception follows a different, more complex psycho-acoustical model than the perception of most natural instruments. This poses a challenging problem for the synthesis of virtual voices at least for the synthesis of semantic phrases as is the case for text-to-speech applications.

The production and perception of sounds is studied in phonetics, a research field of linguistics. Articulatory phonetics specifically deals with the sounds that are produced through articulation of the voice production physiology. Manners of articulation include vowels and consonants such as stops, fricatives, affricates and approximants [2, p. 18]. Vowels usually are produced by an open vocal tract while consonants are produced by partial to complete obstruction of the vocal tract for instance via the lips or tongue. Phonology deals with the organization of these units of sounds, referred to as phonemes. The structure of words and their internal components is studied in morphology while syntax covers the structure of phrases and sentences. Finally, the literal meaning of language is the topic of semantics. An overview of the research fields in linguistics is shown in Table 1.1.

Field	Topics
Semantics	Literal meaning of words, phrases and sentences.
Syntax	Study of structure of sentences and phrases.
Morphology	Study of words and their internal structure.
Phonology	Organization of sounds or phonemes.
Phonetics	Physical production and perception of speech.

Table 1.1: Research fields within linguistics ordered by abstraction from high (semantics) to low (phonetics) [2].

Inevitably, most research fields of linguistics from phonetics to semantics play a role in the synthesis of spoken or sung phrases or sentences. In contrast, the production of vowels is mainly a topic of phonetics and thus poses a less complex problem. By focusing on vowels, recent advances in mathematical optimization and machine learning can hopefully be combined with a decent understanding on vowel production, resulting in a convincing, real-time capable vowel synthesis method.

1.1 Research Goal and Questions

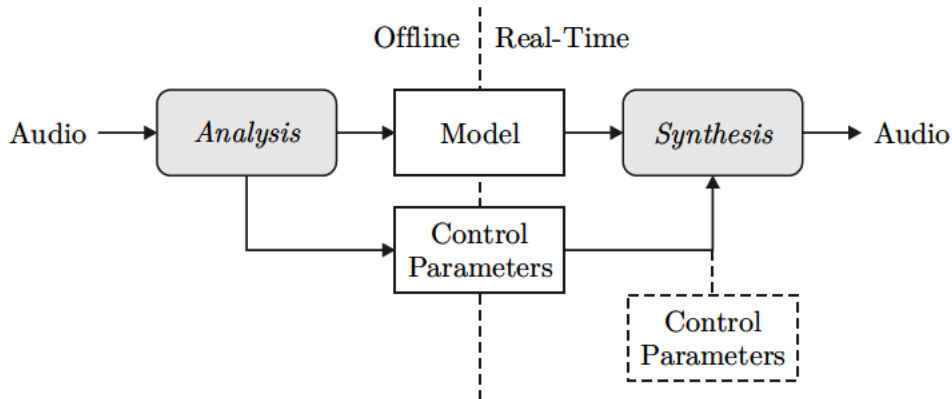


Figure 1.1: During analysis, an audio sample is separated into a singer model, capturing the singers voice qualities, and performance aspects represented by control parameter trajectories. During synthesis, the singer model can be used with new control parameters (dotted) to produce an audio signal in real-time.

This thesis explores possibilities for the real-time synthesis of sung vowels. The primary intention is to develop a synthesis method capable of producing convincing sung vowels in real-time. During analysis, a singer model, capturing vowel and voice qualities, and associated control parameters such as pitch and intensity may be derived from an audio excerpt. During synthesis, new control parameters may be used in conjunction with the virtual singer model to synthesize new audio excerpts. An overview of the intended method is shown in Figure 1.1. This thesis aims to address the following research questions.

- What synthesis method is capable of convincingly reproduce the human voice while maintaining real-time capabilities?
- How can digital signal processing (DSP) and machine learning be combined for voice analysis and synthesis?
- How can such a synthesis model be parameterized from audio samples to capture vowel and the singers voice qualities?
- What methods can be used to evaluate the perceived quality and shortcomings of the analysis and synthesis method?

As a main contribution of this thesis, a novel analysis-by-synthesis parameter estimation method is presented. The method uses a harmonic spectral representation of the target signal and estimates synthesis parameters by minimizing a reconstruction error between the target signal and the synthesized signal.

1.1.1 Thesis Scope

The proposed method is intended to synthesize audio from control parameter such as pitch and intensity. It is important to distinguish between these intermediate level control parameters and high-level parameters. For a vocal performance, a musical score would usually be used to describe text, melody, timing and expression. This description is then interpreted by the singer to create a performance. The performance itself may best be described by the phonetic articulation and timing, pitch and intensity. These variables more directly describe the sound that the vocal physiology produces (Figure 1.2). One example that illustrates this distinction between high level and intermediate level parameters is pitch. A musical score may describe the pitch of individual notes and whether or not vibrato should be used to sing them. It is up to the singer to interpret the score and produce the single pitch trajectory of the performance that captures both note pitch and vibrato.

In voice synthesis, these two tasks may be referred to as parameter modeling and synthesis. Parameter modeling derives intermediate level parameters such as pitch and intensity from high level inputs such as a musical score as discussed for instance in [3, 4]. During synthesis, these intermediate parameters may be converted to low level coefficients which directly drive the synthesis engine. This thesis focuses on the synthesis from intermediate level control parameters, specifically pitch.

1.1.2 Application Cases

The proposed method could see application in the music software industry. The method could be used to develop software emulations of the human voice production similar to existing emulations of natural instruments such as acoustic guitars or pianos. During development, the method could be used to generate virtual singer models that capture vowel and voice qualities of the analysed source sample. The user could then chose

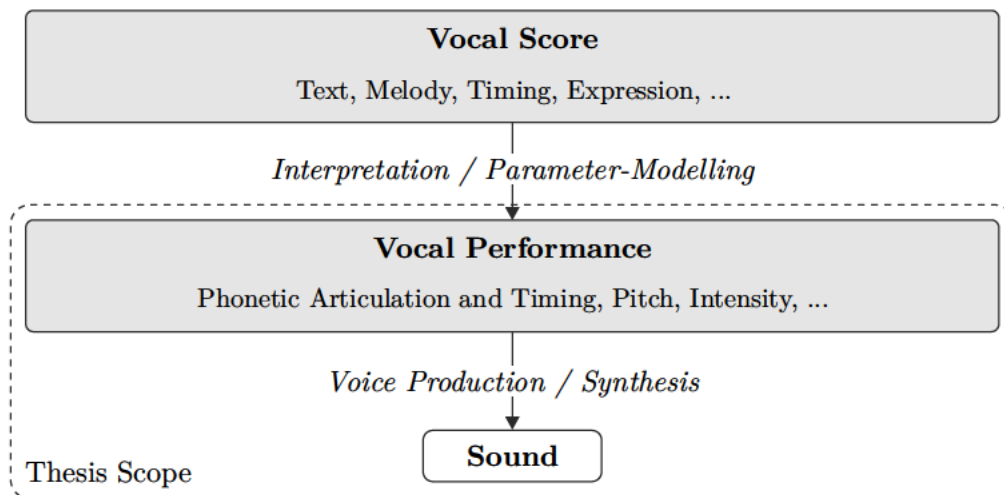


Figure 1.2: A simplified diagram showing the relation between musical vocal score, vocal performance and the generated sound.

between vowels and voice qualities and control pitch in real time to produce new sung excerpts similar to how instrument emulations of virtual synthesizers are used. Further developments could extend this application case by allowing the user to blend between models and as a result morph different vowels and voice qualities. More sophisticated real-time control could be provided for voice intensity, roughness, breathiness and other voice qualities. However, it is worth mentioning again that the proposed method is not intended to produce or reproduce semantic singing voice or speech excerpts as is the case in Text-to-Speech applications.

1.2 Structure of the Thesis

This thesis continues by introducing relevant fundamentals in Chapter 2, including voice physiology and source-filter theory, voice synthesis models, source-filter analysis and machine learning in voice synthesis. Afterwards, the proposed method is described in Chapter 3, separated in the four components that make out the method, namely synthesis model, harmonic analysis, synthesis model parameter estimation and parameter predictor network. Chapter 4 describes the process used to evaluate the performance of the proposed method, starting with an introduction of the dataset used in this thesis followed by an description of the two conducted experiments. Finally, the thesis is concluded in Chapter 5.

2 Fundamentals

This chapter provides an introduction to the field of singing voice analysis and synthesis and some other relevant subjects. First, the human voice production mechanism is described together with a theoretical representation thereof. Afterwards, an overview is given on a selection of DSP methods used in voice synthesis followed by a section on parameter estimation methods used for voice analysis. Section 2.4 discusses some relevant contributions from the field of machine learning. This chapter concludes with short introductions into B-Splines and automatic differentiation.

2.1 Physiology and Source-Filter Theory

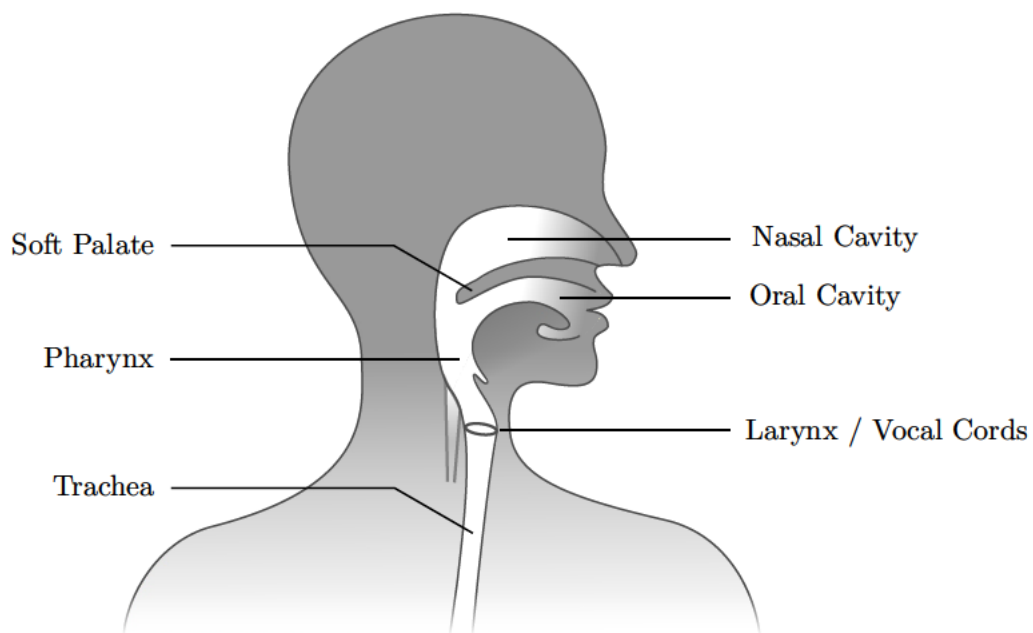


Figure 2.1: The human voice production system.

The production of the human voice involves both physics and physiology [5]. Most organs of the human respiratory system are also responsible for voice production, such as our lungs, the trachea, the vocal folds housed in the larynx and the organs that make up the vocal tract; pharynx, oral cavity, nasal cavity, nostrils and lips. An air flow produced by the lungs passes through the glottis, the opening between the vocal folds. In *voiced speech*, this air flow forces the vocal folds to oscillate, producing an mostly periodic opening and closing of the glottis. The resulting air wave passes through the pharynx,

oral and nasal cavity and radiates out at the lips and nostrils. In *unvoiced speech*, an air wave is not produced by vibrating vocal folds but instead by turbulent airflow introduced by friction effects along the larynx and vocal tract boundaries. As this work focuses on the production of voiced vowels, the generations of fricatives and consonants are left undiscussed. An overview of the organs responsible for voice production is shown in Figure 2.1.

2.1.1 Glottal Flow

The periodic opening and closing of the glottis produce what is commonly referred to as the glottal flow. A single cycle is known as a glottal pulse. The vibration is a result of physical forces, primarily an elastic force and the Bernoulli force, acting on the vocal folds tissue surrounding the vocalis muscles [5, p. 38]. The vocal folds can be modeled as a spring-mass system, that exhibits some elastic force on the vocal folds while out of their natural position. Thus, while shut, the elastic force will move them apart and while widely open, the force will draw them together. Additionally, when the folds are nearly closed, the high airflow passing through the small glottis area induces a low pressure zone. The resulting Bernoulli force draws the vocal folds together. A simplified representation of these forces can be seen in figure 2.2 [5, p. 38]. In (a), the elastic force draws the vocal folds closer to each other until the resulting increase in air flow introduces the Bernoulli force (b), closing them shut. With no airflow, the elastic force again separates the vocal folds. The separation is slightly delayed at the upper part of the tissue due to an Bernoulli force once again induced by the air flow.

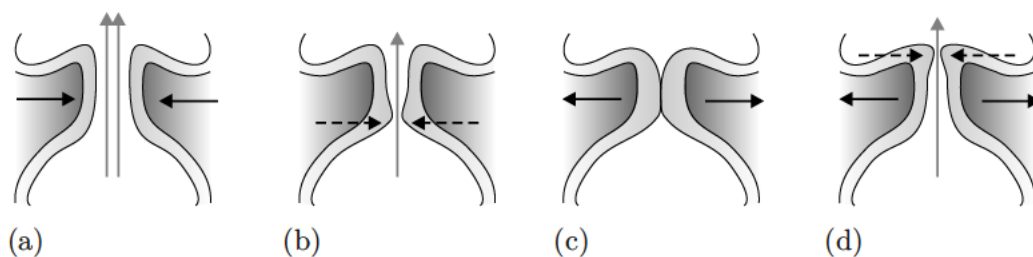


Figure 2.2: A simplified visualization of the oscillation of the vocal folds with the forces acting on the vocal fold tissue, the elastic force (solid line) and the Bernoulli force (dotted line). Graphic simplified from Figure 2.9 in [5, p. 38]

The speaker or singer can control the oscillation frequency and intensity mainly by tensing or relaxing the vocalis muscles and adjusting the pressure in the lungs. One parameter often used to describe the perceived intensity of the glottal flow is the open quotient (OQ), defined as the time between glottal opening instance (GOI) and glottal closure instance (GCI) relative to the glottal wave period T_0 [6]. During speech, but also during singing, random variations in the glottal flow period and shape can occur. These variations are referred to as vocal jitter and shimmer respectively [5, p. 6].

2.1.2 Vocal Tract

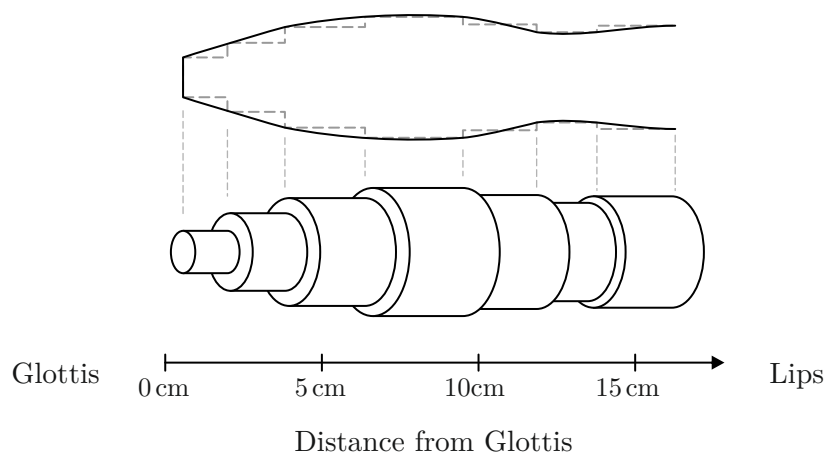


Figure 2.3: A schematic depiction of the vocal tract area function (top) and corresponding tube model consisting of 7 segments (bottom).

The vocal tract acts as a resonator, dampening or amplifying certain frequencies of the glottal flow produced by the vibrating vocal folds. It consists of the larynx, oral cavity and nasal cavity and the connecting pharynx. [5, p. 51] The vocal tract can be modeled as a tube or a series of coupled tube segments, closed at the glottis and open at the lips [7] (Figure 2.3). The tube segments diameter is determined by the cross section area along the vocal tract. For simplicity, the vocal tract is often assumed to be straight and the nasal cavity is ignored. Three exemplary vocal tract area functions for the vowels [a], [i] and [u] are shown in Figure 2.4.

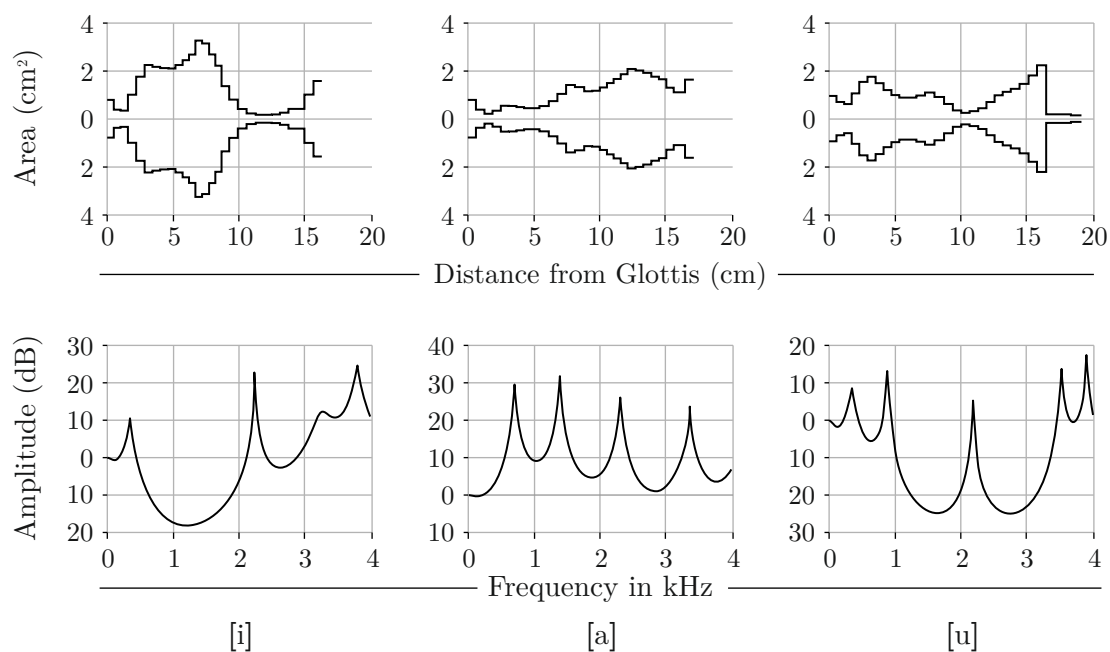


Figure 2.4: Vocal tract area functions for three vowels ([i], [a], [u]) with corresponding vocal tract transfer function [8].

In its most simple form, a tube consisting of a single segment, the model predicts equally spaced poles in the transfer functions. With a length of $l = 14$ cm for a female adult, and a speed of sound of $c = 345$ m/s in the slightly humid air of the vocal tract, the first resonant frequency is predicted at $f_1 = c/(4l) \approx 600$ Hz with following poles at $f_2 = 1800$ Hz, $f_3 = 3000$ Hz, $f_k = f_1(2k - 1)$ Hz. For adult males, the poles are generally positioned slightly lower at 500 Hz, 1500 Hz and 2500 Hz respectively due to a slightly longer vocal tract. By adjusting the vocal tract shape, for instance by moving jaw, tongue, lips or larynx, the pole frequencies can be changed to form vowels. These poles are often referred to as formants. However, in some instances, poles move sufficiently close to each other to form one single perceived formant, as is the case for the so called *singers formant*. In this thesis, these terms are generally used interchangeably. The first two or three formants generally are sufficient to identify a spoken or sung vowel [9] while higher formants mostly contribute to the voice timbre or voice quality.

The nasal cavity is connected to the pharynx through the opening at the velum or soft palate. In nasalization, the velum is opened to couple the nasal cavity with the remaining vocal tract. This coupling drastically increases the complexity of the vocal tract shape with the inclusion of multiple 3-port junctions at the velum and the nostrils and the addition of the paranasal sinuses. This coupling introduces additional poles and zeros, referred to as anti-formants, in the vocal tracts frequency response as discussed in [7, 10]. It is very common to assume that singers don't couple the nasal cavity and such, especially early vocal tract models only considered the main poles created between glottis and lips.

2.1.3 Lip and Nostril Radiation

Finally, the vocal tract waves radiate out through the lips or nostrils. At the lips, this radiation can be modeled with an vibrating piston set in a baffle representing the singers head. This baffle may be approximated as being either infinite or spherical [7, 11]. A common model for the radiation filter H_L at the lips is a simple differentiator [11, 12] with

$$H_L(w) = i\omega \quad (2.1)$$

2.1.4 Source-Filter Theory

The source-filter theory, originally proposed in [7], separates the voice production in two linear systems, modeling the glottal flow and vocal tract including lip radiation. The voice spectrum $Y(\omega)$ is a product of the glottal flow, modeled by the glottal source $S(\omega)$, and the vocal tract, represented by the filter $V(\omega)$ [7, p. 19]. In addition, the lip radiation can be modeled as its own system $L(\omega)$.

$$Y(\omega) = S(\omega)V(\omega)L(\omega) \quad (2.2)$$

This model can further be extended to include aspiration noise $N(\omega)$ which is assumed to be produced at the glottis and thus filtered by the vocal tract. Finally, it is convenient to separate the effects of the source in two systems, one representing a purely flat harmonic impulse train $H_{f_0}(\omega)$ with fundamental f_0 and one modeling the glottal magnitude and phase spectrum $G(\omega)$. This separates the source components into an deterministic part $G(\omega)H_{f_0}(\omega)$ and a non-deterministic part $N(\omega)$. The extended model thus is defined as follows [11].

$$Y(\omega) = \left(G(\omega)H_{f_0}(\omega) + N(\omega) \right) V(\omega)L(\omega) \quad (2.3)$$

The source-filter model is regarded a significant contribution to the understanding of the human voice. While it's widely used and build upon, some drawbacks have to be mentioned. On a perceptual level, it was found that certain timbre or voice quality aspects can't be contributed unambiguously to source or filter [13]. This hints towards a reoccurring issue in singing voice synthesis. As Equation 2.2 suggests, variations in the voice spectrum $Y(\omega)$ may be contributed either to source, filter or radiation. Thus, assumptions have to be made on the specifics of the individual systems in order to separate them. This poses the main challenge in source-filter separation in voice analysis. Moreover, the source-filter model assumes a linear separation of vocal folds and the vocal tract. However, it was found that nonlinear interactions between vocal folds and vocal tract exist [14, 15]. These interactions can be mostly explained by the larynx, which gradually matches the high impedance at the glottis with the lower impedance throughout the vocal tract. This effect can be compared to that of horns in brass instruments.

2.2 Synthesis Models

Voice synthesis models were introduced as early as 1939 with the proposal of the original Vocoder [16]. The presented model consists of an analysis stage that extracts pitch and spectral information from an electric speech signal and a synthesis stage that recreates said signal with an oscillator filtered by a filter bank. While not being mentioned by name, the Vocoder thus implements the source-filter theory which was formally introduced 20 years later by Gunnar Fant [7]. An alternative to the source-filter model is spectral modeling synthesis (SMS) [17]. In SMS, the signal is not separated in source and filter but is instead decomposed in a deterministic component and a stochastic component. The deterministic part consists of R quasi-sinusoidals with amplitude $A_r(t)$ and phase trajectories $\theta_r(t)$ while the stochastic component models the residual $e(t)$ between the original signal $s(t)$ and the deterministic part.

$$s(t) = \sum_{r=1}^R A_r(t) \cos(\theta_r(t)) + e(t) \quad (2.4)$$

By estimating amplitudes and phases of sinusoidal signal components, the residual signal can be calculated by subtracting the deterministic component from the original signal. During synthesis, the stochastic component is synthesized from a white noise signal filtered with an impulse response estimated from $e(t)$. While SMS can be used in various application cases, source-filter models and vocoders are more commonly seen in voice synthesis due to their ability to closely mimic the separation of voice qualities in pitch and intensity (source) and vowel and timbre (filter). For this reason, this section focuses on the description of glottal source and vocal tract models.

2.2.1 Source Models

Various source models have been proposed that approximate the deterministic part $G(\omega)H_{f_0}(\omega)$ of the glottal flow [18, 19, 12, 20]. A comprehensive review and comparison of these models can be found in [11]. Three of these, LF [12], LF- R_d [20], and CALM [21] are discussed in more detail below.

Liljencrants-Fant (LF)

The LF model was introduced in [12] and uses four timing parameter t_p , t_e , T_a and t_c with glottal flow period T_0 to describe the glottal flow and glottal flow derivative in the time domain. A representation in the spectral domain is presented in [22, 23]. As was described in Section 2.1.1, the vocal folds don't close instantaneously. An improvement of the LF model over previous models was the introduction of a return phase that models the gradual closing of the vocal folds. The start of the return phase, is referred to as glottal closure instance and modeled by t_e . Return phase length and decay are represented by t_c and T_a respectively. The time of highest glottal flow is marked by t_p .

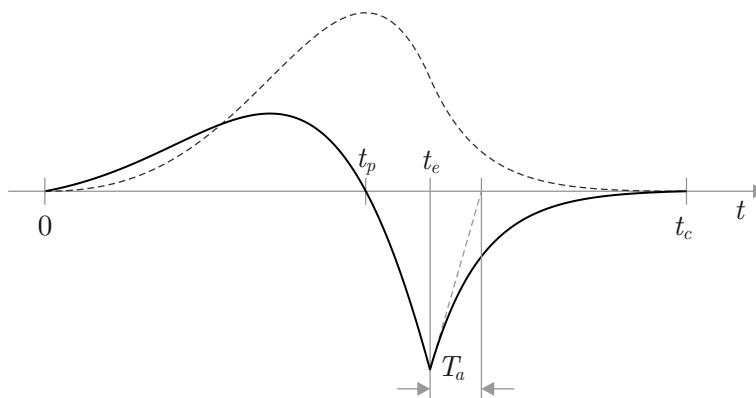


Figure 2.5: Parameters of the LF-model [12]

Figure 2.5 shows the LF model glottal pulse (dashed line) and its derivative (solid line) for parameters $t_p = 0.4$, $t_e = 0.6$, $t_c = 1$, $T_a = 0.08$. A wider range of LF glottal pulses is shown in Figure 2.6.

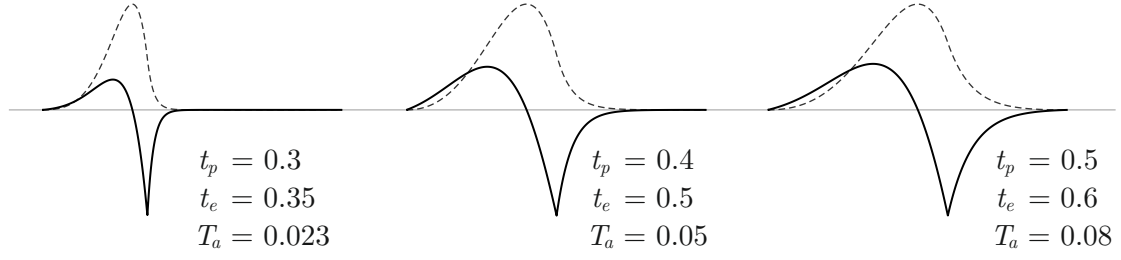


Figure 2.6: Three LF glottal flow (continuous) and derivative (dashed) curve with various values for t_p , t_e , and T_a with fixed $t_c = T_0 = 1$.

The LF-model defines the glottal pulse derivative piece-wise as follows

$$g_{LF}(t) = \begin{cases} E_0 e^{\alpha t} \sin(t/t_p) & 0 < t < t_e \\ \frac{-1}{\epsilon t_a} \left(e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)} \right) & t_e < t < t_c \\ 0 & t_c < t \end{cases}$$

The variables α , ϵ and E_0 are defined implicitly [12] and need to be determined using numerical methods or approximations. By defining the pulse derivative, the LF model directly includes the lip radiation model $L(s) = i\omega$ which is assumed to be static.

LF- R_d

In [20], Fant revisited the LF model and proposed the LF- R_d model with the intention of covering a high voice quality range with a single parameter. R_d is used to define the LF parameters t_p , t_e and T_a , with $t_c = T_0 = 1$. The conversion is given as follows.

$$\begin{aligned} R_{ap} &= (-1 + 4.8R_d)/100 \\ R_{kp} &= (22.4 + 11.8R_d)/100 \\ R_{gp} &= 1/(4((0.11R_d/(1/2 + 1.2R_{kp})) - R_{ap})/R_{kp}) \end{aligned} \tag{2.5}$$

$$\begin{aligned} t_p &= \frac{1}{2R_{gp}} \\ t_e &= t_p(R_{kp} + 1) \\ T_a &= R_{ap} \end{aligned} \tag{2.6}$$

The LF- R_d model is defined for a range of $0.3 \leq R_d \leq 2.7$, the time domain glottal pulse derivative and associated magnitude spectra of the first 40 harmonics can be seen in Figure 2.7. R_d simultaneously controls the timing of glottal closure instance t_e and the general spectral tilt which correlates with short T_a and $t_e - t_p$. Notably, the spectral peak shifts away from the fundamental at high R_d values to the second and third harmonic at low values. Additionally, ripples in the magnitude spectrum can be observed especially for $R_d > 2$.

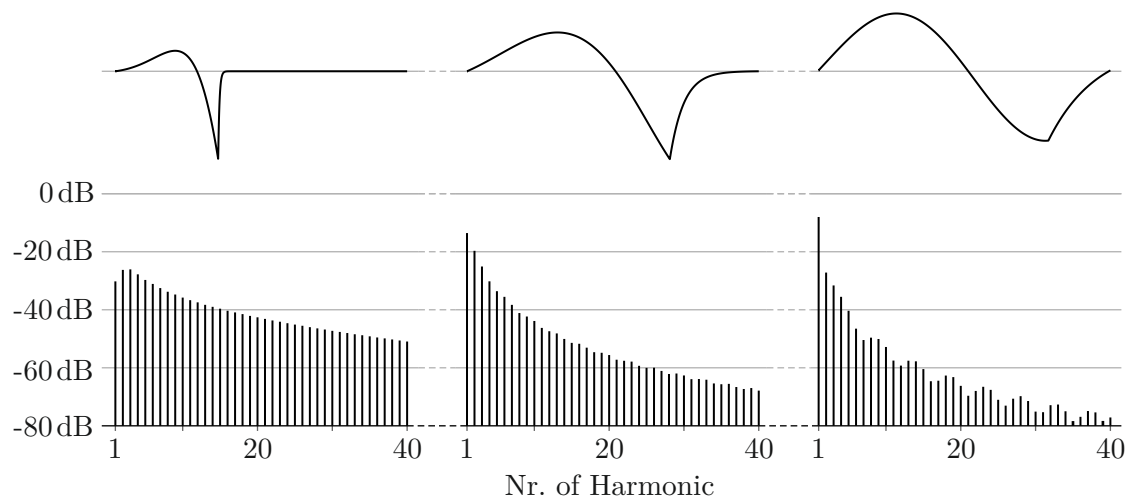


Figure 2.7: LF- R_d pulse derivative and associated spectra for the first 40 harmonics for $R_d = 0.3$, $R_d = 1.2$ and $R_d = 2.7$ from left to right.

Causal-Anticausal Linear Model (CALM)

The causal-anticausal linear model (CALM) describes the glottal pulse as a system consisting of a real pole and a complex conjugate anti-causal pole pair [21]. Considering, that the glottal pulse derivative is excited by an impulse, the return phase $t > t_e$ of the LF model can be described as a truncated impulse response of a first order pole, an exponential decay. Likewise, the open phase $t < t_e$ can be described as a truncated impulse response of a anti-causal complex pole pair [21]. This leads to an approximation of the glottal flow derivative spectrum as consisting of an 2nd order band-pass associated with the open phase and mostly determined by the open quotient and a 1st order low-pass filter associated with the return phase, mostly controlled by T_a in the LF model. The band-pass produces what is commonly referred to as the glottal formant at frequency F_g [21] and determines the spectral peak that can be observed between the first three harmonics in Figure 2.7. Additionally, the causal real pole of the return phase can directly be approximated as an additional -6 dB per octave low-pass drop that is usually found at higher frequencies F_c and, in combination with the glottal formant, produces an overall spectral slope of -12 dB per octave above $f > F_c$.

2.2.2 Filter Models

Over the years, many models for the vocal tract filter response have been presented. In [7], Fant proposes the use of all-pole infinite impulse response (IIR) filters. An all-pole filter consists only of poles with coefficients a_k in the form

$$H(z) = \frac{1}{\sum_{k=0}^K a_k z^{-k}} \quad (2.7)$$

or alternatively, pole radius r_n , pole phase $\omega_k = 2\pi f_k/f_s$ and gain g with

$$H(z) = g \frac{1}{\prod_{k=1}^K (z - r_k e^{i\omega_k})} \quad (2.8)$$

As discussed in Section 2.1, nasal coupling introduces additional poles and zeros, the latter can't be modeled with an all-pole filter. Moreover, interaction effects between vocal tract and vocal folds [14] might require the use of a more flexible approach to approximate nonlinearities that affect the vocal tract transfer function. For these reasons, pole-zero filters might be used instead of all-pole filters. In product form, the transfer function of a pole-zero IIR filter is given by

$$H(z) = g \frac{\prod_{j=1}^J (z - q_j)}{\prod_{k=1}^K (z - p_k)} \quad (2.9)$$

with poles p_k , zeros q_j and gain g . Pole-zero filters can be implemented in various structures. To prevent instability issues due to numerical precision limitations, it is advisable to implement a high order IIR filter in a series of second order sections. To further improve computational efficiency, serial filter banks can be converted to an equivalent parallel filter bank to utilize the performance of modern computer architectures, which utilize parallel processing for instance using single instruction, multiple data (SIMD).

AR, ARX and ARMA

In voice synthesis and analysis, the term autoregressive (AR) is often used to describe processes that are modeled with an all-pole filter as follows [24]

$$y[n] = \sum_{k=1}^K a_k y[n-k] + \sigma \varepsilon[n] \quad (2.10)$$

with gain σ and white Gaussian noise $\varepsilon[n]$. In extension, autoregressive with exogenous input (ARX) describes an autoregressive process with additional deterministic input $\mu[n]$ [24] with

$$y[n] = \sum_{k=1}^K a_k y[n-k] + \mu[n] + \sigma \varepsilon[n] \quad (2.11)$$

Finally, pole-zero filters are represented in this family of models by the autoregressive-moving-average (ARMA) model [25], with

$$\begin{aligned}
y[n] = & \sum_{j=1}^J b_j (\mu[n-j] + \sigma \varepsilon[n-j]) \\
& + \sum_{k=1}^K a_k y[n-k] + \mu[n] + \sigma \varepsilon[n]
\end{aligned}
\tag{2.12}$$

The definition of the ARMA model here is extended to include the exogenous input in addition to the Gaussian noise. For consistency, this model should be termed ARMAX. Both ARX and ARMA are often seen throughout voice synthesis literature due to their resemblance of the source-filter model with the vocal tract represented by a linear filter, the glottal flow being represented by the exogenous input $\mu[n]$ and residual $\varepsilon[n]$.

Physical Modeling

Physical modeling aims to model sound synthesis or manipulation by simulating the underlying physical mechanisms of an instrument, effect or the human voice. A benefit of physical modeling is that the models parameter space relates closely to that of the physical counterpart. For instance, instead of describing the vocal tract with complex pole positions, it can be expressed by its length and cross-sectional area function.

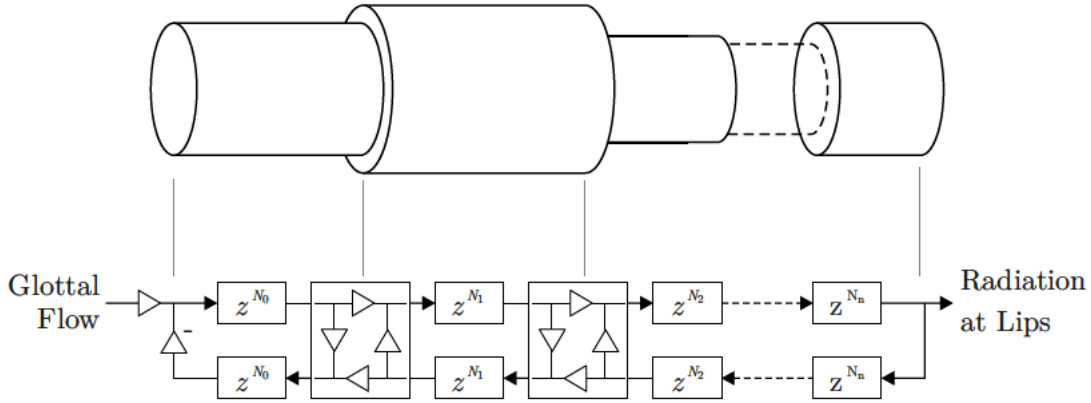


Figure 2.8: A simplified waveguide model for a vocal tract. Tube segments are modeled as discrete signal delays. Tube connections are modeled as Kelly-Lochbaum junctions.

The Kelly-Lochbaum model describes a discrete implementation of the tube model of a vocal tract using waveguides [26, 27]. A waveguide consists of two parallel delay lines that emulate the bidirectional wave propagation in strings or tubes. The vocal tract is modeled as a series of coupled tube segments, each with a fixed diameter. Wave propagation through the tubes is implemented with discrete signal delays. At segment intersections, the change in impedance can be modeled by a two-port junction. The Kelly-Lochbaum junction partially reflects and inverts incoming waves at both sides according to the impedance of the connected tube segments. A simplified schematic of a vocal tract waveguide and the theoretical modeled segmented tube can be seen in Figure 2.8. Improvements including fractional delay lines, conical tube sections and the inclusion of loss effects were presented in [28, 29, 30]. Three-port junctions have been proposed in

[31] to incorporate nasal coupling. A two-dimensional waveguide model has been proposed in [32]. An alternative to the waveguide approach is the Chain-Matrix model [33, 34]. Chain matrices can be used to calculate the Laplace domain impedance transfer function of a vocal tract tube model. By sampling the frequency response, a discrete impulse response can be calculated to model the vocal tract as a finite impulse response (FIR) filter.

2.3 Source-Filter Analysis

From Equation 2.3, it is clear that inverse filtering can be used to separate source and filter once one of them is known. With a vocal tract filter estimate $V(\omega)$ and lip radiation filter $L(\omega)$, the source spectrum $S(\omega) = G(\omega)H_{f_0}(\omega) + N(\omega)$ can be extracted by inverse filtering as follows [11, p. 57]

$$S(\omega) = \frac{Y(\omega)}{V(\omega)L(\omega)} \quad (2.13)$$

This method is referred to as glottal inverse filtering (GIF). A similar approach can be taken when the vocal tract is to be estimated from a known source spectrum. However, due to the harmonic structure of the source spectrum and the presence of noise, inverse filtering by $S(\omega)$ is impractical. Instead, spectral envelope estimation methods can be used. The vocal tract transfer function can be approximated with

$$\hat{V}(\omega) = \mathcal{E} \left(\frac{Y(\omega)}{G(\omega)L(\omega)} \right) \quad (2.14)$$

where \mathcal{E} estimates the spectral envelope of a harmonic or otherwise complex spectrum [11, p. 59] which will be discussed in section 2.3.1. In this case, the stochastic component $N(\omega)$ is neglected and the harmonic impulse train $H_{f_0}(\omega)$ usually has little impact on the spectral envelope estimation and thus can be ignored as well.

In practice, the separation of source and filter poses a challenging problem mainly for two reasons. First, without assumptions on the specifics of source $S(\omega)$ and filter $V(\omega)$, no unique separation of source and filter can be achieved through inverse filtering. To address this, various models for both source and filter have been proposed to describe both magnitude and phase of source and filter. Additional assumptions can be made, for instance about the temporal variations of source or filter parameters. Secondly, due to the harmonic structure and the presence of noise in the signal, reliable estimations of the vocal tract can only be made at the frequencies which are "sampled" by the harmonics of the glottal source. This introduces an issue of undersampling especially with high fundamental frequencies. For instance, with a fundamental of $f_0 = 500$ Hz, the frequency band up to 2 kHz is only sampled by four harmonics even though this frequency range is critical to estimate frequency and resonance of both the first two vocal tract formants and the glottal formant.

2.3.1 Spectral Envelope Estimation

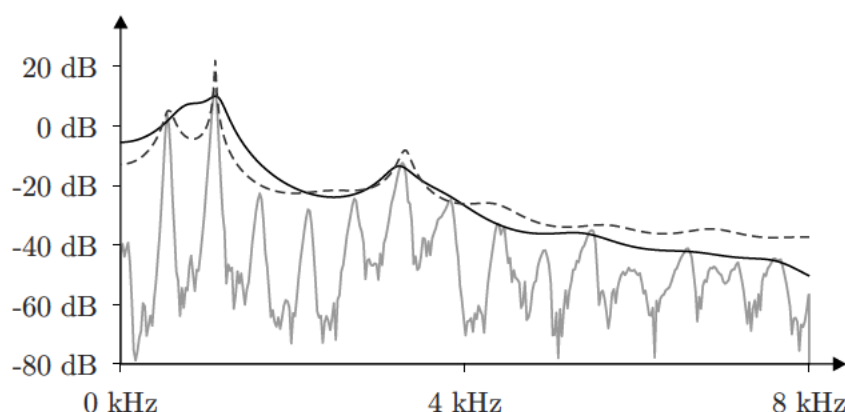


Figure 2.9: Two spectral envelope estimations for a sung vowel [a] ($f_0 \approx 540$ Hz) with LPC (dotted line) and true envelope (solid line).

The spectral envelope describes a smooth envelope that follows the spectral peaks of a harmonic or otherwise complex spectrum. Spectral envelope estimation is particularly useful in voice analysis as the vocal tract can only be reliably observed at the harmonics of the glottal source. Linear predictive coding (LPC) models an observed signal as an all-pole excited by a flat excitation signal [35, 36]. Under this assumption, the observed signal $y[n]$ can be described as

$$y[n] = gx[n] - \sum_{k=1}^K a_k y[n-k] \quad (2.15)$$

where g is a gain factor, $x[n]$ the excitation and a_k the K coefficients describing the all-pole filter. These coefficients can be estimated by the methods of least squares, minimizing the error $e[n]$ with [36]

$$e[n] = y[n] - \hat{y}[n] = y[n] + \sum_{k=1}^K a_k y[n-k] \quad (2.16)$$

where

$$\hat{y}[n] = - \sum_{k=1}^K a_k y[n-k] \quad (2.17)$$

LPC is widely used in speech synthesis and telecommunication though it has some limitations. Most notably, the method has the tendency to overfit by fitting individual harmonic peaks with poles especially in signals with high fundamental frequencies [37]. One possible solution is to smear the spectrum to reduce notches between harmonic peaks. An iterative method to smear spectrum $X[k] = \mathcal{F}\{x[n]\}$, obtained from a time domain frame $x[n]$ through Fourier transformation \mathcal{F} , is described as follows [38, 37, 39].

Let $A_i[k]$ be the logarithmic absolute representation of smeared spectrum at iteration i with $A_0[k] = \log(|X[k]|)$. Iteratively, $A_i[k]$ is updated as follows

$$A_i[k] = \max(A_0[k], \mathcal{C}(A_{i-1}[k])) \quad (2.18)$$

where \mathcal{C} is a cepstral smoothing operation performed by liftering in the cepstral domain (analogous to filtering in the spectral domain) and defined as follows

$$\mathcal{C}(A_i[k]) = \mathcal{F} \left\{ W[n] \mathcal{F}^{-1} \{ A_i[k] \} \right\} \quad (2.19)$$

with cepstral low-pass filter $W[n]$. After an abort criteria is reached, for instance after $\max |A_i[n] - A_{i-1}[n]|$ falls below a predefined threshold, the smeared magnitude spectrum $C[k]$ of $X[k]$ can be obtained by $C[k] = e^{A_i[k]}$.

In true envelope (TE) [38, 37, 39], the described spectral smearing is combined with LPC by estimating the all-pole coefficients of the smeared magnitude spectrum $C[k]$. An advantage of TE is the ability to precisely control the smearing bandwidth according to the signal to be analysed. Additionally, the method prevents overfitting of harmonic peaks through poles as can often be observed with LPC. An example of 40th order LPC and TE estimations for a vowel [o] of a female singer at roughly 540 Hz is shown in Figure 2.9. Noteworthy is the absence of harmonic overfitting of the first two harmonics with TE.

In [40], an alternative approach for estimating the spectral envelope over multiple audio frames was proposed. One common issue in spectral envelope estimation is that the envelope can only be estimated reliably at harmonic peaks. Especially with relatively high harmonic frequencies, this results in a low resolution of the spectral envelope. One method to circumvent this low resolution is to observe the signal over multiple frames during pitch modulation such as vibrato. As discussed in [40], for a sung vowel at 200 Hz with a vibrato ranging ± 50 cents, half a semitone, a continuous sampling of the spectral envelope over one vibrato period can be achieved from approximately 3.3 kHz upwards. However, this approach assumes that, over one vibrato period, neither the excitation signal nor the filter changes. Early tests with the dataset used in this thesis showed that this assumption rarely held true to a degree that multi-frame envelope estimation could be utilized.

2.3.2 Iterative Adaptive Inverse Filtering

The Iterative Adaptive Inverse Filtering (IAIF) [41] family of estimation methods, including the recent contributions iterative optimal preemphasis-IAIF (IOP-IAIF) [42] and glottal flow model-IAIF (GFM-IAIF) [43], are based on assumptions about spectral characteristics of source and filter to perform a separation and joint estimation of both. Summaries and comparisons of all three methods can be found in [44] and [45]. All three models assume, that the spectral tilt in speech signals is produced mainly by the

glottal flow and lip radiation while the vocal tract can be assumed to contain no to little spectral tilt over a large frequency range. From this consideration, the methods iteratively perform LPC estimations and inverse filtering to separate glottal flow, vocal tract and lip radiation.

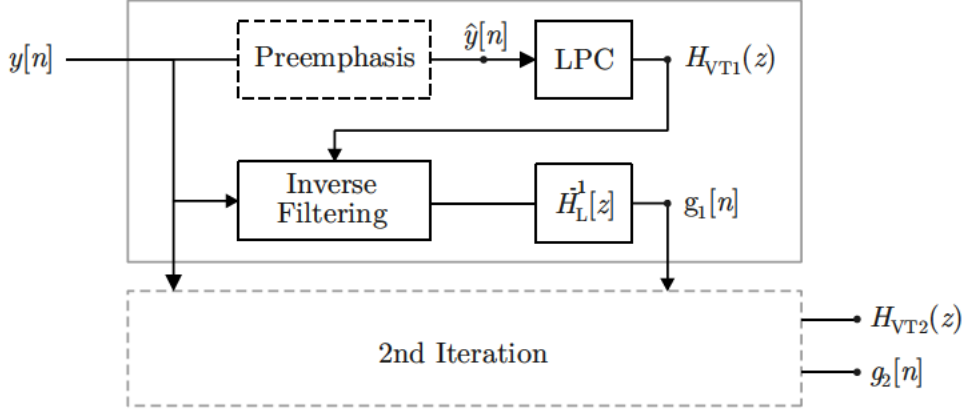


Figure 2.10: An overview of IAIF methods [44]. The first iteration produces a crude estimate of the glottal flow signal $g_1[n]$ which is used in the second iteration to refine both the glottal flow estimate $g_2[n]$ and the all-pole vocal tract model $H_{VT2}[n]$

Generally, all three methods share the same general structure as shown in figure 2.10. First, after an initial preemphasis of the observed spectrum $y[n]$ to remove the spectral tilt of glottal flow and lip radiation, LPC is used to estimate an all-pole vocal tract transfer function $H_{VT1}[z]$. Using inverse filtering, the glottal flow plus radiation signal can be extracted from the original signal. Finally, the glottal flow signal $g_1[n]$ is extracted by removing the lip radiation effect through inverse filtering of the lip radiation $H_L[z]$ which is assumed to be a differentiator. With the initial glottal flow estimate, the process can be repeated in a second iteration, this time by using the glottal flow estimate $g_1[n]$ as the basis of the preemphasis. The three methods differ mostly in the preemphasis step. While the original IAIF uses a first order LPC analysis with following inverse filtering to remove the spectral tilt with a maximum of -6 dB / octave, both IOP-IAIF and GFM-IAIF practically use a cascade of 1st order all-pole filters, each obtained by iterative LPC analysis and inverse filtering, to more reliably remove any spectral tilt. During early tests with IAIF-based methods, it was decided to not use these methods in this thesis mainly for three reasons.

- IAIF methods inherit the harmonic overfitting issues of LPC as described above. Though it should be mentioned that these issues might be solvable by using TE instead.
- Many audio samples in the dataset include observable notches in the spectrum which are assumed to correspond to zeros in the vocal tract transfer function. These zeros might be a result of nasal coupling or other effects. Due to their predominance in nearly every analyzed audio sample, it was decided to use pole-zero filters instead of all-pole filters to possibly model these zeros.

- The fundamental assumption of IAIF-based methods, that the vocal tract does not include any spectral tilt, seems unrealistic. This is suggested in many publications, especially from the physical modeling domain, which describe the various effects of energy loss and damping that can be observed in the vocal tract [46, 30, 47, 48].

2.3.3 Analysis-by-Synthesis

Analysis-by-Synthesis refers to the use of optimization methods to estimate synthesis parameters for both vocal tract and source by minimizing a reconstruction loss. While these methods have been around since the 1960s [49], they recently grew in popularity due to advances in global optimization and machine learning [50, 51, 52]. While the specifics of the method vary with the used synthesis model, optimization method and reconstruction loss, the principle stays similar. With frame based methods, a target spectrum $Y_n(\omega)$ at frame n is to be reconstructed with a synthesis method \mathcal{S} from parameters θ_n with

$$\hat{Y}_n(\omega) = \mathcal{S}(\theta_n, \omega) \quad (2.20)$$

where $\hat{Y}_n(\omega)$ is the reconstruction of $Y_n(\omega)$. With the reconstruction error $L(Y_n(\omega), \hat{Y}_n(\omega))$, the optimization can be described as

$$\arg \min_{\theta_n} L(Y_n(\omega), \hat{Y}_n(\omega)) \quad (2.21)$$

The previously proposed analysis-by-synthesis methods differ mainly in three aspects. The first is the choice of synthesis model \mathcal{S} . Within source-filter models, filters can be based on all-pole (ARX) models [53, 54], pole-zero (ARMA / ARMAX) models [55], Kelly-Lochbaum based tube models [30] or chain matrices [34, 56]. The glottal source can be approximated using the four-parameter LF model [57, 54], the Rosenberg-Klatt model [58], or any other deterministic glottal flow model such as LF-Rd. The second choice is regarding the optimization methods. To name a few, methods have been proposed based on differential evolution [57], quasi-Newton methods [34], state-space methods and Kalman filtering [58, 53], particle filtering [59, 60, 56], automatic differentiation [52] and various custom optimization schemes. Finally, a choice has to be made about the reconstruction loss L which calculates the distance between the target spectrum and its reconstruction. The loss could be calculated in time domain or spectral domain with or without consideration of the phase component or in the cepstral domain. Additionally, methods could optimize audio frames individually or considering temporal dependencies either by restricting or punishing variations in parameter trajectories or by choosing optimization approaches that implicitly consider temporal dependencies.

2.3.4 Minimum Phase Assumption

In order to fully utilize the models for glottal flow and vocal tract, one can incorporate phase observations into source filter estimation methods. As described in the CALM model in Section 2.2.1, the glottal pulse can be approximated by an anti-causal complex conjugate pole pair in the open phase and a causal real pole in the return phase. While models for the vocal tract generally predict a minimum-phase system, the anti-causal pole pair in the glottal flow introduces a maximum-phase component [11]. This is especially helpful for the distinction between glottal formant from vocal tract formants as they both appear as second order poles in the magnitude spectrum. Under this mixed-phase assumption, phase-based source-filter estimation methods separate the speech spectrum in a minimum-phase system representing the vocal tract and a maximum-phase system representing the glottal pulse [11].

2.4 Machine Learning in Voice Synthesis

In recent years, various methods for voice synthesis using machine learning methods have been proposed. Specific attention is given to WaveNet [61] and differentiable digital signal processing (DDSP) [52]. Other notable methods for machine-learning-based voice synthesis include WGANsing [62], WaveRNN [63] and a convolutional neural network approach presented in [64].

2.4.1 WaveNet

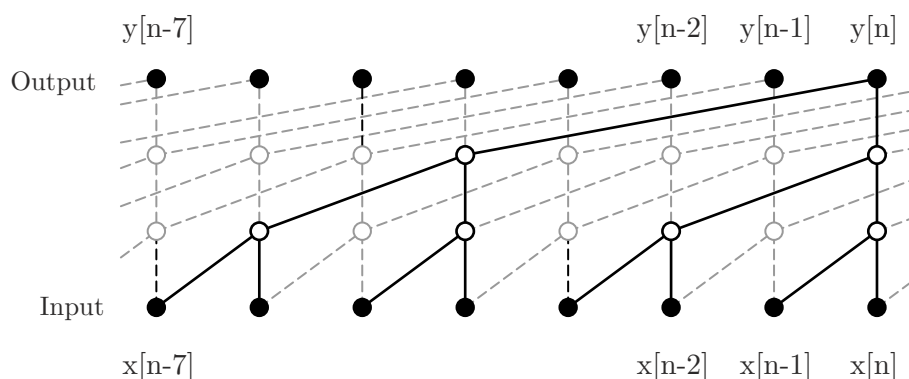


Figure 2.11: The structure of the fully convolution, dilated WaveNet [61]. Every node in the output sequence $y[n]$ depends on input samples $x[n]$ to $x[n - 2^O - 1]$ where O is the number of hidden layers in the network. [61]

In 2016, WaveNet was presented as a method to generate raw audio using dilated fully convolutional neural networks as shown in Figure 2.11 [61]. A benefit of the proposed structure is that, depending on the order O of the network, every output node is connected to every input node up to 2^O samples in the past by exactly one path. This structure reduces the complexity of the network compared to other convolutional structures. The

network is trained on a sequence $x[n]$ and tries to predict the sequences next sample with $x[n + 1] \triangleq y[n]$. During generation, the previous prediction $y[n]$ is used as the next sample in the input sequence.

$$x[n + 1] \triangleq y[n] \quad (2.22)$$

The result is that the network generates a new sequence following the patterns it learned during training. By introducing conditioning signals, such as speaker or phoneme identities, new signals can be created in a controlled manner. In [61], a WaveNet was used for text-to-speech applications. In [51], WaveNet is used to predict control parameters for the vocoder WORD [65]. The proposed method can be controlled with high level control parameters, a musical score and phenom timing tables to produce a fully phrased singing voice.

2.4.2 Differentiable Digital Signal Processing (DDSP)

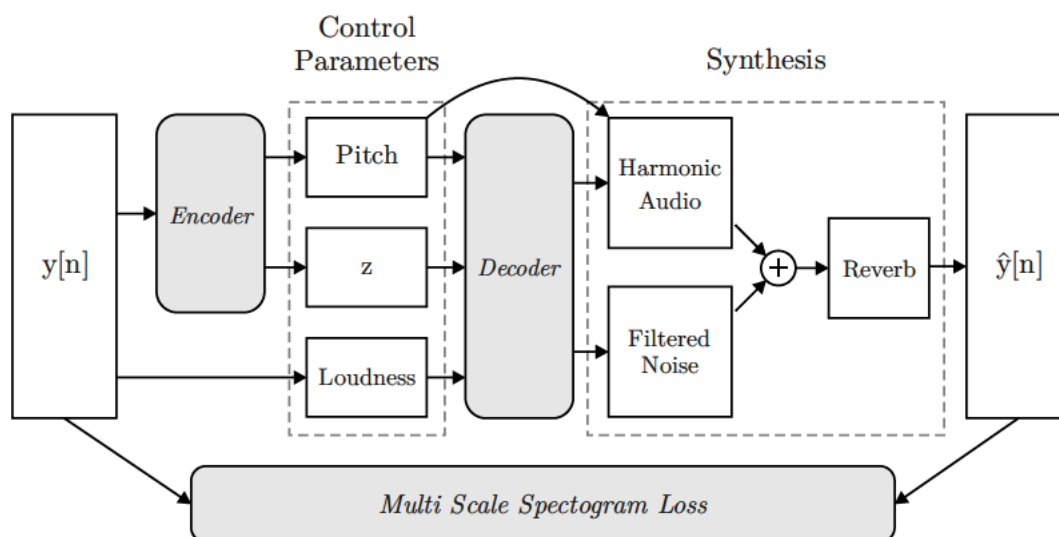


Figure 2.12: The neural network of DDSP. Control parameters pitch and loudness are extracted from the source audio $y[n]$. The decoder trains to predict synthesis parameters for the synthesis layer to reconstruct the source audio as $\hat{y}[n]$. Graphic altered from [52, p. 6]

Recently, DDSP [52] was presented as a method to integrate DSP algorithms such as oscillators or filters in an End-to-End neural network for analysis and synthesis of various audio signals. During training, control parameters such as pitch and loudness are extracted from a source signal $y[n]$. These control parameters are fed into a decoder that estimates synthesis parameters for the following integrated synthesis stage, producing $\hat{y}[n]$. Finally, the training loss is defined as the multi-scale spectrogram loss between the synthesized signal $\hat{y}[n]$ and the source signal $y[n]$. The network can be seen in Figure 2.12. Once trained, the model can be used to synthesize new signals from new control parameter trajectories. The synthesis method used relies on additive synthesis and

FIR filters. As these methods are differentiable, they can be implemented as part of a neural network while allowing loss errors to propagate back to the estimated synthesis parameters during training. This approach has two advantages. First, DSP methods like oscillators and linear filters are well understood and can be monitored and controlled easily, unlike neural networks which often are considered black boxes. Second, DSP methods can be implemented computationally efficiently on various platforms.

For voice synthesis, one drawback of the presented method is the lack of a support for IIR filters. As already discussed, IIR filters are often used in voice synthesis due to their ability to efficiently model the physical behavior of the vocal tract. While IIR filters can be implemented as part of neural networks, their use drastically increases the networks complexity and thus decreases training performance. This is caused by the theoretically infinite impulse response length characterizing IIR filters which produce large backpropagation trees in neural networks. Very recently, the lack of IIR support in DDSP was addressed in [66] with the presentation of a method for including IIR filter inside deep-learning networks. Additionally, a method for equalizer parameter estimation using neural networks was presented in [67]. Even though both papers share similarities with the method presented in this thesis, they were published after development of the proposed method was already concluded.

2.5 Automatic Differentiation

Automatic differentiation or automatic gradient computation refers to the use of APIs or other code to automatically calculate symbolic derivatives. For instance, TensorFlow's automatic gradient computation can be used to calculate derivatives $\delta y / \delta x$ for some function $y = f(x)$ where $f(x)$ might include a series of complex arithmetic operations [68]. Let's consider an exemplary pole-zero filter parameterized with gain g , poles $\mathbf{p} = (p_j)$ and zeros $\mathbf{q} = (q_j)$. The filters frequency response at complex frequencies $z = e^{i2\pi f / f_s}$ is given with

$$V(z) = g \prod_{j=1}^J \frac{(z - q_j)(z - \bar{q}_j)}{(z - p_j)(z - \bar{p}_j)} \quad (2.23)$$

where poles and zeros are further separated into radii $\mathbf{r}_\mathbf{p}$ and $\mathbf{r}_\mathbf{q}$ as well as frequencies $\mathbf{f}_\mathbf{p}$ and $\mathbf{f}_\mathbf{q}$ with

$$p = r_p e^{i2\pi f_p / f_s}, \quad q = r_q e^{i2\pi f_q / f_s} \quad (2.24)$$

with sample rate f_s . The filters frequency response can be calculated in TensorFlow as follows

```

def response(z, p, q, g):
    # frequency response numerator and denominator
    H_num = (z - q) * (z - tf.math.conj(q))
    H_den = (z - p) * (z - tf.math.conj(p))

    # calculate transfer function
    V = g * tf.reduce_prod(H_num / H_den, 1)

```

where p , q , g and z correspond to \mathbf{p} , \mathbf{q} , g and z . This function first calculates the nominator H_num and denominator H_den of the frequency response for each complex conjugate pole and zero pair and afterwards calculates the product itself with `tf.reduce_prod` before applying the gain factor. By implementing the function using TensorFlow's type system, automatic differentiation can be used to calculate derivatives, for instance $\delta V / \delta g$, $\delta V / \delta \mathbf{r}_p$, $\delta V / \delta \mathbf{f}_p$ and so on.

In the example above, the symbolic derivatives may be calculated using the function `tf.gradients()`. Let's assume automatic differentiation is used to update g , \mathbf{r}_p , \mathbf{r}_q , \mathbf{f}_p and \mathbf{f}_q so that V approximates some target frequency response V_{target} . Using TensorFlow, a gradient descent scheme might then be implemented as follows.

```

while it < it_max:
    // calculate poles and zeros from real parameters
    p = r_p * tf.exp(tf.complex(0., 2 * np.pi * f_p / f_s))
    q = r_q * tf.exp(tf.complex(0., 2 * np.pi * f_q / f_s))

    // calculate V
    V = response(z, p, q, g)

    // calculate loss, e.g. mean squared error
    L = loss(V, V_target)

    // gradient calculating for all optimized parameters
    dL_drp = tf.gradients(L, r_p)
    dL_drq = tf.gradients(L, r_q)
    dL_dfp = tf.gradients(L, f_p)
    dL_dfq = tf.gradients(L, f_q)
    dL_dg = tf.gradients(L, g)

    // update all real parameters
    r_p -= learning_rate * dL_drp
    r_q -= learning_rate * dL_drq
    f_p -= learning_rate * dL_dfp
    f_q -= learning_rate * dL_dfq
    g -= learning_rate * dL_dg

```

The use of automatic differentiation drastically simplifies optimization problems for which calculating the symbolic derivative is not trivial. It is used to incorporate DSP methods in neural networks in DDSP [52] and it is used in this work to solve multiple optimization problems.

2.6 B-Spline Modeling

This section gives a brief overview of B-splines as they are used in the method proposed in this thesis. B-splines are commonly used for parameter regression or curve fitting. B-splines are piecewise polynomial basis functions described by their order o and knot sequence s [69; 70, p. 34]. In this thesis, $o = 1$, $o = 2$ and $o = 3$ are used to refer to constant, linear and quadratic basis functions respectively. L basis functions $b_l(t)$ can be linearly combined to form a smooth curve $u(t)$ with

$$u(t) = \sum_{l=1}^L c_l b_l(t) \quad (2.25)$$

with coefficients c_l determining the amplitude of each basis function.

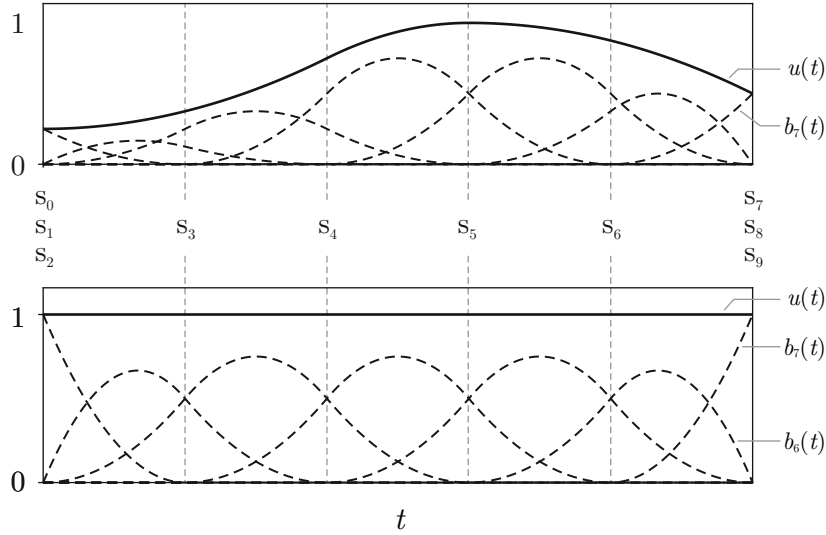


Figure 2.13: A 3rd order B-spline curve consisting of 7 basis function $b_l(t)$, $1 \leq l \leq 7$ (dotted) to form a smooth spline $u(t)$ (solid). The lower graph shows a curve with coefficients $c_l = 1$. The upper graph shows some exemplary curve with different c_l for every basis function.

Figure 2.13 shows two B-spline curves $u(t)$ with their associated basis functions $b_l(t)$. The lower graph shows a curve with $c_l = 1$, in which case the linear combination $u(t) = 1$. The graph above shows an example with individually adjusted weights and the resulting curve. Given a B-spline with L coefficients $\mathbf{c} = c_l \in \mathbb{R}^{L \times 1}$ and matrix representing L discrete basis functions with M samples $\mathbf{B} \in \mathbb{R}^{M \times L}$, a discrete B-spline $\mathbf{u} = u_m \in \mathbb{R}^{M \times 1}$ can be calculated with

$$\mathbf{B}\mathbf{c} = \mathbf{u} \quad (2.26)$$

B-spline modeling can be used to upsample signals. When $M/L \in \mathcal{N}$, $M/L > 1$ the basis functions b_l acts as an interpolation filter kernel at least for the inner section of the spline where the basis functions b_l are identical. In the example shown in Figure 2.13, this is the case for the basis functions b_3 to b_5 .

3 Method

Goal of this thesis was the design and development of a voice analysis and synthesis method. During analysis, a source audio sample would be separated in performance characteristics, represented by control parameter trajectories such as pitch and intensity, and a voice model capturing vowel and voice qualities. During synthesis, new control parameter trajectories could be used in combination with the voice model to create new audio samples. An overview of the method is shown in Figure 1.1.

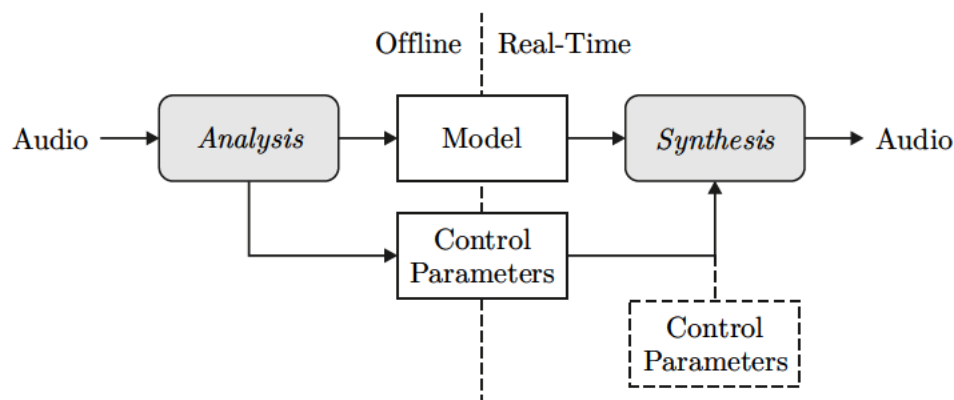


Figure 1.1: The intended method separates an audio sample into voice model and control parameters. During synthesis, the control parameters can be replaced with new ones to produce new sounds. Repeated from Page 2.

Particular attention was given on the real-time capabilities of the synthesis part. For this reason, it was decided to rely on well known DSP methods to model the voice production. The proposed synthesis model estimation method is inspired by the recently proposed DDSP [52] method which incorporated familiar DSP methods as part of a neural network. In this work, DSP methods are incorporated in an gradient descent optimizer in a similar manner.

3.1 Overview

The proposed method consists of four analysis and synthesis steps, namely, harmonic analysis \mathcal{A} , synthesis model parameter estimation \mathcal{E} , prediction network \mathcal{P} , and synthesis \mathcal{S} . First, pitch and harmonic partials are estimated from a source audio sample using the harmonic analysis method \mathcal{A} . Afterwards, synthesis parameters are estimated from the harmonic partials and pitch in the novel synthesis model estimation method \mathcal{E} . As the last step of the analysis, the synthesis parameter predictor network \mathcal{P} is trained to

predict synthesis parameters from the control parameter pitch. During synthesis, the trained network can then be used to predict synthesis parameters from a new input pitch trajectory. Afterwards, the predicted parameters are used in the synthesis method \mathcal{S} to generate audio.

In practice, such a method would be used to train a singer model from one or more source audio samples. The singer model would capture sound characteristics such as vocal tract articulation and voice qualities from the singer as a function of the control parameters, such as vowel, pitch or intensity. In this thesis, the predictor \mathcal{P} only captures the vocal tract articulation and glottal intensity as a function of the single control parameter pitch. To differentiate the proposed simplified model from more complex singer models, it is instead referred to as a voice model or as the predictor.

The outlined approach was chosen for multiple reasons. In order to achieve a high flexibility, it was decided to separate the method into multiple steps. As a result, individual stages can easily be modified or extended. For instance, the predictor network \mathcal{P} can be replaced with different neural network structures and can be scaled up or down to balance computational performance and accuracy of the prediction. Or the harmonic analysis could be replaced with alternative methods for tracing magnitude and phase of harmonic partials. In order for the method to be used in real-time applications, it was decided to rely on classic DSP methods for the synthesis \mathcal{S} .

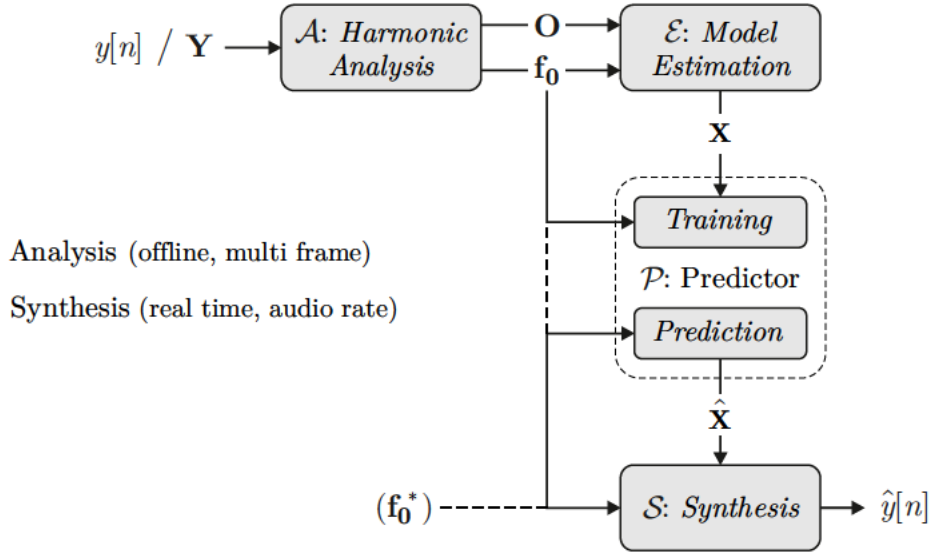


Figure 3.2: An overview of the proposed voice analysis and synthesis method. The method can be used either to resynthesize the source audio sample by reusing \mathbf{f}_0 to predict $\hat{\mathbf{X}}$ and synthesize $\hat{y}[n]$. Or to synthesize a new audio signal by using any other pitch trajectory indicated here with \mathbf{f}_0^* .

An overview of the proposed method is shown in Figure 3.2. The analysis is performed offline. An audio sample $y[n]$ is first separated into M audio frames $\mathbf{Y} \in \mathbb{R}^{N_B \times M}$ with frame size $N_B = 2048$ and hop size $N_H = 64$. Since all M frames are analyzed simultaneously and interdependently, the approach can be described as a multi-frame analysis. First, the harmonic analysis method jointly estimates pitch $\mathbf{f}_0 \in \mathbb{R}^M$ and

harmonic partials $\mathbf{O} \in \mathbb{C}^{K \times M}$ representing magnitude and phase for K harmonics and M frames.

Both pitch $\mathbf{f_0}$ and harmonic partials \mathbf{O} are then used to estimate synthesis parameters $\mathbf{X} = (\chi_m) \in \mathbb{C}^{2+2J \times M}$ for M frames by fitting the synthesis model to the previously extracted harmonic partials. This task is referred to as synthesis model parameter estimation \mathcal{E} . An exact description of χ will be given in Section 3.2.

The parameter predictor network \mathcal{P} links analysis and synthesis. During analysis, the predictor is trained to reconstruct the previously estimated synthesis parameters \mathbf{X} from pitch trajectory $\mathbf{f_0}$ and thus capturing vocal characteristics of the source audio sample in dependency of pitch.

In order to resynthesize the source audio sample $y[n]$, the predictor network \mathcal{P} is used to reconstruct synthesis parameters $\hat{\mathbf{X}} = (\hat{\chi}_m)$ from pitch trajectory $\mathbf{f_0}$ extracted during analysis before they are used with the synthesis model \mathcal{S} to synthesize the audio sample $\hat{y}[n]$.

A more likely application case would be to synthesize new audio signals in real-time, in which case the predictor network would be used to generate synthesis parameters from a freely controllable pitch signal, indicated in Figure 3.2 with $(\mathbf{f_0}^*)$. The synthesis parameters are then again used with the synthesis model \mathcal{S} . The synthesis model directly produces continuous time-domain signals in audio rate with no latency. Its parameters can be updated either in audio-rate or, to reduce the computational load, in regular intervals in which case the synthesis parameters should be interpolated between updates to prevent artifacts.

3.2 Synthesis Model \mathcal{S}

The synthesis model was chosen with real-time applications in mind with the focus being put on the computational efficiency. The approach is based on the source-filter theory and separates the voice signal $Y(z)$ into a combined glottal source and lip radiation model $S(z)$ and a vocal tract filter $V(z)$.

$$Y(z) = V(z)S(z) \quad (3.1)$$

Random aspects such as aspiration noise or vocal jitter and shimmer were neglected. The synthesis is implemented in the time domain. The source is modeled using a variable-gain wavetable oscillator and produces the signal $s[n]$. The source signal passes through a serial filter bank $V(z)$ modeling the vocal tract to produce the output signal $y[n]$. An overview is shown in Figure 3.3.

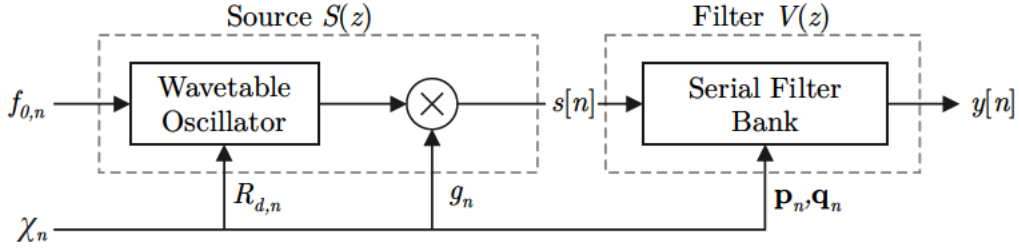


Figure 3.3: The synthesis method implements a source-filter model in the time domain. Parameters R_d and $g = 10^{g_{dB}/20}$ parameterize the source which consists of a wavetable oscillator and a gain adjustment. The filter, parameterized with \mathbf{p} and \mathbf{q} is implemented as a serial filter bank consisting of J second order sections.

Every sample, the synthesis produces a new audio sample $y[n]$ and updates its state θ_n from the current pitch $f_{0,n}$, current synthesis parameters χ_n and previous synthesis state θ_{n-1} . Thus, the method can be described with

$$y[n], \theta_n = \mathcal{S}\{f_{0,n}, \chi_n, \theta_{n-1}\} \quad (3.2)$$

The synthesis parameters χ_n at sample n are defined with

$$\chi = [R_d \ g_{dB} \ \mathbf{p} \ \mathbf{q}] \in \mathbb{C}^{2+2J} \quad (3.3)$$

where gain g_{dB} and R_d parameterize the source model and poles $\mathbf{p} = (p_j) \in \mathbb{C}^J$ and zeros $\mathbf{q} = (q_j) \in \mathbb{C}^J$ model the vocal tract. The state θ is defined with

$$\theta = [\varphi \ \mathbf{s}_1 \ \mathbf{s}_2] \quad (3.4)$$

where φ stores the phase of the wavetable oscillator and $\mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^J$ store the state the J second order sections making up the vocal tract filter bank.

3.2.1 Source Model $S(z)$

The LF- R_d model was chosen to synthesize the glottal flow mainly because of its simplicity. The glottal flow and lip radiation $H_L(\omega) = i\omega$ are modeled by the time domain LF- R_d glottal flow derivative, parameterized with R_d and amplitude g_dB . As the wave shape only depends on R_d , the source can easily be implemented as a two-dimensional wavetable oscillator [71, 72]. The oscillator implemented here uses a matrix, the wavetable, to store single cycles of the glottal flow derivative for a range of discrete values of R_d . A phase generator is used to calculate the phase φ of the oscillator from pitch f_0 . The phase and glottal parameter R_d are then used to calculate the indices and fractional positions with which to read from the wavetable using bi-linear interpolation. Finally, the gain factor g_dB is applied to the wavetable output to calculate the source signal $s[n]$. Notably, while the decibel gain factor g_dB is used to parameterize the synthesis model because of the beneficial interpolation characteristics, the linear magnitude factor $g = 10^{g_dB/20}$ is used in per-sample synthesis. A schematic of this process is shown in Figure 3.4.

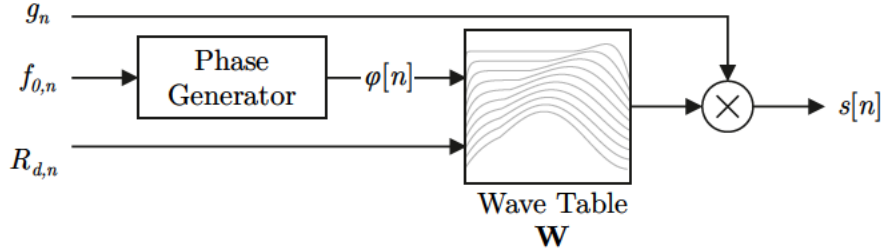


Figure 3.4: The source $s[n]$ is calculated using a phase generator driving a wavetable and a simple gain to adjust the sources amplitude.

The wavetable matrix \mathbf{W} is generated in advance. The first dimension covers the phase with $0 \leq \varphi < 2\pi$ and the second dimension covers R_d in a range of $0.3 \leq R_d \leq 2.7$.

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & \dots & w_{1,M_{R_d}} \\ \vdots & & \vdots \\ w_{M_\varphi,1} & \dots & w_{M_\varphi,M_{R_d}} \end{bmatrix} = (w_{m_\varphi,m_{R_d}}) \in \mathbb{R}^{M_\varphi \times M_{R_d}} \quad (3.5)$$

The glottal flow derivative $g(t)$ of the LF model, parameterized with t_p , t_e and T_a ($t_c = 1$), is defined in Equation 2.2.1. The conversions between R_d and t_p , t_e and T_a is defined in Equations 2.5 and 2.6. From that, the elements for the wavetable are calculated with

$$w_{m_\varphi,m_{R_d}} = g_{R_d} \left((t + t_e) \bmod 1 \right) \quad (3.6)$$

with $t = \varphi/(2\pi)$, glottal closure instance t_e and table indices m_φ and m_{R_d} given with

$$\begin{aligned} m_\varphi &= \lfloor M_\varphi \cdot \varphi/(2\pi) \rfloor \\ m_{R_d} &= \lfloor M_{R_d} (R_d - 0.3)/(2.7 - 0.3) \rfloor \end{aligned} \quad (3.7)$$

The real modulo operation mod wraps t between 0 and 1 and is defined as

$$a \bmod b = a - b[a/b] \quad (3.8)$$

By offsetting the time t with the glottal closure instance t_e as done in Equation 3.6, the main glottal flow discontinuity at t_e is shifted to a fixed position at $t = 0$ or $\varphi = 0$. This is done to reduce aliasing artifacts during interpolation as changing R_d will no longer change the position of the discontinuity. For comparison, the wavetable without this adjustment is shown in Figure 3.5 (b). As can be seen, the position of the main discontinuity varies drastically especially for low R_d values. This improvement has drastically improved the audio quality of the synthesis during R_d parameter fluctuations and allows for more rapid parameter changes without audible artifacts.

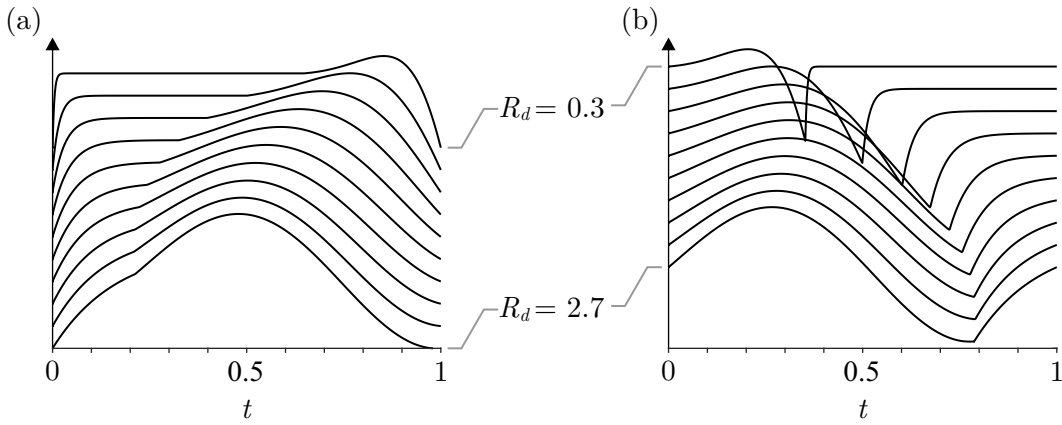


Figure 3.5: Figure (a) shows a glottal pulse derivatives as stored in \mathbf{W} . Figure (b) shows a wavetable $g_{R_d}(t)$ without t_e -offset. ($M_{R_d} = 10$. The glottal flow graphs are offset vertically for readability.)

The harmonic oscillation is produced by a phase generator, that determines the phase φ at which to read from the wavetable. The generator's phase φ is one of the state variables which make up the current state θ_n of the synthesis method. With the fundamental f_0 and sampling rate f_s , the phase is updated per sample as follows

$$\varphi[n] = \left(\frac{2\pi f_0}{f_s} + \varphi[n-1] \right) \bmod (2\pi) \quad (3.9)$$

again, where $(\varphi \bmod 2\pi)$ wraps φ to $0 \leq \varphi < 2\pi$. During synthesis, the phase is updated according to the current pitch f_0 . Afterwards, both φ and R_d are used to calculate the corresponding indices m_φ and m_{R_d} as defined in Equation 3.7. Afterwards, the fractional positions determining the sub-sample position for the wavetable lookup are calculated with

$$\begin{aligned} k_\varphi &= (M_\varphi \varphi / 2\pi) - m_\varphi \\ k_{R_d} &= (M_{R_d} (R_d - 0.3) / (2.7 - 0.3)) - m_{R_d} \end{aligned} \quad (3.10)$$

The wavetable output $s_{wt}[n]$ can be calculated by bi-linear interpolation with

$$s_{wt}[n] = (1 - k_\varphi) \left((1 - k_{R_d})w_{m_\varphi, m_{R_d}} + k_{R_d}w_{m_\varphi, m_{R_d}+1} \right) + k_\varphi \left((1 - k_{R_d})w_{m_\varphi+1, m_{R_d}} + k_{R_d}w_{m_\varphi+1, m_{R_d}+1} \right) \quad (3.11)$$

Finally, the source signal is calculated from the wavetable output and gain factor $g = 10^{g_{dB}/20}$ with

$$s[n] = g s_{wt}[n] \quad (3.12)$$

3.2.2 Vocal Tract Model $V(z)$

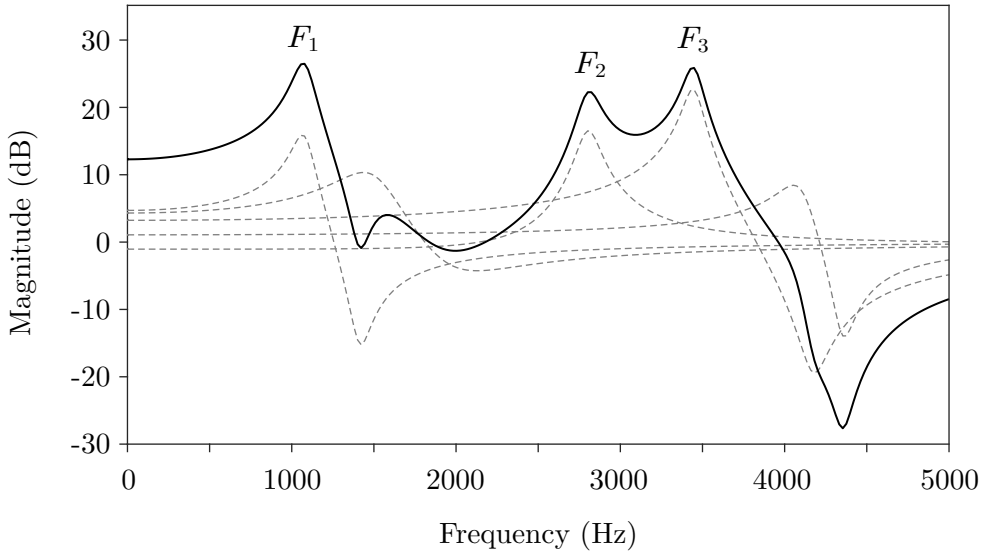


Figure 3.6: The magnitude frequency response of an exemplary vocal tract filter $V(z)$ (solid line) consisting of 5 second order sections (dotted lines). The first three formants are marked with F_1 to F_3 .

In early experiments with the VocalSet dataset [73], it was noticed that the vowel spectra rarely follow ideal all-pole frequency responses. For that reason, it was decided to use pole-zero filters to model the vocal tract [25]. A pole-zero filter with sufficiently high order is flexible enough to model a wide range of vocal tract transfer functions to a sufficient accuracy. The filter is parameterized with equal length list of J complex conjugate poles $\mathbf{p} = (p_j)$ and J complex conjugate zeros $\mathbf{q} = (q_j)$ with

$$\begin{aligned} \mathbf{p} &= (p_j) \in \mathbb{C}^J \\ \mathbf{q} &= (q_j) \in \mathbb{C}^J \end{aligned} \quad (3.13)$$

The transfer function $V(z)$ extends these poles and zeros with their respective complex conjugates \bar{p}_j and \bar{q}_j and is defined as follows

$$V(z) = \frac{Y(z)}{S(z)} = \prod_{j=1}^J \frac{(z - q_j)(z - \bar{q}_j)}{(z - p_j)(z - \bar{p}_j)} \quad (3.14)$$

As a result, the proposed filter model is limited to produce even order IIR filters with complex pole or zero pairs and can't contain single real poles or zeros. The filter is implemented as a series of J second order sections in transposed direct form II given with

$$\begin{aligned} a_{0j} y_j[n] &= b_{0j} x_j[n] + s_{1j}[n-1] \\ s_{1j}[n] &= b_{1j} x_j[n] + s_{2j}[n-1] - a_{1j} y_j[n] \\ s_{2j}[n] &= b_{2j} x_j[n] - a_{2j} y_j[n] \end{aligned} \quad (3.15)$$

where coefficients b and a can be obtained from pole and zero pairs p_j and q_j with

$$\frac{b_{0j} + b_{1j}z^{-1} + b_{2j}z^{-2}}{a_{0j} + a_{1j}z^{-1} + a_{2j}z^{-2}} = \frac{(z - q_j)(z - \bar{q}_j)}{(z - p_j)(z - \bar{p}_j)} \quad (3.16)$$

The filters are concatenated in series with $x_{j+1}[n] = y_j$ with output and $y[n] = y_J[n]$ and input $x_1[n] = s[n]$. The state of the j th section is stored in two variables $\mathbf{s}_1 = (s_{1j}) \in \mathbb{R}^J$ and $\mathbf{s}_2 = (s_{2j}) \in \mathbb{R}^J$. In practice, a value of $J = 10$ was chosen as a good compromise between quality and complexity. An exemplary vocal tract frequency response with $J = 5$ is shown in Figure 3.6.

3.2.3 Summary

The synthesis model proposed in this thesis is based on the source-filter theory and is implemented using basic DSP methods in the time domain. The glottal source and lip radiation are modeled using a two-dimensional wavetable oscillator with parameters f_0 and R_d . The vocal tract is represented by a filter bank consisting of J second order sections in series.

In its current form, the synthesis model uses two parameter, R_d and g_{dB} , to control the glottal source while most of it's flexibility stems from the pole-zero model used for the vocal tract which is parameterized with J poles and zeros. This produces some discrepancy in terms of flexibility or degrees of freedom between glottal source and vocal tract which affects how the model can be fitted to audio samples.

Modeling the glottal source with a two-parameter model, one controlling the glottal formant and the other controlling the spectral tilt, might produce better results as it allows the source to model a wider variety of real glottal flow signals.

The filter model is parameterized with J complex poles \mathbf{p} and zeros \mathbf{q} , for which the conjugate pairs are added during synthesis. While this approach allows makes the synthesis model in its current form very flexible, it also produces some redundancy. For

instance, two sets of parameters χ could produce identical signals when the orders of otherwise identical poles and zeros in the filter bank are scrambled. In addition, there exists no clear association between poles and zeros in the filter model and formants and anti-formants in the targeted vocal tract transfer function. The formants and antiformants produced by the vocal tract depend on its articulation. This negatively impacts the model's ability to morph between two parameter sets.

Ideally, when interpolating between two parameter sets of a vocal tract model, the interpolated poles and zeros should follow the formants and anti-formants of the associated morphs between two vocal tract articulations. This can be accomplished by some tube-based physical models. With the proposed vocal tract model however, this interpolation behavior can't be guaranteed. Instead, it is likely that a direct interpolation between two parameter sets \mathbf{p}_1 , \mathbf{q}_1 , and \mathbf{p}_2 , \mathbf{q}_2 , even after ordering the individual poles and zeros by frequency, produces unconvincing results and noticeable spectral artifacts.

To improve this characteristic of the proposed vocal tract model, one would need to take inspiration from the existing physical models which are often parameterized using a representation of the vocal tract articulation instead of parameterizing the frequency response it produces.

Finally, in order to improve the computational efficiency, the vocal tract filter might be implemented as a parallel filter bank instead of a serial filter bank. Parallel filter banks can take advantage from modern computer CPUs, which allow second order sections to be processed simultaneously in groups of 4 or 8 using SIMD operations. However, it should be tested whether or not this performance gain outweighs the additional overhead that comes from calculating the coefficients for a equivalent parallel filter bank from the filter poles \mathbf{p} and zeros \mathbf{q} .

3.3 Harmonic Analysis \mathcal{A}

A custom harmonic analysis method \mathcal{A} was developed that jointly estimates pitch $\mathbf{f}_0 = (f_{0m}) \in \mathbb{R}^M$ and harmonic partials \mathbf{O} from audio frames \mathbf{Y} .

$$\mathbf{O}, \mathbf{f}_0 = \mathcal{A}\{\mathbf{Y}\} \quad (3.17)$$

The complex harmonic partial $o_{k,m}$ represent magnitude and phase of the k -th partial at frequency kf_0 in frame m . The first $1 \leq k \leq K$ partials of the harmonic signal for M frames are stored in \mathbf{O} with

$$\mathbf{O} = \begin{bmatrix} o_{1,1} & \dots & o_{1,M} \\ \vdots & & \vdots \\ o_{K,1} & \dots & o_{K,M} \end{bmatrix} = (o_{k,m}) \in \mathbb{C}^{K \times M} \quad (3.18)$$

The audio frames are stored in \mathbf{Y} with M frames of lengths N_B each with

$$\mathbf{Y} = \begin{bmatrix} y_{0,1} & \dots & y_{0,M} \\ \vdots & & \vdots \\ y_{N_B-1,1} & \dots & y_{N_B-1,M} \end{bmatrix} = (y_{n,m}) \in \mathbb{R}^{N_B \times M} \quad (3.19)$$

This section describes the approach to estimate pitch and harmonic partials for any audio sample $x[n]$ though in practice, the method would analyze the audio frames simultaneously. Given an audio frame $x[n]$, one can separate the frame into a deterministic component $\hat{x}[n]$ and a residual component $r[n]$ as is discussed in [74].

$$x[n] = \hat{x}[n] + r[n] \quad (3.20)$$

The harmonic analysis methods proposed here is based on the assumption that, for voiced speech and sung vowels, most energy within one audio frame $x[n]$ is stored in the harmonic signal component $\hat{x}[n]$. By minimizing the difference between $x[n]$ and $\hat{x}[n]$, one can estimate the pitch f_0 , pitch slope D and harmonic partials (o_k) . Here, the pitch slope is defined as the constant change of pitch over time with $D = 100$ corresponding to a change of 100 Hz per second. The estimation of pitch f_0 , pitch slope D to minimize the loss L between $x[n]$ and $\hat{x}[n]$ for a single frame poses an optimization problem with

$$\arg \min_{f_0, D} L(x[n], \hat{x}[n]) \quad (3.21)$$

Afterwards, the harmonic partials $(o_k) = [o_1, \dots, o_K]$ can be derived from $x[n]$, f_0 and D . This optimization problem is solved for M frames simultaneously using a gradient descent method, an overview of which is shown in figure 3.7.

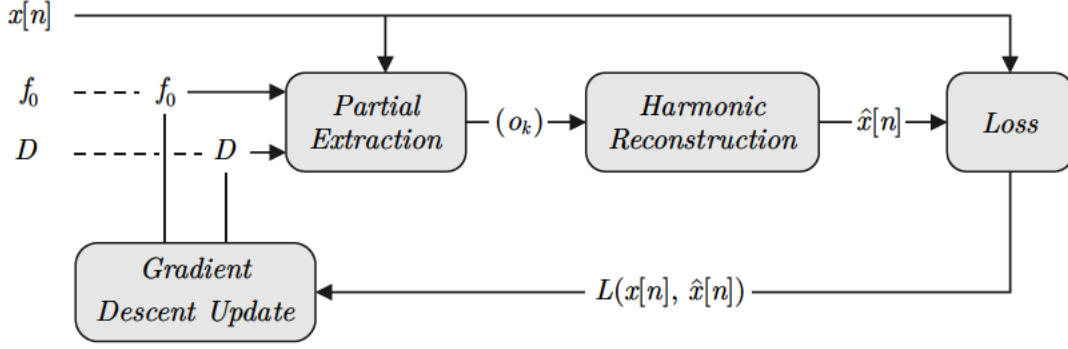


Figure 3.7: The proposed harmonic analysis estimates pitch f_0 , pitch slope D and harmonic partials (o_k) by minimizing the loss between frame $x[n]$ and harmonic reconstruction $\hat{x}[n]$. In addition to $x[n]$, the method requires some initial values for pitch and pitch slope, implied here by the dotted lines.

3.3.1 Partial Extraction and Harmonic Reconstruction

Assuming a frame $x[n]$ consists of a harmonic signal composed of harmonic partials (o_k) with associated partial frequencies $f_k = kf_0$, the k th harmonic partial can be extracted using the discrete time Fourier transformation

$$o_k = \frac{1}{N_B} \sum_{n=0}^{N_B-1} x[n] e^{-i2\pi n f_k / f_s} \quad (3.22)$$

where N_B is the block size and f_s is the sampling rate. However, early results have shown that, with $N_B = 2048$ and $f_s = 44100$, the fundamental frequency f_0 is under significant movement even within one frame. As a result, the frequency of each harmonic partial is under modulation and the energy of each partial is spread to neighboring frequency bins of the Fourier transformed spectrum. Figure 3.8 shows two magnitude spectra reconstructed from the same analyzed audio frame with the currently described method. As can be seen, the harmonic signal under modulation ($D = 823 \text{ Hz/s}$) drops in magnitude especially for high harmonics as their energy is spread across a wider frequency band within the frame.

To circumvent this issue, it was decided to compensate for this by modeling pitch as a linear function instead of a constant. The frequency within one frame is defined by the constant frequency f_0 and the linear frequency change D , referred to as frequency slope. The phase φ_k of the k th harmonic partial is then defined as a quadratic expression

$$\varphi_k[n] = 2\pi k (f_0 t[n] + 0.5 D t[n]^2) \quad (3.23)$$

with time $t[n]$ ranging from $t[0] = -0.5N_B/f_s$ to $t[N_B - 1] = 0.5N_B/f_s$. This time range was chosen as it positions the time at which the linear frequency is equal to the constant term f_0 to the center of the frame.

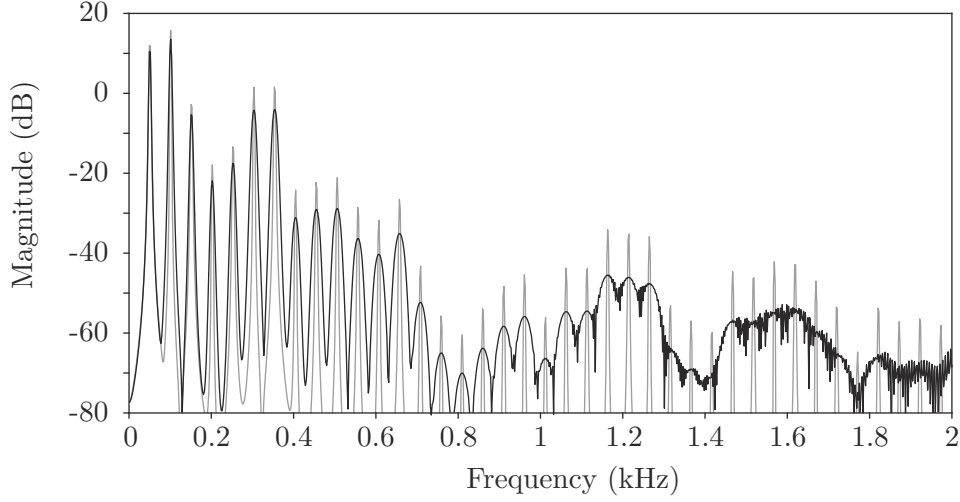


Figure 3.8: Magnitude spectrum of a vowel *a* reconstructed from extracted harmonic partials with original pitch $f_0 = 505$ Hz with fixed pitch slope $D = 0$ Hz/s (gray) and $D = 823$ Hz/s (black).

With the quadratic phase term, the k th partial o_k can be extracted from frame $x[n]$ with

$$o_k = \frac{1}{N_B} \sum_{n=0}^{N_B-1} w[n]x[n]e^{-i\varphi_k[n]} \quad (3.24)$$

where $w[n]$ is the Hann window function. From this, the harmonic reconstruction $\hat{x}[n]$ of frame $x[n]$ can be calculated with

$$\hat{x}[n] = \sum_{k=1}^K \left(e^{i\varphi_k[n]} o_k + e^{-i\varphi_k[n]} \bar{o}_k \right) \quad (3.25)$$

3.3.2 Joint Optimization

Based on the previously described harmonic reconstruction method, an optimization method can be used to minimize the loss L , the root-mean-squared difference between the original frame $x[n]$ and the harmonic reconstruction $\hat{x}[n]$ with

$$L = \sqrt{\frac{1}{N_B} \sum_{n=0}^{N_B-1} (w[n]\hat{x}[n] - w[n]x[n])^2} \quad (3.26)$$

again where $w[n]$ is the Hann window function. In practice, a first estimation for f_0 and D used to compute $\hat{x}[n]$ is necessary, after which a gradient descent method [75] is sufficient to minimize L . The gradient descent method iteratively calculates the loss and updates the estimates for f_0 and D at iteration i with

$$f_{0,i+1} = f_{0,i} - r_{f_0} \frac{\delta L_i}{\delta f_{0,i}}, \quad D_{i+1} = D_i - r_D \frac{\delta L_i}{\delta D_i} \quad (3.27)$$

where $r_{f_0,i}$ and $r_{D,i}$ are learning rates for parameters f_0 and D which decrease exponentially with every iteration with

$$r_{f_0,i} = c_r r_{f_0,i-1}, \quad r_{D,i} = c_r r_{D,i-1} \quad (3.28)$$

where $c_r < 1$. For an initial estimate of $f_{0,i=0}$, the CREPE pitch estimator [76] is used while the initial pitch slope is set to $D_{i=0} = 0$. An exemplary estimate for both pitch and pitch slope estimate as well as the original CREPE pitch estimate is shown in Figure 3.9.

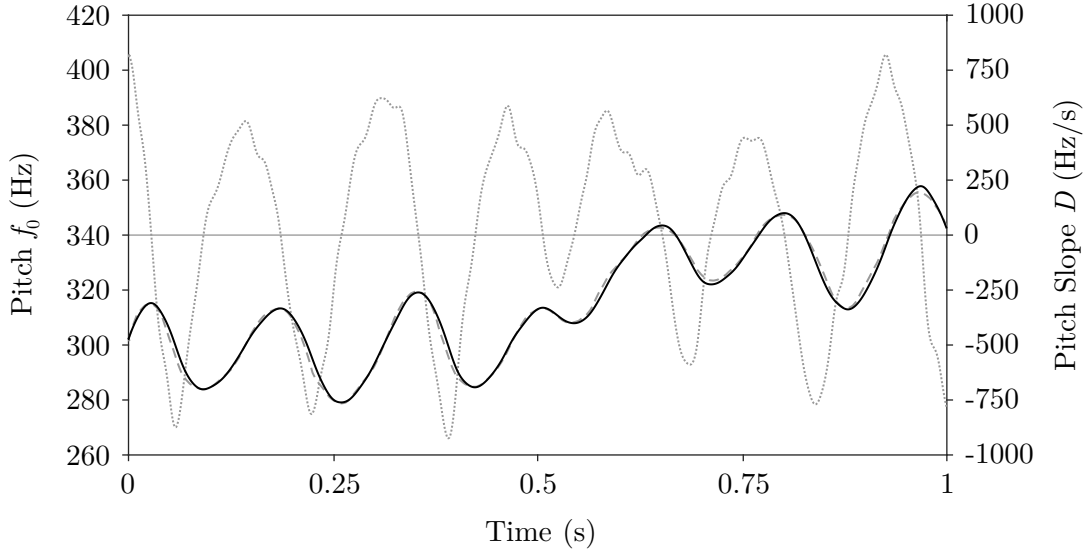


Figure 3.9: This Figure shows the pitch estimate (solid black) and CREPE pitch estimate (dashed gray) as well as the pitch slope estimate (dotted gray) over a duration of one second.

The gradient descent optimization was implemented using TensorFlow [68] with the help of automatic differentiation. All operations shown in Figure 3.7 are implemented within TensorFlow’s type system so that symbolic derivatives $\delta L/\delta f_0$ and $\delta L/\delta D$ can easily be calculated. In addition, the optimization is performed simultaneously for a batch of $M_B < M$ frames to take advantage of parallelization capabilities on modern GPUs. In this work, a NVIDIA GeForce RTX 2080 Super was used to simultaneously estimate pitch and pitch slope in batches of $M_B = 4096$ frames over 600 gradient descent iterations.

3.3.3 Summary

In summary, the harmonic analysis method is based on the assumption that most energy within one frame $x[n]$ is stored in the harmonic signal component $\hat{x}[n]$. To separate the harmonic component from the remaining residual $r[n]$ and simultaneously estimating both pitch f_0 and pitch slope D , a gradient descent optimization method is used to minimize the root-mean-squared difference between original frame $x[n]$ and harmonic reconstruction $\hat{x}[n]$. With estimated pitch and pitch slope, the harmonic partials (o_k) can be calculated by complex division following Equation 3.22. The approach requires

an initial estimate for both pitch and pitch slope of a frame which is supplied by the CREPE pitch estimator [76]. An exemplary frequency and magnitude estimation for the first few harmonics of a spectrum is shown in figure 3.10. As can be seen, the magnitude of the estimated partials diverge slightly from the underlying magnitude peak of the analyzed spectrum due to the energy bleeding to neighboring bins as discussed above.

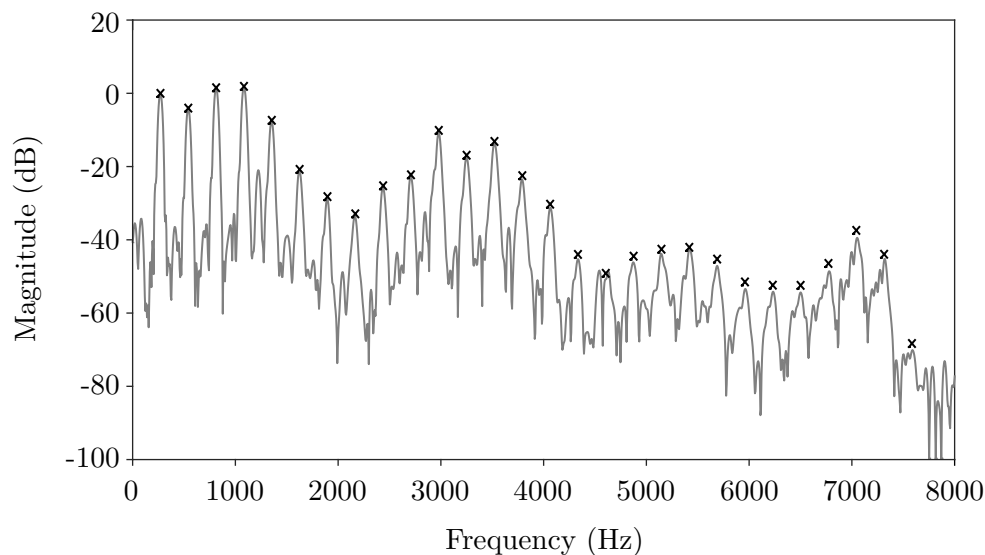


Figure 3.10: This figure shows the estimated frequency and magnitude of the first few harmonics (crosses) together with the original spectrum.

One advantage of the proposed method is its ability to model pitch within one frame as a linear function over time instead of a constant. This practically removes the impact of linear pitch modulation from the estimation of the harmonic partials and drastically simplifies the following analysis steps. This approach could also easily be extended with a quadratic term to allow for more complex pitch trajectories to be modeled within a frame.

A downside of the method has to do with the base assumption that harmonic partials are located at perfect multiples of the fundamental frequency, $f_k = kf_0$. It was noticed that the assumption does not hold true in practice. Especially high frequencies diverge from their ideal harmonic frequency under modulation. This might be explained by the frequency specific phase delay of the vocal tract or by physical properties of the vocal folds vibration. Furthermore, the proposed overtone extraction method fails in the presence of vocal jitter or shimmer, in which case the glottal source can't be considered harmonic.

3.4 Synthesis Model Parameter Estimation \mathcal{E}

To estimate synthesis parameters \mathbf{X} from harmonic partials \mathbf{O} and pitch \mathbf{f}_0 , the synthesis model parameter estimation method \mathcal{E} is used. Goal of the method is to find a set of synthesis parameters \mathbf{X} with which the previously extracted harmonic partial trajectories \mathbf{O} can be best approximated using the synthesis model \mathcal{S} . The method is outlined as follows

$$\mathbf{X} = \mathcal{E}\{\mathbf{O}, f_{0,m}\} \quad (3.29)$$

This task poses an optimization problem, in which the synthesis parameters \mathbf{X} have to be optimized to minimize the loss L , described as the difference between the previously estimated harmonic partials \mathbf{O} and the partials $\hat{\mathbf{O}}$ synthesized from \mathbf{X} .

$$\arg \min_{\mathbf{C}} L \quad (3.30)$$

where parameters \mathbf{X} are derived from \mathbf{C} as will be described in Section 3.4.2. Note that in practice, the loss L takes additional factors into consideration as will be discussed in Section 3.4.4. A gradient descent optimizer is used to solve this problem.

As described in Section 3.3, the harmonic partials $\mathbf{O} = (o_{k,m})$ define magnitude and phase of K harmonics for M frames. The synthesis parameter matrix $\mathbf{X} = (\chi_m)$ describes the synthesis parameter set for M frames with

$$\mathbf{X} = \begin{bmatrix} R_{d1} & \dots & R_{dM} \\ g_{dB1} & \dots & g_{dB M} \\ p_{1,1} & \dots & p_{1,M} \\ \vdots & & \vdots \\ p_{J,1} & \dots & p_{J,M} \\ q_{1,1} & \dots & q_{1,M} \\ \vdots & & \vdots \\ q_{J,1} & \dots & q_{J,M} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_d \\ \mathbf{g}_{dB} \\ \mathbf{P} \\ \mathbf{Q} \end{bmatrix} = (\chi_m) \in \mathbb{C}^{2+2J \times M} \quad (3.31)$$

with

$$\begin{aligned} \mathbf{R}_d &= (R_{dm}) \in \mathbb{R}^{1 \times M} \\ \mathbf{g}_{dB} &= (g_{dBm}) \in \mathbb{R}^{1 \times M} \\ \mathbf{P} &= (p_{j,m}) \in \mathbb{C}^{J \times M} \\ \mathbf{Q} &= (q_{j,m}) \in \mathbb{C}^{J \times M} \end{aligned} \quad (3.32)$$

As described in Section 3.2, the synthesis parameter set consists of two parameters, R_d and g_{dB} , modeling the glottal source and $2J$ parameters, \mathbf{p} and \mathbf{q} modeling the vocal

tract. Since the method simultaneously estimates the parameters for M frames, these parameters trajectories are represented in \mathbf{X} as vectors with M elements. In Equation 3.31, the block matrix notation is used to describe \mathbf{X} by its component parameter vectors or matrices.

3.4.1 Method Overview

The proposed method takes full advantage of TensorFlow's automatic differentiation by implementing the synthesis method \mathcal{S} as part of a fully differentiable network. In short, the network is used to adjust variables \mathbf{C} , described later in more detail, to minimize some loss L_{dB} between the previously extracted harmonic partial magnitudes $\mathbf{O}_{dB} = 20 \log_{10} |\mathbf{O}|$ and the synthesized magnitudes $\hat{\mathbf{O}}_{dB}$. As will be discussed in Section 3.4.4, the loss L which is minimized additionally includes a temporal regularization loss term L_T calculated from \mathbf{C} . The synthesis is implemented in spectral domain by calculating decibel magnitudes of the synthesized harmonic partials instead of in time domain. An overview of the network is shown in Figure 3.11.

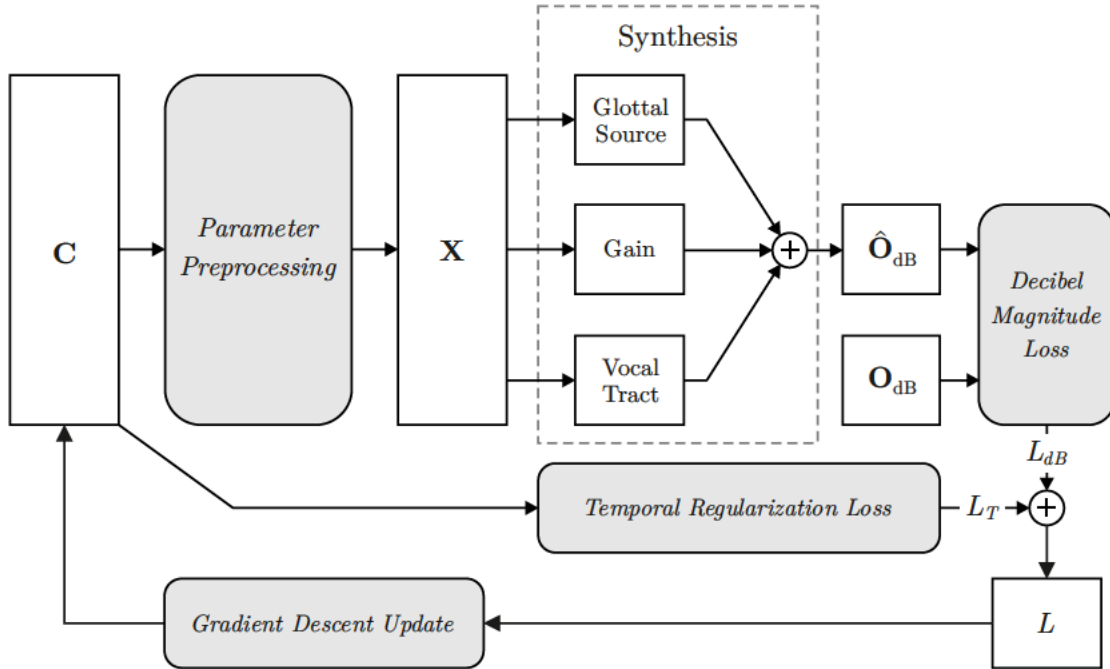


Figure 3.11: The general structure of the proposed synthesis model parameter estimation method.

Starting with variables \mathbf{C} being optimized, the synthesis parameters \mathbf{X} are first obtained from \mathbf{C} through parameter preprocessing. The synthesis layer calculates the per-frame harmonic partials in decibel magnitude $\hat{\mathbf{O}}_{dB}$ from \mathbf{X} . The loss term L is composed of a decibel magnitude loss L_{dB} , calculated from the difference between $\hat{\mathbf{O}}_{dB}$ and \mathbf{O}_{dB} , and a temporal regularization loss term L_T calculated from \mathbf{C} . Using TensorFlow's automatic differentiation, symbolic derivatives $\delta L / \delta \mathbf{C}$ are automatically calculated with which \mathbf{C} are adjusted following a gradient descent method to decrease the loss in the next iteration.

3.4.2 Parameter Preprocessing

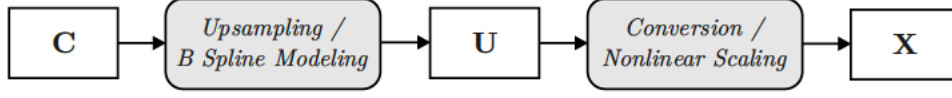


Figure 3.12: Optimized variables \mathbf{C} are first upsampled to obtain \mathbf{U} and then converted or scaled nonlinearly to obtain synthesis parameters \mathbf{X}

Instead of optimizing the parameters in \mathbf{X} directly, a different representation $\mathbf{C} \in \mathbb{R}^{2+4J \times T}$ is used, storing $2 + 4J$ variables over T frames. The synthesis parameters \mathbf{X} are obtained in two steps. First, the representation is upscaled from T frames to M frames using B-Spline modeling to obtain $\mathbf{U} \in \mathbb{R}^{2+4J \times M}$. Afterwards, the variables are scaled and converted from a normalized real representation in \mathbf{U} to their non-normalized complex representation in $\mathbf{X} \in \mathbb{C}^{2+2J \times M}$. The first step, upscaling, is done to reduce the number of parameters which are simultaneously optimized and to prevent temporal overfitting as will be described later on. The second step, nonlinear scaling or conversion of parameters, is done to restrict the parameters to stable ranges and to optimize a more uniform representation in terms of value ranges and perceived impact on the produced output $\hat{\mathbf{O}}_{dB}$.

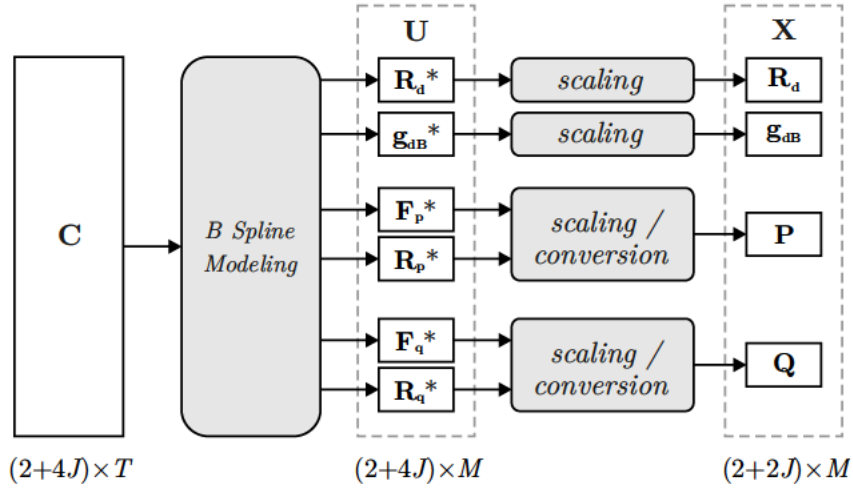


Figure 3.13: Parameter preprocessing is done in two steps. First, B-Spline modeling is used to upsample the representation \mathbf{C} to acquire \mathbf{U} . Afterwards, parameters are scaled nonlinearly to derive \mathbf{X} .

An overview of the performed steps is shown in figure 3.13. The optimized parameter representation \mathbf{C} is given with

$$\mathbf{C} \in \mathbb{R}^{2+4J \times T} \quad (3.33)$$

with $2 + 4J$ parameters and T frames per parameter. \mathbf{U} is given with

$$\mathbf{U} = \begin{bmatrix} \mathbf{R}_d^* \\ g_{dB}^* \\ \mathbf{F}_p^* \\ \mathbf{R}_p^* \\ \mathbf{F}_q^* \\ \mathbf{R}_q^* \end{bmatrix} \in \mathbb{R}^{2+4J \times M} \quad (3.34)$$

where $\mathbf{F}_p^* = (f_{p_{j,m}}^*) \in \mathbb{R}^{J \times M}$ describes pole frequencies and $\mathbf{R}_p^* = (r_{p_{j,m}}^*) \in \mathbb{R}^{J \times M}$ describes pole radii for M frames and J poles. Similarly, \mathbf{F}_q^* and \mathbf{R}_q^* describe zero frequency and radii. Here, the notation x^* is used to describe a nonlinearly scaled representation of which x is derived with a function $x = f(x^*)$. Parameters in \mathbf{U} are scaled and pole frequencies and radii combined to derive \mathbf{X} .

B-Spline Modeling

The optimized parameter representation \mathbf{C} uses a resolution of T frames per parameter which is upsampled to a resolution of M frames per parameter used in \mathbf{U} and \mathbf{X} . In practice, a ratio of $M = 10T$ was found to be suitable for this application case. With a hopsize of $N_H = 64$ samples at a sample rate of $f_s = 44100$ kHz used in \mathbf{O} , \mathbf{X} and \mathbf{U} , this corresponds to a resolution of approximately $44100/(10 \cdot 64) \approx 68.9$ frames per second to model parameter trajectories in \mathbf{C} . For upscaling, B-Spline modeling is used to interpolate each parameter trajectory. For this, every parameter trajectory in \mathbf{C} is treated as a vector of B-spline coefficients with which a spline is calculated using Equation 2.26 with a pre-calculated table for the basis functions \mathbf{B} . For the basis splines are of order $o = 3$ (quadratic) and the B-spline is defined with the knot sequence

$$\mathbf{s} = (s_t) = [0, 0, 0, 1, 2, \dots, T-5, T-4, T-3, T-3, T-3] \quad (3.35)$$

where the t th basis function is defined using the sequence $s_t \dots s_t + 3$. Since the matrix \mathbf{B} is calculated in advance, B-spline modeling only requires one matrix multiplication for every parameter to upsample the T -value representation in \mathbf{C} to an M -value parameter trajectory in \mathbf{U} .

Parameter Conversion and Nonlinear Scaling

Parameter trajectories are converted or scaled nonlinearly to linearize their perceptive impact on the output and balance their parameter ranges for the gradient descent optimization. Additionally, scaling is done to restrict parameter to stable ranges especially for poles and zeros which might otherwise produce unstable filter responses. These conversions are performed to obtain the parameters in \mathbf{X} (see Equation 3.31) from \mathbf{U} (see Equation 3.34). For R_d and g_{dB} the relations are given with

$$\begin{aligned} R_d &= (2.7 - 0.3) \operatorname{sig}(R_d^*) + 0.3 \\ g_{dB} &= 100g_{dB}^* - 100 \end{aligned} \quad (3.36)$$

Where $\operatorname{sig}(x) = 1/(1 + e^{-x})$ describes the sigmoid function. Pole and zeros are represented with a pair of real numbers r_p^* and f_p^* or r_q^* and f_q^* controlling pole radius and frequency or zero radius and frequency respectively. This representation using two real numbers has to be converted to a representation using one complex number. For poles, the complex pole position is given with

$$\begin{aligned} p &= r_p e^{i2\pi f_p/f_s} \\ f_p &= 8000 \operatorname{sig}(f_p^*) \\ r_p &= 1 - 10^{-90 \operatorname{sig}(r_p^*)/20} \end{aligned} \quad (3.37)$$

The equivalent is used to calculate complex zero positions q from f_q^* and r_q^* .

$$\begin{aligned} q &= r_q e^{i2\pi f_q/f_s} \\ f_q &= 8000 \operatorname{sig}(f_q^*) \\ r_q &= 1 - 10^{-90 \operatorname{sig}(r_q^*)/20} \end{aligned} \quad (3.38)$$

3.4.3 Synthesis

One challenge in implementing IIR filters as part of a fully differentiable optimization network is the dependency between time-domain output samples and filter coefficients. As the impulse response of IIR filters is theoretically infinite, filter coefficients at time n_0 will affect every output sample at $n \geq n_0$. While this doesn't render the use of IIR filters in neural networks or optimization algorithms impossible, it drastically increases their computational complexity. In order to incorporate IIR filters efficiently, the proposed synthesis model parameter estimation method approximates the synthesis in the spectral-domain. Per frame, the method calculates the harmonic partials of the glottal source and applies the gain and vocal tract filter response by complex multiplication. The synthesis model can be expressed in the spectral domain with

$$\hat{o}_k = G(\omega_k) g V(\omega_k) e^{i\sigma_k} \quad (3.39)$$

with glottal source spectrum $G(\omega_k)$, gain $g = 10^{g_{dB}/20}$, vocal tract filter response $V(\omega_k)$ and phase offset σ_k at $\omega_k = 2\pi f_k$ with $f_k = k f_0$ for $1 \leq k \leq K$. There are two observations which can be made about the vocal tract filter that allows this approximation to be sufficiently accurate for the described method. Firstly, a vocal tract filter response is generally assumed to be minimum phase [11] and experiments as part of this thesis have suggested, that the vocal tract related impulse response decays rather quickly within a few periods of the glottal source. Secondly, the vocal tract articulation generally only

varies slowly over time, especially when restricted to vowels. To remove the impact of phase, the differentiable synthesis model is expressed in decibel magnitude with

$$\hat{o}_{dB,k} = 20 \log_{10} |G(\omega_k)| + g_{dB} + 20 \log_{10} |V(\omega_k)| \quad (3.40)$$

The glottal source $G(\omega_k)$ is based on the spectral representation of the LF model described in [22] and implemented in [23]. The vocal tract transfer function is calculated in the z-domain as described in Equation 3.14 which allows for an efficient implementation using TensorFlow's automatic differentiation application programming interface (API).

3.4.4 Optimization Loss

The optimization method minimizes two additive losses. The first loss L_{dB} represents the decibel magnitude difference between observed harmonics \mathbf{O}_{dB} and synthesized harmonics $\hat{\mathbf{O}}_{dB}$. The second loss L_T acts as a regularization loss and penalizes temporal overfitting of the parameters \mathbf{C} . The combined loss L is given by

$$L = L_{dB} + L_T \quad (3.41)$$

Decibel Magnitude Loss L_{dB}

The decibel magnitude loss can best be described as a mean squared error loss with weighting of the harmonic partials. The loss is given with

$$L_{dB} = \frac{1}{MK} \sum_k^K \sum_m^M (w_{k,m} (20 \log_{10}(o_{k,m}) - 20 \log_{10}(\hat{o}_{k,m})))^2 \quad (3.42)$$

where $o_{k,m}$ and $\hat{o}_{k,m}$ represents the original and synthesized harmonic partials respectively and $w_{k,m}$ represents the weighting matrix intended to emphasize low frequencies over high frequencies. The weighting matrix was designed as a low-pass filter with variable cutoff and order. The low-pass filter is designed in the Laplace domain with

$$H_w(s) = \left(\frac{1}{1 + \frac{s}{2\pi f_c}} \right)^o \quad (3.43)$$

with order o and cutoff f_c . From there, the weights are calculated from the absolute weighting filter response at the harmonic frequencies $s = i2\pi k f_{0,m}$ and normalized with the average weighting over all harmonics.

$$w_{k,m} = \frac{|H_w(f_{k,m})|}{\frac{1}{K} \sum_k^K |H_w(f_{k,m})|} \quad (3.44)$$

Temporal Regularization Loss L_T

It is expected that the proposed synthesis model isn't capable of reproducing every characteristic of the human voice such as nonlinear behaviors, aspiration noise or nonlinear interactions between glottal oscillation and the vocal tract. For that reason, optimizing synthesis parameters to reduce the magnitude loss L_{dB} for every frame may easily produce overfitted results. For example, without the inclusion of the temporal regularization term, the vocal tract parameters \mathbf{p} and \mathbf{q} were often fitted to track the harmonic peaks of the partials instead of the hidden vocal tract frequency response. Even though the proposed method restricts temporal fluctuations to some degree by optimizing a undersampled representation \mathbf{C} from which \mathbf{X} is derived using upscaling as discussed in Section 3.4.2, further measures were necessary to prevent this kind of temporal overfitting. It was decided to include a form of temporal regularization loss L_T to the loss term which penalizes high temporal fluctuations of parameters. The regularization loss is calculated from \mathbf{C} by highpass filtering parameter trajectories over all frames using convolution

$$c_{HP}(r, t) = \sum_{n=-4}^{+4} c[r, t - n] \cdot b_n \quad (3.45)$$

with $\mathbf{C}_{HP} = (c_{HP}[r, t]) \in \mathbb{R}^{2+2J, T}$. Here, the subscripts are put into square brackets for better readability, i.e $c[r, t] = c_{r, t}$. In Equation 3.45, the last valid values are used at the edges with $c[r, (t - n) < 1] = c[r, 1]$ and $c[r, (t + n) > T] = c[r, T]$. The filter kernel b_n is symmetric around $n = 0$ with $-4 \leq n \leq 4$ and is derived from

$$b_{-4} + b_{-3}z^{-1} \dots b_3z^{-7} + b_4z^{-8} = (-0.25 + 0.5z - 0.25z^2)^4 \quad (3.46)$$

The derived filter coefficients are shown in Figure 3.14.

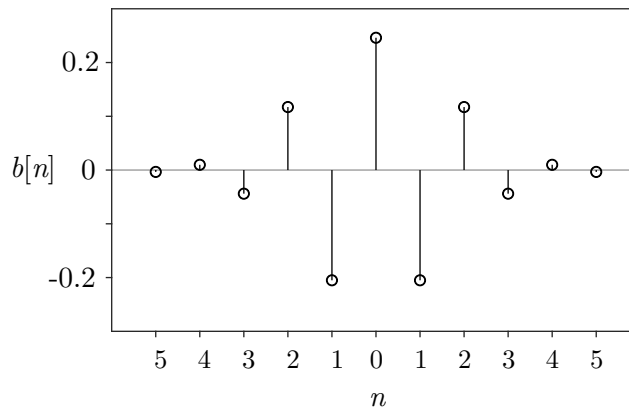


Figure 3.14: Filter coefficients for the temporal regularization loss filter.

From the highpass filtered matrix \mathbf{C}_{HP} , the mean square over all T frames is calculated to determine the temporal loss $L_{T,r}$ for the r th parameter.

$$L_{T,r} = \frac{1}{T} \sum_{t=1}^T c_{HP}[r,t]^2 \quad (3.47)$$

Finally, the total temporal loss term L_T is calculated by summing the weighted parameter-specific temporal losses $L_{T,r}$ with

$$L_T = \sum_{r=1}^{2+4J} w_r L_{T,r} \quad (3.48)$$

where $(w_r) \in \mathbb{R}^{2+4J}$ is introduced as a vector of weighting variables to adjust the impact of the specific parameters on the temporal regularization loss L_T . With this weighting variable, temporal fluctuations of different parameters can be penalized more or less.

For parameters which are constant over all frames, the mean square of the highpassed parameter trajectory and thus the parameter-specific temporal loss $L_{T,r}$ would be zero and thus no additional loss would be added to the loss term. The stronger the fluctuation between frames, indicated by more high frequency content in the parameter trajectory, the bigger the additional loss $L_{T,r}$. Figure 3.15 shows two hypothetical parameter trajectories and their highpassed and squared highpassed versions. The lower trajectory approximately contains half the bandwidth of the upper trajectory and thus produces a lower $L_{T,r}$. In practice, a higher weight for vocal tract parameters were used, following the assumption that the vocal tract articulation and thus the formant and anti-formant frequencies and resonances change rather slowly compared to the parameters R_d and g_{dB} which mainly describe the glottal flow intensity.

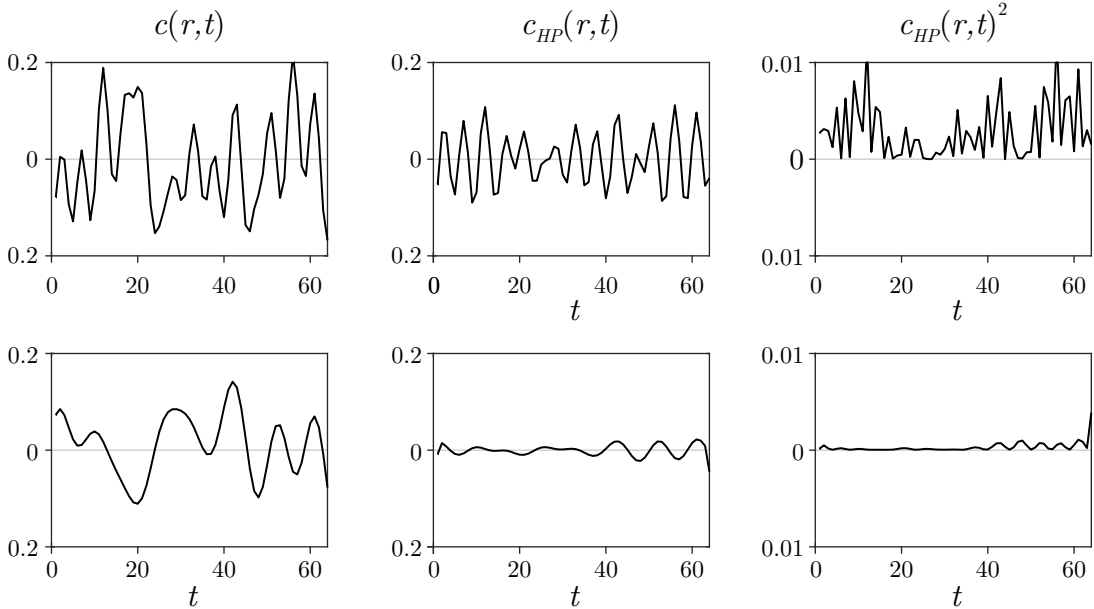


Figure 3.15: Two parameter trajectories $(c(r,t))$ and their high-passed versions $c_{HP}(r,t)$ as well as squared highpass trajectory $c_{HP}(r,t)^2$. The upper parameter trajectory includes more high frequency fluctuations compared to the lower and thus would produce a larger regularization loss $L_{T,r}$.

3.4.5 Optimization and Implementation

The optimization network described above is implemented using the TensorFlow Keras API. The Adam [77] optimizer is used which implements a form of gradient descent optimization. During testing, it was noticed that lower values for the loss L could be achieved by first optimizing using a higher loss weights w_r for the temporal regularization loss L_T for pole and zero parameters before reducing the weight. This can be described as first finding a general minimum for which the pole and zero frequencies and magnitudes appear static for the duration of the source audio sample before allowing the optimizer to adjust pole and zero frequencies and magnitudes more freely to fit subtle variations in the vocal tract articulation once the temporal regularization loss weights w_r is reduced. The implementation supports the use of GPU devices. On a NVIDIA GeForce RTX 2080, the optimization takes around 20 minutes to estimate the parameters of a 10 second audio source samples with $M = 7730$ frames, $K = 40$ harmonic partials and $J = 10$ poles and zeros.

3.4.6 Summary

The proposed synthesis model parameter estimation method implements the synthesis model as part of gradient descent optimizer. Synthesis parameters are thus estimated by minimizing the difference between synthesized harmonic partials $\hat{\mathbf{O}}$ and with the harmonic analysis estimated partials \mathbf{O} . In addition, a temporal regularization loss term is included in the loss calculation to prevent temporal overfitting especially of the poles and zeros by penalizing high fluctuations in the parameter trajectories.

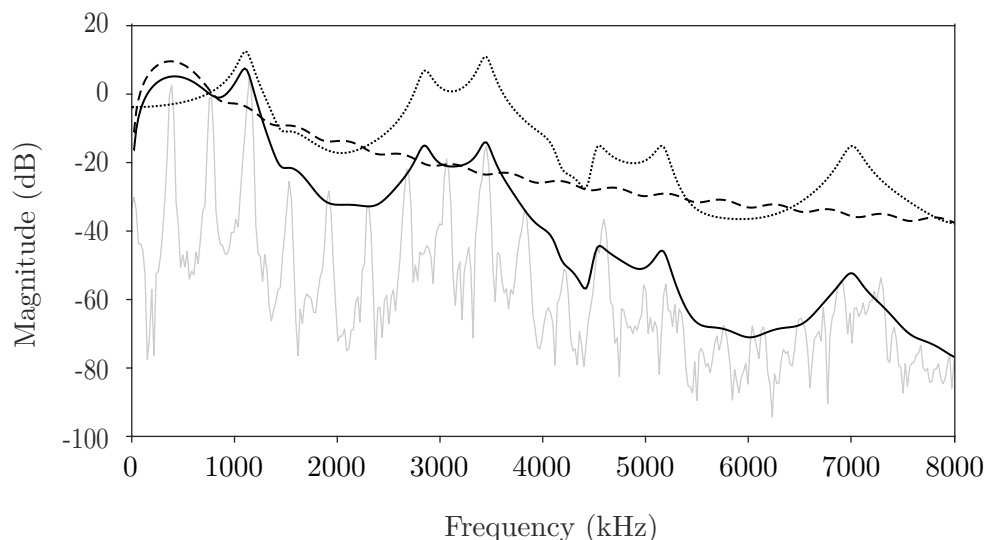


Figure 3.16: This graph shows the magnitude spectrum of the source sample (grey) together with the reconstruction (solid black), consisting of glottal source spectrum (dashed), vocal tract spectrum (dotted) and gain. The source audio sample belongs to a female singer, singing the vowel [o] at 380 Hz.

Figure 3.16 shows the magnitude spectrum of a single frame together with the reconstructed spectral envelope of glottal source model, vocal tract model and their combination.

Subjective tests have shown that the approach is capable of estimating parameter trajectories that closely predict the harmonic partials of the source sample. More importantly, by restricting temporal fluctuations of parameters using the temporal regularization loss term, the method tends to naturally find feasible solutions for the vocal tract filter even though the filter model is very flexible and not specifically tailored to model a vocal tract. However, after using the method to analyze various audio samples in the dataset, a list of shortcomings were noticed.

The biggest problem that this method had to solve is the separation of source and filter. The method tends to make estimation errors below 2 kHz. As discussed in Chapter 2, only few harmonics are present in this frequency range, which leads to what could be considered a lack of information. In contrast, multiple variables have a drastic impact on the harmonic partials in this frequency range, mainly the first few vocal tract formants and anti-formants and the glottal formant. As a result, the parameter estimation method tends to not clearly separate the effect of glottal source and vocal tract, for instance by modulating the source parameter R_d to reproduce the first vocal tract formant. There are various possible solutions to address this issue. First, as discussed in 2.2.1, the spectral description of the anti-causal glottal formant differs from a causal vocal tract formant mainly in phase. For that reason, ignoring phase in the loss term L_{dB} might discard additional information which could be useful to distinguish between vocal tract formant and glottal formant. On the same subject, another possible issue might be the simplified glottal source model that, in practice, approximates the two main qualities, glottal formant and spectral tilt, with only one parameter. Thus, changing to a two or three parameter glottal flow model might improve estimation results. Finally, in order to improve the vocal tract formant, restrictions of pole and zero frequencies and resonance might be derived from the tube model of the vocal tract in order to prevent unlikely estimations of the vocal tract transfer function.

3.5 Parameter Predictor Network \mathcal{P}

The predictor network \mathcal{P} presented in this section bridges the gap between analysis and synthesis. It is trained from the previously estimated synthesis parameters $\mathbf{X} = (\chi_m)$ and pitch trajectory $\mathbf{f}_0 = (f_{0,m})$. During synthesis, it is used predicts synthesis parameters $\hat{\chi}$ from any pitch f_0 with

$$\hat{\chi} = \mathcal{P}\{f_0\} \quad (3.49)$$

where $\hat{\chi}$ is used to refer to the predicted set of parameters as opposed to the estimated parameters $\mathbf{X} = (\chi_m)$ obtained from the synthesis model parameter estimation. For control parameters, it was decided to solely rely on pitch f_0 . As the focus of this thesis was the synthesis model parameter estimation \mathcal{E} , a fully connected neural network was used for \mathcal{P} with the option to evaluate more complex approaches, such as recurrent or convolutional networks, in the future.

3.5.1 Network

The neural network consists of a core of fully connected (FC, dense) layers with custom pre-scaling and post-scaling layers. An overview is shown in figure 3.17.

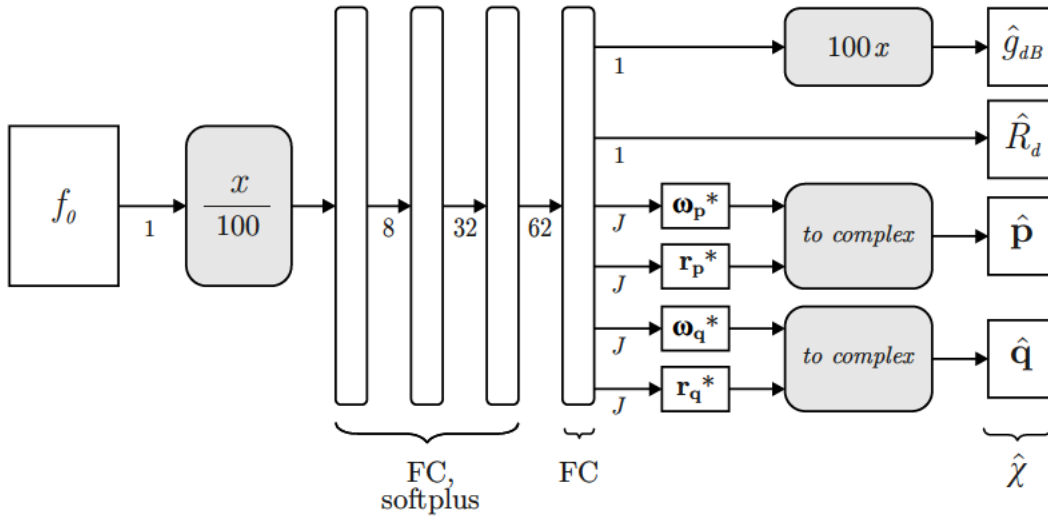


Figure 3.17: The predictor network \mathcal{P} consists of a f_0 pre-scaling layer, 4 fully connected (dense) layers, the first three using the softplus activation function, and a post-scaling and complex conversion layer.

Pre- and Post-Scaling

The pre-scaling of control parameter f_0 scales the parameter down by a factor of 100. Similarly, the gain parameter \hat{g}_{dB} is scaled up by a factor of 100. The fully connected layers output radii r_p^* , r_q^* and phases ω_p^* , ω_q^* to describe poles \hat{p} and zeros \hat{q} respectively. These are combined as follows to calculate the complex pole or zero positions.

$$\begin{aligned}
\hat{\mathbf{p}} &= \mathbf{r}_{\mathbf{p}} e^{i\omega_{\mathbf{p}}} & \hat{\mathbf{q}} &= \mathbf{r}_{\mathbf{q}} e^{i\omega_{\mathbf{q}}} \\
\omega_{\mathbf{p}} &= \pi \operatorname{sig}(\omega_{\mathbf{p}}^*) & \omega_{\mathbf{q}} &= \pi \operatorname{sig}(\omega_{\mathbf{q}}^*) \\
\mathbf{r}_{\mathbf{p}} &= 1 - 10^{-80 \operatorname{sig}(\mathbf{r}_{\mathbf{p}}^*)/20} & \mathbf{r}_{\mathbf{q}} &= 1 - 10^{-80 \operatorname{sig}(\mathbf{r}_{\mathbf{q}}^*)/20}
\end{aligned} \tag{3.50}$$

This scaling also prevents poles and zeros from moving outside of the unit circle thus preventing in stable filters. Additionally, the scaling limits the pole and zero phase to $0 < \omega < \pi$.

Fully Connected Layers

The first three dense layers shape the output to a size of 8, 32 and 62 nodes respectively and use the softplus activation function $f(x)$, the derivative of which $f'(x)$ is equal to the sigmoid activation function with

$$\begin{aligned}
f(x) &= \ln(1 + e^x) \\
f'(x) &= \frac{1}{1 + e^{-x}}
\end{aligned} \tag{3.51}$$

The last dense layer shapes the output to $2 + 4J$ nodes and uses the identity function. The specific configuration of the proposed neural networks was obtained from subjective tests with various numbers and sizes of dense layers and different activation functions.

3.5.2 Loss

As the network has multiple outputs, two scalars R_d and g_{dB} and two complex vectors \mathbf{p} and \mathbf{q} , parameter-specific loss terms were used with the overall loss L being defined as

$$\begin{aligned}
L &= w_{R_d} L_{R_d} + w_{g_{dB}} L_{g_{dB}} \\
&+ \sum_{j=1}^J (w_{\omega} L_{\arg(p_j)} + w_r L_{|p_j|}) \\
&+ \sum_{j=1}^J (w_{\omega} L_{\arg(q_j)} + w_r L_{|q_j|})
\end{aligned} \tag{3.52}$$

The loss for the complex poles \mathbf{p} and zeros \mathbf{q} is calculated separately for their phases $\arg(p_j)$, $\arg(q_j)$ and magnitudes $|p_j|$, $|q_j|$. The impact of each parameter can be adjusted by scaling the associated weighting variables, w_{R_d} for R_d , w_g for g_{dB} and w_{ω} for the phases of all poles and zeros as well as w_r for the magnitudes (or radii) of all poles and zeros. For g_{dB} , R_d and pole and zero phases, the loss is calculated using the mean squared error (MSE) with

$$\operatorname{MSE}(x, \hat{x}) = \frac{1}{M} \sum_m^M (x - \hat{x})^2 \tag{3.53}$$

where M represents the number of frames over which the squared error is averaged. the losses L_{R_d} , $L_{g_{dB}}$, $L_{\arg(p_j)}$ and $L_{\arg(q_j)}$ are given with

$$\begin{aligned} L_{R_d} &= \text{MSE}(R_d, \hat{R}_d) \\ L_{g_{dB}} &= \text{MSE}(g_{dB}, \hat{g}_{dB}) \\ L_{\arg(p_j)} &= \text{MSE}(\arg(p_j), \arg(\hat{p}_j)) \\ L_{\arg(q_j)} &= \text{MSE}(\arg(q_j), \arg(\hat{q}_j)) \end{aligned} \quad (3.54)$$

Notably, the phases of the poles and zeros are restricted to a range of 0 to π in the post-scaling layer. For that reason, the mean squared error is sufficient to calculate the phase error. The pole and zero magnitudes $|p_j|$, $|q_j|$ are first converted to a decibel range so that the calculated error roughly corresponds to the perceived difference between two filters with different pole or zero radii. The conversion is based on the frequency response of a single zero filter evaluated at the angle of the zero. For a complex zero c , the response for such a filter is given with

$$\begin{aligned} H(z) &= 1 - cz^{-1} \\ H(e^{i \arg(c)}) &= 1 - \frac{|c|e^{i \arg(c)}}{e^{i \arg(c)}} \\ H(e^{i \arg(c)}) &= 1 - |c| \\ 20 \log_{10}(H(e^{i \arg(c)})) &= 20 \log_{10}(1 - |c|) \end{aligned} \quad (3.55)$$

Notably, the decibel magnitude response for the corresponding single pole filter is simply the inverse of the single zero filter. Thus the loss for the pole and zero magnitudes $|p_j|$, $|q_j|$ is given with

$$\begin{aligned} L_{|p_j|} &= \text{MSE}(20 \log_{10}(1 - |p_j|), 20 \log_{10}(1 - |\hat{p}_j|)) \\ L_{|q_j|} &= \text{MSE}(20 \log_{10}(1 - |q_j|), 20 \log_{10}(1 - |\hat{q}_j|)) \end{aligned} \quad (3.56)$$

3.5.3 Implementation and Training

The prediction network is trained with M frames with input \mathbf{f}_0 and output \mathbf{X} obtained from the harmonic analysis and synthesis model parameter estimation respectively. The network was implemented using the TensorFlow Keras API [68]. The optimization algorithm Adam [77] was used for training. As the data originated from a sequence and consecutive frames can't be assumed to be independent, no validation tests were performed and instead, the network was trained and tested with the full sequences.

3.5.4 Summary

The proposed predictor \mathcal{P} bridges the gap between control parameter pitch f_0 and synthesis parameters χ . The network consists of four fully connected layers with softplus

activation function and pre- and post-scaling layers. The latter is used to restrict parameters to stable ranges which is especially important for the parameters R_d , poles \mathbf{p} and zeros \mathbf{q} . The network was implemented and trained using TensorFlow's Keras API.

The regression problem for which the network is used for poses a challenge, as the input \mathbf{f}_0 and output \mathbf{X} originate from one sequence and thus, frames m can't be considered independent observations. A more sophisticated network would train to predict parameters for multiple singers and vowels and use recurrent or convolutional neural networks to capture temporal dependencies within the sequences. However, as discussed in Section 3.2.3, two estimated synthesis parameter sets may not be interpolated and still produce an convincing vocal tract frequency response. For this reason, training on estimated synthesis parameters from multiple source samples isn't straight forward and would require a revised synthesis model or parameter estimation method \mathcal{E} .

4 Evaluation

Goal of the evaluation was to get a better understanding of the proposed method. As previously described, the proposed singing voice synthesis method consists of multiple successive components, namely, harmonic analysis \mathcal{A} , synthesis model parameter estimation \mathcal{E} , predictor network \mathcal{P} and synthesis \mathcal{S} . Each component depends on the previous. To obtain a better understanding on how the method performs, it is crucial to assess how these component perform relatively to another. Furthermore, early subjective tests with the proposed method suggested that the quality of synthesized audio samples depends strongly on the kind of source audio sample used in the analysis. Effects were noticed for instance for the gender of the singer from the source audio sample or the sung vowel. Finally, as the proposed singing voice method focuses on some aspects such as the vocal tract filter while neglecting others like aspiration noise, it is assumed that the proposed method does well in per serving some timbre qualities from the source sample while being unable to preserve other qualities. From these ideas, the evaluation was intended to answer three core research question.

1. Which component of the proposed method outlined in section 3.1 hat the biggest impact in terms of perceived audio quality of synthesized audio samples?
2. How do different source audio samples such as samples from singers of different genders or different sung vowels, affect the performance in terms of overall perceived audio quality?
3. How well are timbre qualities such as breathiness or roughness preserved by the proposed method when resynthesizing samples?

To answer the first and second question, a multi stimulus with hidden reference and anchor (MUSHRA) test was conducted. In this experiment, participants were asked to rate the perceived quality of different variants of the same source audio sample, referred to as conditions. In addition to a reference sample and an anchor sample, these conditions were reconstructions of the reference generated at various stages of the proposed multi-stage method as will be described in Section 4.1. Using the results from the MUSHRA test, one can evaluate differences in perceived quality between these conditions to assess which component of the proposed method is responsible for the biggest loss of sound quality. Furthermore, by taking additional independent variables such as the gender of the singer or the sung vowel into account, one can evaluate the impact of these variables on how well the proposed method preserves sound quality in the reconstruction.

To answer the third question, a pairwise comparison listening test was conducted in which participants were asked to rate the perceived difference between two stimuli in regards to four timbre qualities; naturalness, breathiness, brightness and roughness. The two stimuli were a reference sample and it’s reconstruction using the proposed singing voice synthesis method.

It should be noted that, since both experiments aim to evaluate the perceived quality difference between some reference sample and various reconstructions of it, one aspect that wasn’t taken into account in this evaluation is the ability of the proposed method to produce new audio samples from a trained predictor network and a new pitch trajectory. There are two reasons why this aspect was neglected in the evaluation. First, since the predictor \mathcal{P} predicts synthesis parameters from a single control parameter, namely pitch, it is assumed that predicting synthesis parameters using a pitch trajectory not used during training shouldn’t introduce significant artifacts. Listening tests like MUSHRA are intended to be used with different conditions of one reference sample. When the intention is to compare two uncorrelated audio samples, for instance two synthesized samples using the same trained predictor but different pitch trajectories, a different experiment design would have been necessary.

4.1 Dataset and Stimuli Generation

The VocalSet dataset [73] was used to generate the stimuli for the evaluation. VocalSet consists of a large set of singing excerpts including scales, arpeggios and individual notes of five vowels in different keys. In total, 9 female and 11 male singers were asked to sing in each of these contexts, using a variety of vocal techniques. A breakdown of the different combinations of contexts, vocal techniques and vowels is shown in Table 4.1.

Context	Technique	Vowel
Long Tones	Vibrato, Straight, Forte, Pianissimo, Trill, Trillo, Inhaled, Messa di voce	a, e, i, o, u
Arpeggios, Scales	Straight, Belt, Breathy, Fast Forte, Fast Piano, Slow Forte, Slow Piano, Vibrato, Vocal Fry, Lip Trill	

Table 4.1: A breakdown of the different contexts, vocal techniques and vowels in the VocalSet.

From this dataset, a subset of 30 samples was chosen to be included in the two experiments. During experimentation with the dataset and the proposed method, it was noticed that the method performs well with source audio samples consisting of a single vowel, covering a wide range of pitches, sung in legato. For this reason, it was decided to use audio samples from the scales context sung in slow forte as these audio samples included no breaks between the individual notes forming the scales. Three vowels, a, i and o, were chosen to be included to reduce the overall number of audio

samples while keeping a large variety of tonally distinct vowels. Five different singers from each gender were chosen to be represented in both experiments of this evaluation resulting in a total number of 30 audio samples, grouped in 2 genders, 3 vowels and 5 singers.

C	Method
C_R	Unaltered sample from the dataset VocalSet.
C_H	Harmonic reconstruction of C_R following the harmonic analysis \mathcal{A} .
C_E	Synthesized from estimated parameters using synthesis model parameter estimation \mathcal{E} .
C_P	Synthesized from predicted parameters using predictor network \mathcal{P} .
C_A	Anchor, 3.5 kHz low pass filtered reference.

Table 4.2: A breakdown of the different stimuli or conditions used in the evaluation.

As the intention of the first experiment was to evaluate the impact off each components of the proposed singing voice method on the perceived quality of the synthesized sample, multiple variations of the reference sample were generated which are here referred to as conditions in accordance with the ITU-R BS.1534-3 recommendation for MUSHRA tests. These conditions $C \in \{C_R, C_H, C_E, C_P, C_A\}$ include the reference stimuli C_R , the harmonic reconstruction C_H , the synthesis from estimated parameters C_E , the synthesis from predicted parameters C_P and an anchor condition C_A as shown in Table 4.2. These conditions except C_R and C_A could be considered as the reconstructions of C_R at different stages of the proposed method as shown in Figure 4.1.

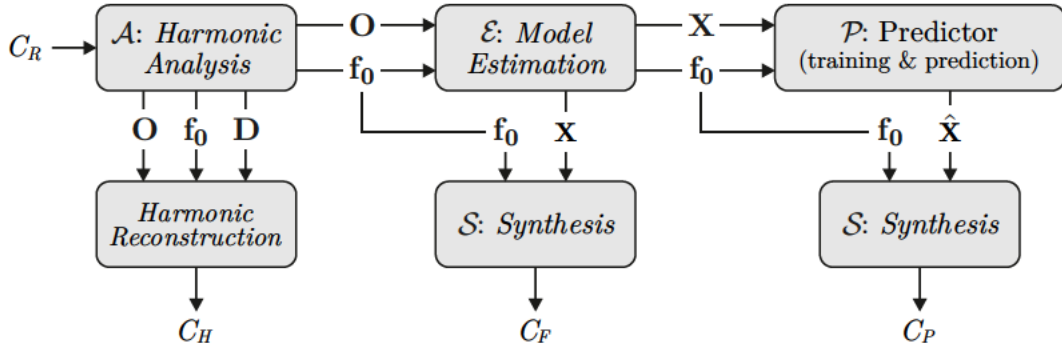


Figure 4.1: From a reference sample C_R , C_H is generated from the results of the harmonic analysis, C_E is synthesized using the estimated synthesis parameters and C_P is synthesized using the predicted synthesis parameters.

The harmonic reconstruction C_H is generated using the results from the harmonic analysis method \mathcal{A} described in Section 3.3. After estimating pitch f_0 and pitch slope D as well as harmonic partials O from the reference sample C_R , Equation 3.25 is used to reconstruct individual frames. With a frame size of $N_B = 2048$ and hop size of $N_H = 64$, the reconstructed frames were combined using a method similar to overlap-add with

$$y[n] = \frac{2N_H}{N_B} \sum_{m=1}^M \left(\sum_{n=0}^{N_B-1} w[n] \hat{x}_m[n - mN_H] \right) \quad (4.1)$$

where $w[n]$ is the Hann window of length N_B . The factor $2N_H/N_B$ adjusts the amplitude to compensate for the overlap of frames and the applied Hann window. The m th audio frame $\hat{x}_m[n]$ was reconstructed using Equation 3.25.

Both C_E and C_P were reconstructed from the reference sample C_R using a set of synthesis parameters and the synthesis method \mathcal{S} described in Section 3.2. The condition C_E is generated from the estimated parameters \mathbf{X} and pitch \mathbf{f}_0 (both a product of the synthesis model parameter estimation \mathcal{E} described in Section 3.4) using the synthesis method \mathcal{S} . Again, a frame size of $N_B = 2048$ and hop size of $N_H = 64$ were used for the synthesis model parameter estimation. Parameters \mathbf{X} were linearly interpolated to upsample the parameter trajectory from a frame rate with $N_H = 64$ to the audio rate necessary for the synthesis method \mathcal{S} .

For C_P , the predictor \mathcal{P} was trained using \mathbf{f}_0 and estimated synthesis parameters \mathbf{X} . Afterwards, the predictor was used to generate $\hat{\mathbf{X}}$ from the same pitch trajectory \mathbf{f}_0 . Afterwards, the predicted synthesis parameters were used to produce the condition C_P .

For the first experiment, an anchor condition C_A is used in accordance with the ITU-R BS.1534-3 recommendation for MUSHRA tests. The anchor can be obtained by low pass filtering the reference condition. For that, the 3.5 kHz lowpass filter implementation from webMSURHA [78] is used. The filter coefficients are calculated as a windowed sinc impulse with

$$b[n] = w_\alpha[n] \frac{\sin(2\pi(f_c/f_s)(n - 0.5N + 0.5))}{(n - 0.5N + 0.5)\pi} \quad (4.2)$$

with $0 \leq n < N$, $N = 192$, cutoff frequency $f_c = 3750$ Hz and sample rate $f_s = 44100$ Hz. For the window $w_\alpha[n]$, a Kaiser window is used with $\alpha = 1.0345$. The filters frequency response is shown in figure Figure 4.2.

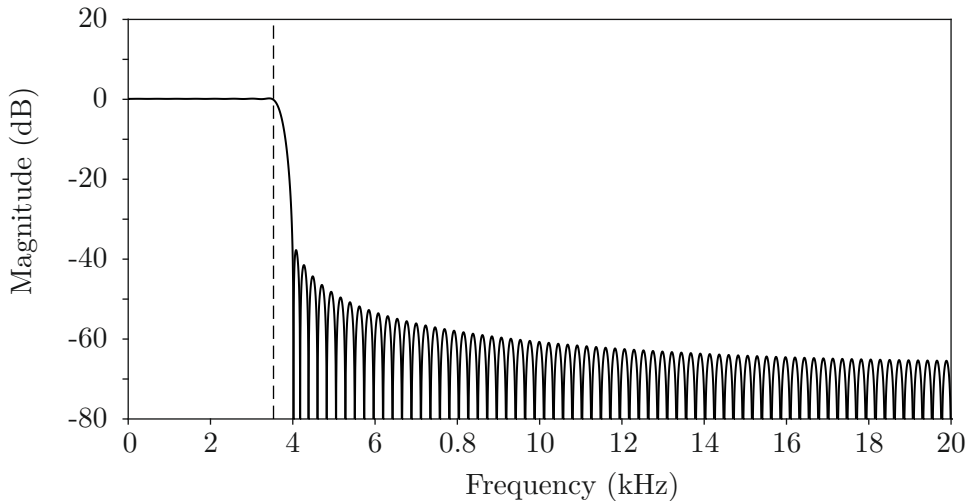


Figure 4.2: The frequency response of the anchor condition filter as implemented in webMUSHRA [78] The targeted cutoff frequency f_c is marked with a dotted line.

4.2 Experiment 1

The goal of the first experiment, a MUSHRA test [79], was to evaluate, which component of the proposed method for singing voice synthesis has the biggest impact on the perceived audio quality of the synthesized samples. In order to answer this question, 3 synthesized samples corresponding to 3 different components of the method were presented together with the reference stimuli and the anchor. As every component of the proposed method builds on top of the previous, it is hypothesized that these conditions are rated significantly different from another with the rating $R(C)$ of condition C being ordered $R(C_R) > R(C_H) > R(C_E) > R(C_P)$. Further statements can then be made about the impact of the singers gender $G \in \{G_F, G_M\}$ for female and male singers and the sung vowel $V \in \{V_A, V_O, V_I\}$ for vowels a, i and o on the perceived quality.

4.2.1 Method

Participants

At the end of the online survey, participants were asked to fill out a short questionnaire to collect demographic information of the participants. In total, 33 subjects participated in the experiment, 6 of which stated that they identified as female, 26 identified as male and 1 person gave no information about their gender. Participants' ages ranged from 22 to 60 ($M = 30.94$, $SD=9.10$). At least 21 of the participants were students at TU Berlin. Other participants came from authors workplace or from the authors family and circle of friends.

Stimuli

The experiment consisted of a total of 30 tests, each presenting a set of 5 unlabeled stimuli together with a labeled reference. As described in Section 4.1, reference stimuli were taken from the VocalSet dataset [73]. The selected subset of reference stimuli included a total of 30 samples containing scales from 5 singers of 2 genders, singing 3 vowels. The 5 unlabeled stimuli presented in each test included the conditions C_H , C_E , C_P , an anchor C_A and a hidden reference C_R and were displayed in random order.

Procedure

The experiment was conducted as an online survey using the webMUSHRA API [78]. After going through a welcoming page detailing the length and structure of the two experiments, participants were first introduced to the MUSHRA experiment. Participants were advised to use headphones for the experiment and use a desktop or laptop computer with the Google Chrome web browser in cases of technical difficulties. The participants were instructed as follows.

- The goal of this experiment is to evaluate the perceived quality of various sounds.
- In this listening test, you will be presented with a reference sound as well as a collection of 5 altered versions (conditions) of this reference. Please rate the quality of each sound in relation to the reference.

On the following page, participants were presented a sound example and a volume slider and were instructed to adjust the volume using the slider and avoid further adjusting the volume throughout the survey.

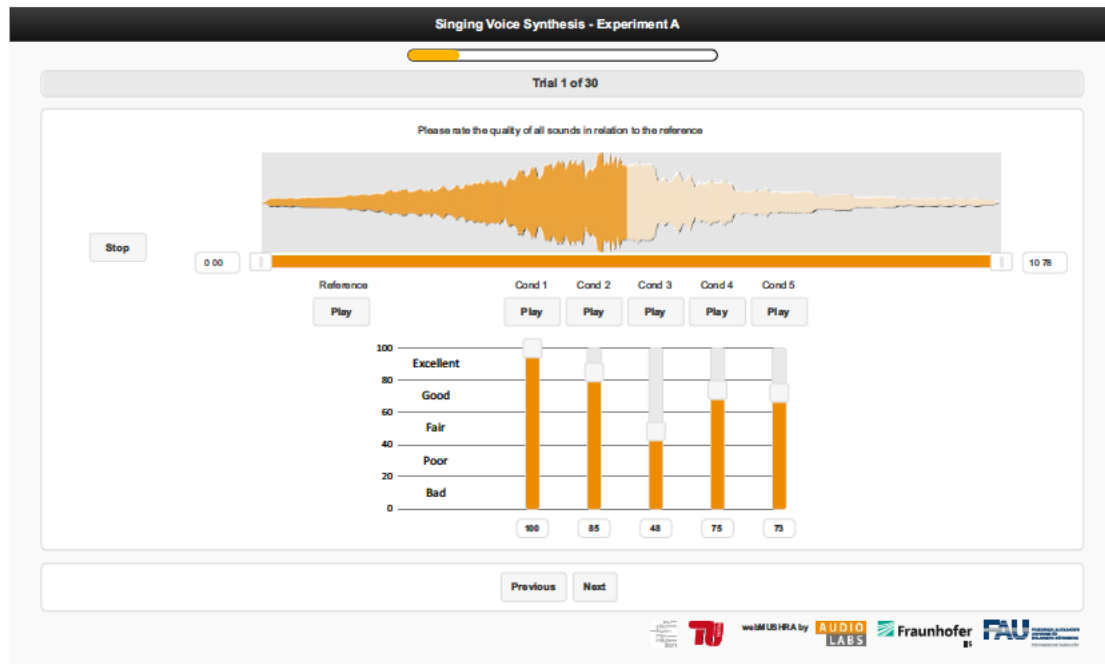


Figure 4.3: A MUSHRA trial page as used in experiment 1. The participants were instructed to rate the quality of the five conditions in relation to the reference. Participants could switch between conditions by clicking on the corresponding play buttons. Additionally, participants could select a range of the sample to loop and jump in time by clicking on the waveform.

On the following pages, the participants were introduced to the trial pages and the audio stimuli by presenting four exemplary trial pages such as the one shown in Figure 4.3. Afterwards, the participants were asked to go through the 30 trial pages making up the experiment. The order of these pages as well as the order of the unlabeled conditions on each page was randomized for every participant.

After the 30 trial pages, participants were asked to fill out a small survey. In addition to the participants age and gender, the questionnaire asked whether the participant used headphones, loudspeakers or any other playback device. Additionally, students of TU Berlin were asked for their student identification number.

4.2.2 Results

In total, the 33 participants each rated a total of 150 stimuli, resulting in 4950 ratings. First, a Friedman test was conducted to test the hypothesis, that ratings R for the four conditions C are ordered $C_R > C_H > C_E > C_P$. Results from this test were also used to draw conclusions about the impacts of each component of the proposed singing voice synthesis method on the perceived audio quality. Afterwards, a mixed model was used draw conclusions about how the singers gender and the sung vowel affected the rating of each stimuli.

Friedman Test

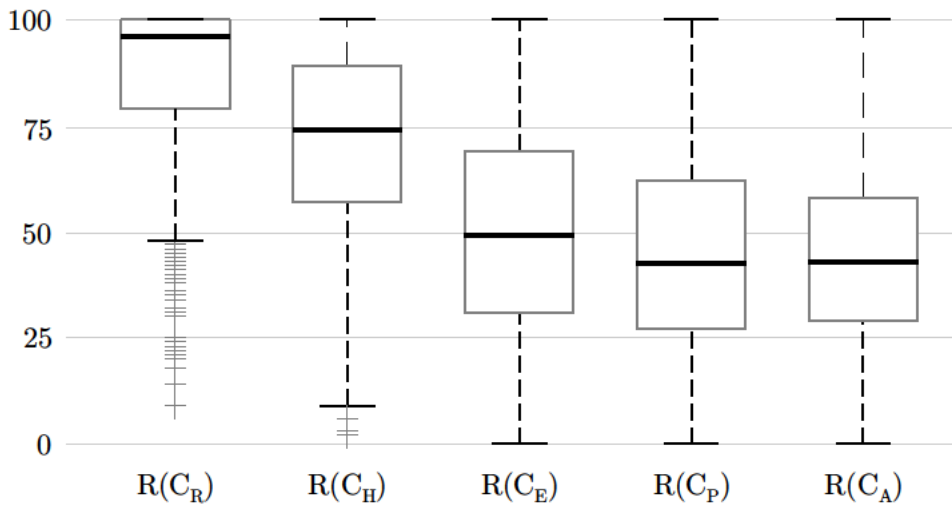


Figure 4.4: A boxplot showing the MUSHRA ratings for the five tested conditions.

Since the ratings were not normally distributed, the Friedman tests, implemented in Matlab [80, 81, 82], was chosen for analysis. The Friedman test is a non-parametric alternative to the analysis of variance (ANOVA) and thus can be used in this application. A boxplot of the ratings for the five tested conditions is shown in Figure 4.4. The median ratings $\text{median}(R(C_R)) = 96$, $\text{median}(R(C_H)) = 74$, $\text{median}(R(C_E)) = 49$, $\text{median}(R(C_P)) = 42.5$, $\text{median}(R(C_A)) = 43$. A Friedman Test was chosen to evaluate the results of the MUSHRA test as recommended in [83]. The test showed that the rating differences between the conditions C_R , C_H , C_E , C_P and C_A are significant with $X^2(4, N = 33) = 1.86 \cdot 10^3$, $p < 0.001$. Post hoc tests using Wilcoxon signed-rank test [80, 82, 84] with Bonferroni-adjusted alpha level of $\alpha = 0.05/10 = 0.005$ showed that differences between all but one paired comparison, that between C_P and C_A , were significant. The non-parametric Wilcoxon signed-rank test was chosen as the data was not normally distributed. The Bonferroni correction specifies that the alpha level for post hoc tests needs to be adjusted with α_0/n where n is the number of paired comparisons and $\alpha_0 = 0.05$ is the targeted significant level. The results are shown in Table 4.3.

Test	Wilcoxon T	Wilcoxon Z	p ($\alpha = 0.05$)
$C_R - C_H$	$3.3 \cdot 10^5$	18.79	< 0.001
$C_R - C_E$	$4.5 \cdot 10^5$	25.08	< 0.001
$C_R - C_P$	$4.6 \cdot 10^5$	25.72	< 0.001
$C_R - C_A$	$4.7 \cdot 10^5$	25.78	< 0.001
$C_H - C_E$	$4.1 \cdot 10^5$	20.71	< 0.001
$C_H - C_P$	$4.2 \cdot 10^5$	22.57	< 0.001
$C_H - C_A$	$4.4 \cdot 10^5$	22.65	< 0.001
$C_E - C_P$	$2.7 \cdot 10^5$	7.87	< 0.001
$C_E - C_A$	$2.9 \cdot 10^5$	6.29	< 0.001
$C_P - C_A$	$2.4 \cdot 10^5$	0.46	0.64

Table 4.3: Results of the post hoc Wilcoxon signed-rank test.

Mixed Model

A mixed model was trained on the data from experiment 1 to acquire insight into how the singers gender G and the sung vowel V affect the performance of the proposed method. Mixed models or mixed effects models are used to test the impact of independent variables of an dependent variable and pose an alternative to ANOVA tests especially for non-normally distributed data. The term mixed effects refers to the use of fixed effects and random effects. Fixed effects are primarily used to model observed effects such as the singers gender or vowel. Random effects instead model sources of random levels such as the subjects preference in rating stimuli [85]. A mixed model is mainly defined by it's Wilkinson notation describing which variables are included either as fixed or random effects and by the levels of each effect. Once fitted to the data, the fixed effects can be further analyzed to draw conclusions about their impact on the dependent variable

The model used to fit the data from the MUSHRA experiment includes fixed effect variables condition $C \in \{C_R, C_H, C_E, C_P, C_A\}$, gender $G \in \{G_F, G_M\}$ for female and male singers and vowel $V \in \{V_A, V_O, V_I\}$ for vowels a, i and o as well as the a random effect variable $I_{ID} \in \mathbb{N}$, $1 \leq I_{ID} \leq 33$ for the 33 participants. The model's Wilkinson notation is given with

$$R \sim 1 + C + G + V + C:V + C:G + V:G + G:C:V + (1|I_{ID}) \quad (4.3)$$

The model includes the main fixed effects C , G and V as well as the 2-way interaction effects $C:G$, $C:V$, $G:V$ and 3-way effects $C:G:V$. Interaction effects consist of all possible commentary levels of their component variables. For instance, the 2-way interaction between vowel V and gender G would consist of $C : G = \{V_A G_F, V_I G_F, V_O G_F, V_A G_M, V_I G_M, V_O G_M\}$. In addition, the model includes the random effect $I|I_{ID}$ and the fixed intercept (denoted as 1 in the Wilkinson notation) which is best described as a constant offset for the dependent variable R . The mixed effects model was generated with the statistics software Jamovi [86, 87] using the linear model library GAMLj [88]. An overview is given in Table 4.4.

Info	
Estimate	Linear mixed model fit by Restricted Maximum Likelihood
Call	$R \sim 1 + C + V + G + C:V + C:G + V:G + G:V:C + (1 ID)$
AIC	43325.002
BIC	43472.936
LogLikelihood	-21600.354
R-squared Marginal	0.383
R-squared Conditional	0.542

Table 4.4: General information on the mixed model as reported by GAMLj.

Fixed Effects An omnibus test was conducted to determine the significance of all fixed effects. A significant test suggests that, for a fixed effect, at least two of its groups are significantly different. The test confirmed, that between C , V and G , all fixed main effects, 2-way and 3-way effects were significant as shown in Table 4.5. For variables C and G , simple coding was chosen with reference levels C_R and G_F respectively. For variable V , deviation coding was chosen. In total, the model consists of 8 main effect parameters including the intercept, 14 2-way effect parameters and 8 3-way effect parameters, 30 fixed effects parameters in total as can be seen in Table 4.6. Notably, not all fixed effect parameters are significant ($p \geq 0.05$).

Effect	F	Num df	Den df	p ($\alpha = 0.05$)
Condition	967.24	4	4888	< 0.001
Vowel	4.64	2	4888	0.010
Gender	14.40	1	4888	< 0.001
Condition * Vowel	3.51	8	4888	< 0.001
Condition * Gender	33.72	4	4888	< 0.001
Vowel * Gender	13.99	2	4888	< 0.001
Condition * Vowel * Gender	6.88	8	4888	< 0.001

Table 4.5: Mixed model omnibus test results with F-score, degrees of freedom (Num df), degrees of freedom denominator (Den df), and p-value (p) as reported by GAMLj.**Table 4.6:** Fixed effects parameter estimates of the mixed model. ($\alpha = 0.05$)

Effect	Estimate	SE	df	t	p
(Intercept)	59.531	1.959	32.0	30.396	< 0.001
$C_H - C_R$	-15.759	0.851	4888.0	-18.523	< 0.001
$C_E - C_R$	-36.639	0.851	4888.0	-43.066	< 0.001
$C_P - C_R$	-41.607	0.851	4888.0	-48.905	< 0.001
$C_A - C_R$	-42.458	0.851	4888.0	-49.905	< 0.001
$V_I - (V_A, V_I, V_O)$	0.578	0.380	4888.0	1.520	0.129
$V_O - (V_A, V_I, V_O)$	-1.159	0.380	4888.0	-3.047	0.002
$G_M - G_F$	-2.040	0.538	4888.0	-3.791	< 0.001
$C_H - C_R * V_I - (V_A, V_I, V_O)$	0.995	1.203	4888.0	0.827	0.408
$C_E - C_R * V_I - (V_A, V_I, V_O)$	1.497	1.203	4888.0	1.244	0.213
$C_P - C_R * V_I - (V_A, V_I, V_O)$	3.228	1.203	4888.0	2.683	0.007
$C_A - C_R * V_I - (V_A, V_I, V_O)$	-2.121	1.203	4888.0	-1.763	0.078
$C_H - C_R * V_O - (V_A, V_I, V_O)$	-1.963	1.203	4888.0	-1.631	0.103
$C_E - C_R * V_O - (V_A, V_I, V_O)$	-2.818	1.203	4888.0	-2.342	0.019
$C_P - C_R * V_O - (V_A, V_I, V_O)$	-2.978	1.203	4888.0	-2.475	0.013
$C_A - C_R * V_O - (V_A, V_I, V_O)$	1.094	1.203	4888.0	0.909	0.363

Table 4.6 Continued: Fixed effects parameter estimates of the mixed model. ($\alpha = 0.05$)

Effect	Estimate	SE	df	t	p
$C_H - C_R * G_M - G_F$	3.840	1.702	4888.0	2.257	0.024
$C_E - C_R * G_M - G_F$	-6.762	1.702	4888.0	-3.974	< 0.001
$C_P - C_R * G_M - G_F$	-7.731	1.702	4888.0	-4.544	< 0.001
$C_A - C_R * G_M - G_F$	8.705	1.702	4888.0	5.116	< 0.001
$V_I - (V_A, V_I, V_O) * G_M - G_F$	-1.710	0.761	4888.0	-2.248	0.025
$V_O - (V_A, V_I, V_O) * G_M - G_F$	-2.301	0.761	4888.0	-3.023	0.003
$C_H - C_R * V_I - (V_A, V_I, V_O) * G_M - G_F$	-6.313	2.406	4888.0	-2.624	0.009
$C_E - C_R * V_I - (V_A, V_I, V_O) * G_M - G_F$	-0.408	2.406	4888.0	-0.170	0.865
$C_P - C_R * V_I - (V_A, V_I, V_O) * G_M - G_F$	0.974	2.406	4888.0	0.405	0.686
$C_A - C_R * V_I - (V_A, V_I, V_O) * G_M - G_F$	4.186	2.406	4888.0	1.740	0.082
$C_H - C_R * V_O - (V_A, V_I, V_O) * G_M - G_F$	2.717	2.406	4888.0	1.129	0.259
$C_E - C_R * V_O - (V_A, V_I, V_O) * G_M - G_F$	-6.214	2.406	4888.0	-2.582	0.010
$C_P - C_R * V_O - (V_A, V_I, V_O) * G_M - G_F$	-10.020	2.406	4888.0	-4.164	< 0.001
$C_A - C_R * V_O - (V_A, V_I, V_O) * G_M - G_F$	-2.511	2.406	4888.0	-1.044	0.297

Random Components The model includes two random components, namely, the random effect $1|I_{ID}$ (SD=11.1, Var= 124) and the residual (SD=18.9, Var=358). The random effect $1|I_{ID}$ was used to model the difference in rating behavior between each participant (SD= 11.1, Var= 124). A likelihood ratio test determined that this random effect is significant (LRT= 1306, $p < 0.001$).

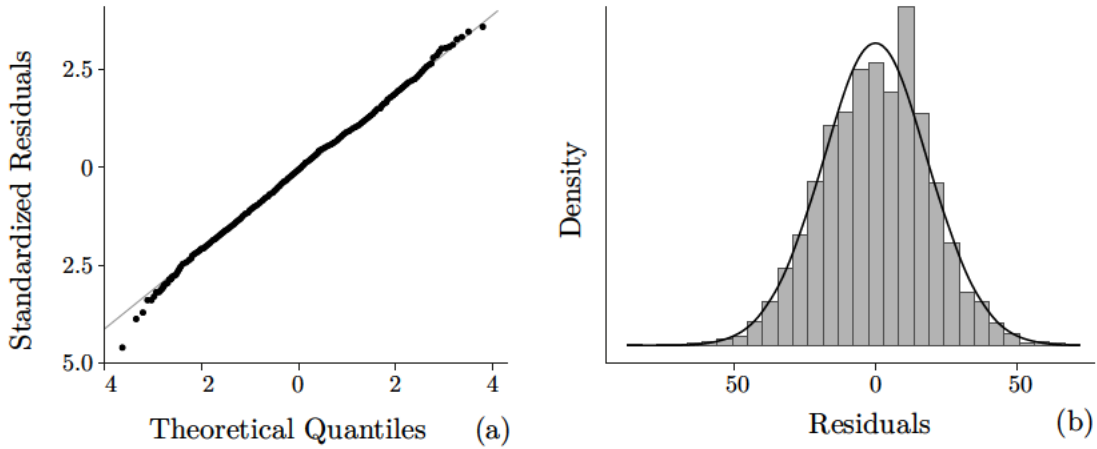


Figure 4.5: (a) The Q-Q plot shows that the residuals are mostly distributed normally with only some deviation at the lower quantiles. (b) The histogram of the residuals show some artifacts in form of an spike around a value of 10 but otherwise mostly resembles the histogram of a normal distribution.

Assumption Checks Linear mixed models require the residuals, the differences between estimated and observed ratings, to be normal distributed. Both the histogram of the residuals and the Quantile-Quantile (Q-Q) plot (Figure 4.5) suggest that this assumption holds true for the model. Most discrepancies from a normal distribution of the residuals can probably be explained with the how the experiment was conducted as the ratings, especially for C_R and C_H are strongly shifted and clipped towards $R = 100$. Especially

the Q-Q plot suggest that with minor variations at the lower quantiles, the distribution of the residuals closely follow a normal distribution.

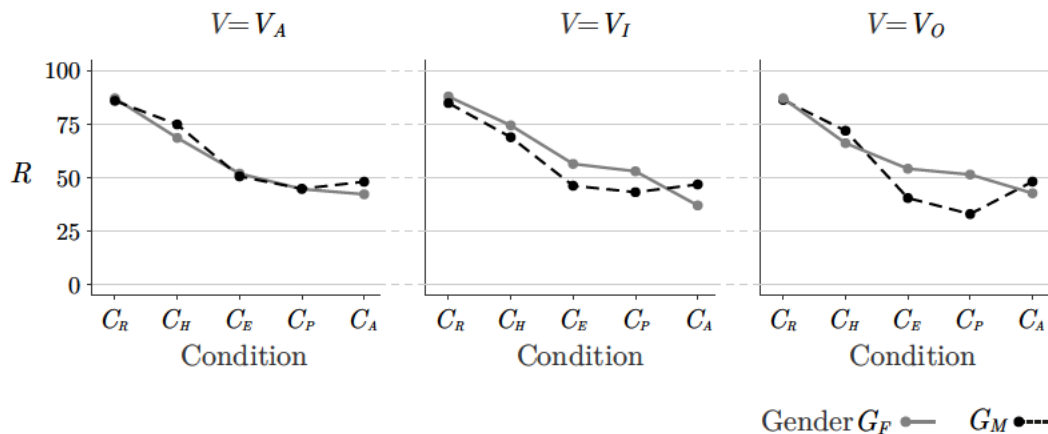


Figure 4.6: Estimates for rating R of the mixed model for conditions C separated by vowel V (separate plots) and gender G (separate lines).

Estimates & Post Hoc Tests Figure 4.6 shows the estimates of the ratings R for conditions C , genders G and vowels V . Post Hoc Tests with Bonferroni correction were performed for all main, 2-way and 3-way factors. An excerpt from the results is shown in Table 4.7. All post hoc results are shown in Appendix 5. The tests confirm that in general, ratings for all conditions C are significantly distinct with the exception $C_P - C_A$. For the 2-way effect Condition * Vowel, only the difference $V_I - V_O$ for condition C_P is significant. For the 2-way effect Condition * Gender, the ratings for the genders are significantly different for conditions C_E , C_P and C_A but not for C_R and C_H . The difference between male and female singers for C_E and C_P additionally varies between vowels where a strong difference can be seen for V_O , a less pronounced but still significant difference can be seen for V_I while no significant difference was found for the vowel V_A . These differences, especially that between conditions C and between genders G for C_E and C_P and different vowels are also visible in Figure 4.6.

Table 4.7: Post Hoc comparison results for fixed effects (extract). p-values are Bonferroni corrected. For full list see Appendix 5.

Comparison			Difference	SE	t	df	p ($\alpha = 0.05$)
C_H	-	C_P	25.848	0.851	30.382	4888	< 0.001
C_H	-	C_A	26.699	0.851	31.382	4888	< 0.001
C_H	-	C_E	20.881	0.851	24.543	4888	< 0.001
C_P	-	C_A	0.851	0.851	1.000	4888	1.000
C_R	-	C_H	15.759	0.851	18.523	4888	< 0.001
C_R	-	C_P	41.607	0.851	48.905	4888	< 0.001
C_R	-	C_A	42.458	0.851	49.905	4888	< 0.001
C_R	-	C_E	36.639	0.851	43.066	4888	< 0.001
C_E	-	C_P	4.968	0.851	5.839	4888	< 0.001
C_E	-	C_A	5.818	0.851	6.839	4888	< 0.001
$C_R V_I$	-	$C_R V_O$	-0.3152	1.47	-0.2139	4888	1.000
$C_R V_A$	-	$C_R V_I$	0.1091	1.47	0.0740	4888	1.000

Table 4.7 Continued: Post Hoc comparison results for fixed effects (extract). p-values are Bonferroni corrected. For full list see Appendix 5.

Comparison			Difference	SE	t	df	p ($\alpha = 0.05$)				
C_R	V_A	-	C_R	V_O	-0.2061	1.47	-0.1398	4888	1.000		
C_H	V_I	-	C_H	V_O	2.6424	1.47	1.7932	4888	1.000		
C_H	V_A	-	C_H	V_I	0.0818	1.47	0.0555	4888	1.000		
C_H	V_A	-	C_H	V_O	2.7242	1.47	1.8487	4888	1.000		
C_E	V_I	-	C_E	V_O	4.0000	1.47	2.7145	4888	0.699		
C_E	V_A	-	C_E	V_I	-0.0667	1.47	-0.0452	4888	1.000		
C_E	V_A	-	C_E	V_O	3.9333	1.47	2.6692	4888	0.801		
C_P	V_I	-	C_P	V_O	5.8909	1.47	3.9977	4888	0.007		
C_P	V_A	-	C_P	V_I	-3.3697	1.47	-2.2867	4888	1.000		
C_P	V_A	-	C_P	V_O	2.5212	1.47	1.7109	4888	1.000		
C_A	V_I	-	C_A	V_O	-3.5303	1.47	-2.3957	4888	1.000		
C_A	V_A	-	C_A	V_I	3.2576	1.47	2.2107	4888	1.000		
C_A	V_A	-	C_A	V_O	-0.2727	1.47	-0.1851	4888	1.000		
C_H	G_F	-	C_H	G_M	-2.190	1.20	-1.820	4888	1.000		
C_P	G_F	-	C_P	G_M	9.382	1.20	7.798	4888	< 0.001		
C_R	G_F	-	C_R	G_M	1.651	1.20	1.372	4888	1.000		
C_A	G_F	-	C_A	G_M	-7.055	1.20	-5.863	4888	< 0.001		
C_E	G_F	-	C_E	G_M	8.412	1.20	6.992	4888	< 0.001		
C_R	V_A	G_F	-	C_R	V_A	G_M	1.1576	2.08	0.5555	4888	1.000
C_R	V_I	G_F	-	C_R	V_I	G_M	3.0485	2.08	1.4628	4888	1.000
C_R	V_O	G_F	-	C_R	V_O	G_M	0.7455	2.08	0.3577	4888	1.000
C_H	V_A	G_F	-	C_H	V_A	G_M	-6.2788	2.08	-3.0129	4888	1.000
C_H	V_I	G_F	-	C_H	V_I	G_M	5.5212	2.08	2.6494	4888	1.000
C_H	V_O	G_F	-	C_H	V_O	G_M	-5.8121	2.08	-2.7890	4888	1.000
C_E	V_A	G_F	-	C_E	V_A	G_M	1.2970	2.08	0.6224	4888	1.000
C_E	V_I	G_F	-	C_E	V_I	G_M	10.2182	2.08	4.9033	4888	< 0.001
C_P	V_I	G_F	-	C_P	V_I	G_M	9.8061	2.08	4.7055	4888	0.001
C_E	V_O	G_F	-	C_E	V_O	G_M	13.7212	2.08	6.5842	4888	< 0.001
C_P	V_A	G_F	-	C_P	V_A	G_M	-0.1576	2.08	-0.0756	4888	1.000
C_P	V_O	G_F	-	C_P	V_O	G_M	18.4970	2.08	8.8759	4888	< 0.001
C_A	V_A	G_F	-	C_A	V_A	G_M	-5.8727	2.08	-2.8181	4888	1.000
C_A	V_I	G_F	-	C_A	V_I	G_M	-9.8424	2.08	-4.7230	4888	0.001
C_A	V_O	G_F	-	C_A	V_O	G_M	-5.4485	2.08	-2.6145	4888	1.000

4.2.3 Discussion

Goal of experiment 1 was to determine which component of the proposed multi-stage singing voice synthesis has a high impact on the perceived quality of synthesized samples and how different source samples affect the performance of the proposed method. To answer this question, a MUSHRA test was conducted with 33 participants. Five conditions were generated for 30 source samples each, taken from the dataset VocalSet [73]. In addition to the source sample, labeled as condition C_R , conditions included a harmonic reconstruction C_H , a resynthesis from estimated parameters C_E , a resynthesis from predicted parameters C_P and an anchor stimuli C_A . After a brief introduction, participants

were asked to comparatively rate the quality of these five conditions in 30 MUSHRA trials. The results were analyzed using the Friedman test with post hoc Wilcoxon signed-rank test and a mixed model. The hypothesis, that the quality ratings are ordered $R(C_R) > R(C_H) > R(C_E) > R(C_P)$, could be confirmed both with the Friedman test and the mixed model. Additional conclusions can be made from the individual ratings of each condition. As can be seen in Table 4.7, the difference between conditions C_H and C_R is minor, hinting that the harmonic model of the reference signal does not degrade the perceived audio quality much. The biggest difference can be seen between C_H and C_E . This suggests that currently, synthesis model parameter estimation produces the largest drop in terms of perceived audio quality. While the difference between C_P and C_E was significant in both the Friedman test and the mixed model Post Hoc test, it's substantially smaller compared to that between C_E and C_H , hinting that the training and subsequent prediction of parameters using the predictor network \mathcal{P} currently does not produce a large degradation of the audio quality.

From the mixed model, further statements can be made about the models performance for specific vowels or genders. In both experiments conducted as part of this thesis, only 5 female and 5 male singers were included. For this reason, some caution needs to be taken especially when drawing conclusion about 2-way and 3-way interactions. Post Hoc tests were done for all main, 2-way and 3-way effects. The results for post hoc comparisons discussed in this section can be found in Table 4.7. Starting with conditions C_R and C_H , no significant 2-way differences could be found, meaning that ratings for C_R or C_H are not significantly different between genders or vowels. For the anchor condition C_A , a significant difference was found between the two genders where a higher rating was found for male singers. This is likely a result of the difference in pitch range between male and female singers as female singers sung one octave higher than male singers. Because of this, more partials were cut off by the fixed lowpass filter used to create the anchor stimuli.

Some significant differences could be found between the ratings for female and male singers both for C_E and C_P where the quality rating for female singers was higher than that for male singers. From the 3-way post hoc tests, it can be taken that this difference varies between vowels where no significant difference can be observed for vowel $V = V_A$ while the largest difference can be observed for vowel $V = V_O$. It is important to note that these differences can be contributed to differences in voice qualities such as formant frequencies or to differences in the pitch range present in the samples as female singers sung one octave higher than male singers. Because of this and the apparent differences of this effect between vowels, it is hard to make assumptions about why the proposed synthesis model parameter estimation performs better for female singers, especially for the vowels i and o. A likely explanation might be that the synthesis model parameter estimation performs better for samples with higher pitches and the resulting lower density of harmonic partials. Another explanation might be that on average, formant frequencies for females are slightly higher for those of male singers as the vocal tract is usually shorter.

The first two research questions discussed at the start of this chapter can be answered as follows.

1. Which component of the proposed method outlined in section 3.1 has the biggest impact in terms of perceived audio quality of synthesized samples? It is strongly suggested by the results of this experiment, that synthesis model parameter estimation \mathcal{E} produces the biggest drop in perceived quality followed by the harmonic analysis \mathcal{A} . The smallest impact was observed for the parameter predictor \mathcal{P} .
2. How do different source samples such as samples from singers of different genders or different sung vowels, affect the performance in terms of overall perceived audio quality? The singers gender seems to affect the perceived audio quality of the synthesized sample strongly. A notable difference in quality of resynthesized samples (C_P) was found between male and female singers. Separated by vowels, this effect is insignificant for the vowel a, but significant for the vowels i and o. Averaged over both genders, the vowel however has no impact on the perceived quality of the resynthesized sample with the exception of a small significant difference between vowels i and o for C_P .

4.3 Experiment 2

The goal of the second experiment, a pair-wise timbre comparison test, was to evaluate the question, how the proposed method affects different timbre qualities. In order to answer the question, participants were asked to rate the relative presence of four timbre qualities, breathiness, roughness, brightness and naturalness, between two stimuli A and B, one being the hidden reference C_R and one being the resynthesized stimuli C_P as described in Section 4.1. A bipolar rating scale was used which measured the relative presence of these qualities between the two stimuli.

With this, it was possible to evaluate how well the proposed method preserves certain timbre qualities. A neutral rating would suggest that the proposed method preserves the tested timbre qualities well for the tested stimuli pair assuming that the specific quality was noticeably present in the reference sample. Any other rating would suggest that the method either emphasizes or de-emphasizes certain timbre qualities.

The four tested qualities, breathiness, roughness, brightness and naturalness, were selected mainly for two reasons. First, it was assumed that there is a common understanding among participants of these terms. As discussed in [89], the interpretation of semantics descriptors for acoustic qualities can vary across different languages and cultures. Second, the qualities are assumed to likely be affected by the proposed methods. Roughness and Breathiness both are commonly used to describe voice qualities [90, 91] while naturalness and brightness are widely used in the field of timbre semantics. A comprehensive discussion on that topic can be found in [89]. In the study, the following definitions were used for the timbre qualities.

- **Naturalness:** A voice sounds natural when it's in perceived accordance with the listeners expectation of the human voice.
- **Breathiness:** A breathy voice is characterized by an open sound with low intensity and audible noise, similar to a whispery voice.
- **Brightness:** Bright sounds are characterized by an emphasis on high frequency components as opposed to low frequency components.
- **Roughness:** A rough voice is characterized by an irregularity oscillation which can be perceived as scratchy and intense and can be considered opposite of a soft voice.

4.3.1 Method

Participants

At the end of the online survey, participants were asked to fill out a short questionnaire to collect demographic information of the participants. In total, 32 subjects participated in the second experiment, 11 of which stated that they identified as female and 21 identified as male. Participants' ages ranged from 21 to 60 ($M = 30.63$, $SD=9.11$). At least 17 of the participants were students at TU Berlin. Other participants came from the authors workplace or from the authors family and circle of friends.

Stimuli

The experiment consisted of a total of 30 tests, each presenting a pair of stimuli. As described in Section 4.1, stimuli were taken from the VocalSet dataset [73]. The selected subset included a total of 30 samples including sung scales from 5 singers of 2 genders, singing 3 vowels. In this experiment, one of the two stimuli was the hidden reference (condition C_R) taken directly from the VocalSet while the other stimulus was reconstructed from said reference using the proposed method (C_P). Stimuli were randomly labeled A or B .

Procedure

The second experiment was conducted as an online survey using the webMUSHRA API [78]. The API was modified to allow for a custom trial page design. After going through a general welcoming page detailing the length and structure of the two conducted experiments, participants were first introduced to the second experiment. Participants were advised to use headphones for the experiment and use a desktop or laptop machine with the Google Chrome online browser in cases of technical difficulties. The participants were instructed as follows.

- The goal of this experiment is to rate the perceived quality difference between two audio samples.
- You will be presented with two sounds, A and B and are asked to rate the difference regarding specific voice qualities aspects in 9 increments.
- The relevant voice quality terms are explained in the following audio pages using audio examples.

As with experiment A, participants are then forwarded to a page where they can adjust their volume and are asked to avoid adjusting the volume during the experiment.

On the following pages, the four timbre quality terms are defined as described at the start of Section 4.3. For breathiness, brightness and roughness, two exemplary stimuli were presented which were selected to represent either a low or high level of breathiness, brightness and roughness respectively.

After that, the participants had to go through 30 trial pages such as the one shown in Figure 4.7. Again, the order of these pages were randomized for each participant. Additionally, the position of the two conditions C_R and C_P , either A or B, on each page was randomized for every participant. The precise question asked was "Which of the two samples sound more breathy / more rough / brighter / more natural?". A 9-point scale was used for rating. The lowest (-4) and highest values (+4) would indicate that a timbre quality was perceived to be more prominent for stimulus A or B respectively while the center value (0) would indicate that a timbre quality was not perceived to be more prominent for either of the two stimuli.

Trial 1 of 30

Please answer the questions below

0:00 7:08

Play A Stop Play B

4 3 2 1 0 +1 +2 +3 +4
(strongly) (neutral) (strongly)

Breathiness: Which of the two samples sounds more breathy? A B

Roughness: Which of the two samples sounds more rough? A B

Brightness: Which of the two samples sounds brighter? A B

Naturalness: Which of the two samples sounds more natural? A B

Previous Next

TU weBMUSHR by AUDIO LABS Fraunhofer FAU

Figure 4.7: A trial page for experiment 2. The participants were instructed to rate which of the two presented stimuli sound more breathy / rough / bright / neutral.

After the 30 trial pages, participants were asked to fill out a small survey. In addition to the participants age and gender, the questionnaire asked whether the participant used headphones, loudspeakers or any other playback device. Additionally, students of TU Berlin were asked for their student identification number.

4.3.2 Results

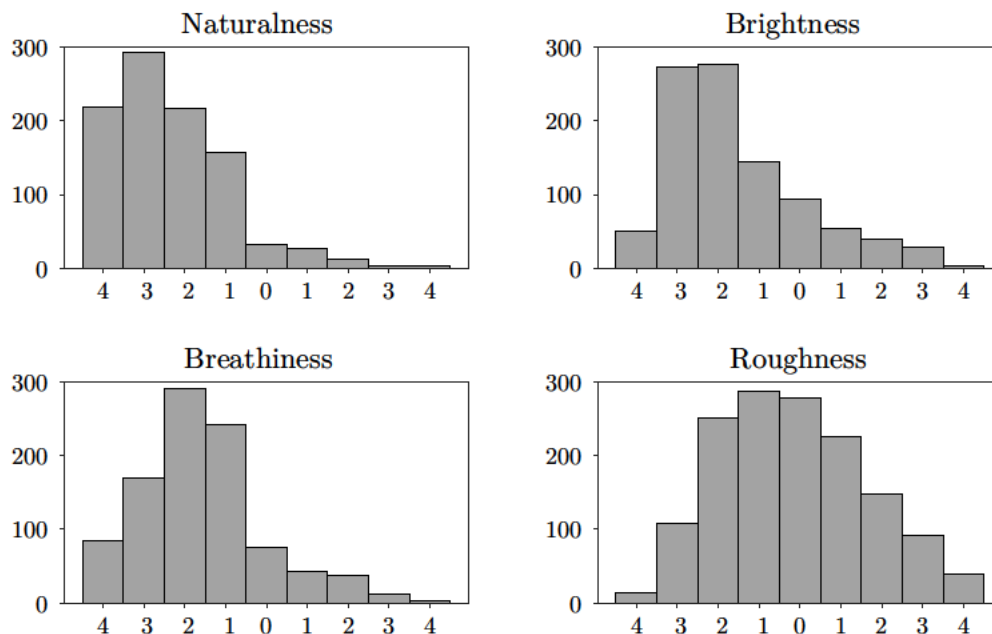


Figure 4.8: The histograms of the four ratings of naturalness, breathiness, brightness and roughness.

The histograms for the four tested timbre qualities naturalness R_N , breathiness R_{Bre} , brightness R_{Bri} and roughness R_R are shown in Figure 4.8. Before analyzing the data, the ratings were adjusted so that the reference stimuli would always correspond to stimuli A and to ratings < 0 .

Wilcoxon Signed-Rank Test

For all four qualities, the primary hypothesis is that $R < 0$, indicating that the relative qualities were rated in favor of the reference. The null hypothesis to reject thus is given with $R \geq 0$. To reject the null hypothesis and test the alternative hypothesis, the one-sided, one-sample Wilcoxon signed rank test was used [80, 82, 84]. Results from the test can be seen in Table 4.8.

Quality	median	SD	Wilcoxon Z	p ($\alpha = 0.05$)
Naturalness	-2	1.45	-25.16	< 0.001
Breathiness	-2	1.56	-21.27	< 0.001
Brightness	-2	1.70	-20.13	< 0.001
Roughness	0	1.81	-2.76	0.003

Table 4.8: Results of the Wilcoxon signed-rank test.

4.3.3 Discussion

The second experiment was conducted to investigate how well the proposed singing voice synthesis method preserves timbre qualities present in source samples when resynthesizing these. After a brief introduction, 32 participants were asked to rate the relative naturalness, breathiness, brightness and roughness between two stimuli for 30 stimuli pairs. The pairs included an unlabeled reference stimulus, taken from the VocalSet dataset [73], and its reconstruction using the proposed method. The results were analyzed using the Wilcoxon signed-rank test.

For all four tested voice qualities, naturalness, breathiness, brightness and roughness, the Wilcoxon signed-rank test returned significant results with strong deviations from a neutral rating for naturalness, brightness and breathiness and a small but significant deviation for roughness. For all four timbre qualities, the ratings shifted towards the reference stimulus, suggesting that participants felt that the reference signal sounded more natural, breathy, bright or rough respectively. Interestingly, the ratings for roughness showed the largest standard deviation and the smallest effect. In addition, some participants gave verbal feedback that roughness was hard to rate. There are various possible explanations for why the rating of roughness shows such a large standard deviation. For instance, the selected reference samples might not have included samples with a noticeably rough voice. Alternatively, the proposed singing voice synthesis method preserved the roughness and such there was no noticeable difference between reference and synthesized stimulus in terms of roughness. Finally, participants might have had issues associating the term roughness with some audible expectation and such couldn't consistently rate it.

The third research questions introduced at the start of this chapter can be answered as follows.

3. How well are timbre qualities such as breathiness or roughness preserved by the proposed method when resynthesizing samples? In general, a significant loss in naturalness, breathiness and brightness could be confirmed when comparing a reference stimulus with its reconstruction. Additional studies are likely necessary to obtain a more comprehensive understanding on how the method affects different timbres.

5 Conclusion

This thesis proposed a singing voice analysis and synthesis method suitable for real time applications. Using a multi-stage analysis approach, the method trains a voice model capturing vocal characteristics, such as vocal tract articulation or glottal intensity, from a source audio sample. During synthesis, said voice model is used in conjunction with robust DSP methods to efficiently synthesize new audio signals in real-time.

The proposed method can be separated into four stages, namely, harmonic analysis \mathcal{A} , synthesis model parameter estimation \mathcal{E} , predictor network \mathcal{P} and synthesis model \mathcal{S} as shown in figure 3.2. During analysis, pitch and harmonic partials \mathbf{O} are first extracted from audio frames of the source audio sample. Subsequently, the synthesis parameters \mathbf{X} are estimated by fitting the synthesis model to reconstruct the previously extracted harmonic partials. Finally, the neural network \mathcal{P} is trained to predict synthesis parameters $\hat{\mathbf{X}}$ from either from the original pitch trajectory \mathbf{f}_0 or from any alternative trajectory \mathbf{f}_0^* before they are used to generate audio using the synthesis model \mathcal{S} .

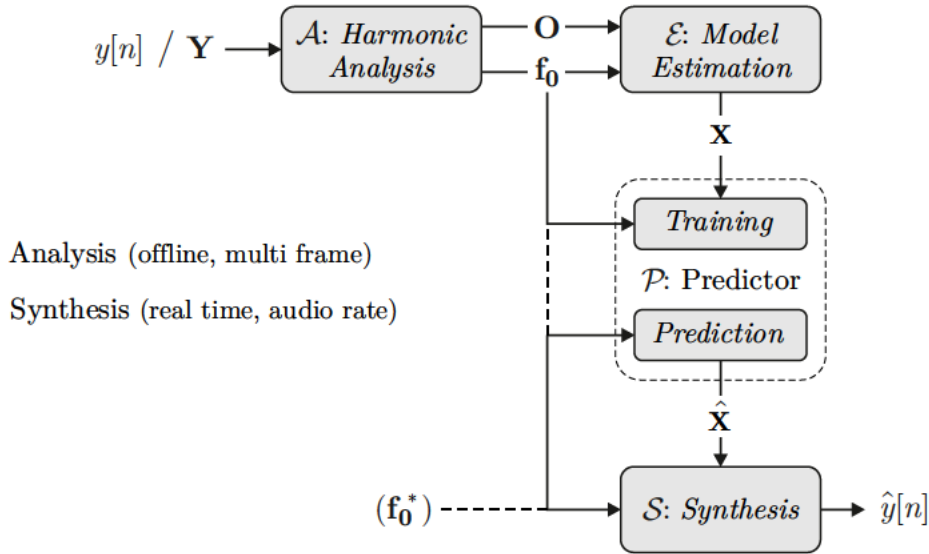


Figure 3.2: An overview of the proposed voice analysis and synthesis method. Repeated from Page 26.

While the synthesis was designed to be real-time capable, no such requirement was put on the analysis. Thus, the analysis is performed offline and takes full advantage of recent developments in the field of optimization and machine learning. Specifically, both the harmonic analysis \mathcal{A} and synthesis model parameter estimation \mathcal{E} utilize automatic differentiation to efficiently solve their respective optimization problems.

The synthesis on the other hand prioritizes computational efficiency and robustness to support real-time application cases. The synthesis model is based on well known DSP methods such as second order section filters and wavetable oscillators not only because of their computational efficiency but also because of their stability and robustness. The predictor \mathcal{P} , the link between analysis and synthesis, is implemented as a neural network which can also be implemented very efficiently. Thus, while only a prototype of the proposed synthesis was implemented, a real-time capable implementation, comparable to commercial software synthesizer in terms of computational efficiency, is straightforward. Such a real-time implementation would not only support desktop or laptop systems but also mobile or embedded devices.

To evaluate the performance of the proposed method especially in terms of perceived quality of synthesized audio samples, an empirical study consisting of two experiments was conducted. Goal of the first experiment was to evaluate what impact each component of the proposed analysis method has on the audio quality and to determine how different source samples, such as samples from different singer genders or different vowels, affect the audio quality. In this experiment, 33 participants rated the perceived quality of five stimuli, three of which corresponding to different components of the multi-stage analysis, in 30 MUSHRA tests. The results from this experiment suggests that the synthesis model parameter estimation \mathcal{E} currently produces the largest drop in terms of perceived audio quality. This effect was found to be less significant for female singers.

In a second experiment, participants were asked to rate the presence of four timbre qualities, naturalness, breathiness, brightness and roughness, between a reference sample and its resynthesis using the proposed method. Results showed that the 32 participants noticed a significant difference between the two stimuli for all four timbre qualities.

Multiple conclusions can be drawn from the evaluation. Firstly, to improve the overall performance in terms of sound quality, attention should be given to the synthesis model parameter estimation as it currently has the biggest impact on perceived audio quality. Secondly, additional work is necessary to determine how the pitch range present in the source audio samples affects the sound quality. As of now, it is expected that the proposed method performs better for source material containing higher pitches. Finally, the impact, including interaction effects, of vowel and gender on the methods performance should be evaluated in more detail as the first experiment suggests that the sound quality depends on both the vowel and gender.

The results from the first experiments confirmed existing suspicions about the performance of the proposed method. A difference in perceived audio quality between the reference sample and its harmonic reconstruction using harmonic analysis \mathcal{A} can likely be explained by the lack of inharmonic signal components such as aspiration noise or glottal jitter. These vocal qualities are neglected by the proposed method. While the synthesis could be adapted to model both aspiration noise and various kinds of aperiodic behavior of the glottal source, estimating these kinds of effects from a source audio sample would require substantial changes to the synthesis parameter estimation.

Improving the synthesis parameter estimation poses a more challenging problem. An initial intention of this thesis was to propose a model that would allow for morphing or blending between singer models or vowels during synthesis. As the design and development of the synthesis model parameter estimation required considerably more effort than expected, and the proposed filter model doesn't directly support interpolation between parameter sets as discussed in Section 3.2.3, this goal could not be achieved. In order to adjust the method to allow for morphing capabilities, the vocal tract filter model would need to be revised to achieve a closer correlation between filter parameters and vocal tract articulation or formant position.

In subjective listening tests, it was noticed that the optimization approach used in \mathcal{E} repeatedly was not able to successfully separate the effects of glottal source and vocal tract filter, especially in the lower frequency region below 2 kHz. This is likely explained by the flexibility of the synthesis model and the nature of the interaction between glottal source and vocal tract. As discussed in Section , separating a voice signal into glottal source and vocal tract filter poses a challenging issue mainly because, unless the filter model is specifically tailored to only reproduce vocal tract frequency responses, multiple possible separations of source and filter may explain an observed frequency spectrum. To combat this issue, the parameter estimation method \mathcal{E} restricts temporal fluctuations of the vocal tract parameters. However, even with this restriction, the parameter estimation at times produced unrealistic vocal tract parameterizations. For instance, the vocal tract was often found to fit poles to individual harmonic partials instead of the underlying, hidden, vocal tract formants. It is assumed that the high model complexity or flexibility of the vocal tract model is likely one of the main problems of the proposed method in its current form.

One solution to this problem is to reduce the flexibility of the vocal tract model to prevent it from producing unrealistic frequency responses. For instance, a more sophisticated vocal tract model could take inspiration from the Kelly-Lochbaum model or the Chain-Matrix model. A second approach would require the phase components of the observed harmonic partials to be included in the parameter estimation.

In contrast to the vocal tract model, the one parameter glottal flow model used in this work is assumed to be too restrictive. As described in Section 2.2.1, the glottal source may best be described by the position and magnitude of the glottal formant in the lower frequencies and the overall spectral slope of the glottal pulse spectrum. Thus, a two parameter glottal source model may represent a good compromise.

There are various possible future projects which can build on the findings of this work. As mentioned above, different alternative vocal tract and glottal source models may be tested in conjunction with the proposed analysis method to hopefully improve the robustness of the synthesis parameter estimation. Furthermore, a second attempt could be made to design a neural network, which practically would incorporate the estimation of synthesis parameter as part of the predictor network \mathcal{P} . Such a network could then be trained on multiple singers simultaneously and include additional control parameters

such as the singers gender or the vowel. This network could also be implemented as an end-to-end auto encoder as shown in Figure 5.2. In this case, additional latent space control parameters could be trained from source audio samples.

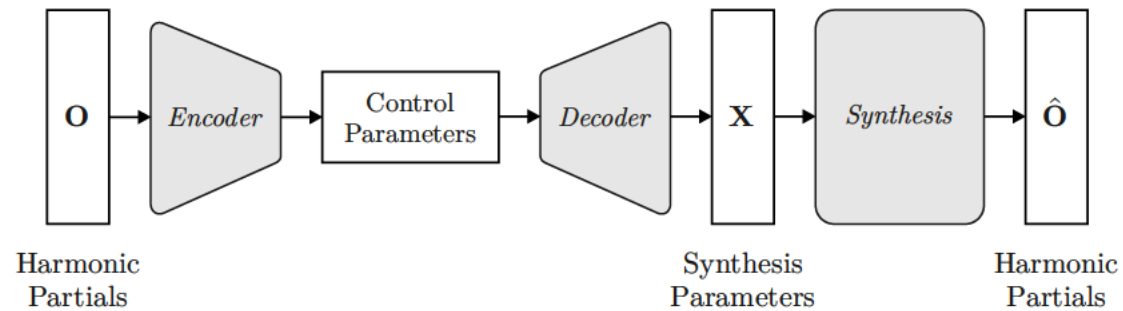


Figure 5.2: An alternative approach for the proposed singer model would include parameter estimation as part of a neural network.

Some aspects of the proposed method couldn't be assessed in the evaluation. Notably, the predictors ability to generalize, i.e. predict synthesis parameters from a previously unseen pitch trajectory, was not tested empirically. Some subjective tests with a hand full of synthesized audio samples confirmed that the predictor is generally capable of producing equally convincing results when predicting from a new pitch trajectory as when predicting from the source samples original pitch trajectory. This is likely explained with the limited input of the predictor as the network solely relies on pitch to estimate synthesis parameters. However, an empirical study would be required to confirm these preliminary results.

In conclusion, while this work couldn't propose a definitive solution to the problem of singing voice source filter separation, some findings can contribute to future research projects. Most importantly, the incorporation of DSP into machine learning and optimization methods using automatic differentiation seems to be a worthwhile endeavor for singing voice analysis and synthesis application cases. And further, IIR filters can in some application cases be integrated into gradient descent based methods by approximating their response in the frequency domain rather than calculating their precise but theoretically infinite impulse responses in the time domain.

Literature

- [1] Schweinberger, Stefan R.; Hideki Kawahara; Adrian P. Simpson; Verena G. Skuk; and Romi Zäske (2014): “Speaker perception”. In: *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(1), pp. 15–25. doi:10.1002/wcs.1261.
- [2] Fasold, Ralph W. and Jeff Connor-Linton (Eds.) (2006): *An introduction to language and linguistics*. Cambridge, UK: Cambridge University Press.
- [3] Hua, Kanru (2018): “Modeling Singing F0 With Neural Network Driven Transition-Sustain Models”. In: *arXiv:1803.04030 [cs, eess]*.
- [4] Blaauw, Merlijn and Jordi Bonada (2017): “A Neural Parametric Singing Synthesizer”. In: *INTERSPEECH 2017*. pp. 4001–4005. doi:10.21437/Interspeech.2017-1420.
- [5] Chen, C Julian (2016): *Elements of Human Voice*. 1. World Scientific.
- [6] Henrich, Nathalie; Christophe d’Alessandro; Boris Doval; and Michèle Castellengo (2005): “Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency”. In: *The Journal of the Acoustical Society of America*, 117(3), pp. 1417–1430. doi:10.1121/1.1850031.
- [7] Fant, Gunnar (1960): *Acoustic Theory of Speech Production*. Mouton, The Hague.
- [8] Boë, Louis-Jean; et al. (2019): “Which way to the dawn of speech?: Reanalyzing half a century of debates and data in light of speech science”. In: *Science Advances*, 5(12). doi:10.1126/sciadv.aaw3916.
- [9] Strange, Winifred (1989): “Evolving Theories of Vowel Perception”. In: *The Journal of the Acoustical Society of America*, 85(5).
- [10] Feng, Guanping and Elisa Castelli (1996): “Some acoustic features of nasal and nasalized vowels: a target for vowel nasalization.”. In: *The Journal of the Acoustical Society of America*, 99(6). doi:10.1121/1.414967.
- [11] Degottex, Gilles (2010): *Glottal source and vocal-tract separation*. Ph.D. thesis, Université Pierre-et-Marie-Curie.
- [12] Fant, G.; J. Liljencrants; and Qiguang Lin (1985): “A Four-Parameter Model of Glottal Flow”. In: *STL-QPSR*, 4.
- [13] Erickson, Molly (2016): “Acoustic Properties of the Voice Source and the Vocal Tract: Are They Perceptually Independent?”. In: *Journal of Voice*, 30. doi:10.1016/j.jvoice.2015.11.010.

- [14] Döllinger, Michael; David Berry; and Douglas Montequin (2006): “The influence of epilarynx area on vocal fold dynamics”. In: *Otolaryngology–Head and Neck Surgery*, 135, pp. 724–729. doi:10.1016/j.otohns.2006.04.007.
- [15] Kaburagi, Tokihiko; Momoyo Ando; and Yasufumi Uezu (2019): “Source-filter interaction in phonation: A study using vocal-tract data of a soprano singer”. In: *Acoustical Science and Technology*, 40(5), pp. 313–324. doi:10.1250/ast.40.313.
- [16] Dudley, Homer (1939): “The Vocoder”. In: *Bell Labs Rec*, 18(2), pp. 122–126.
- [17] Serra, Xavier and Julius Smith (1990): “Spectral Modeling Synthesis”. In: *Computer Music Journal*, 14.
- [18] Rosenberg, A. E. (1971): “Effect of Glottal Pulse Shape on the Quality of Natural Vowels”. In: *The Journal of the Acoustical Society of America*, 49(2B), pp. 583–590. doi:10.1121/1.1912389.
- [19] Klatt, D. H. and Lothar Klatt (1990): “Analysis, synthesis, and perception of voice quality variations among female and male talkers.”. In: *The Journal of the Acoustical Society of America*, 87(2). doi:10.1121/1.398894.
- [20] Fant, G; J Liljencrants; and Qiguang Lin (1995): “The LF-model revisited. Transformations and frequency domain analysis”. In: *STL-QPSR*, 2(3), pp. 119–156.
- [21] Doval, Boris (2003): “The voice source as a causal/anticausal linear filter”. In: *VOQUAL’03*.
- [22] Doval, Boris and C. d’Alessandro (1997): “Spectral correlates of glottal waveform models: an analytic study”. In: *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. doi:10.1109/ICASSP.1997.596183.
- [23] Degottex, Gilles; John Kane; Thomas Drugman; Tuomo Raitio; and Stefan Scherer (2014): “COVAREP: A Collaborative Voice Analysis Repository for Speech Technologies”. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*. doi:10.1109/ICASSP.2014.6853739.
- [24] Berezina-Greene, Maria; Daniel Rudoy; and Patrick Wolfe (2010): “Autoregressive modeling of voiced speech”. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 5042–5045. doi:10.1109/ICASSP.2010.5495058.
- [25] Walker, Jacqueline and Peter Murphy (2007): “A Review of Glottal Waveform Analysis”. In: *Progress in Nonlinear Speech Processing*. Berlin, Heidelberg: Springer-Verlag, pp. 1–21.
- [26] Kelly, J. L. and C. C. Lochbaum (1962): “Speech Synthesis”. In: *Proceedings of the Fourth International Congress on Acoustics*.
- [27] Smith, Julius O. (2010): *Physical Audio Signal Processing*. 2010 edition. W3K Publishing, <http://books.w3k.org/>,. URL <http://ccrma.stanford.edu/~jos/pasp/>.

- [28] Välimäki, Vesa and Matti Karjalainen (1994): “Improving the kelly-lochbaum vocal tract model using conical tube sections and fractional delay filtering techniques”. In: *ICSLP 94*.
- [29] Välimäki, Vesa; M. Karjalainen; and T. Kuisma (1994): “Articulatory control of a vocal tract model based on fractional delay waveguide filters”. In: *Proceedings of ICSPNN '94*. doi:10.1109/SIPNN.1994.344847.
- [30] Schnell, Karl and Arild Lacroix (2007): “Joint Analysis of Speech Frames for Synthesis Based on Lossy Tube Models”. In: *The 6th ISCA Speech Synthesis Workshop*.
- [31] Välimäki, Vesa (1995): *Discrete-time Modeling of Acoustic Tubes Using Fractional Delay Filters*. Ph.D. thesis.
- [32] Mullen, Jack (2006): *Physical Modelling of the Vocal Tract with the 2D Digital Waveguide Mesh*. Ph.D. thesis.
- [33] Panchapagesan, Sankaran and Abeer Alwan (2008): “Vocal tract inversion by cepstral analysis-by-synthesis using chain matrices”. In: *INTERSPEECH 2008*.
- [34] Panchapagesan, Sankaran and Abeer Alwan (2011): “A study of acoustic-to-articulatory inversion of speech by analysis-by-synthesis using chain matrices and the Maeda articulatory model”. In: *The Journal of the Acoustical Society of America*, 129(4). doi:10.1121/1.3514544.
- [35] Atal, B S and S L Hanauer (1971): “Speech Analysis and Synthesis by Linear Prediction of the Speech Wave”. In: *The Journal of the Acoustical Society of America*, 50(637).
- [36] Makhoul, John (1975): “Linear prediction: A tutorial review”. In: *Proceedings of the IEEE*, 63(4), pp. 561–580. doi:10.1109/PROC.1975.9792.
- [37] Roebel, Axel and Xavier Rodet (2005): “Efficient Spectral Envelope Estimation and its application to pitch shifting and envelope preservation”. In: *Proceedings of the 8th International Conference on Digital Audio Effects*.
- [38] Imai, Satoshi and Yoshihiro Abe (1979): “Spectral envelope extraction by improved cepstral method”. In: *Electronics and Communicatio*, 62-A(4), pp. 10–17.
- [39] Villavicencio, F.; A. Robel; and X. Rodet (2006): “Improving LPC Spectral Envelope Extraction Of Voiced Speech By True-Envelope Estimation”. In: *2006 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.
- [40] Degottex, Gilles; Luc Ardaillon; and Axel Roebel (2016): “Multi-Frame Amplitude Envelope Estimation for Modification of Singing Voice”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7), pp. 1242–1254. doi:10.1109/TASLP.2016.2551863.

- [41] Alku, Paavo (1992): “Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering”. In: *Speech Communication*, 11(2), pp. 109–118. doi:10.1016/0167-6393(92)90005-R.
- [42] Mokhtari, Parham and Hiroshi Ando (2017): “Iterative Optimal Preemphasis for Improved Glottal-Flow Estimation by Iterative Adaptive Inverse Filtering”. In: *INTERSPEECH 2017*. doi:10.21437/Interspeech.2017-79.
- [43] Perrotin, Olivier and Ian Vince McLoughlin (2017): “On the Use of a Spectral Glottal Model for the Source-filter Separation of Speech”. In: *arXiv:1712.08034 [eess.AS]*.
- [44] Mokhtari, Parham; Brad Story; Paavo Alku; and Hiroshi Ando (2018): “Estimation of the glottal flow from speech pressure signals: Evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production”. In: *Speech Communication*, 104, pp. 24–38. doi:10.1016/j.specom.2018.09.005.
- [45] Perrotin, Olivier and Ian McLoughlin (2019): “A Spectral Glottal Flow Model for Source-filter Separation of Speech”. In: *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 7160–7164. doi:10.1109/ICASSP.2019.8682625.
- [46] Schnell, K and A Lacroix (2003): “Analysis of Lossy Vocal Tract Models for Speech Production”. In: *INTERSPEECH 2003*.
- [47] Rasilo, Heikki; Unto Laine; and Okko Räsänen (2010): “Estimation of vocal tract shape trajectory using lossy Kelly-Lochbaum model”. In: *INTERSPEECH 2010*.
- [48] Choudhury, Aayush (2019): *A time domain vocal tract model with wall damping and visco-thermal losses*. Ph.D. thesis.
- [49] Bell, Chester Gordon; Hiroya Fujisaki; J. M. Heinz; K. N. Stevens; and A. S. House (1961): “Reduction of Speech Spectra by Analysis-by-Synthesis Techniques”. In: *The Journal of the Acoustical Society of America*, 33(12), pp. 1725–1736. doi:10.1121/1.1908556.
- [50] Shi, Liming; Jesper Kjær Nielsen; Jesper Rindom Jensen; and Mads Græsbøll Christensen (2017): “A Variational EM Method for Pole-Zero Modeling of Speech with Mixed Block Sparse and Gaussian Excitation”. In: *arXiv:1706.07927 [cs]*.
- [51] Blaauw, Merlijn and Jordi Bonada (2017): “A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs”. In: *Applied Sciences*, 7(12).
- [52] Engel, Jesse; Lamtharn Hantrakul; Chenjie Gu; and Adam Roberts (2020): “DDSP: Differentiable Digital Signal Processing”. In: *International Conference on Learning Representations 2020*.
- [53] Alzamendi, Gabriel A. and Gastón Schlotthauer (2017): “Modeling and joint estimation of glottal source and vocal tract filter by state-space methods”. In: *Biomedical Signal Processing and Control*, 37, pp. 5–15. doi:10.1016/j.bspc.2016.12.022.

- [54] Li, Yongwei; Ken-Ichi Sakakibara; and Masato Akagi (2019): “Simultaneous Estimation of Glottal Source Waveforms and Vocal Tract Shapes from Speech Signals Based on ARX-LF Model”. In: *Journal of Signal Processing Systems*. doi:10.1007/s11265-019-01510-4.
- [55] Marelli, Damián and Peter Balazs (2010): “On Pole-Zero Model Estimation Methods Minimizing a Logarithmic Criterion for Speech Analysis”. In: *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2), pp. 237–248. doi:10.1109/TASL.2009.2025544.
- [56] Venkataramani, Adarsh Akkshai (2018): *Time-Varying Modeling of Glottal Source and Vocal Tract and Sequential Bayesian Estimation of Model Parameters for Speech Synthesis*. Master’s thesis.
- [57] Schleusing, Olaf; Tomi Kinnunen; Brad Story; and Jean-Marc Vesin (2013): “Joint Source-Filter Optimization for Accurate Vocal Tract Estimation Using Differential Evolution”. In: *IEEE Transactions on Audio, Speech, and Language Processing*, 21(8), pp. 1560–1572. doi:10.1109/TASL.2013.2255275.
- [58] Jinachitra, P. and J.O. Smith (2005): “Joint estimation of glottal source and vocal tract for vocal synthesis using Kalman smoothing and EM algorithm”. In: *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. pp. 327–330. doi:10.1109/ASPAA.2005.1540235.
- [59] Wang, Yebin and Heming Zhao (2008): “Vocal tract resonances tracking by auxiliary particle filters”. In: *2008 International Conference on Audio, Language and Image Processing*. pp. 1131–1134. doi:10.1109/ICALIP.2008.4590124.
- [60] Kalgaonkar, Kaustubh and Mark Clements (2008): “Vocal tract area based formant tracking using particle filter”. In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. pp. 3405–3408. doi:10.1109/ICASSP.2008.4518382.
- [61] Oord, Aaron; et al. (2016): “WaveNet: A Generative Model for Raw Audio”. In: *The 9th ISCA Speech Synthesis Workshop*.
- [62] Chandna, Pritish; Merlijn Blaauw; Jordi Bonada; and Emilia Gomez (2019): “WGANSing: A Multi-Voice Singing Voice Synthesizer Based on the Wasserstein-GAN”. In: *2019 27th European Signal Processing Conference*.
- [63] Kalchbrenner, Nal; et al. (2018): “Efficient Neural Audio Synthesis”. In: *Proceedings of Machine Learning Research*, vol. 80. pp. 2410–2419.
- [64] Nakamura, Kazuhiro; Kei Hashimoto; Keiichiro Oura; Yoshihiko Nankaku; and Keiichi Tokuda (2019): “Singing voice synthesis based on convolutional neural networks”. In: *INTERSPEECH 2019*.
- [65] Morise, Masanori; Fumiya Yokomori; and Kenji Ozawa (2016): “WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications”. In: *IEICE Transactions on Information and Systems*, E99.D(7), pp. 1877–1884.

- [66] Kuznetsov, Boris; Julian D. Parker; and Fabián Esqueda (2020): “Differentiable IIR filters for machine learning applications”. In: *Proceedings of the 23rd International Conference on Digital Audio Effects*.
- [67] Nercessian, Shahan (2020): “Neural Parametric Equalizer Matching Using Differentiable Biquads”. In: *Proceedings of the 23rd International Conference on Digital Audio Effects*.
- [68] Abadi, Martín; et al. (2016): “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”. In: *arXiv:1603.04467 [cs]*.
- [69] de Boor, Carl (2001): *A Practical Guide to Splines*, vol. 27 of *Applied Mathematical Sciences*. New York: Springer-Verlag.
- [70] Hahn, Henrik (2015): *Expressive sampling synthesis. Learning extended source-filter models from instrument sound databases for expressive sample manipulations*. PhD Thesis.
- [71] Bristow-Johnson, Robert (1996): “Wavetable Synthesis 101, A Fundamental Perspective”. In: *Journal of the Audio Engineering Society*, vol. 101.
- [72] Massie, Dana C. (2002): “Wavetable Sampling Synthesis”. In: Mark Kahrs and Karlheinz Brandenburg (Eds.) *Applications of Digital Signal Processing to Audio and Acoustics*. Boston: Kluwer Academic Publishers, pp. 311–341.
- [73] Wilkins, Julia; Prem Seetharaman; Alison Wahl; and Bryan Pardo (2018): “VocalSet: A Singing Voice Dataset”. In: *ISMIR 2018*. pp. 468–474.
- [74] Serra, Xavier (1997): “Musical Sound Modeling with Sinusoids plus Noise”. In: *Musical Signal Processing*. pp. 91–122.
- [75] Ruder, Sebastian (2017): “An overview of gradient descent optimization algorithms”. In: *arXiv:1609.04747 [cs]*.
- [76] Kim, Jong Wook; Justin Salamon; Peter Li; and Juan Pablo Bello (2018): “CREPE: A Convolutional Representation for Pitch Estimation”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*.
- [77] Kingma, Diederik P. and Jimmy Ba (2015): “Adam: A Method for Stochastic Optimization”. In: *ICLR 2015*.
- [78] Schoeffler, Michael; Bernd Edler; Fabian-Robert Stöter; and Jürgen Herre (2015): “Towards the Next Generation of Web-based Experiments: A Case Study Assessing Basic Audio Quality Following the ITU-R Recommendation BS.1534 (MUSHRA)”. In: *Proceedings of 1st Web Audio Conference*.
- [79] ITU-R Recommendation BS.1534 (2003): *Method for the subjective assessment of intermediate quality level of coding systems*. Tech. rep., ITU.
- [80] (2020): *MATLAB*, vol. version 9.8.0 (R2020a). Natick, Massachusetts: The MathWorks Inc.

- [81] Hollander, Myles; Douglas A. Wolfe; and Eric Chicken (1999): *Nonparametric Statistical Methods*. Hoboken, NJ: John Wiley & Sons, Inc.
- [82] Hogg, Robert V. and Johannes Ledolter (1987): *Engineering Statistics*. New York: MacMillan.
- [83] Mendonça, Catarina and Symeon Delikaris-Manias (2018): “Statistical tests with MUSHRA data”. In: *Journal of the Audio Engineering Society*, vol. 144.
- [84] Gibbons, Jean D. and Subhabrata Chakraborti (2011): *Nonparametric Statistical Inference*. 5. Boca Raton, FL: Chapman & Hall/CRC Press, Taylor & Francis Group.
- [85] Seltman, Howard (2018): *Experimental Design and Analysis*. Carnegie Mellon University.
- [86] The jamovi project (2020): *jamovi*. URL <https://www.jamovi.org>.
- [87] R Core Team (2020): *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing. URL <https://www.R-project.org/>.
- [88] Gallucci, M. (2019): *GAMLj: General analyses for linear models*. URL <https://gamlj.github.io/>.
- [89] Saitis, Charalampos and Stefan Weinzierl (2019): “The Semantics of Timbre”. In: Kai Siedenburg; Charalampos Saitis; Stephen McAdams; Arthur N. Popper; and Richard R. Fay (Eds.) *Timbre: Acoustics, Perception, and Cognition*, vol. 69 of *Springer Handbook of Auditory Research*. Cham: Springer International Publishing, pp. 119–149.
- [90] Wendler, J.; A. Rauhut; and H. Krüger (1986): “Classification of voice qualities”. In: *Journal of Phonetics*, 14(3-4), pp. 483–488. doi:10.1016/S0095-4470(19)30694-1.
- [91] Henrich, Nathalie; et al. (2008): “Towards a Common Terminology to Describe Voice Quality in Western Lyrical Singing: Contribution of a Multidisciplinary Research Group”. In: *Journal of Interdisciplinary Music Studies*, 2.

This page is intentionally left blank.

Appendix A

Mixed Model Post Hoc Test Results

Comparison			Difference	SE	t	df	<i>p</i> _{bonferroni}
C_H	-	C_P	25.848	0.851	30.382	4888	< 0.001
C_H	-	C_A	26.699	0.851	31.382	4888	< 0.001
C_H	-	C_E	20.881	0.851	24.543	4888	< 0.001
C_P	-	C_A	0.851	0.851	1.000	4888	1.000
C_R	-	C_H	15.759	0.851	18.523	4888	< 0.001
C_R	-	C_P	41.607	0.851	48.905	4888	< 0.001
C_R	-	C_A	42.458	0.851	49.905	4888	< 0.001
C_R	-	C_E	36.639	0.851	43.066	4888	< 0.001
C_E	-	C_P	4.968	0.851	5.839	4888	< 0.001
C_E	-	C_A	5.818	0.851	6.839	4888	< 0.001
V_I	-	V_O	1.73758	0.659	2.63667	4888	0.025
V_A	-	V_I	0.00242	0.659	0.00368	4888	1.000
V_A	-	V_O	1.74000	0.659	2.64035	4888	0.025
G_F	-	G_M	2.04	0.538	3.79	4888	< 0.001

Table A.1: Post hoc comparisons for main effects.

Comparison				Difference	SE	t	df	<i>p</i> _{bonferroni}
V_I	G_M	-	V_O G_M	2.033	0.932	2.181	4888	0.438
V_I	G_F	-	V_I G_M	3.750	0.932	4.024	4888	< 0.001
V_I	G_F	-	V_O G_M	5.783	0.932	6.205	4888	< 0.001
V_I	G_F	-	V_O G_F	1.442	0.932	1.548	4888	1.000
V_I	G_F	-	V_A G_M	0.887	0.932	0.952	4888	1.000
V_O	G_F	-	V_I G_M	2.308	0.932	2.476	4888	0.200
V_O	G_F	-	V_O G_M	4.341	0.932	4.657	4888	< 0.001
V_O	G_F	-	V_A G_M	-0.555	0.932	-0.596	4888	1.000
V_A	G_M	-	V_I G_M	2.863	0.932	3.072	4888	0.032
V_A	G_M	-	V_O G_M	4.896	0.932	5.253	4888	< 0.001
V_A	G_F	-	V_I G_M	0.892	0.932	0.957	4888	1.000
V_A	G_F	-	V_I G_F	-2.858	0.932	-3.067	4888	0.033
V_A	G_F	-	V_O G_M	2.925	0.932	3.138	4888	0.026
V_A	G_F	-	V_O G_F	-1.416	0.932	-1.519	4888	1.000
V_A	G_F	-	V_A G_M	-1.971	0.932	-2.115	4888	0.517

Table A.2: Post hoc comparisons for Vowel * Gender

Table A.3: Post hoc comparisons for Condition * Vowel

Comparison					Difference	SE	t	df	<i>P</i> _{bonferroni}
C_H	V_I	-	C_H	V_O	2.6424	1.47	1.7932	4888	1.000
C_H	V_I	-	C_P	V_I	23.6152	1.47	16.0257	4888	< 0.001
C_H	V_I	-	C_P	V_O	29.5061	1.47	20.0234	4888	< 0.001
C_H	V_I	-	C_R	V_O	-15.0788	1.47	-10.2328	4888	< 0.001
C_H	V_I	-	C_A	V_I	29.8152	1.47	20.2332	4888	< 0.001
C_H	V_I	-	C_A	V_O	26.2848	1.47	17.8374	4888	< 0.001
C_H	V_I	-	C_E	V_I	20.3788	1.47	13.8295	4888	< 0.001
C_H	V_I	-	C_E	V_O	24.3788	1.47	16.5439	4888	< 0.001
C_H	V_O	-	C_P	V_O	26.8636	1.47	18.2302	4888	< 0.001
C_H	V_O	-	C_A	V_O	23.6424	1.47	16.0442	4888	< 0.001
C_H	V_O	-	C_E	V_O	21.7364	1.47	14.7507	4888	< 0.001
C_H	V_A	-	C_H	V_I	0.0818	1.47	0.0555	4888	1.000
C_H	V_A	-	C_H	V_O	2.7242	1.47	1.8487	4888	1.000
C_H	V_A	-	C_P	V_I	23.6970	1.47	16.0812	4888	< 0.001
C_H	V_A	-	C_P	V_O	29.5879	1.47	20.0789	4888	< 0.001
C_H	V_A	-	C_P	V_A	27.0667	1.47	18.3680	4888	< 0.001
C_H	V_A	-	C_R	V_I	-14.6818	1.47	-9.9634	4888	< 0.001
C_H	V_A	-	C_R	V_O	-14.9970	1.47	-10.1772	4888	< 0.001
C_H	V_A	-	C_A	V_I	29.8970	1.47	20.2887	4888	< 0.001
C_H	V_A	-	C_A	V_O	26.3667	1.47	17.8930	4888	< 0.001
C_H	V_A	-	C_A	V_A	26.6394	1.47	18.0780	4888	< 0.001
C_H	V_A	-	C_E	V_I	20.4606	1.47	13.8850	4888	< 0.001
C_H	V_A	-	C_E	V_O	24.4606	1.47	16.5995	4888	< 0.001
C_H	V_A	-	C_E	V_A	20.5273	1.47	13.9302	4888	< 0.001
C_P	V_I	-	C_H	V_O	-20.9727	1.47	-14.2325	4888	< 0.001
C_P	V_I	-	C_P	V_O	5.8909	1.47	3.9977	4888	0.007
C_P	V_I	-	C_R	V_O	-38.6939	1.47	-26.2585	4888	< 0.001
C_P	V_I	-	C_A	V_I	6.2000	1.47	4.2074	4888	0.003
C_P	V_I	-	C_A	V_O	2.6697	1.47	1.8117	4888	1.000
C_P	V_I	-	C_E	V_O	0.7636	1.47	0.5182	4888	1.000
C_P	V_O	-	C_A	V_O	-3.2212	1.47	-2.1860	4888	1.000
C_P	V_A	-	C_H	V_I	-26.9848	1.47	-18.3125	4888	< 0.001
C_P	V_A	-	C_H	V_O	-24.3424	1.47	-16.5193	4888	< 0.001
C_P	V_A	-	C_P	V_I	-3.3697	1.47	-2.2867	4888	1.000
C_P	V_A	-	C_P	V_O	2.5212	1.47	1.7109	4888	1.000
C_P	V_A	-	C_R	V_I	-41.7485	1.47	-28.3314	4888	< 0.001
C_P	V_A	-	C_R	V_O	-42.0636	1.47	-28.5452	4888	< 0.001
C_P	V_A	-	C_A	V_I	2.8303	1.47	1.9207	4888	1.000
C_P	V_A	-	C_A	V_O	-0.7000	1.47	-0.4750	4888	1.000
C_P	V_A	-	C_A	V_A	-0.4273	1.47	-0.2900	4888	1.000
C_P	V_A	-	C_E	V_I	-6.6061	1.47	-4.4830	4888	< 0.001
C_P	V_A	-	C_E	V_O	-2.6061	1.47	-1.7685	4888	1.000
C_R	V_I	-	C_H	V_I	14.7636	1.47	10.0189	4888	< 0.001
C_R	V_I	-	C_H	V_O	17.4061	1.47	11.8121	4888	< 0.001
C_R	V_I	-	C_P	V_I	38.3788	1.47	26.0446	4888	< 0.001
C_R	V_I	-	C_P	V_O	44.2697	1.47	30.0423	4888	< 0.001
C_R	V_I	-	C_R	V_O	-0.3152	1.47	-0.2139	4888	1.000
C_R	V_I	-	C_A	V_I	44.5788	1.47	30.2521	4888	< 0.001
C_R	V_I	-	C_A	V_O	41.0485	1.47	27.8563	4888	< 0.001
C_R	V_I	-	C_E	V_I	35.1424	1.47	23.8484	4888	< 0.001
C_R	V_I	-	C_E	V_O	39.1424	1.47	26.5628	4888	< 0.001
C_R	V_O	-	C_H	V_O	17.7212	1.47	12.0260	4888	< 0.001
C_R	V_O	-	C_P	V_O	44.5848	1.47	30.2562	4888	< 0.001
C_R	V_O	-	C_A	V_O	41.3636	1.47	28.0702	4888	< 0.001

Table A.3 Continued: Post Hoc Comparisons - Condition * Vowel

Comparison				Difference	SE	t	df	<i>P</i> _{bonferroni}	
C_R	V_O	-	C_E	V_O	39.4576	1.47	26.7767	4888	< 0.001
C_R	V_A	-	C_H	V_I	14.8727	1.47	10.0929	4888	< 0.001
C_R	V_A	-	C_H	V_O	17.5152	1.47	11.8861	4888	< 0.001
C_R	V_A	-	C_H	V_A	14.7909	1.47	10.0374	4888	< 0.001
C_R	V_A	-	C_P	V_I	38.4879	1.47	26.1187	4888	< 0.001
C_R	V_A	-	C_P	V_O	44.3788	1.47	30.1163	4888	< 0.001
C_R	V_A	-	C_P	V_A	41.8576	1.47	28.4054	4888	< 0.001
C_R	V_A	-	C_R	V_I	0.1091	1.47	0.0740	4888	1.000
C_R	V_A	-	C_R	V_O	-0.2061	1.47	-0.1398	4888	1.000
C_R	V_A	-	C_A	V_I	44.6879	1.47	30.3261	4888	< 0.001
C_R	V_A	-	C_A	V_O	41.1576	1.47	27.9304	4888	< 0.001
C_R	V_A	-	C_A	V_A	41.4303	1.47	28.1154	4888	< 0.001
C_R	V_A	-	C_E	V_I	35.2515	1.47	23.9224	4888	< 0.001
C_R	V_A	-	C_E	V_O	39.2515	1.47	26.6369	4888	< 0.001
C_R	V_A	-	C_E	V_A	35.3182	1.47	23.9676	4888	< 0.001
C_A	V_I	-	C_H	V_O	-27.1727	1.47	-18.4400	4888	< 0.001
C_A	V_I	-	C_P	V_O	-0.3091	1.47	-0.2098	4888	1.000
C_A	V_I	-	C_R	V_O	-44.8939	1.47	-30.4659	4888	< 0.001
C_A	V_I	-	C_A	V_O	-3.5303	1.47	-2.3957	4888	1.000
C_A	V_I	-	C_E	V_O	-5.4364	1.47	-3.6892	4888	0.024
C_A	V_A	-	C_H	V_I	-26.5576	1.47	-18.0225	4888	< 0.001
C_A	V_A	-	C_H	V_O	-23.9152	1.47	-16.2293	4888	< 0.001
C_A	V_A	-	C_P	V_I	-2.9424	1.47	-1.9968	4888	1.000
C_A	V_A	-	C_P	V_O	2.9485	1.47	2.0009	4888	1.000
C_A	V_A	-	C_R	V_I	-41.3212	1.47	-28.0414	4888	< 0.001
C_A	V_A	-	C_R	V_O	-41.6364	1.47	-28.2553	4888	< 0.001
C_A	V_A	-	C_A	V_I	3.2576	1.47	2.2107	4888	1.000
C_A	V_A	-	C_A	V_O	-0.2727	1.47	-0.1851	4888	1.000
C_A	V_A	-	C_E	V_I	-6.1788	1.47	-4.1931	4888	0.003
C_A	V_A	-	C_E	V_O	-2.1788	1.47	-1.4786	4888	1.000
C_E	V_I	-	C_H	V_O	-17.7364	1.47	-12.0363	4888	< 0.001
C_E	V_I	-	C_P	V_I	3.2364	1.47	2.1963	4888	1.000
C_E	V_I	-	C_P	V_O	9.1273	1.47	6.1940	4888	< 0.001
C_E	V_I	-	C_R	V_O	-35.4576	1.47	-24.0622	4888	< 0.001
C_E	V_I	-	C_A	V_I	9.4364	1.47	6.4037	4888	< 0.001
C_E	V_I	-	C_A	V_O	5.9061	1.47	4.0080	4888	0.007
C_E	V_I	-	C_E	V_O	4.0000	1.47	2.7145	4888	0.699
C_E	V_O	-	C_P	V_O	5.1273	1.47	3.4795	4888	0.053
C_E	V_O	-	C_A	V_O	1.9061	1.47	1.2935	4888	1.000
C_E	V_A	-	C_H	V_I	-20.4455	1.47	-13.8747	4888	< 0.001
C_E	V_A	-	C_H	V_O	-17.8030	1.47	-12.0815	4888	< 0.001
C_E	V_A	-	C_P	V_I	3.1697	1.47	2.1510	4888	1.000
C_E	V_A	-	C_P	V_O	9.0606	1.47	6.1487	4888	< 0.001
C_E	V_A	-	C_P	V_A	6.5394	1.47	4.4378	4888	< 0.001
C_E	V_A	-	C_R	V_I	-35.2091	1.47	-23.8936	4888	< 0.001
C_E	V_A	-	C_R	V_O	-35.5242	1.47	-24.1075	4888	< 0.001
C_E	V_A	-	C_A	V_I	9.3697	1.47	6.3585	4888	< 0.001
C_E	V_A	-	C_A	V_O	5.8394	1.47	3.9627	4888	0.008
C_E	V_A	-	C_A	V_A	6.1121	1.47	4.1478	4888	0.004
C_E	V_A	-	C_E	V_I	-0.0667	1.47	-0.0452	4888	1.000
C_E	V_A	-	C_E	V_O	3.9333	1.47	2.6692	4888	0.801

Table A.4: Post hoc comparisons for Condition * Gender

Comparison					Difference	SE	t	df	<i>p</i> _{bonferroni}
C_H	G_M	-	C_P	G_M	31.634	1.20	26.292	4888	< 0.001
C_H	G_M	-	C_A	G_M	24.267	1.20	20.169	4888	< 0.001
C_H	G_M	-	C_E	G_M	26.182	1.20	21.761	4888	< 0.001
C_H	G_F	-	C_H	G_M	-2.190	1.20	-1.820	4888	1.000
C_H	G_F	-	C_P	G_M	29.444	1.20	24.472	4888	< 0.001
C_H	G_F	-	C_P	G_F	20.063	1.20	16.675	4888	< 0.001
C_H	G_F	-	C_R	G_M	-16.028	1.20	-13.322	4888	< 0.001
C_H	G_F	-	C_A	G_M	22.077	1.20	18.349	4888	< 0.001
C_H	G_F	-	C_A	G_F	29.131	1.20	24.212	4888	< 0.001
C_H	G_F	-	C_E	G_M	23.992	1.20	19.941	4888	< 0.001
C_H	G_F	-	C_E	G_F	15.580	1.20	12.949	4888	< 0.001
C_P	G_M	-	C_A	G_M	-7.368	1.20	-6.124	4888	< 0.001
C_P	G_F	-	C_H	G_M	-22.253	1.20	-18.495	4888	< 0.001
C_P	G_F	-	C_P	G_M	9.382	1.20	7.798	4888	< 0.001
C_P	G_F	-	C_R	G_M	-36.091	1.20	-29.996	4888	< 0.001
C_P	G_F	-	C_A	G_M	2.014	1.20	1.674	4888	1.000
C_P	G_F	-	C_A	G_F	9.069	1.20	7.537	4888	< 0.001
C_P	G_F	-	C_E	G_M	3.929	1.20	3.266	4888	0.049
C_R	G_M	-	C_H	G_M	13.838	1.20	11.502	4888	< 0.001
C_R	G_M	-	C_P	G_M	45.473	1.20	37.794	4888	< 0.001
C_R	G_M	-	C_A	G_M	38.105	1.20	31.671	4888	< 0.001
C_R	G_M	-	C_E	G_M	40.020	1.20	33.262	4888	< 0.001
C_R	G_F	-	C_H	G_M	15.489	1.20	12.873	4888	< 0.001
C_R	G_F	-	C_H	G_F	17.679	1.20	14.693	4888	< 0.001
C_R	G_F	-	C_P	G_M	47.123	1.20	39.166	4888	< 0.001
C_R	G_F	-	C_P	G_F	37.741	1.20	31.368	4888	< 0.001
C_R	G_F	-	C_R	G_M	1.651	1.20	1.372	4888	1.000
C_R	G_F	-	C_A	G_M	39.756	1.20	33.042	4888	< 0.001
C_R	G_F	-	C_A	G_F	46.810	1.20	38.906	4888	< 0.001
C_R	G_F	-	C_E	G_M	41.671	1.20	34.634	4888	< 0.001
C_R	G_F	-	C_E	G_F	33.259	1.20	27.642	4888	< 0.001
C_A	G_F	-	C_H	G_M	-31.321	1.20	-26.032	4888	< 0.001
C_A	G_F	-	C_P	G_M	0.313	1.20	0.260	4888	1.000
C_A	G_F	-	C_R	G_M	-45.160	1.20	-37.534	4888	< 0.001
C_A	G_F	-	C_A	G_M	-7.055	1.20	-5.863	4888	< 0.001
C_A	G_F	-	C_E	G_M	-5.139	1.20	-4.272	4888	< 0.001
C_E	G_M	-	C_P	G_M	5.453	1.20	4.532	4888	< 0.001
C_E	G_M	-	C_A	G_M	-1.915	1.20	-1.592	4888	1.000
C_E	G_F	-	C_H	G_M	-17.770	1.20	-14.769	4888	< 0.001
C_E	G_F	-	C_P	G_M	13.865	1.20	11.523	4888	< 0.001
C_E	G_F	-	C_P	G_F	4.483	1.20	3.726	4888	0.009
C_E	G_F	-	C_R	G_M	-31.608	1.20	-26.271	4888	< 0.001
C_E	G_F	-	C_A	G_M	6.497	1.20	5.400	4888	< 0.001
C_E	G_F	-	C_A	G_F	13.552	1.20	11.263	4888	< 0.001
C_E	G_F	-	C_E	G_M	8.412	1.20	6.992	4888	< 0.001

Table A.5: Post hoc comparisons for Gender * Vowel * Condition

Comparison					Difference	SE	t	df	<i>p</i> _{bonferroni}		
G_M	V_I	C_H	-	G_M	V_I	C_P	25.7576	2.08	12.3600	4888	< 0.001
G_M	V_I	C_H	-	G_M	V_I	C_A	22.1333	2.08	10.6208	4888	< 0.001
G_M	V_I	C_H	-	G_M	V_I	C_E	22.7273	2.08	10.9058	4888	< 0.001
G_M	V_I	C_H	-	G_M	V_O	C_H	-3.0242	2.08	-1.4512	4888	1.000

Table A.5 Continued: Post Hoc Comparisons - Gender * Vowel * Condition

Comparison				Difference	SE	t	df	<i>P</i> _{bonferroni}
$G_M V_I C_H$	-	$G_M V_O C_P$		35.9939	2.08	17.2719	4888	< 0.001
$G_M V_I C_H$	-	$G_M V_O C_A$		20.8000	2.08	9.9810	4888	< 0.001
$G_M V_I C_H$	-	$G_M V_O C_E$		28.4788	2.08	13.6657	4888	< 0.001
$G_M V_I C_H$	-	$G_M V_A C_P$		24.1455	2.08	11.5864	4888	< 0.001
$G_M V_I C_H$	-	$G_M V_A C_A$		20.8606	2.08	10.0101	4888	< 0.001
$G_M V_I C_H$	-	$G_M V_A C_E$		18.3333	2.08	8.7974	4888	< 0.001
$G_M V_I C_H$	-	$G_F V_I C_P$		15.9515	2.08	7.6544	4888	< 0.001
$G_M V_I C_H$	-	$G_F V_I C_A$		31.9758	2.08	15.3438	4888	< 0.001
$G_M V_I C_H$	-	$G_F V_I C_E$		12.5091	2.08	6.0026	4888	< 0.001
$G_M V_I C_H$	-	$G_F V_O C_H$		2.7879	2.08	1.3378	4888	1.000
$G_M V_I C_H$	-	$G_F V_O C_P$		17.4970	2.08	8.3960	4888	< 0.001
$G_M V_I C_H$	-	$G_F V_O C_A$		26.2485	2.08	12.5955	4888	< 0.001
$G_M V_I C_H$	-	$G_F V_O C_E$		14.7576	2.08	7.0815	4888	< 0.001
$G_M V_I C_H$	-	$G_F V_A C_P$		24.3030	2.08	11.6620	4888	< 0.001
$G_M V_I C_H$	-	$G_F V_A C_A$		26.7333	2.08	12.8282	4888	< 0.001
$G_M V_I C_H$	-	$G_F V_A C_E$		17.0364	2.08	8.1750	4888	< 0.001
$G_M V_I C_P$	-	$G_M V_I C_A$		-3.6242	2.08	-1.7391	4888	1.000
$G_M V_I C_P$	-	$G_M V_O C_P$		10.2364	2.08	4.9120	4888	< 0.001
$G_M V_I C_P$	-	$G_M V_O C_A$		-4.9576	2.08	-2.3789	4888	1.000
$G_M V_I C_P$	-	$G_M V_A C_A$		-4.8970	2.08	-2.3498	4888	1.000
$G_M V_I C_P$	-	$G_F V_I C_A$		6.2182	2.08	2.9838	4888	1.000
$G_M V_I C_P$	-	$G_F V_O C_P$		-8.2606	2.08	-3.9639	4888	0.033
$G_M V_I C_P$	-	$G_F V_O C_A$		0.4909	2.08	0.2356	4888	1.000
$G_M V_I C_P$	-	$G_F V_A C_A$		0.9758	2.08	0.4682	4888	1.000
$G_M V_I C_R$	-	$G_M V_I C_H$		16.0000	2.08	7.6777	4888	< 0.001
$G_M V_I C_R$	-	$G_M V_I C_P$		41.7576	2.08	20.0377	4888	< 0.001
$G_M V_I C_R$	-	$G_M V_I C_A$		38.1333	2.08	18.2985	4888	< 0.001
$G_M V_I C_R$	-	$G_M V_I C_E$		38.7273	2.08	18.5836	4888	< 0.001
$G_M V_I C_R$	-	$G_M V_O C_H$		12.9758	2.08	6.2265	4888	< 0.001
$G_M V_I C_R$	-	$G_M V_O C_P$		51.9939	2.08	24.9497	4888	< 0.001
$G_M V_I C_R$	-	$G_M V_O C_R$		-1.4667	2.08	-0.7038	4888	1.000
$G_M V_I C_R$	-	$G_M V_O C_A$		36.8000	2.08	17.6587	4888	< 0.001
$G_M V_I C_R$	-	$G_M V_O C_E$		44.4788	2.08	21.3435	4888	< 0.001
$G_M V_I C_R$	-	$G_M V_A C_H$		10.0182	2.08	4.8073	4888	< 0.001
$G_M V_I C_R$	-	$G_M V_A C_P$		40.1455	2.08	19.2641	4888	< 0.001
$G_M V_I C_R$	-	$G_M V_A C_A$		36.8606	2.08	17.6878	4888	< 0.001
$G_M V_I C_R$	-	$G_M V_A C_E$		34.3333	2.08	16.4751	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_I C_H$		10.4788	2.08	5.0283	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_I C_P$		31.9515	2.08	15.3322	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_I C_A$		47.9758	2.08	23.0215	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_I C_E$		28.5091	2.08	13.6803	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_O C_H$		18.7879	2.08	9.0155	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_O C_P$		33.4970	2.08	16.0738	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_O C_R$		-2.2121	2.08	-1.0615	4888	1.000
$G_M V_I C_R$	-	$G_F V_O C_A$		42.2485	2.08	20.2732	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_O C_E$		30.7576	2.08	14.7592	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_A C_H$		16.2970	2.08	7.8202	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_A C_P$		40.3030	2.08	19.3397	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_A C_A$		42.7333	2.08	20.5059	4888	< 0.001
$G_M V_I C_R$	-	$G_F V_A C_E$		33.0364	2.08	15.8527	4888	< 0.001
$G_M V_I C_A$	-	$G_M V_O C_A$		-1.3333	2.08	-0.6398	4888	1.000
$G_M V_I C_A$	-	$G_F V_O C_A$		4.1152	2.08	1.9747	4888	1.000
$G_M V_I C_E$	-	$G_M V_I C_P$		3.0303	2.08	1.4541	4888	1.000
$G_M V_I C_E$	-	$G_M V_I C_A$		-0.5939	2.08	-0.2850	4888	1.000
$G_M V_I C_E$	-	$G_M V_O C_P$		13.2667	2.08	6.3661	4888	< 0.001

Table A.5 Continued: Post Hoc Comparisons - Gender * Vowel * Condition

Comparison				Difference	SE	t	df	<i>P</i> _{bonferroni}
$G_M V_I C_E$	-	$G_M V_O C_A$		-1.9273	2.08	-0.9248	4888	1.000
$G_M V_I C_E$	-	$G_M V_O C_E$		5.7515	2.08	2.7599	4888	1.000
$G_M V_I C_E$	-	$G_M V_A C_P$		1.4182	2.08	0.6805	4888	1.000
$G_M V_I C_E$	-	$G_M V_A C_A$		-1.8667	2.08	-0.8957	4888	1.000
$G_M V_I C_E$	-	$G_F V_I C_P$		-6.7758	2.08	-3.2514	4888	0.503
$G_M V_I C_E$	-	$G_F V_I C_A$		9.2485	2.08	4.4380	4888	0.004
$G_M V_I C_E$	-	$G_F V_O C_P$		-5.2303	2.08	-2.5098	4888	1.000
$G_M V_I C_E$	-	$G_F V_O C_A$		3.5212	2.08	1.6897	4888	1.000
$G_M V_I C_E$	-	$G_F V_O C_E$		-7.9697	2.08	-3.8243	4888	0.058
$G_M V_I C_E$	-	$G_F V_A C_P$		1.5758	2.08	0.7561	4888	1.000
$G_M V_I C_E$	-	$G_F V_A C_A$		4.0061	2.08	1.9223	4888	1.000
$G_M V_O C_H$	-	$G_M V_I C_P$		28.7818	2.08	13.8112	4888	< 0.001
$G_M V_O C_H$	-	$G_M V_I C_A$		25.1576	2.08	12.0720	4888	< 0.001
$G_M V_O C_H$	-	$G_M V_I C_E$		25.7515	2.08	12.3570	4888	< 0.001
$G_M V_O C_H$	-	$G_M V_O C_P$		39.0182	2.08	18.7231	4888	< 0.001
$G_M V_O C_H$	-	$G_M V_O C_A$		23.8242	2.08	11.4322	4888	< 0.001
$G_M V_O C_H$	-	$G_M V_O C_E$		31.5030	2.08	15.1170	4888	< 0.001
$G_M V_O C_H$	-	$G_M V_A C_P$		27.1697	2.08	13.0376	4888	< 0.001
$G_M V_O C_H$	-	$G_M V_A C_A$		23.8848	2.08	11.4613	4888	< 0.001
$G_M V_O C_H$	-	$G_M V_A C_E$		21.3576	2.08	10.2486	4888	< 0.001
$G_M V_O C_H$	-	$G_F V_I C_P$		18.9758	2.08	9.1057	4888	< 0.001
$G_M V_O C_H$	-	$G_F V_I C_A$		35.0000	2.08	16.7950	4888	< 0.001
$G_M V_O C_H$	-	$G_F V_I C_E$		15.5333	2.08	7.4538	4888	< 0.001
$G_M V_O C_H$	-	$G_F V_O C_P$		20.5212	2.08	9.8472	4888	< 0.001
$G_M V_O C_H$	-	$G_F V_O C_A$		29.2727	2.08	14.0467	4888	< 0.001
$G_M V_O C_H$	-	$G_F V_O C_E$		17.7818	2.08	8.5327	4888	< 0.001
$G_M V_O C_H$	-	$G_F V_A C_P$		27.3273	2.08	13.1132	4888	< 0.001
$G_M V_O C_H$	-	$G_F V_A C_A$		29.7576	2.08	14.2794	4888	< 0.001
$G_M V_O C_H$	-	$G_F V_A C_E$		20.0606	2.08	9.6262	4888	< 0.001
$G_M V_O C_P$	-	$G_M V_I C_A$		-13.8606	2.08	-6.6511	4888	< 0.001
$G_M V_O C_P$	-	$G_M V_O C_A$		-15.1939	2.08	-7.2909	4888	< 0.001
$G_M V_O C_P$	-	$G_M V_A C_A$		-15.1333	2.08	-7.2618	4888	< 0.001
$G_M V_O C_P$	-	$G_F V_I C_A$		-4.0182	2.08	-1.9282	4888	1.000
$G_M V_O C_P$	-	$G_F V_O C_A$		-9.7455	2.08	-4.6764	4888	0.001
$G_M V_O C_P$	-	$G_F V_A C_A$		-9.2606	2.08	-4.4438	4888	0.004
$G_M V_O C_R$	-	$G_M V_I C_H$		17.4667	2.08	8.3815	4888	< 0.001
$G_M V_O C_R$	-	$G_M V_I C_P$		43.2242	2.08	20.7415	4888	< 0.001
$G_M V_O C_R$	-	$G_M V_I C_A$		39.6000	2.08	19.0023	4888	< 0.001
$G_M V_O C_R$	-	$G_M V_I C_E$		40.1939	2.08	19.2873	4888	< 0.001
$G_M V_O C_R$	-	$G_M V_O C_H$		14.4424	2.08	6.9303	4888	< 0.001
$G_M V_O C_R$	-	$G_M V_O C_P$		53.4606	2.08	25.6534	4888	< 0.001
$G_M V_O C_R$	-	$G_M V_O C_A$		38.2667	2.08	18.3625	4888	< 0.001
$G_M V_O C_R$	-	$G_M V_O C_E$		45.9455	2.08	22.0472	4888	< 0.001
$G_M V_O C_R$	-	$G_M V_A C_H$		11.4848	2.08	5.5111	4888	< 0.001
$G_M V_O C_R$	-	$G_M V_A C_P$		41.6121	2.08	19.9679	4888	< 0.001
$G_M V_O C_R$	-	$G_M V_A C_A$		38.3273	2.08	18.3916	4888	< 0.001
$G_M V_O C_R$	-	$G_M V_A C_E$		35.8000	2.08	17.1789	4888	< 0.001
$G_M V_O C_R$	-	$G_F V_I C_H$		11.9455	2.08	5.7321	4888	< 0.001
$G_M V_O C_R$	-	$G_F V_I C_P$		33.4182	2.08	16.0359	4888	< 0.001
$G_M V_O C_R$	-	$G_F V_I C_A$		49.4424	2.08	23.7253	4888	< 0.001
$G_M V_O C_R$	-	$G_F V_I C_E$		29.9758	2.08	14.3841	4888	< 0.001
$G_M V_O C_R$	-	$G_F V_O C_H$		20.2545	2.08	9.7193	4888	< 0.001
$G_M V_O C_R$	-	$G_F V_O C_P$		34.9636	2.08	16.7775	4888	< 0.001
$G_M V_O C_R$	-	$G_F V_O C_A$		43.7152	2.08	20.9770	4888	< 0.001
$G_M V_O C_R$	-	$G_F V_O C_E$		32.2242	2.08	15.4630	4888	< 0.001

Table A.5 Continued: Post Hoc Comparisons - Gender * Vowel * Condition

Comparison				Difference	SE	t	df	<i>P</i> _{bonferroni}
$G_M V_O C_R$	-	$G_F V_A C_H$		17.7636	2.08	8.5240	4888	< 0.001
$G_M V_O C_R$	-	$G_F V_A C_P$		41.7697	2.08	20.0435	4888	< 0.001
$G_M V_O C_R$	-	$G_F V_A C_A$		44.2000	2.08	21.2097	4888	< 0.001
$G_M V_O C_R$	-	$G_F V_A C_E$		34.5030	2.08	16.5565	4888	< 0.001
$G_M V_O C_E$	-	$G_M V_I C_P$		-2.7212	2.08	-1.3058	4888	1.000
$G_M V_O C_E$	-	$G_M V_I C_A$		-6.3455	2.08	-3.0449	4888	1.000
$G_M V_O C_E$	-	$G_M V_O C_P$		7.5152	2.08	3.6062	4888	0.137
$G_M V_O C_E$	-	$G_M V_O C_A$		-7.6788	2.08	-3.6847	4888	0.101
$G_M V_O C_E$	-	$G_M V_A C_P$		-4.3333	2.08	-2.0794	4888	1.000
$G_M V_O C_E$	-	$G_M V_A C_A$		-7.6182	2.08	-3.6556	4888	0.113
$G_M V_O C_E$	-	$G_F V_I C_P$		-12.5273	2.08	-6.0113	4888	< 0.001
$G_M V_O C_E$	-	$G_F V_I C_A$		3.4970	2.08	1.6780	4888	1.000
$G_M V_O C_E$	-	$G_F V_O C_P$		-10.9818	2.08	-5.2697	4888	< 0.001
$G_M V_O C_E$	-	$G_F V_O C_A$		-2.2303	2.08	-1.0702	4888	1.000
$G_M V_O C_E$	-	$G_F V_A C_P$		-4.1758	2.08	-2.0038	4888	1.000
$G_M V_O C_E$	-	$G_F V_A C_A$		-1.7455	2.08	-0.8376	4888	1.000
$G_M V_A C_H$	-	$G_M V_I C_H$		5.9818	2.08	2.8704	4888	1.000
$G_M V_A C_H$	-	$G_M V_I C_P$		31.7394	2.08	15.2304	4888	< 0.001
$G_M V_A C_H$	-	$G_M V_I C_A$		28.1152	2.08	13.4913	4888	< 0.001
$G_M V_A C_H$	-	$G_M V_I C_E$		28.7091	2.08	13.7763	4888	< 0.001
$G_M V_A C_H$	-	$G_M V_O C_H$		2.9576	2.08	1.4192	4888	1.000
$G_M V_A C_H$	-	$G_M V_O C_P$		41.9758	2.08	20.1424	4888	< 0.001
$G_M V_A C_H$	-	$G_M V_O C_A$		26.7818	2.08	12.8514	4888	< 0.001
$G_M V_A C_H$	-	$G_M V_O C_E$		34.4606	2.08	16.5362	4888	< 0.001
$G_M V_A C_H$	-	$G_M V_A C_P$		30.1273	2.08	14.4568	4888	< 0.001
$G_M V_A C_H$	-	$G_M V_A C_A$		26.8424	2.08	12.8805	4888	< 0.001
$G_M V_A C_H$	-	$G_M V_A C_E$		24.3152	2.08	11.6678	4888	< 0.001
$G_M V_A C_H$	-	$G_F V_I C_H$		0.4606	2.08	0.2210	4888	1.000
$G_M V_A C_H$	-	$G_F V_I C_P$		21.9333	2.08	10.5249	4888	< 0.001
$G_M V_A C_H$	-	$G_F V_I C_A$		37.9576	2.08	18.2142	4888	< 0.001
$G_M V_A C_H$	-	$G_F V_I C_E$		18.4909	2.08	8.8730	4888	< 0.001
$G_M V_A C_H$	-	$G_F V_O C_H$		8.7697	2.08	4.2082	4888	0.011
$G_M V_A C_H$	-	$G_F V_O C_P$		23.4788	2.08	11.2665	4888	< 0.001
$G_M V_A C_H$	-	$G_F V_O C_A$		32.2303	2.08	15.4659	4888	< 0.001
$G_M V_A C_H$	-	$G_F V_O C_E$		20.7394	2.08	9.9519	4888	< 0.001
$G_M V_A C_H$	-	$G_F V_A C_P$		30.2848	2.08	14.5324	4888	< 0.001
$G_M V_A C_H$	-	$G_F V_A C_A$		32.7152	2.08	15.6986	4888	< 0.001
$G_M V_A C_H$	-	$G_F V_A C_E$		23.0182	2.08	11.0454	4888	< 0.001
$G_M V_A C_P$	-	$G_M V_I C_P$		1.6121	2.08	0.7736	4888	1.000
$G_M V_A C_P$	-	$G_M V_I C_A$		-2.0121	2.08	-0.9655	4888	1.000
$G_M V_A C_P$	-	$G_M V_O C_P$		11.8485	2.08	5.6856	4888	< 0.001
$G_M V_A C_P$	-	$G_M V_O C_A$		-3.3455	2.08	-1.6053	4888	1.000
$G_M V_A C_P$	-	$G_M V_A C_A$		-3.2848	2.08	-1.5763	4888	1.000
$G_M V_A C_P$	-	$G_F V_I C_P$		-8.1939	2.08	-3.9319	4888	0.037
$G_M V_A C_P$	-	$G_F V_I C_A$		7.8303	2.08	3.7574	4888	0.076
$G_M V_A C_P$	-	$G_F V_O C_P$		-6.6485	2.08	-3.1903	4888	0.622
$G_M V_A C_P$	-	$G_F V_O C_A$		2.1030	2.08	1.0092	4888	1.000
$G_M V_A C_P$	-	$G_F V_A C_A$		2.5879	2.08	1.2418	4888	1.000
$G_M V_A C_R$	-	$G_M V_I C_H$		17.0545	2.08	8.1837	4888	< 0.001
$G_M V_A C_R$	-	$G_M V_I C_P$		42.8121	2.08	20.5437	4888	< 0.001
$G_M V_A C_R$	-	$G_M V_I C_R$		1.0545	2.08	0.5060	4888	1.000
$G_M V_A C_R$	-	$G_M V_I C_A$		39.1879	2.08	18.8046	4888	< 0.001
$G_M V_A C_R$	-	$G_M V_I C_E$		39.7818	2.08	19.0896	4888	< 0.001
$G_M V_A C_R$	-	$G_M V_O C_H$		14.0303	2.08	6.7325	4888	< 0.001
$G_M V_A C_R$	-	$G_M V_O C_P$		53.0485	2.08	25.4557	4888	< 0.001

Table A.5 Continued: Post Hoc Comparisons - Gender * Vowel * Condition

Comparison				Difference	SE	t	df	<i>P</i> _{bonferroni}
$G_M V_A C_R$	-	$G_M V_O C_R$		-0.4121	2.08	-0.1978	4888	1.000
$G_M V_A C_R$	-	$G_M V_O C_A$		37.8545	2.08	18.1648	4888	< 0.001
$G_M V_A C_R$	-	$G_M V_O C_E$		45.5333	2.08	21.8495	4888	< 0.001
$G_M V_A C_R$	-	$G_M V_A C_H$		11.0727	2.08	5.3133	4888	< 0.001
$G_M V_A C_R$	-	$G_M V_A C_P$		41.2000	2.08	19.7701	4888	< 0.001
$G_M V_A C_R$	-	$G_M V_A C_A$		37.9152	2.08	18.1939	4888	< 0.001
$G_M V_A C_R$	-	$G_M V_A C_E$		35.3879	2.08	16.9811	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_I C_H$		11.5333	2.08	5.5344	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_I C_P$		33.0061	2.08	15.8382	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_I C_R$		-1.9939	2.08	-0.9568	4888	1.000
$G_M V_A C_R$	-	$G_F V_I C_A$		49.0303	2.08	23.5275	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_I C_E$		29.5636	2.08	14.1863	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_O C_H$		19.8424	2.08	9.5215	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_O C_P$		34.5515	2.08	16.5798	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_O C_R$		-1.1576	2.08	-0.5555	4888	1.000
$G_M V_A C_R$	-	$G_F V_O C_A$		43.3030	2.08	20.7793	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_O C_E$		31.8121	2.08	15.2653	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_A C_H$		17.3515	2.08	8.3262	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_A C_P$		41.3576	2.08	19.8457	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_A C_A$		43.7879	2.08	21.0119	4888	< 0.001
$G_M V_A C_R$	-	$G_F V_A C_E$		34.0909	2.08	16.3588	4888	< 0.001
$G_M V_A C_A$	-	$G_M V_I C_A$		1.2727	2.08	0.6107	4888	1.000
$G_M V_A C_A$	-	$G_M V_O C_A$		-0.0606	2.08	-0.0291	4888	1.000
$G_M V_A C_A$	-	$G_F V_I C_A$		11.1152	2.08	5.3337	4888	< 0.001
$G_M V_A C_A$	-	$G_F V_O C_A$		5.3879	2.08	2.5854	4888	1.000
$G_M V_A C_E$	-	$G_M V_I C_P$		7.4242	2.08	3.5626	4888	0.161
$G_M V_A C_E$	-	$G_M V_I C_A$		3.8000	2.08	1.8235	4888	1.000
$G_M V_A C_E$	-	$G_M V_I C_E$		4.3939	2.08	2.1085	4888	1.000
$G_M V_A C_E$	-	$G_M V_O C_P$		17.6606	2.08	8.4746	4888	< 0.001
$G_M V_A C_E$	-	$G_M V_O C_A$		2.4667	2.08	1.1836	4888	1.000
$G_M V_A C_E$	-	$G_M V_O C_E$		10.1455	2.08	4.8684	4888	< 0.001
$G_M V_A C_E$	-	$G_M V_A C_P$		5.8121	2.08	2.7890	4888	1.000
$G_M V_A C_E$	-	$G_M V_A C_A$		2.5273	2.08	1.2127	4888	1.000
$G_M V_A C_E$	-	$G_F V_I C_P$		-2.3818	2.08	-1.1429	4888	1.000
$G_M V_A C_E$	-	$G_F V_I C_A$		13.6424	2.08	6.5464	4888	< 0.001
$G_M V_A C_E$	-	$G_F V_I C_E$		-5.8242	2.08	-2.7948	4888	1.000
$G_M V_A C_E$	-	$G_F V_O C_P$		-0.8364	2.08	-0.4013	4888	1.000
$G_M V_A C_E$	-	$G_F V_O C_A$		7.9152	2.08	3.7981	4888	0.064
$G_M V_A C_E$	-	$G_F V_O C_E$		-3.5758	2.08	-1.7159	4888	1.000
$G_M V_A C_E$	-	$G_F V_A C_P$		5.9697	2.08	2.8646	4888	1.000
$G_M V_A C_E$	-	$G_F V_A C_A$		8.4000	2.08	4.0308	4888	0.025
$G_F V_I C_H$	-	$G_M V_I C_H$		5.5212	2.08	2.6494	4888	1.000
$G_F V_I C_H$	-	$G_M V_I C_P$		31.2788	2.08	15.0093	4888	< 0.001
$G_F V_I C_H$	-	$G_M V_I C_A$		27.6545	2.08	13.2702	4888	< 0.001
$G_F V_I C_H$	-	$G_M V_I C_E$		28.2485	2.08	13.5552	4888	< 0.001
$G_F V_I C_H$	-	$G_M V_O C_H$		2.4970	2.08	1.1982	4888	1.000
$G_F V_I C_H$	-	$G_M V_O C_P$		41.5152	2.08	19.9213	4888	< 0.001
$G_F V_I C_H$	-	$G_M V_O C_A$		26.3212	2.08	12.6304	4888	< 0.001
$G_F V_I C_H$	-	$G_M V_O C_E$		34.0000	2.08	16.3151	4888	< 0.001
$G_F V_I C_H$	-	$G_M V_A C_P$		29.6667	2.08	14.2358	4888	< 0.001
$G_F V_I C_H$	-	$G_M V_A C_A$		26.3818	2.08	12.6595	4888	< 0.001
$G_F V_I C_H$	-	$G_M V_A C_E$		23.8545	2.08	11.4468	4888	< 0.001
$G_F V_I C_H$	-	$G_F V_I C_P$		21.4727	2.08	10.3038	4888	< 0.001
$G_F V_I C_H$	-	$G_F V_I C_A$		37.4970	2.08	17.9932	4888	< 0.001
$G_F V_I C_H$	-	$G_F V_I C_E$		18.0303	2.08	8.6520	4888	< 0.001

Table A.5 Continued: Post Hoc Comparisons - Gender * Vowel * Condition

Comparison				Difference	SE	t	df	<i>P</i> _{bonferroni}
$G_F V_I C_H$	-	$G_F V_O C_H$		8.3091	2.08	3.9872	4888	0.030
$G_F V_I C_H$	-	$G_F V_O C_P$		23.0182	2.08	11.0454	4888	< 0.001
$G_F V_I C_H$	-	$G_F V_O C_A$		31.7697	2.08	15.2449	4888	< 0.001
$G_F V_I C_H$	-	$G_F V_O C_E$		20.2788	2.08	9.7309	4888	< 0.001
$G_F V_I C_H$	-	$G_F V_A C_P$		29.8242	2.08	14.3114	4888	< 0.001
$G_F V_I C_H$	-	$G_F V_A C_A$		32.2545	2.08	15.4776	4888	< 0.001
$G_F V_I C_H$	-	$G_F V_A C_E$		22.5576	2.08	10.8244	4888	< 0.001
$G_F V_I C_P$	-	$G_M V_I C_P$		9.8061	2.08	4.7055	4888	0.001
$G_F V_I C_P$	-	$G_M V_I C_A$		6.1818	2.08	2.9664	4888	1.000
$G_F V_I C_P$	-	$G_M V_O C_P$		20.0424	2.08	9.6175	4888	< 0.001
$G_F V_I C_P$	-	$G_M V_O C_A$		4.8485	2.08	2.3266	4888	1.000
$G_F V_I C_P$	-	$G_M V_A C_A$		4.9091	2.08	2.3557	4888	1.000
$G_F V_I C_P$	-	$G_F V_I C_A$		16.0242	2.08	7.6893	4888	< 0.001
$G_F V_I C_P$	-	$G_F V_O C_P$		1.5455	2.08	0.7416	4888	1.000
$G_F V_I C_P$	-	$G_F V_O C_A$		10.2970	2.08	4.9411	4888	< 0.001
$G_F V_I C_P$	-	$G_F V_A C_A$		10.7818	2.08	5.1737	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_I C_H$		19.0485	2.08	9.1405	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_I C_P$		44.8061	2.08	21.5005	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_I C_R$		3.0485	2.08	1.4628	4888	1.000
$G_F V_I C_R$	-	$G_M V_I C_A$		41.1818	2.08	19.7614	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_I C_E$		41.7758	2.08	20.0464	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_O C_H$		16.0242	2.08	7.6893	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_O C_P$		55.0424	2.08	26.4125	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_O C_R$		1.5818	2.08	0.7590	4888	1.000
$G_F V_I C_R$	-	$G_M V_O C_A$		39.8485	2.08	19.1216	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_O C_E$		47.5273	2.08	22.8063	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_A C_H$		13.0667	2.08	6.2701	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_A C_P$		43.1939	2.08	20.7269	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_A C_A$		39.9091	2.08	19.1507	4888	< 0.001
$G_F V_I C_R$	-	$G_M V_A C_E$		37.3818	2.08	17.9379	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_I C_H$		13.5273	2.08	6.4912	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_I C_P$		35.0000	2.08	16.7950	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_I C_A$		51.0242	2.08	24.4843	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_I C_E$		31.5576	2.08	15.1431	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_O C_H$		21.8364	2.08	10.4783	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_O C_P$		36.5455	2.08	17.5366	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_O C_R$		0.8364	2.08	0.4013	4888	1.000
$G_F V_I C_R$	-	$G_F V_O C_A$		45.2970	2.08	21.7361	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_O C_E$		33.8061	2.08	16.2221	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_A C_H$		19.3455	2.08	9.2831	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_A C_P$		43.3515	2.08	20.8025	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_A C_A$		45.7818	2.08	21.9687	4888	< 0.001
$G_F V_I C_R$	-	$G_F V_A C_E$		36.0848	2.08	17.3156	4888	< 0.001
$G_F V_I C_A$	-	$G_M V_I C_A$		-9.8424	2.08	-4.7230	4888	0.001
$G_F V_I C_A$	-	$G_M V_O C_A$		-11.1758	2.08	-5.3628	4888	< 0.001
$G_F V_I C_A$	-	$G_F V_O C_A$		-5.7273	2.08	-2.7483	4888	1.000
$G_F V_I C_E$	-	$G_M V_I C_P$		13.2485	2.08	6.3574	4888	< 0.001
$G_F V_I C_E$	-	$G_M V_I C_A$		9.6242	2.08	4.6183	4888	0.002
$G_F V_I C_E$	-	$G_M V_I C_E$		10.2182	2.08	4.9033	4888	< 0.001
$G_F V_I C_E$	-	$G_M V_O C_P$		23.4848	2.08	11.2694	4888	< 0.001
$G_F V_I C_E$	-	$G_M V_O C_A$		8.2909	2.08	3.9785	4888	0.031
$G_F V_I C_E$	-	$G_M V_O C_E$		15.9697	2.08	7.6632	4888	< 0.001
$G_F V_I C_E$	-	$G_M V_A C_P$		11.6364	2.08	5.5838	4888	< 0.001
$G_F V_I C_E$	-	$G_M V_A C_A$		8.3515	2.08	4.0075	4888	0.027
$G_F V_I C_E$	-	$G_F V_I C_P$		3.4424	2.08	1.6519	4888	1.000

Table A.5 Continued: Post Hoc Comparisons - Gender * Vowel * Condition

Comparison				Difference	SE	t	df	<i>P</i> _{bonferroni}
$G_F V_I C_E$	-	$G_F V_I C_A$		19.4667	2.08	9.3412	4888	< 0.001
$G_F V_I C_E$	-	$G_F V_O C_P$		4.9879	2.08	2.3935	4888	1.000
$G_F V_I C_E$	-	$G_F V_O C_A$		13.7394	2.08	6.5929	4888	< 0.001
$G_F V_I C_E$	-	$G_F V_O C_E$		2.2485	2.08	1.0790	4888	1.000
$G_F V_I C_E$	-	$G_F V_A C_P$		11.7939	2.08	5.6594	4888	< 0.001
$G_F V_I C_E$	-	$G_F V_A C_A$		14.2242	2.08	6.8256	4888	< 0.001
$G_F V_O C_H$	-	$G_M V_I C_P$		22.9697	2.08	11.0222	4888	< 0.001
$G_F V_O C_H$	-	$G_M V_I C_A$		19.3455	2.08	9.2831	4888	< 0.001
$G_F V_O C_H$	-	$G_M V_I C_E$		19.9394	2.08	9.5681	4888	< 0.001
$G_F V_O C_H$	-	$G_M V_O C_H$		-5.8121	2.08	-2.7890	4888	1.000
$G_F V_O C_H$	-	$G_M V_O C_P$		33.2061	2.08	15.9342	4888	< 0.001
$G_F V_O C_H$	-	$G_M V_O C_A$		18.0121	2.08	8.6432	4888	< 0.001
$G_F V_O C_H$	-	$G_M V_O C_E$		25.6909	2.08	12.3280	4888	< 0.001
$G_F V_O C_H$	-	$G_M V_A C_P$		21.3576	2.08	10.2486	4888	< 0.001
$G_F V_O C_H$	-	$G_M V_A C_A$		18.0727	2.08	8.6723	4888	< 0.001
$G_F V_O C_H$	-	$G_M V_A C_E$		15.5455	2.08	7.4596	4888	< 0.001
$G_F V_O C_H$	-	$G_F V_I C_P$		13.1636	2.08	6.3167	4888	< 0.001
$G_F V_O C_H$	-	$G_F V_I C_A$		29.1879	2.08	14.0060	4888	< 0.001
$G_F V_O C_H$	-	$G_F V_I C_E$		9.7212	2.08	4.6648	4888	0.001
$G_F V_O C_H$	-	$G_F V_O C_P$		14.7091	2.08	7.0583	4888	< 0.001
$G_F V_O C_H$	-	$G_F V_O C_A$		23.4606	2.08	11.2577	4888	< 0.001
$G_F V_O C_H$	-	$G_F V_O C_E$		11.9697	2.08	5.7437	4888	< 0.001
$G_F V_O C_H$	-	$G_F V_A C_P$		21.5152	2.08	10.3242	4888	< 0.001
$G_F V_O C_H$	-	$G_F V_A C_A$		23.9455	2.08	11.4904	4888	< 0.001
$G_F V_O C_H$	-	$G_F V_A C_E$		14.2485	2.08	6.8372	4888	< 0.001
$G_F V_O C_P$	-	$G_M V_I C_A$		4.6364	2.08	2.2248	4888	1.000
$G_F V_O C_P$	-	$G_M V_O C_P$		18.4970	2.08	8.8759	4888	< 0.001
$G_F V_O C_P$	-	$G_M V_O C_A$		3.3030	2.08	1.5850	4888	1.000
$G_F V_O C_P$	-	$G_M V_A C_A$		3.3636	2.08	1.6141	4888	1.000
$G_F V_O C_P$	-	$G_F V_I C_A$		14.4788	2.08	6.9477	4888	< 0.001
$G_F V_O C_P$	-	$G_F V_O C_A$		8.7515	2.08	4.1995	4888	0.012
$G_F V_O C_P$	-	$G_F V_A C_A$		9.2364	2.08	4.4321	4888	0.004
$G_F V_O C_R$	-	$G_M V_I C_H$		18.2121	2.08	8.7392	4888	< 0.001
$G_F V_O C_R$	-	$G_M V_I C_P$		43.9697	2.08	21.0992	4888	< 0.001
$G_F V_O C_R$	-	$G_M V_I C_A$		40.3455	2.08	19.3601	4888	< 0.001
$G_F V_O C_R$	-	$G_M V_I C_E$		40.9394	2.08	19.6451	4888	< 0.001
$G_F V_O C_R$	-	$G_M V_O C_H$		15.1879	2.08	7.2880	4888	< 0.001
$G_F V_O C_R$	-	$G_M V_O C_P$		54.2061	2.08	26.0112	4888	< 0.001
$G_F V_O C_R$	-	$G_M V_O C_R$		0.7455	2.08	0.3577	4888	1.000
$G_F V_O C_R$	-	$G_M V_O C_A$		39.0121	2.08	18.7202	4888	< 0.001
$G_F V_O C_R$	-	$G_M V_O C_E$		46.6909	2.08	22.4050	4888	< 0.001
$G_F V_O C_R$	-	$G_M V_A C_H$		12.2303	2.08	5.8688	4888	< 0.001
$G_F V_O C_R$	-	$G_M V_A C_P$		42.3576	2.08	20.3256	4888	< 0.001
$G_F V_O C_R$	-	$G_M V_A C_A$		39.0727	2.08	18.7493	4888	< 0.001
$G_F V_O C_R$	-	$G_M V_A C_E$		36.5455	2.08	17.5366	4888	< 0.001
$G_F V_O C_R$	-	$G_F V_I C_H$		12.6909	2.08	6.0898	4888	< 0.001
$G_F V_O C_R$	-	$G_F V_I C_P$		34.1636	2.08	16.3937	4888	< 0.001
$G_F V_O C_R$	-	$G_F V_I C_A$		50.1879	2.08	24.0830	4888	< 0.001
$G_F V_O C_R$	-	$G_F V_I C_E$		30.7212	2.08	14.7418	4888	< 0.001
$G_F V_O C_R$	-	$G_F V_O C_H$		21.0000	2.08	10.0770	4888	< 0.001
$G_F V_O C_R$	-	$G_F V_O C_P$		35.7091	2.08	17.1353	4888	< 0.001
$G_F V_O C_R$	-	$G_F V_O C_A$		44.4606	2.08	21.3347	4888	< 0.001
$G_F V_O C_R$	-	$G_F V_O C_E$		32.9697	2.08	15.8207	4888	< 0.001
$G_F V_O C_R$	-	$G_F V_A C_H$		18.5091	2.08	8.8817	4888	< 0.001
$G_F V_O C_R$	-	$G_F V_A C_P$		42.5152	2.08	20.4012	4888	< 0.001

Table A.5 Continued: Post Hoc Comparisons - Gender * Vowel * Condition

Comparison				Difference	SE	t	df	<i>P</i> _{bonferroni}
$G_F V_O C_R$	-	$G_F V_A C_A$		44.9455	2.08	21.5674	4888	< 0.001
$G_F V_O C_R$	-	$G_F V_A C_E$		35.2485	2.08	16.9142	4888	< 0.001
$G_F V_O C_A$	-	$G_M V_O C_A$		-5.4485	2.08	-2.6145	4888	1.000
$G_F V_O C_E$	-	$G_M V_I C_P$		11.0000	2.08	5.2784	4888	< 0.001
$G_F V_O C_E$	-	$G_M V_I C_A$		7.3758	2.08	3.5393	4888	0.176
$G_F V_O C_E$	-	$G_M V_O C_P$		21.2364	2.08	10.1904	4888	< 0.001
$G_F V_O C_E$	-	$G_M V_O C_A$		6.0424	2.08	2.8995	4888	1.000
$G_F V_O C_E$	-	$G_M V_O C_E$		13.7212	2.08	6.5842	4888	< 0.001
$G_F V_O C_E$	-	$G_M V_A C_P$		9.3879	2.08	4.5048	4888	0.003
$G_F V_O C_E$	-	$G_M V_A C_A$		6.1030	2.08	2.9286	4888	1.000
$G_F V_O C_E$	-	$G_F V_I C_P$		1.1939	2.08	0.5729	4888	1.000
$G_F V_O C_E$	-	$G_F V_I C_A$		17.2182	2.08	8.2623	4888	< 0.001
$G_F V_O C_E$	-	$G_F V_O C_P$		2.7394	2.08	1.3145	4888	1.000
$G_F V_O C_E$	-	$G_F V_O C_A$		11.4909	2.08	5.5140	4888	< 0.001
$G_F V_O C_E$	-	$G_F V_A C_P$		9.5455	2.08	4.5805	4888	0.002
$G_F V_O C_E$	-	$G_F V_A C_A$		11.9758	2.08	5.7467	4888	< 0.001
$G_F V_A C_H$	-	$G_M V_I C_H$		-0.2970	2.08	-0.1425	4888	1.000
$G_F V_A C_H$	-	$G_M V_I C_P$		25.4606	2.08	12.2175	4888	< 0.001
$G_F V_A C_H$	-	$G_M V_I C_A$		21.8364	2.08	10.4783	4888	< 0.001
$G_F V_A C_H$	-	$G_M V_I C_E$		22.4303	2.08	10.7633	4888	< 0.001
$G_F V_A C_H$	-	$G_M V_O C_H$		-3.3212	2.08	-1.5937	4888	1.000
$G_F V_A C_H$	-	$G_M V_O C_P$		35.6970	2.08	17.1294	4888	< 0.001
$G_F V_A C_H$	-	$G_M V_O C_A$		20.5030	2.08	9.8385	4888	< 0.001
$G_F V_A C_H$	-	$G_M V_O C_E$		28.1818	2.08	13.5232	4888	< 0.001
$G_F V_A C_H$	-	$G_M V_A C_H$		-6.2788	2.08	-3.0129	4888	1.000
$G_F V_A C_H$	-	$G_M V_A C_P$		23.8485	2.08	11.4439	4888	< 0.001
$G_F V_A C_H$	-	$G_M V_A C_A$		20.5636	2.08	9.8676	4888	< 0.001
$G_F V_A C_H$	-	$G_M V_A C_E$		18.0364	2.08	8.6549	4888	< 0.001
$G_F V_A C_H$	-	$G_F V_I C_H$		-5.8182	2.08	-2.7919	4888	1.000
$G_F V_A C_H$	-	$G_F V_I C_P$		15.6545	2.08	7.5119	4888	< 0.001
$G_F V_A C_H$	-	$G_F V_I C_A$		31.6788	2.08	15.2013	4888	< 0.001
$G_F V_A C_H$	-	$G_F V_I C_E$		12.2121	2.08	5.8601	4888	< 0.001
$G_F V_A C_H$	-	$G_F V_O C_H$		2.4909	2.08	1.1953	4888	1.000
$G_F V_A C_H$	-	$G_F V_O C_P$		17.2000	2.08	8.2535	4888	< 0.001
$G_F V_A C_H$	-	$G_F V_O C_A$		25.9515	2.08	12.4530	4888	< 0.001
$G_F V_A C_H$	-	$G_F V_O C_E$		14.4606	2.08	6.9390	4888	< 0.001
$G_F V_A C_H$	-	$G_F V_A C_P$		24.0061	2.08	11.5195	4888	< 0.001
$G_F V_A C_H$	-	$G_F V_A C_A$		26.4364	2.08	12.6857	4888	< 0.001
$G_F V_A C_H$	-	$G_F V_A C_E$		16.7394	2.08	8.0325	4888	< 0.001
$G_F V_A C_P$	-	$G_M V_I C_P$		1.4545	2.08	0.6980	4888	1.000
$G_F V_A C_P$	-	$G_M V_I C_A$		-2.1697	2.08	-1.0411	4888	1.000
$G_F V_A C_P$	-	$G_M V_O C_P$		11.6909	2.08	5.6100	4888	< 0.001
$G_F V_A C_P$	-	$G_M V_O C_A$		-3.5030	2.08	-1.6810	4888	1.000
$G_F V_A C_P$	-	$G_M V_A C_P$		-0.1576	2.08	-0.0756	4888	1.000
$G_F V_A C_P$	-	$G_M V_A C_A$		-3.4424	2.08	-1.6519	4888	1.000
$G_F V_A C_P$	-	$G_F V_I C_P$		-8.3515	2.08	-4.0075	4888	0.027
$G_F V_A C_P$	-	$G_F V_I C_A$		7.6727	2.08	3.6818	4888	0.102
$G_F V_A C_P$	-	$G_F V_O C_P$		-6.8061	2.08	-3.2659	4888	0.478
$G_F V_A C_P$	-	$G_F V_O C_A$		1.9455	2.08	0.9335	4888	1.000
$G_F V_A C_P$	-	$G_F V_A C_A$		2.4303	2.08	1.1662	4888	1.000
$G_F V_A C_R$	-	$G_M V_I C_H$		18.2121	2.08	8.7392	4888	< 0.001
$G_F V_A C_R$	-	$G_M V_I C_P$		43.9697	2.08	21.0992	4888	< 0.001
$G_F V_A C_R$	-	$G_M V_I C_R$		2.2121	2.08	1.0615	4888	1.000
$G_F V_A C_R$	-	$G_M V_I C_A$		40.3455	2.08	19.3601	4888	< 0.001
$G_F V_A C_R$	-	$G_M V_I C_E$		40.9394	2.08	19.6451	4888	< 0.001

Table A.5 Continued: Post Hoc Comparisons - Gender * Vowel * Condition

Comparison				Difference	SE	t	df	<i>P</i> _{bonferroni}
$G_F V_A C_R$	-	$G_M V_O C_H$		15.1879	2.08	7.2880	4888	< 0.001
$G_F V_A C_R$	-	$G_M V_O C_P$		54.2061	2.08	26.0112	4888	< 0.001
$G_F V_A C_R$	-	$G_M V_O C_R$		0.7455	2.08	0.3577	4888	1.000
$G_F V_A C_R$	-	$G_M V_O C_A$		39.0121	2.08	18.7202	4888	< 0.001
$G_F V_A C_R$	-	$G_M V_O C_E$		46.6909	2.08	22.4050	4888	< 0.001
$G_F V_A C_R$	-	$G_M V_A C_H$		12.2303	2.08	5.8688	4888	< 0.001
$G_F V_A C_R$	-	$G_M V_A C_P$		42.3576	2.08	20.3256	4888	< 0.001
$G_F V_A C_R$	-	$G_M V_A C_R$		1.1576	2.08	0.5555	4888	1.000
$G_F V_A C_R$	-	$G_M V_A C_A$		39.0727	2.08	18.7493	4888	< 0.001
$G_F V_A C_R$	-	$G_M V_A C_E$		36.5455	2.08	17.5366	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_I C_H$		12.6909	2.08	6.0898	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_I C_P$		34.1636	2.08	16.3937	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_I C_R$		-0.8364	2.08	-0.4013	4888	1.000
$G_F V_A C_R$	-	$G_F V_I C_A$		50.1879	2.08	24.0830	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_I C_E$		30.7212	2.08	14.7418	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_O C_H$		21.0000	2.08	10.0770	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_O C_P$		35.7091	2.08	17.1353	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_O C_R$		1.56e-14	2.08	7.48e-15	4888	1.000
$G_F V_A C_R$	-	$G_F V_O C_A$		44.4606	2.08	21.3347	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_O C_E$		32.9697	2.08	15.8207	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_A C_H$		18.5091	2.08	8.8817	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_A C_P$		42.5152	2.08	20.4012	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_A C_A$		44.9455	2.08	21.5674	4888	< 0.001
$G_F V_A C_R$	-	$G_F V_A C_E$		35.2485	2.08	16.9142	4888	< 0.001
$G_F V_A C_A$	-	$G_M V_I C_A$		-4.6000	2.08	-2.2073	4888	1.000
$G_F V_A C_A$	-	$G_M V_O C_A$		-5.9333	2.08	-2.8472	4888	1.000
$G_F V_A C_A$	-	$G_M V_A C_A$		-5.8727	2.08	-2.8181	4888	1.000
$G_F V_A C_A$	-	$G_F V_I C_A$		5.2424	2.08	2.5156	4888	1.000
$G_F V_A C_A$	-	$G_F V_O C_A$		-0.4848	2.08	-0.2327	4888	1.000
$G_F V_A C_E$	-	$G_M V_I C_P$		8.7212	2.08	4.1849	4888	0.013
$G_F V_A C_E$	-	$G_M V_I C_A$		5.0970	2.08	2.4458	4888	1.000
$G_F V_A C_E$	-	$G_M V_I C_E$		5.6909	2.08	2.7308	4888	1.000
$G_F V_A C_E$	-	$G_M V_O C_P$		18.9576	2.08	9.0969	4888	< 0.001
$G_F V_A C_E$	-	$G_M V_O C_A$		3.7636	2.08	1.8060	4888	1.000
$G_F V_A C_E$	-	$G_M V_O C_E$		11.4424	2.08	5.4907	4888	< 0.001
$G_F V_A C_E$	-	$G_M V_A C_P$		7.1091	2.08	3.4113	4888	0.283
$G_F V_A C_E$	-	$G_M V_A C_A$		3.8242	2.08	1.8351	4888	1.000
$G_F V_A C_E$	-	$G_M V_A C_E$		1.2970	2.08	0.6224	4888	1.000
$G_F V_A C_E$	-	$G_F V_I C_P$		-1.0848	2.08	-0.5206	4888	1.000
$G_F V_A C_E$	-	$G_F V_I C_A$		14.9394	2.08	7.1688	4888	< 0.001
$G_F V_A C_E$	-	$G_F V_I C_E$		-4.5273	2.08	-2.1724	4888	1.000
$G_F V_A C_E$	-	$G_F V_O C_P$		0.4606	2.08	0.2210	4888	1.000
$G_F V_A C_E$	-	$G_F V_O C_A$		9.2121	2.08	4.4205	4888	0.004
$G_F V_A C_E$	-	$G_F V_O C_E$		-2.2788	2.08	-1.0935	4888	1.000
$G_F V_A C_E$	-	$G_F V_A C_P$		7.2667	2.08	3.4870	4888	0.214
$G_F V_A C_E$	-	$G_F V_A C_A$		9.6970	2.08	4.6532	4888	0.001