TU WIEN Informatics

Estimating Vocal Tract Resonances of Synthesized High-Pitched Vowels Using CNN

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Biomedical Engineering

by

Ivana Mikušová, BSc.
Registration Number 01630329

to the Faculty of Informatics

at the TU Wien

Advisor: Associate Prof. Dipl.-Ing. Dr.techn. Peter Knees
Assistance: Bodo Maass, MA
Ass.-Prof. Priv.-Doz. Mag. art. Christian T. Herbst, PhD

Vienna, 14th November, 2021

Ivana Mikušová
Digitally signed by Ivana Mikusova
DN: CN = Ivana Mikusova (Signature),
C = SK, SN = Mikusova, GN = Ivana,
serialNumber = 96061753048

Peter Knees
Peter Knees

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Informatics

Estimating Vocal Tract Resonances of Synthesized High-Pitched Vowels Using CNN

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Biomedical Engineering

eingereicht von

Ivana Mikušová, BSc.
Matrikelnummer 01630329

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Associate Prof. Dipl.-Ing. Dr.techn. Peter Knees
Mitwirkung: Bodo Maass, MA
Ass.-Prof. Priv.-Doz. Mag. art. Christian T. Herbst, PhD

Wien, 14. November 2021

Ivana Mikušová

Digitally signed by Ivana Mikusova
DN: CN = Ivana Mikusova (Signature),
C = SK, SN = Mikusova, GN = Ivana,
serialNumber = 96061753048

Peter Knees

Peter Knees

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Erklärung zur Verfassung der Arbeit

Ivana Mikušová, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 14. November 2021

Ivana Mikušová

Digitally signed by Ivana Mikusova
DN: CN = Ivana Mikusova (Signature),
C = SK, SN = Mikusova, GN = Ivana,
serialNumber = 96061753048

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Bibliothek Your knowledge hub

Acknowledgements

I would like to thank everyone who has made this thesis possible.

My advisor, Associate Prof. Dipl.-Ing. Dr.techn. Peter Knees, for his advice and guidance. Bodo Maass, for letting me be part of his doctorate, providing countless batches of synthesized data, and his thorough help at every step of the process. Ass.-Prof. Priv.-Doz. Mag. art. Christian Herbst, for his role in organizing the entire arrangement and the lab experiments, and for the valuable wisdom he shared.

Josipa Bainac, for her help and company while preparing the lab experiments with bursts of creativity. Philipp Prinzinger, for providing 3D printed vocal tract models. Richard Vogl, for managing the escher server and his help in accessing it. Andreas Rauber, for introducing me to Peter Knees.

My family and friends for their support. Especially Simon for his endless patience and being there every day. My father and brother for the thematic discussions and being my inspiration in programming. My mom and grandparents for feeding me and keeping me on my toes with their regular inquiries of progress.

Thank you for believing in me.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Abstract

In speaking or singing, a source sound coming from the larynx is filtered by the vocal tract. Formants, the peaks of the resulting spectrum, determine the vowel and the timbre of the voice. At speech frequencies, between 100 Hz - 400 Hz, the harmonics of the source sound are spaced densely, so the peaks of the output spectrum largely correspond to the resonance frequencies of the vocal tract filter. At higher fundamental frequencies, like in singing or child speech, there are regions with no or low acoustic energy between the harmonics. The peaks of the output spectrum are then determined more by the location of the harmonics than of the filter resonance frequencies. Traditional methods for formant estimation, LPC and cepstrum, only use information from the spectral envelope. They perform well at speech frequencies, but at higher fundamental frequencies, they are not able to find the resonance frequencies of the vocal tract and determine the harmonics instead. Information about the location of the resonances is however still present in the sound and its spectrum, even if not in the spectral envelope. Breathiness, due to incomplete vocal fold closure, produces noise in the sound source, which, after filtering, corresponds to the shape of the vocal tract transfer function. Vibrato, periodic modulation of the pitch, produces frequency sweeps that hold information on the location of nearby resonance frequencies. A method able to extract this information at high frequencies could solve the current lack of an in vivo ground truth and would be applicable for singing training, language learning and some types of speech therapy, such as gender conversion.

In this thesis, a convolutional neural network was trained on synthetic data. It is able to predict the resonance frequencies of the filter accurately for new data with the same parameter settings. The mean absolute error is 23 Hz for 6 resonances and a fundamental frequency range of 100 Hz - 1000 Hz. The performance is stable throughout the frequency range and consistently better than that of the LPC algorithm implemented by the software Praat. The architecture and hyperparameters were selected in a systematic investigation. The influence of the parameters breathiness, vibrato, and resonance spacing on the performance has proven to be very important. The real-life applicability was tested with an additional dataset filtered by plastic tubes and a 3D printed vocal tract model. Finally, recommendations were formulated for perfecting this network, by incorporating recorded sounds and various well-designed parameter values in the training data.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Kurzfassung

Beim Sprechen oder Singen wird ein vom Kehlkopf kommender Schall durch den Vokaltrakt gefiltert. Formanten, die Maxima des resultierenden Spektrums, bestimmen den Vokal und die Stimmfarbe. Bei Sprachfrequenzen liegen die Obertöne der Schallquelle dicht beieinander, so dass die Maxima des Ausgangsspektrums weitgehend mit den Resonanzfrequenzen des Vokaltraktfilters übereinstimmen. Bei höheren Grundfrequenzen, wie bei Gesang oder Kindersprache, werden die Maxima des Ausgangsspektrums eher durch die Lage der Obertöne als durch die Resonanzfrequenzen bestimmt. Die üblichen Verfahren zur Formantschätzung, LPC und Cepstrum, basieren auf der spektralen Hüllkurve. Sie funktionieren gut bei Sprachfrequenzen, aber bei höheren Grundfrequenzen bestimmen sie die Obertöne statt die Resonanzfrequenzen. Informationen über die Lage der Resonanzen sind jedoch immer noch im Klang vorhanden, z. B. in der Behauchung und im Vibrato. Eine Methode, die in der Lage ist, diese Informationen bei hohen Frequenzen zu erkennen, würde das derzeitige Fehlen einer in vivo-Ground-Truth beheben und wäre für Anwendungen wie das Gesangstraining, das Erlernen von Fremdsprachen oder manche Arten der Sprachtherapie, wie z. B. bei der Geschlechtsumwandlung, geeignet.

In dieser Arbeit wurde ein konvolutionelles neuronales Netz trainiert, das 6 Resonanzen mit einem mittleren absoluten Fehler von 23 Hz bestimmen kann. Die Leistung ist im Grundfrequenzbereich von 100 Hz - 1000 Hz stabil und besser als die des von der Software Praat implementierten LPC-Algorithmus. Der Einfluss der Parameter Behauchung, Vibrato und Resonanzabstand hat sich als sehr wichtig erwiesen. Die Praxisanwendbarkeit wurde mit einem zusätzlichen Datensatz getestet, der mit Kunststoffröhren und einem 3D-gedruckten Vokaltraktmodell gefiltert wurde. Es wurden Empfehlungen für die Perfektionierung des Netzwerks formuliert, indem aufgenommene Klänge und verschiedene gut entworfene Parameterwerte in die Trainingsdaten einbezogen wurden.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Contents

Abstract **ix**

Kurzfassung **xi**

Contents **xiii**

1 Introduction **1**
1.1 Motivation and problem statement 1
1.2 Research questions 3
1.3 Structure of the work 5

2 Theoretical Background **7**
2.1 Sound 7
2.2 Voice function and physiology 9

3 State of the Art **19**
3.1 Speech recognition 19
3.2 Estimation of formants and resonances 20
3.3 Deep learning in speech recognition and formant estimation 23
3.4 High pitch formant estimation 25

4 Methods **29**
4.1 Neural networks 29
4.2 Data synthesis 36
4.3 Lab experiments 40
4.4 Evaluation 45
4.5 Implementation 46

5 Results and Discussion **47**
5.1 Loss function 47
5.2 Optimizer 51
5.3 Network architecture 53
5.4 Performance of the network on different datasets 61

5.5 Network trained with lab recorded sounds and tested on the vocal tract model . . . 71

6 Summary and Future Work **75**

Appendices **77**

Scatterplots synthesized data **79**

Scatterplot recorded data **85**

Resonances of lab recorded sounds **95**

Dependencies of the conda environment **97**

List of Figures **101**

List of Tables **109**

Bibliography **111**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

Introduction

1.1 Motivation and problem statement

Formant estimation in speech is a well developed field of research, necessary for speech recognition and synthesis. The description of phonation is often based on the linear source-filter model. The source is the glottis that vibrates with a fundamental frequency. The harmonics, integer multiples of the fundamental frequency, decrease in intensity with increasing frequency. The vocal tract acts as a filter that amplifies the harmonics according to their distance to the filter resonances. The peaks of the resulting spectrum are called formants and determine the perceived vowel. The low fundamental frequencies of speech (about 100 Hz for men and 200 Hz for women) have harmonics that are spaced more closely than the vocal tract resonances and allow estimation of both resonances and formants.

For higher fundamental frequencies, that in soprano singing can exceed 1000 Hz, the vocal tract resonances become undetectable with the classical methods for formant estimation such as Linear Predictive Coding (LPC) or Cepstral Analysis. The fundamental frequency may surpass the first vocal tract resonance and the harmonics are so sparsely spaced that the transfer function of the vocal tract is undersampled, as there may be no acoustic energy close to the resonances. The peaks of the resulting spectrum in this case do not correspond to the vocal tract resonances, as is illustrated in figures 1.1 and 1.2. Figure 1.1 is an example of a low frequency vowel with sufficiently densely spaced harmonics, so the output spectrum corresponds with the vocal tract transfer function. Figure 1.2 illustrates a vowel at 784 Hz, where the transfer function is undersampled and the peaks of the output spectrum do not correspond to the vocal tract resonances. In the latter case, reverse filtering fails and none of the classical methods are able to accurately determine the vocal tract resonances.

This is a well-known problem in the field and several adjustments to the classical methods have been proposed to solve it. They however still limit their definition of high pitch to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

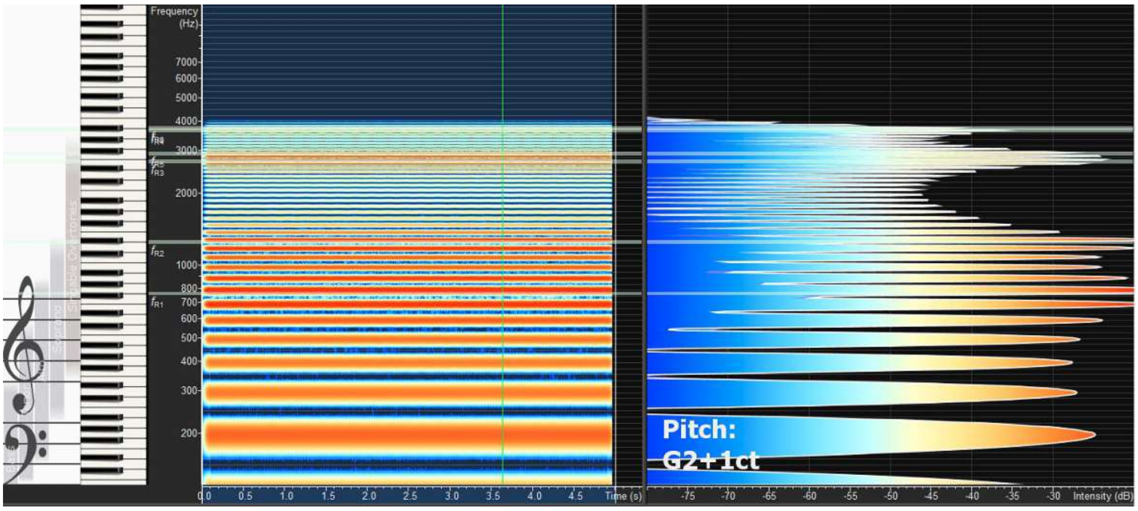


Figure 1.1: *Spectrogram of a vowel A at a frequency of 98 Hz (G2). The right part of the image shows the spectrum at the time selected by the green cursor. The vocal tract resonances are shown by the rulers and correspond with the peaks of the output spectrum.*

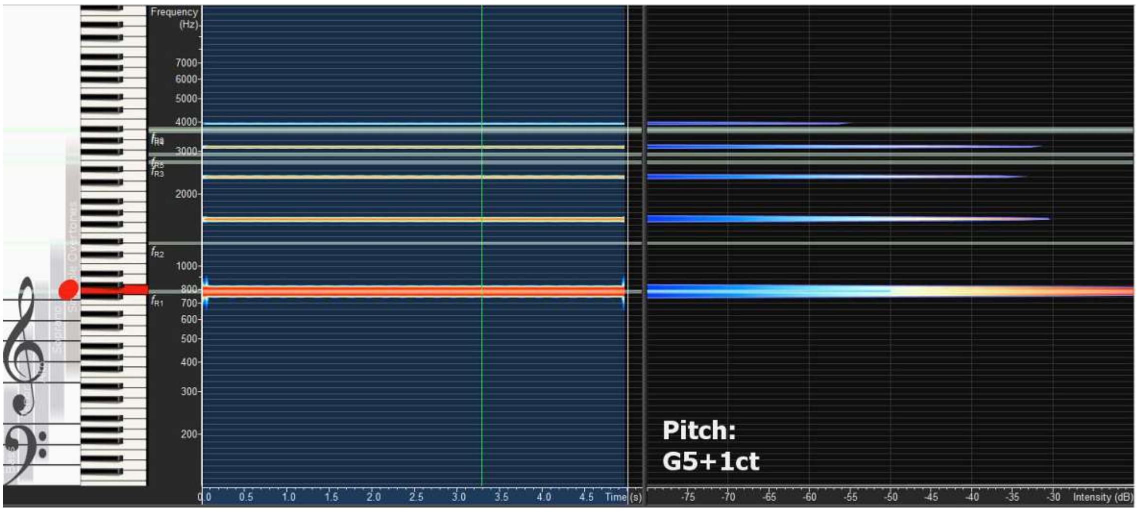


Figure 1.2: *Spectrogram of a vowel A at a frequency of 784 Hz (G5). The resulting spectrum consists only of the fundamental frequency and its harmonics. The vocal tract resonances are shown by the rulers and influence the amplitude of the harmonics, as the third harmonic has a higher amplitude than the second.*

speech-relevant fundamental frequencies, with ranges of up to 400-500 Hz. For speech recognition, also a limited number of 2 or 3 formants is sufficient, unlike for voice timbre manipulation. A regression using neural networks may provide a more reliable solution, by modelling the relationship between the resulting sound and the known ground truth resonances. This way, also features can be taken into account that would not be considered in methods based on the output spectrum, like breathiness (would be filtered out as noise) and vibrato (unstable pitch). This deep learning method provides a solution for estimating vocal tract resonances with an accuracy and in a fundamental frequency range that other methods are not able to at this time.

Applications requiring accurate knowledge about the vocal tract resonances include singing training, language learning and transgender voice training. Singers regulate their sound, the exact vowel and the colour of their tone by fine-tuning their vocal tract. In singing training, pedagogues use abstract, descriptive language to instruct their students. Different singing styles have very specific aesthetic requirements and visualization of both the student's sound and the desired result in a real-time feedback tool would be greatly valuable. In language learning, learners often have to learn new vowels that do not exist in their language and in transgender voice therapy, patients try to achieve a new voice timbre. Understanding and visualizing the influence of specific changes in vocal tract shape would significantly facilitate the learning process.

1.2 Research questions

The aim of this work was to train a neural network that would be able to estimate at least the first six vocal tract resonances with an error lower than the current state-of-art, in the fundamental frequency range of 300 – 1000 Hz, in order to reconstruct the transfer function of the vocal tract. In cooperation with Christian Herbst and Bodo Maass from the University for Music and Performing Arts Vienna, static, mid-vowel utterances with a wide range of fundamental frequencies were synthesized as training data, so the vocal tract resonances are known exactly. This is in contrast with existing work, where the ground truth resonances are often obtained by other methods, that themselves suffer from accuracy limitations in determining the underlying resonances. There is currently no sufficiently accurate method to estimate the vocal tract resonances based on solely the output spectrum, especially in the considered frequency range.

Inspired by Dissen [12], his approach from 2019 was applied for estimation: a convolutional neural network with raw spectrogram input. The pre-processing of the data was minimized, in order not to lose information that may be useful for the network. Different network architectures with different numbers of convolutional layers were examined, and the convolutional layers proved their applicability. Tracking, for example using recurrent neural networks, was outside the scope of this work, however, if the implementation of the resulting network is fast enough, it could still be applied in a real-time feedback tool. Features that were considered specifically are breathiness and vibrato, illustrated in figures 1.3 and 1.4. They were both systematically varied in the input data, with the hypothesis that data with more breathiness or vibrato would yield more accurate results. Breathiness is white noise in the glottal signal, that after filtering in the vocal tract shows local maxima of acoustic energy at the resonance frequencies, even where there are no adjacent harmonics of the fundamental frequency. The information that it holds is not used by methods based on the output spectrum, as those often filter out

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

breathiness as noise. Vibrato is a periodic fluctuation of the fundamental frequency and synchronously all harmonics. These fluctuations show higher acoustic energy closer to a nearby resonance frequency, aiding in determining its location.

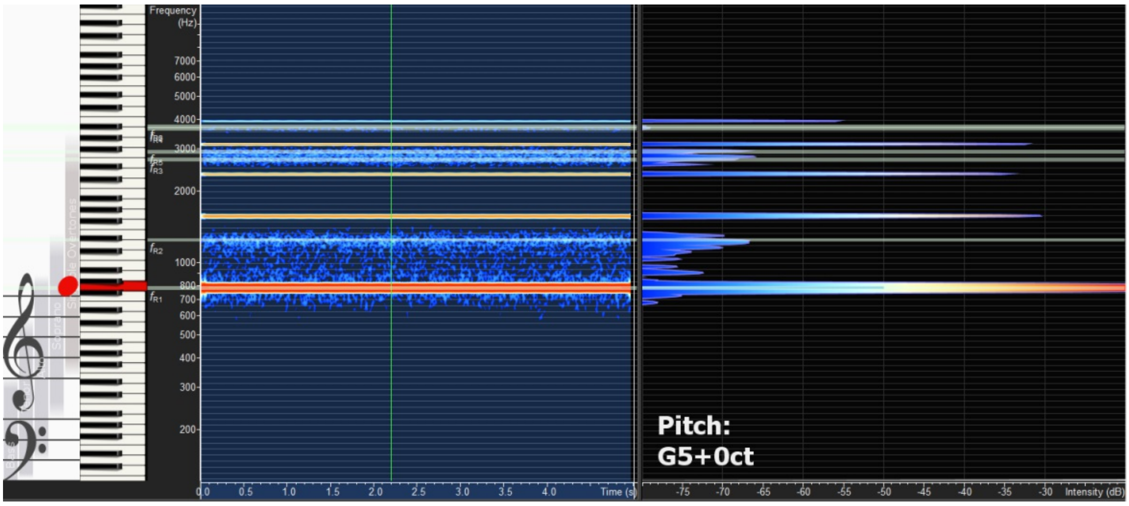
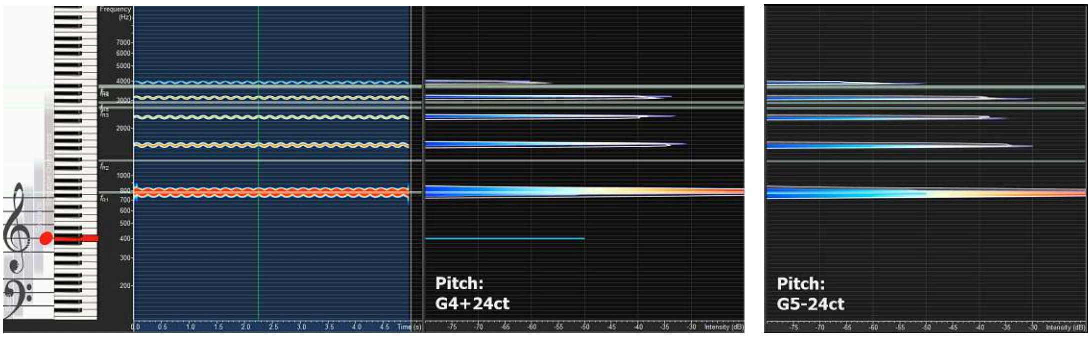


Figure 1.3: *Spectrogram of a vowel A at a frequency of 784 Hz (G5) with 1% breathiness. The white noise filtered by the vocal tract follows the shape of the transfer function at a low amplitude. The peaks of this low amplitude spectrum correspond with the vocal tract resonances.*



(a) (b)

Figure 1.4: *Spectrogram of a vowel A at a frequency of 784 Hz (G5) with 10 cent, 5 Hz vibrato. The amplitudes of the frequency fluctuations change according to their distance to nearby resonances. Left: crest and right: trough of the oscillation.*

Different loss functions and optimizers were used for fitting the model and their influence was examined. Finally, the following questions were answered:

- What network performs best in the task of reconstructing the vocal tract transfer function of synthesized mid-vowel utterances?

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

- How does this performance relate to other methods, both deep learning and other?
- Does the performance decrease for higher resonances?
- Does error increase with increasing fundamental frequency?
- Which features prove most useful in the estimation of the resonances?
- Are the resonances estimated with a higher accuracy for breathier vowels?
- Does vibrato aid in resonance estimation?
- Is this approach successful in addressing the shortcomings of current methods in the considered frequency range?

1.3 Structure of the work

This work is structured as follows: Chapter 1 is the introduction, including the motivation, problem statement, aim of the work and research questions. Chapter 2 describes the theoretical background, focusing on the physics of sound, and voice anatomy and function. In Chapter 3, the currently used methods are explained, and an overview is presented of the state of the art in formant and resonance estimation for the speaking and the singing voice, and of the applications of deep learning that try to solve these questions. In Chapter 4, the experiment is defined and the used methods are explained. Chapter 5 presents and discusses the results. Finally, Chapter 6 formulates a conclusion and an outlook for future work.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 2

Theoretical Background

2.1 Sound

This section briefly describes sound, sine waves, the Fourier transform and discrete Fourier transform, filtering and psychoacoustic perception of sound. The main source is [18].

Sound is a pressure wave through a medium, such as air, with frequencies in the audible range from 20 – 20 000 Hz. A wave is described by its wavelength and frequency, related to each other in the following formula:

$$f = \frac{c}{\lambda}$$

where f is the frequency in Hertz (Hz), λ the wavelength in meters (m) and c the velocity of the wave. For sound in dry air at 20^oC, this velocity is 343 m/s. The unit of frequency, Hertz, stands for oscillations per second. The reciprocal of frequency is the period T in seconds.

$$T = \frac{1}{f}$$

The period is the time needed to complete one entire oscillation. The displacement amplitude corresponds to the largest displacement of a single point from its equilibrium position. The pressure amplitude is the maximal pressure difference from the equilibrium pressure. Amplitudes can be peak-to-peak, peak, or RMS. The peak-to-peak amplitude is the distance from the crest (highest pressure) to the through (lowest pressure), and the peak amplitude from equilibrium to crest. The RMS amplitude is the root mean square of all discretely sampled values of the wave, or the amplitude of a constant signal with the same energy. For a continuous sine wave, it is the root mean square of the peak amplitude, that is $\frac{A}{\sqrt{2}}$. The sound pressure level is the relationship of the sound pressure at a certain point in space to a reference value by the formula

$$SPL = 20log_{10}\left(\frac{p}{p_0}dB\right)$$

where the reference pressure is most often set to the threshold of hearing, or $p_0 = 20\mu Pa$. The phase in radians (or degrees) describes the angle of a sine or cosine wave at moment t, corresponding to the current location in the oscillation. A full oscillation is a full circle of 2π radians, so a quarter of an oscillation is $\frac{\pi}{2}$. The phase is for example important for the summation of waves of different frequencies, as their phase difference decides whether their interference is constructive or destructive.

A basic sine-wave with frequency ω, amplitude A, and phase ϕ is described by the equation

$$y(t) = Asin(2\pi * (\omega t - \phi))$$

More complicated waves can be described as a sum of simple sine waves of different frequencies, amplitudes and phases. The Fourier transform is used to extract those different frequencies, amplitudes and phases, by conversion of the signal to the frequency domain. The formula for a continuous signal f(t) is [34]

$$\hat{f}(\omega) = \int_{t\in\mathbb{R}} f(t)exp(-2\pi i\omega t)dt$$

or expressed in the complex form,

$$\hat{f}(\omega) = \int_{t\in\mathbb{R}} f(t)cos(-2\pi\omega t)dt + i\int_{t\in\mathbb{R}} f(t)sin(-2\pi\omega t)dt$$

Based on the polar coordinates of $\hat{f}(\omega)$; $|\hat{f}(\omega)|$ and γ_ω, the magnitudes d_ω and phases ϕ_ω can be calculated of the sine waves, with different frequencies ω, that constitute the signal: [34]

$$d_\omega = \sqrt{2}|\hat{f}(\omega)|$$
$$\phi_\omega = -\frac{\gamma_\omega}{2\pi}$$

For digital representation of sound, discretization is necessary. The continuous signal is expressed as discrete, regularly taken samples. The sound level is sampled at a number of points per second – that frequency is called the sampling rate. An important theorem regarding the sampling rate, is the Nyquist-Shannon sampling theorem [39]. It states that only frequencies up to half the sampling rate can be correctly represented and reconstructed. Otherwise aliasing occurs, in which case lower frequency components are erroneously added, as they cannot be distinguished from the ones above the Nyquist frequency or folding frequency.

The Fourier transform also needs to be adjusted for discrete data, resulting in the discrete Fourier transform (DFT):

$$X(k) = \hat{x}(k/N) = \sum_{n=0}^{N-1} x(n)exp(-2\pi ikn/N) \text{ for } k \in [0 : N-1]$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

A short-time Fourier transform (STFT) considers the signal only for a certain window in time. Different window functions are used, resulting in different weights of the included samples. The number of samples that the window is shifted by in successive time steps, is called the hoplength. A function f(u), windowed by function g(u), results in a STFT at time t of

$$\hat{f}_{g,t}(\omega) = \int_{u\in\mathbb{R}} f(u)g(u-t)exp(-2\pi i\omega u)du$$

2.1.1 Filtering

When a sound passes through a filter, its spectrum is changed. The influence it has on specific frequencies is a characteristic of the filter, expressed in the filter transfer function. The transfer function is convoluted with the input spectrum to obtain the output spectrum. By transforming it to the frequency domain, the convolution is expressed by a multiplication. Fft filters, which are applied here, do indeed first perform a Fourier transform on the input, multiply it with their transfer function, and inverse transform back to the time domain to return the output. Other types of digital filters exist, but in this thesis, the simplification of fft filters is used. [39]

2.1.2 Psychoacoustic counterparts

It is important to differentiate between the physical quantities as described above, and human perception. By humans, most characteristics are perceived non-linearly and dependent on other parameters. Sound pressure level is perceived as loudness, but their relationship is not linear and depends on the frequency of the sound. Similarly, fundamental frequency is perceived as pitch, but the perceived pitch may also be dependent on the spectrum or other parameters. The sound spectrum influences the perceived timbre or colour of the sound. Although these characteristics do have a clear connection to accurately measurable physical quantities, their exact perception cannot be quantified because of differences between individual listeners and also for one individual listener in different times and circumstances.

2.2 Voice function and physiology

In this section, an overview of the anatomy and function of the human voice is given, with a specific focus on singing. Its main sources are [50], [24], and [21]. First, the anatomy of the entire system is described and illustrated. Then the functional mechanisms are explained that allow us to speak and sing, focusing on the three separate parts; the power source or respiratory system, the sound source or larynx, including the vocal folds, and the vocal tract, acting as a sound modifier or filter. Finally, the interactions between these separate systems are mentioned.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

2.2.1 Anatomical overview

The human voice originates in the throat, inside the larynx, which is part of the trachea, and in front of the esophagus. An overview is shown in Figure 2.1.

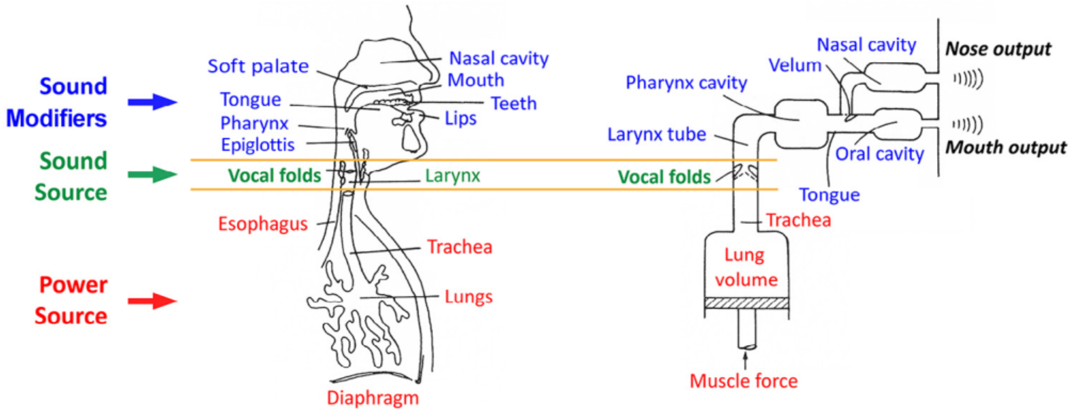


Figure 2.1: *A schematic overview of the human body parts, contributing to voice production. Based on [36], modified by [22].*

Respiratory system

The organs used for breathing and gas exchange form the respiratory system. These organs are the nose, the nasal cavities, pharynx, larynx, trachea, lungs, and the respiratory muscles. The trachea is the tube through which inhaled air travels from the nose to the lungs. At the bifurcation, it splits into the right and left bronchi, leading to the right and left lungs. The primary bronchi split up into smaller bronchi and then into bronchioles, finally leading to alveoli. These little vesicles with thin walls are the location of gas exchange between the air and the blood. Blood becomes oxygenated and disposes of CO_2. The lungs are located in the thorax, where they are protected by the ribcage from the sides and by the diaphragm from below. The diaphragm, a smooth planar muscle, is the boundary between the thorax and the abdominal cavity. When relaxed, it has the shape of a cupola, and when it is active, it flattens out, increasing the thoracic volume. The intercostal and stomach muscles are also part of the respiratory system.

Larynx [50]

The larynx consists of several cartilages, ligaments, and muscles [50]. It is loosely embedded in the surrounding tissue and not attached to the skeleton. The only bone connected to the entire structure is the hyoid bone, also a floating bone. The larynx can move several centimeters under influence of muscles or neighbouring structures, such as food passing through the esophagus.

The cricoid cartilage is a ring-shaped cartilage with a wider back than front, right above the upper ring of the trachea, forming the base of the larynx. Above it is the irregularly

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

shaped thyroid cartilage. Its notch, the connecting place of two wide frontal plates in the front, forms the protruding Adam's apple in men. In women and children, the connecting angle is wider and therefore less visible, and overall, the larynx is smaller. Laterally on the posterior side, the thyroid cartilage has an inferior cornu, connecting to the cricoid cartilage, and a superior cornu, connecting to the hyoid bone via a ligament and the thyrohyoid membrane. The cricoid and thyroid cartilage are also connected by the cricothyroid muscle, important for pitch control.

In the back, behind the thyroid cartilage and on top of the cricoid cartilage, are two mirrored arytenoid cartilages. The anteriorly located vocal processes of these cartilages are where the vocal ligaments are attached, and the muscular process in the back connects them to the cricoid cartilage via the lateral and posterior cricoarytenoid muscles. On their anterior long sides, the thyroarytenoid muscle attaches them to the thyroid cartilage. The cricoarytenoid joint allows the flexible movements necessary for vocal adduction and abduction and pitch control. The lateral and posterior cricoarytenoid muscles act as vocal fold adductor and abductor respectively. They are both attached to the muscular processes of the arytenoid cartilages. Finally, the arytenoid cartilages are connected by the interarytenoid muscle, which also adducts the vocal folds.

Besides the described muscles, all part of the larynx itself, also extrinsic muscles connect it to surrounding structures, like the sternum and the hyoid bone. Those are able to move the larynx vertically over several centimeters.

The vocal folds themselves are the medial thyroarytenoid muscle, also called the thyrovocalis, and the mucous membranes covering it. These membranes consist of the same layers as other mucosa, with an outer epithelium and underneath lamina propria. Lamina propria have three layers, of which the outer is more elastic due to the high elastin content, and the inner is more rigid as it consists of collagen. The opening between the vocal folds is called the glottis.

Cavities

The pharynx is the space behind the nose and mouth, divided into the nasopharynx, oropharynx and laryngopharynx. The latter is the place where the trachea and esophagus split, with the epiglottis, a flat cartilage, covering the trachea during swallowing. These cavities, together with the nasal cavity and the mouth, contribute to the filtering of the sound. The shape of the mouth can be influenced substantially by opening the jaw, moving the tongue or lifting the soft palate, the soft roof of the mouth in the back.

2.2.2 Sound production

When describing sound production using the source-filter theory, the previously presented structures are named after their roles:

- the respiratory system, providing flowing air needed to start vocal fold vibration, is called the power source

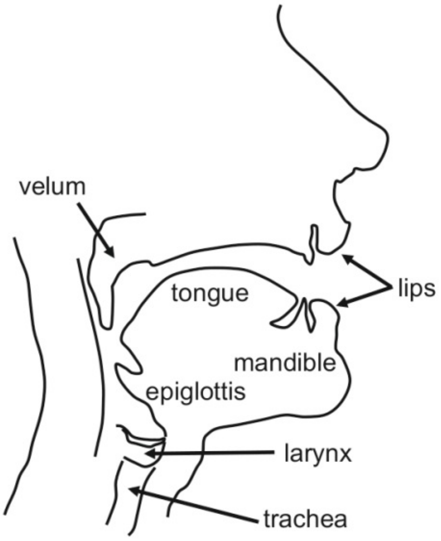


Figure 2.2: *A schematic overview of the articulatory organs. From [42].*

- the vocal folds themselves, or the larynx as a whole is called the sound source
- the supraglottal vocal tract, or the cavities above the glottis up to the mouth and nose are called the sound modifiers or the sound filter.

The source-filter theory is linear, assuming no interactions between these separate parts. That means that the sound coming from the source, the pulsating air shaped by the glottis, is assumed to be independent of what comes after it - the filter does not influence the source. The source sound consists of a certain fundamental frequency and its harmonics, decreasing with a certain spectral slope. Expressed in dB/octave, that slope describes the reduction of sound energy of these harmonics with increasing frequency, causing the spectrum of the source sound. The filter then shapes that spectrum, amplifying some frequencies and attenuating others. Assuming linearity in the source-filter theory, that shaping is a superposition - a simple multiplication in the frequency domain. The assumption of linearity is also made in this work, although it does often not hold physiologically, especially in the higher frequency ranges that are considered here. [48]

Power source

Breathing is a natural process that most of the time occurs subconsciously. It is driven by the contraction of the intercostal muscles and the diaphragm, which expands the ribcage. This creates an underpressure in the lungs, so air is sucked in through the nose until an equilibrium is reached. The relaxation of the breathing muscles pushes the air back out. In singing, the relaxation is gradual and controlled to get a steady flow out.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The breath first flows through the approximately cylindrical trachea, and for speech or singing, the cross-sectional area is diminished at the glottis by adduction of the vocal folds. The breath flow, together with differences in subglottal, interglottal and supraglottal pressure together make self-sustained oscillation of the vocal folds possible. Active muscle contractions can only reach frequencies up to 10 Hz [24], whereas the vocal folds, depending on their length and tension can produce sounds with a frequency of over 1500 Hz.

Sound source

During breathing, the vocal folds are abducted with a larger opening between them. In order to make oscillation possible, they need to be adducted. Two different mechanisms of adduction have been proposed by [25], [21]. For membranous medialization, the thyroarytenoid muscle, which forms the inside of the vocal fold, contracts, thickening and closing the vocal folds. For cartilaginous adduction, the posterior part of the glottis is pulled forward by the lateral cricoarytenoid and the interarytenoid muscles. These two mechanisms can occur together or independently from each other, but membranous medialization is always necessary for phonation.

The myoelastic-aerodynamic theory of phonation was already described by [2] and at that time attributed the self-sustained oscillation of the vocal folds to the increase of subglottal pressure, below the glottis, that opened the vocal folds, and the decrease of subglottal and interglottal pressures, partly due to the Bernouilli effect, that closed them. Since then, based on data obtained with pressure catheters and electroglottographs, the theory was updated [47]. The opening and closing of the vocal folds is caused by the push-pull effect of the intraglottal pressure, influenced by the subglottal, supraglottal and Bernouilli pressures. The push-pull effect means that the intraglottal pressure is larger during opening, and smaller during closing of the vocal folds. This depends on two mechanisms: the Vocal Tract Inertance Mechanism, and the Vertical Phase Differences Mechanisms.

Vocal tract inertance means that the air column in the vocal tract reacts to pressure changes with a delay. In the opening phase, the supraglottal pressure is positive and the airflow through the glottis is smaller. In the closing phase, the supraglottal pressure is negative and the glottal airflow larger. These effects support the push-pull effect on the vocal folds.

The vertical phase difference is based on the fact that the upper part of the vocal folds vibrates with a delay compared to the lower part. That means that during the opening phase, when the lower part is more open than the upper part, the shape of the glottis is convergent. The subglottal pressure then becomes dominant for the intraglottal pressure. During the closing phase, the divergent shape increases the influence of the supraglottal pressure. This contributes to the push-pull effect, because the mean subglottal pressure is larger than the mean supraglottal pressure.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Dividing and modeling the vocal folds with two masses above one another allows to formulate two vibration modes: a translational mode, where both parts vibrate in phase, and a rotational mode, where both vibrate exactly out of phase. Realistic oscillation can then be modeled as a combination of those two modes. Both the adduction mechanisms mentioned before and the vibrational modes play a role in register selection. The translational mode is necessary for both chest and head or falsetto register, but the rotational mode occurs predominantly in the chest register. The thyroarytenoid muscle is active in chest register and relaxed in falsetto or head voice.

The fundamental frequency is controlled by changing the tension of the vocal folds with the cricothyroid muscle, also resulting in lengthening or shortening of the vocal folds. A periodic, sinusoid modulation of the pitch is called vibrato. It is described by a frequency and an amplitude. Realistic values for these characteristics are a peak amplitude range of up to about 100 cents or 1 semitone and a frequency of 4.5-6.5 Hz [50] page 325. That means that an extreme vibrato can have a peak-to-peak amplitude of a whole tone.

The waveform of the glottal flow is depicted in figure 2.3 upper left. It shows a skewed open phase and a straight horizontal closed phase, as closing of the vocal folds cannot be negative, but it is not necessarily zero. The longer this closed phase is, the more and stronger harmonics are needed to construct the waveform. The so-called closed quotient is therefore correlated with the spectral slope, an expression of the decline of sound power in the harmonics in dB/octave. An example can be seen in figure 2.3 lower left. The physiologically realistic range of spectral slope is about -6 dB/octave to -18 dB/octave [24]. A steeper slope and therefore a quick decline of energy at higher frequencies results in a weaker sound that can be described as flutey. A more gradual decline results in more pronouced high frequencies and a more carrying and brassy sound. The spectral slope is therefore crucial in describing the spectrum of the source sound. Importantly, the vocal fold adduction is not always complete. In many speakers and singers, often in women, a gap remains through which air escapes during the closed phase. This posterior glottal chink causes turbulences and noise in the final sound, perceived as breathiness.

Sound modifiers

The source spectrum is then modified before it reaches the mouth. The vocal tract acts as an acoustic filter. An extremely simplified model of the vocal tract would be a cylindric, half-open tube with a mean length of about 15 cm for females and 17.5 cm for males. [42] Such a uniform tube has resonant frequencies, peaks of its transfer function, that can be approximately calculated with the formula

$$f_n = \frac{(2n-1) * c}{4 * L_{tube}}$$

where n is an integer, c is the speed of sound, generally taken to be 343 m/s, as it is in dry air at 20° C, and the length of the tube is expressed in meters. For these values and a vocal tract length of 17 cm, this formula yields results close to 500 Hz, 1500 Hz, 2500 Hz, 3500 Hz etc. [35], as illustrated in figure 2.3 lower middle. More precisely, the values

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

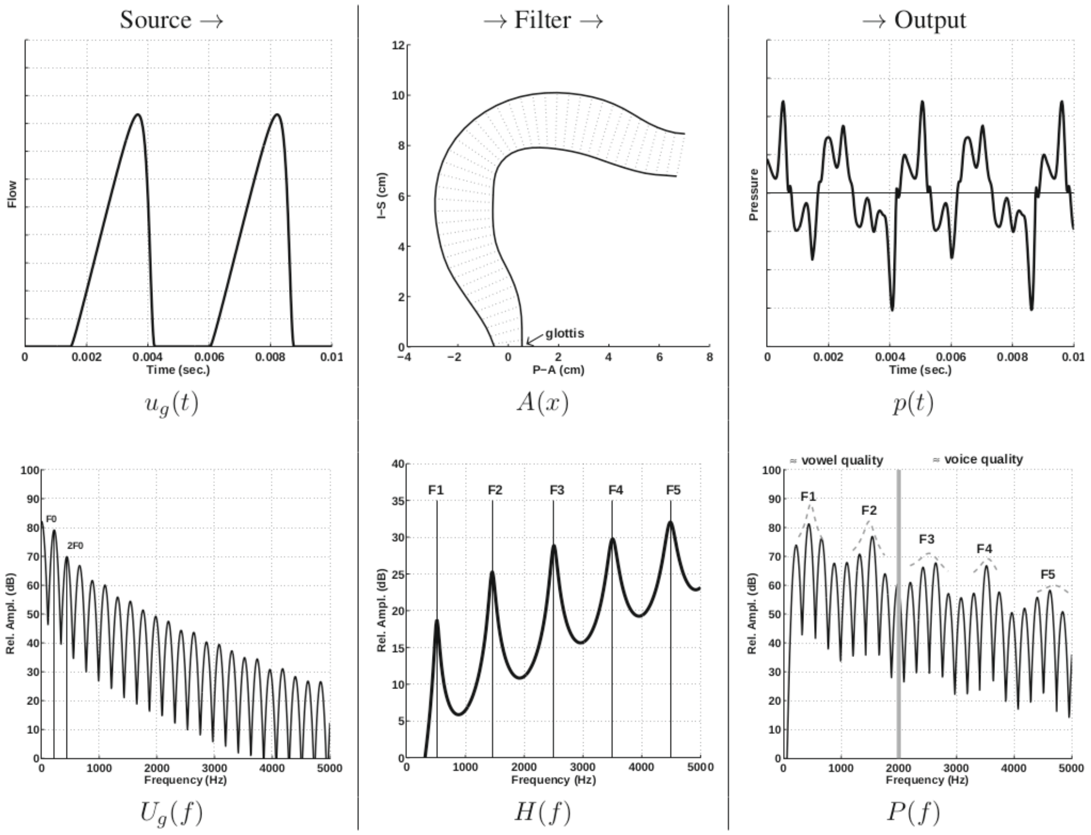
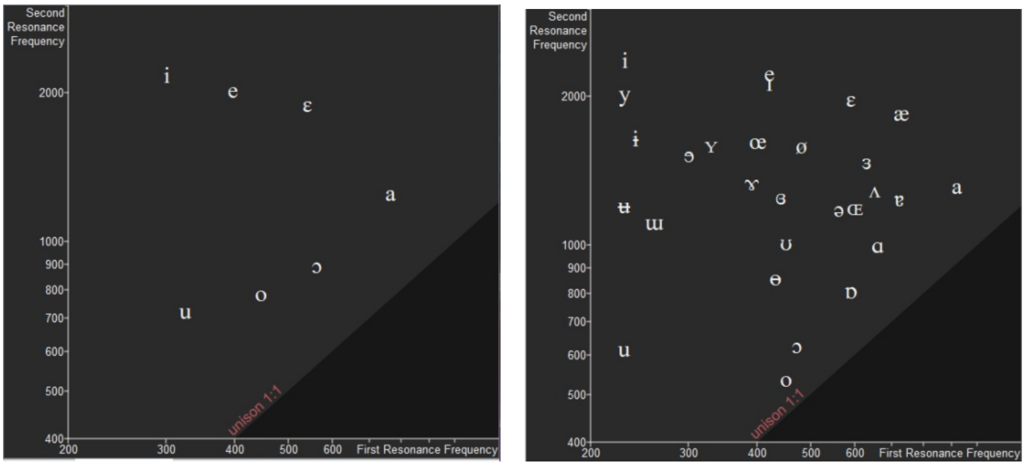


Figure 2.3: *Illustration of glottal flow, spectral slope and filtering by the vocal tract by [42]. Upper left are two cycles of glottal oscillation in time, with a slower opening phase and abrupt closing phase, and zero closed phase. Upper middle is the schematic shape of a neutral vocal tract. Upper right is the output waveform of a glottal flow filtered by such a vocal tract in the time domain. Lower left, the glottal flow is presented in the frequency domain, where all harmonics, multiples of the fundamental frequencies, are represented with their intensity decreasing according to the spectral slope. Lower middle is the transfer function of the neutral vocal tract, with peaks at 500 Hz, 1500 Hz, 2500 Hz, 3500 Hz and 4500 Hz, as is the case for a theoretical 17 cm cylindrical tube. Lower right is the spectrum of the resulting output sound, shaped by both the spectral slope and the vocal tract transfer function. The peaks of the transfer function can be deduced from the output spectrum, as the harmonics are sufficiently closely spaced.*

rounded to two decimals are 504.41 Hz, 1513.24 Hz, 2522.06 Hz, 3530.88 Hz etc. Overall, that means about one resonance for every 1000 Hz for males. These resonant frequencies will be estimated from the output sound in this thesis.

The area of the vocal tract can be consciously varied by adjusting jaw opening, lip, tongue, and velum position or many other neighbouring muscles. This fine-tuning allows for a wide range of resonance frequencies in different combinations, determining the sound colour or voice timbre. The first two resonant frequencies define the vowel and are therefore essential for speech [28]. The next few determine the timbre or colour of the voice, which is also very important for singing. Many singers, especially classical, have a spectral peak around 3000 Hz, called the singer's formant. [46]. It is formed by clustering the nearby resonances together. At the fundamental frequencies of speech, about 100 Hz for males and 200 Hz for females, the resonance frequencies do approximately correspond to the peaks of the output spectrum, called formants. In figure 2.4, a chart is shown of the possible vowels with the frequency of the first resonance on the x axis and of the second resonance on the y axis. On the left, the simple triangle of the Italian vowels is shown, and on the right, the entire scale of the International Phonetic Alphabet (IPA). These vowel charts are from VoceVista. [32]



(a) Italian vowels

(b) International vowels

Figure 2.4: *Vowel charts extracted from VoceVista [32]. The IPA symbols are shown in function of the frequencies of the first and second vocal tract resonance. It should be noted that the frequencies are not defined as exactly and vary with speaker sex and age [28].*

The area function of the vocal tract, expressed in function of the distance from the glottis, can be connected to the resonances using acoustic sensitivity functions, as applied by [41]. This way, the resonances calculated here can be related to vocal tract conformations in later work.

Formants and resonances of the vocal tract

It is important to point out the distinction between formants and resonances. [49] Formants have traditionally been defined as the peaks of the output spectrum. For

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

accuracy it is important to define the resonances separately as the resonance frequencies of the vocal tract filter, or in other words, the peaks of the vocal tract transfer function. At low fundamental frequencies, where harmonics are closely spaced, formants and resonances may be identical, but this is certainly not always the case. At higher fundamental frequencies, as the harmonics are spaced more widely, their distance to the resonances usually increases. In singing, the ability can be trained to colour the vowel so the resonances align with the harmonics, called formant tuning. That way, the filter adapts to the location of the acoustic energy and the sound is amplified.

Following the recommendations of [49], the frequencies of the vocal tract resonances are notated f_{Rx} for the x-th resonance and the formant frequencies, meaning peaks of the spectrum, as f_x for the x-th formant. In many works, this distinction is not stated explicitly or the terms are used interchangeably. From context, it can often be concluded which is meant. Additionally, the fundamental frequency is notated f_o, where the subscript 'o' stands for 'oscillation', also following [49].

Interactions

The linear source-filter theory assumes linear superposition of the source and filter spectra, that means a mathematical convolution in the time domain and multiplication in the frequency domain. This approximation holds for male speech at low frequencies, where the main interaction is skewing of the glottal flow [48], a level 1 interaction. [22]. However, for the ranges considered here, which are relevant in child and female speech and even more in singing, source-filter coupling can be an important phenomenon and more levels of interactions are possible. The coupling strength depends on the ratio of the open area in the glottis and the aria of the vocal tract right above it, on the length of the vocal tract and on the subglottal pressure.

These possible source-filter interactions, as described in [22] and [23] are:

- Level 1: The wave shape of the glottal air pulse can be skewed because the reactance of the vocal tract is positive. That means that the air in the vocal tract will not be exactly in phase with the glottal opening, but transmit the energy a little later.
- Level 2: Changes in the shape of the vocal tract cause changes in its reactance. That may influence the flow and vibration in the glottis.
- Level 3: Biomechanic effects that may explain intrinsic vowel pitch, proposed by [43].

Secondly, the vocal fold adduction has an effect on airflow. The mechanisms mentioned above, cartilaginous adduction and membranous medialization, control singing register and the spectral slope of the source. Adduction does also influence the possible airflow and the subglottal pressure. This is therefore an interaction of the source and the power.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Finally, also the vocal tract and the respiratory system interact. The lungs are connected to the trachea and the diaphragm, so when the latter contracts and lowers, it pulls down the entire system. The larynx, sitting on the trachea, is lowered as well and the vocal tract is lengthened. This change of shape has a direct influence on the resonance frequencies of the filter.

In the current problem, these intricate interactions are not considered. The parameter values in a sample sound are assumed to be the outcome of a steady state of the subsystems. This stable conformation results in the present fundamental frequency, spectral slope, loudness, and resonance frequencies, and the physiological possibility of those combinations is not inspected.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 3

State of the Art

3.1 Speech recognition

Speech recognition is a highly relevant and well-developed field, basing on different methods through the years. The goal of speech recognition is not to extract information about the vocal tract configuration, but about the spoken words, which belong to a certain language, or their underlying meaning. For that reason not only open vowels, but all other phonemes of the language have to be recognized. The output spectrum, as heard by the listener, is more relevant for this goal than the underlying resonances. As the fundamental frequencies of human speech are about 100 Hz for men and 200 Hz for women, undersampling of the harmonic series is not an issue. Moreover, many consonants have no fundamental frequency at all. The relevant information is encoded differently in the acoustic signal than in the current problem, so the pre-processing strategies that have evolved for speech recognition applications will not all serve the current application.

The general pipeline of connecting acoustic signals to words (in English, or characters in Asian languages) consists of the following four steps: [54]

1. Preprocessing the audio signal, noise filtering and feature extraction
2. Acoustic models that calculate acoustic scores
3. Trained language models that estimate word combination probabilities
4. Hypothesis search combines the acoustic and language scores to decide on the most probable word

3.2 Estimation of formants and resonances

The most common methods for envelope-based formant estimation are linear predictive coding (LPC) and cepstral analysis. In the following section, their principles are described, as well as some adjustments applied in recent studies in order to cope with their shortcomings.

3.2.1 LPC

Linear predictive coding is a parametric method for spectrum estimation. The envelope is described as a linear all-pole digital filter, where the n-th sample is expressed as a linear combination of the p previous samples, as described in [31]. Assuming that the signal is stationary in a limited time-frame, the n-th sample in a model of order p becomes:

$$\hat{S}(n) = -\sum_{k=1}^{p} a_k S(n-k)$$

where a_k are the linear coefficients, calculated to minimize the mean squared error. Because of the least squares approach for optimization, the peaks of the spectrum have higher weights in the error criterion, and are modeled more accurately than other, lower energy regions. [1]

The error e(n) between the predicted and the real value is attributed to the glottal source.

$$e(n) = S(n) - \hat{S}(n) = S(n) - \left[-\sum_{k=1}^{p} a_k S(n-k)\right]$$

Taking the Z-transform, that gives

$$E(z) = S(z)\left[1 + \sum_{k=1}^{p} a_k Z^{-k}\right]$$

The Z-transform is comparable to the Laplace transform but for discrete and ideally sampled signals. The unilateral Z-transform is defined as

$$X(z) = \sum_{n=0}^{\infty} x(n) z^{-n}$$

[39] so what would be a convolution in the time domain, will become a multiplication in the Z domain, similarly to the Fourier transformation.

A transfer function is the function that modifies the input to become the output. In mathematical form, where Y(s) is the output, X(s) is the input and H(s) is the transfer function:

$$Y(s) = H(s) * X(s)$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

or

$$H(s) = \frac{Y(s)}{X(s)}$$

In this case, the input is E(z) from the glottal source and the output spectrum is S(z), so the previous expression becomes

$$H(z) = \frac{S(z)}{E(z)}$$

and the transfer function of the vocal tract H(z) can be expressed as [1]

$$H(z) = \frac{1}{[1 + \sum_{k=1}^{p} a_k Z^{-k}]}$$

Because of the 1 in the numerator, this function only has poles and no zeroes. The peaks of the similarly described spectrum are then considered to be the resonances of the filter.

The chosen filter order is an important parameter for successful LPC. The order directly influences the number of formants that the method will return. For a too high order, more will be given than are present, and for too few, formants may be merged. The preferred approach is choosing a higher order rather than a lower one and rejecting formants that do not fit the background knowledge on expected location or bandwidth. [35]

All LP methods aim to predict a sample as a linear combination of previous samples. They find the linear coefficients that minimize the error over a certain window. Different methods are available for that calculation [33]: the covariance method includes the factorization of a covariance matrix. It is more accurate, but may also result in negative bandwidths. The autocorrelation method is more robust, but less accurate. [19]. They are explained in more detail in [5]. The Burg algorithm, the one used by the program Praat, [3] is alternatively called maximum entropy spectral analysis or covariance lattice method. It gets the reflection coefficients recursively from the data [40]. Two different approaches for predicting the sample, forward and backward, are combined into a recursive formula with coefficient k_i, the i-th order reflection coefficient. The partial derivation of the sum of the forward and the backward error in regards to k has to equal zero in this method, allowing solving for k_i for each order i. These k_i are then used to calculate the linear prediction coefficients.

Praat's default settings are optimized for speech: They look for 5 formants in each frame, but the maximum number of formants can be set to any multiple of 0.5. [4] The formant ceiling, or maximal frequency at which to look for formants, is set to correspond with the speaker's vocal tract length. Praat's default value is 5500 Hz, corresponding to an average female speaker. For an average male, the formant ceiling should be set to 5000 Hz and for children up to 8000 Hz. The manual does recommend to experiment with the specific speaker on steady vowels. For the current application however, adjusting the settings to each separate speaker would be suboptimal. The number of formants

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

and the formant ceiling are set for the optimal performance achievable by the algorithm on the data. Praat's approach is as follows: initially, the sound is resampled to twice the formant ceiling. The effects of resampling on these algorithms are described in [33], who recommends not downsampling to frequencies below 16 kHz. The LPC coefficients are then computed using Burg's algorithm, and from those double as many poles as the maximum formants setting are obtained. The LPC algorithm inserts artifacts, very high and low formant values, when the spectral slope does not equal the assumed 6 dB/octave. These artifacts are removed by erasing all formants below 50 Hz or 50 Hz below the formant ceiling. Alternative built-in algorithms and suggestions to tweak the settings to obtain the best results are provided. [4]

A popular adjustment to LPC methods is weighing some samples higher than others in the error criterion. This method is sometimes called Weighted Linear Prediction (WLP). [1] explains that the open glottal phase causes more noise and the strong excitation can disturb the attempt to model only the envelope. Different approaches to weighting and windows have been proposed. Some determine the open and closed phases separately, using electroglottography [1]. [45] sets the open phase to 0 and the closed phase to 1 in a data driven approach for determining these phases using the fundamental frequency.

The LPC method uses a linear scale for frequency, which does not correspond with the quasi logarithmic scale of human perception. This can create disagreement in resonance spacing. [31]

As this method is based on the assumption that the peaks of the spectrum are due to the filter transfer function and not the harmonics, it fails at higher fundamental frequencies. This problem is also called 'pitch locking' [45]. The sparse harmonics hold the sound energy and the influence of the filter resonances is not possible to calculate from the envelope.

3.2.2 Cepstrum

Cepstral methods are based on a specific data transformation. Firstly, a Fourier transform is performed to the frequency domain. There, the effect of the filter on the source sound can be described as a multiplication. By taking the logarithm of the signal, it can be expressed as a sum of the source and filter: [31]

$$log(S * F) = log(S) + log(F)$$

The inverse Fourier transform of this sum does not lead back to the time domain, but to the cepstral domain. Signal analysis terms are changed in this domain by flipping the first few letters: spectrum becomes cepstrum, frequency becomes quefrency, harmonics become rahmonics and filtering becomes liftering. The underlying meaning of cepstrum is the rate of change of spectral magnitude [31] - at low quefrencies, the spectrum has broad, widely spaced peaks, that could describe a filter transfer function. A high quefrency represents closely spaced harmonics or noise.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Based on this understanding, the filter transfer function could be extracted by low-pass liftering. The quefrencies above a selected cut-off quefrency are set to zero and the remaining cepstrum is transformed back to the frequency domain, where the peaks of the spectral curve are inferred to be the filter resonances [44], [55]. This approach also shows directly the problem with higher fundamental frequencies: more widely spaced harmonics have lower quefrencies, possibly below the cut-off frequency. The cepstrum then cannot be clearly separated into the filter and harmonics. A solution implemented by [55] and [58] is rahmonic subtraction. By deriving the fundamental frequency in the frequency domain and deducing the cepstrum of the harmonic components, the cepstrum of the source can be subtracted from the full cepstrum [58]. The influence of the harmonics in the lower quefrencies is thus eliminated and the filter transfer function can be obtained by cepstral liftering and back-transformation as usual.

[31] combines the LPC and cepstrum methods with wavelet transform. As the wavelet transform preserves all information, it can be applied before the other method, either 10-th order LPC or 15 cepstral coefficients. This way, the calculation of the peaks of the spectral envelope can be more accurate.

[10] estimates the vocal tract resonances using Kalman filtering. He uses LPC cepstra (LPCC) instead of LPC coefficients and minimizes approximation with a newly added residual term, trained iteratively and adaptively.

3.3 Deep learning in speech recognition and formant estimation

Current speech recognition methods are widely used by big technology companies to bring applications to end users. Deep learning has boomed thanks to computational power being available, and is proving its relevance in speech processing as well.

The research teams of four institutions; the University of Toronto, Microsoft Research, Google and IBM Research; collaborated on an overview of the state of the art in 2012. [27] In deep learning applications for speech processing, the standard in past decades was a combination of Gaussian mixture models (GMM) with Hidden Markov models (HMM). The raw waveform is usually pre-processed by extracting MFCC or LPC coefficients, both of the original signal and of its derivatives, to limit the information to what is relevant for the problem. Deep neural networks have started to replace the traditional models owing to their ability to extract patterns out of complicated data without prior assumptions, as was necessary with GMM or others. As neural networks are prone to overfitting the data when the training set size is limited, generative pretraining is used, training a few layers separately to extract different features. Those pre-trained layers are then connected to the rest of the network for the main training. That way, some of the prepared features can be readily available for the network, improving performance while limiting computational requirements.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

For the exploratory experiments in the mentioned review [27], and in many others, the TIMIT database is used. The TIMIT corpus is a standard database of spoken sentences by 630 different speakers, each with transcriptions to phonemes and words aligned in time. It is the result of a collaboration of the Massachusetts Institute of Technology (MIT), SRI International (SRI), Texas Instruments, Inc. (TI) and the National Institute of Standards and Technology (NIST). It was first released in 1988 and updated in 1993. Still today, it often serves as a training and evaluation set for new speech recognition models. [17], [16]

Several past works have applied neural networks to estimate the formants of vowels for speech-relevant frequencies. They all used the TIMIT corpus for their experiments. In the following paragraphs, four studies by three authors are described.

Yehoshua Dissen and Joseph Keshet created DeepFormants in 2016. It is a publicly available software package with preprocessing and a pre-trained deep learning model, used to predict formants from recorded sound files. [13], [11] The preprocessing transforms the sound to LPCC coefficients. The model was trained using the data of the TIMIT corpus, with 400 input datapoints from different parameter values of the preprocessing fed in parallel. The fully connected neural network has 3 hidden layers with 1024, 512, and 256 nodes, sigmoid activation in all layers except the last one and was trained with loss MAE and optimizer Adagrad. It predicts 3 formants and is intended for speech. The package also includes a model for tracking using Recurrent Neural Network (RNN).

In 2019, this model was extended with another publication, that proposed an additional model for formant estimation based on a convolutional neural network with raw spectrogram input. This network had 4 convolutional layers and 2 fully connected layers, but the numbers of nodes are not specified. In this study, several datasets were used: the vocal tract resonance (VTR) corpus, which is a part of the TIMIT corpus, the CT corpus [7] and the HGCW dataset [26]. The CT dataset contains sounds from young women exclusively, and the HGCW from men, women and children in approximately equal proportions. In the 2019 study, the influence of higher frequency was addressed by domain adaptation networks: the network, trained on data with limited fundamental frequency, was used as a basis and its output was fed along the original acoustic features into a new adaptation network. This way, not as much high-frequency training data was needed to improve performance for women's and children speech. [35]

Deng had mapped LPCC to vocal tract resonances with Kalman filtering in 2007 [10]. Sai [38] also used LPCC as input, but using a multilayer feed forward (MLFF) neural network. The network was trained with a limited dataset of 324 files from the VTR database, a subset of TIMIT. Two different models were trained with two different ground truth approaches; the hand-picked VTR as included in the TIMIT database, and spectral prominent regions extracted from spectrograms by Dynamic Programming. The network had an input layer with 15 nodes and linear activation, two hidden layers with 50 nodes each and non-linear activation, and an output layer with 3 nodes, corresponding to 3 formants or resonances, with linear activation. Adding of a third hidden layer showed to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

not be an advantage due to the small dataset. The performance of the two models was compared to WaveSurfer and Praat.

In 2019, Dai [9] used gated bilinear networks for formant estimation, also using the VTR dataset, training on 346 utterances with mostly (324) hand-picked ground truth. Analogously to Dissen [12], the input were LPCC of orders 8 to 17 extracted by Matlab, and pitch synchronous cepstrum coefficients, based on Dissen's open source code. In bilinear (BL) networks, the outputs of two different linear transformations are combined with a sigmoid operation, which is preferred to convolutions by the authors due to speed. The model was trained to minimize MAE with optimizer Adam and changing learning rate. In Dissen [12], the last fully connected network layer has the same number of output nodes as is the number of wanted resonances, and the last hidden layer is shared by all resonances. In Dai [8], on the other hand, a multitask output approach is taken. Each resonance output is preceded by a separate dense layer with 256 nodes, as each resonance is connected to its own frequency band.

3.3.1 Image recognition

Using deep learning for image recognition is now a well-estabilished practice. Some applications are image classification, object or face detection and localization. The standard approach are CNNs, and recently Res-Net, with a skip connection, has gained popularity since winning ILSVRC 2015 classification task in 2015. [20]. Current research mainly fine-tunes those by combining different existing methods, or tuning hyperparameters [57], [30], [51], [56]. The latter includes tweaking learning rate of respective layers, freezing and defreezing, or adjusting the number of epochs to fit the current problem. In image classification and object detection, the location of the object on the image is not relevant, unlike in the current application.

3.4 High pitch formant estimation

At common speech fundamental frequencies, the harmonics are spaced sufficiently closely to sample the vocal tract transfer function well, and the spectrum of the output sound corresponds to the shape of this transfer function, including the location of the peaks. The resonance frequencies can be estimated by peak-picking and based on the envelope of the sound.

At higher fundamental frequencies, the information about the location of the resonances is still present, but not necessarily in the envelope. The envelope may demonstrate "pitch locking", where the peaks correspond with the harmonics, multiples of the fundamental frequencies. At other frequencies than the harmonics, no or only little acoustic energy is present to be filtered by the vocal tract. Since the vocal tract cannot add energy in itself, only amplify or attenuate what it already present, the spaces in between the harmonics are lower, if not empty, and not the peaks. The information about the vocal tract transfer function can however still be present at lower amplitudes, for example if

TU Bibliothek WIEN Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

noise is present. [58]. White noise, holding all different frequencies, does follow the shape of the transfer function, but does not influence the envelope and is often even filtered out in speech applications.

Many authors acknowledge the problem at higher fundamental frequencies and different approaches were proposed to solve it. Most of these are variations on the standard LPC and cepstral methods. The definition of "high pitch" varies widely, ranging from 400 Hz to over 1000 Hz. The lower range corresponds to female and child speech. For singing, larger ranges are necessary, up to 1400 Hz in the common opera repertoire. (The Queen of the Night in the opera Die Zauberflöte by W.A. Mozart sings several famous F6, at 1396.91 Hz).

A few studies presenting solutions for higher pitch are presented here. Wang [52] uses fundamental frequencies that change over time, within a single vowel. Sweeping through different pitches reveals which frequencies are amplified and indicates the filter resonances. Although the condition of changing pitch is sometimes present in speech, this approach does not solve the cases where the pitch is high and constant.

Yegnanarayana [53] suggests adjusted spectrogram settings, a combinations of wideband and narrowband, to display formant information better. Liu [29] proposes a variant on weighted linear prediction, extracting the moments of glottal closure automatically based on pitch, but assuming nonstationarity. The previous three authors all limit their considerations of high-pitched speech to 400 Hz.

Alku [1] and Story [44] go up to 500 Hz. Alku [1] uses Weighted Linear Prediction, determining the open glottal phases using electroglottography. Story [44], inspired by cepstral methods, low-pass filters a narrow-band spectrum, where the cutoff quefrency is dependent on the fundamental frequency, and then calculates the peaks of the obtained spectral envelope. The method is tested on synthesized sounds, to have access to a reliable ground truth. Applying rahmonic subtraction, a cepstral method, Zhang [58] and Zarras [55] estimate the vocal tract resonances of synthesized vowels with fundamental frequencies of up to 750 Hz and 800 Hz, respectively.

Drugman [15] developed a technique called Fast Inter-Harmonic Reconstruction (FIHR), to pre-process a speech signal before applying autocorrelation LP peak-picking on the spectral envelope. The FIHR method adds a parametrically set number of inter-harmonics between the real harmonics, artificially lowering the fundamental frequency. The amplitude of the inter-harmonics is dependent on the amplitudes of the neighbouring harmonics and a weight. The calculation is so quick, that real-time processing is possible. The method was tested on synthetic signals with fundamental frequencies up to 1000 Hz.

The discussed approaches are mostly adjustments to the envelope-based LPC and cepstral methods. At higher pitches, the information on the vocal tract resonances is not readily present in the envelope anymore, but hidden in the spectrum. Methods like rahmonic subtraction try to disentangle the relevant information in the envelope from the harmonics. The approach in this work is not envelope-based, but attempting to extract the information from the raw spectrograms. For that reason, the pre-processing of the data is limited,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

to not filter out information that can be relevant, like noise and vibrato. Additionally, because vibrato sweeps over different pitches on a smaller scale, the same mechanism applied by Wang [52] is present in the current data in a realistic amount. The TIMIT database is not used because of the limited frequency range, and because the ground truth of synthesized sounds is more reliable.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 4

Methods

4.1 Neural networks

In deep learning, models are trained to connect input and output data in non-linear models, which are hard to interpret due to their complexity. A neural network is such a model, based on the structure of biological neural networks in the brain. During training, it learns to map input to output based on sufficient training data, in a process that includes readjusting the model parameters repetitively. The training is supervised, so both the input and the ground truth output data are provided.

The basic set-up of a neural network is that of a multi-layer perceptron. It consists of multiple layers of nodes, or neurons, that are connected to the nodes of other layers with weighted connections. The input of a node are the values passed from the previous layer, of which the weighted sum, or another operation, is taken. The result is passed through an activation function, which is often a non-linear operation to suppress or amplify the value, creating the output. If Y_{node} is the output of the current node, g(x) is its activation function, Y_i the output of the i-th previous node and w_i its weight, the formula to represent one node becomes:

$$Y_{node} = g(\sum_{i=1}^{n}(w_i Y_i)) \tag{4.1}$$

4.1.1 Layers

The first layer of a neural networks is called the input layer, the last one the output layer, and the layers inbetween are the hidden layers. They can be interconnected fully, from each node of one layer to each of the other, or in specific patterns that determine the name of the network. The number of nodes in each layer is determined to fit the problem at hand. Adding more nodes, and with it, degrees of freedom, enables the network to model more complex problems. Too many degrees of freedom can however lead to overfitting and increase computational and memory demands. The ideal number of degrees of freedom, ensuring optimal performance with the simplest model possible, is found by experiment.

4.1.2 Convolutional neural network

The network architecture used in the current work is a convolutional neural network (CNN). The first layer or layers are convolutional layers, followed by a flattening layer and some fully connected layers. Convolutional layers consist of filters or convolution kernels of specified sizes, that are multiplied (convoluted) with small, overlapping regions of input, most often images. They are used to extract features from the data. After each convolutional layer, a pooling layer can be used to limit the computational and memory demands, by downsampling the output of the convolution, for example by taking the maximum.

4.1.3 Classification and regression

Neural networks are usually used for two types of tasks: classification and regression. In classification problems, the output layer has as many nodes as there are classes to choose between. The class with the highest value, closest to one, is then considered to be selected by the network. In regression tasks, not the class, but the value of the output is relevant. Several nodes can be still used at the same time and most commonly, a linear activation function is used for the last layer. That is also the case in the current work.

4.1.4 Hyperparameters

Hyperparameters to consider when constructing a network are:

1. Network architecture
    - Number of layers
    - Number of nodes per layer
    - Special architectures like convolutional layers or skip layers
    - Activation function
2. Training process
    - Loss function
    - Optimization and learning rate
    - Number of epochs
    - Batch size

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Below, a list of these hyperparameters is given, along with their formula or a short explanation of their function and possible impact. The mentioned options are those that are readily available in Keras, the used Python library.

Loss

The loss defines the function of the difference between the true and predicted values that should be minimized. It it chosen to fit the current problem and the value distribution. Here, only regression losses are discussed that seem applicable for the current problem of predicting frequencies. [6]

1. **MSE** - Mean squared error

$$\text{MSE} = \frac{\sum_{i=0}^{n}(Y_{i,true} - Y_{i,pred})^2}{n}$$

   MSE is the default loss, assuming a Gaussian distribution. The squaring gives the function the shape of a parabola, where larger errors have a quadratically larger impact than smaller ones.

2. **MAE** - Mean absolute error between predicted and true value.

$$\text{MAE} = \frac{\sum_{i=0}^{n}|Y_{i,true} - Y_{i,pred}|}{n}$$

   If the Gaussian assumption does not hold, the mean absolute error is the simplest option. The linearity of the function causes the slope away from the minimum to be smaller than in MSE, so the convergence is slower.

3. **MAPE** - Mean absolute percent error

$$\text{MAPE} = \frac{\sum_{i=0}^{n}\frac{100*|(Y_{i,true}-Y_{i,pred})|}{Y_{i,true}}}{n}$$

   In this case, the error is related to the value of the ground truth. That means that higher resonance frequencies could be predicted with larger errors. Although the lowest vocal tract resonances are most relevant for vowel identity and colour, the ambition of this work is to accurately predict all, so MAPE is not a fitting approach.

4. **MSLE** - Mean squared logarithmic error

$$\text{MSLE} = \frac{\sum_{i=0}^{n} log^2(Y_{i,true} + 1) - log(Y_{i,pred} + 1)}{n}$$

5. **Cosine similarity**

$$\text{CosSim} = -\sum(||Y_{true}||_2 * ||Y_{pred}||_2)$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

6. **Huber**

$$\text{HuberLoss} = 0.5 * (Y_{true} - Y_{pred})^2 \text{ if } |Y_{true} - Y_{pred}| <= \delta$$

$$\text{HuberLoss} = 0.5 * \delta + \delta * (|Y_{true} - Y_{pred}| - \delta) \text{ if } |Y_{true} - Y_{pred}| > \delta$$

7. **LogCosh** – Logarithm of hyperbolic cosine

$$\text{LogCosh} = log(\frac{(exp(Y_{pred} - Ytrue) + exp(-Y_{pred} + Ytrue)}{2})$$

Optimizer

The goal during network training is to find the network weights for which the loss function is at its global minimum – minimizing the function, parametrized by the network parameters. Several algorithms and strategies are available to solve this problem and avoid pitfalls like being trapped in local maxima or not converging at all. Many of them are based on Stochastic Gradient Descent – SGD. The relevant possibilities readily offered by Keras are presented. [6], [37], [14]

1. **SGD**

The default function is

$$w = w_{t-1} - \eta * g$$

where η is the learning rate and g is the gradient. The weight in the next step is updated to the value of the previous step adjusted with the learning rate times the gradient. The next weight is therefore located downwards along the gradient, closer to the minimum. The learning rate controls how large the steps are. In standard SGD, it is kept constant, but other approaches vary it over the course of training.

SGD also includes non-zero momentum and Nesterov options, which are not applied here, as the other options provide the same benefits, but also combined with learning rate updates.

Momentum improves behaviour in regions where the gradient in one dimension is a lot larger than in the other. The update term then also contains a fraction of the update term of the previous step. It dampens oscillations and goes in the direction of the steepest gradient faster. If v is velocity and m is momentum:

$$w = w_{t-1} + v_t$$

where

$$v_t = m * v_{t-1} - \eta * g$$

Nesterov accelerated gradient adjusts the momentum approach to also include the gradient at the previous step. It limits the speed of unadjusted momentum to prevent missing the minimum.

$$w = w_{t-1} + m * v_t - \eta * g$$

TU Bibliothek Your knowledge hub WIEN
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

and

$$v_t = m * v_{t-1} - \eta * g_{t-1}$$

2. **Adagrad**

   Adagrad - Adaptive gradient - adjusts the learning rate for each parameter separately in every step, based on how much they had to be changed before.

   $$w_t = w_{t-1} - \frac{\eta}{\sqrt{G_t + \epsilon}} * g$$

   ϵ is a smoothing term to prevent division by 0 and G_t is a matrix containing the sum of squares of the gradients of all previous steps, with regards to the separate parameters. That means that if a parameter is updated often, the learning rate decreases to fine-tune it more slowly. The slowing down is permanent – the denominator will not be able to decrease again and the learning rate can only decrease.

3. **Adadelta**

   In Adadelta, to also allow increasing of the learning rate and to not depend on its initial value, the matrix G_t is replaced by a running average of the last gradients and does not hold all previous values. If the adjustment to the weight in the previous step is expressed as Δx_t, the Adadelta method becomes:

   $$w_t = w_{t-1} + \Delta x_t$$

   where

   $$\Delta x_t = -\frac{RMS[\Delta x]_{t-1}}{RMS[g]_t} g_t$$

   The numerator holds the approximation of Δx_t as the RMS over a certain window of previous adjustments. The denominator is the root mean square of the gradients of all previous steps, defined using the expected value of the squared gradients.

   $$RMS[g]_t = \sqrt{E[g^2]_t + \epsilon}$$

   The learning rate is then not expressed in the update rule.

4. **RMSprop**

   Similarly to Adadelta, RMSprop also uses the moving average of the last gradients to approximate the current one, yielding the same denominator. That way, the learning rate can both increase and decrease.

   $$w_t = w_{t-1} - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

5. **Adam**

   Combining the implementation of moving decaying average with moment estimation, Adam uses both the first and the second moment.

   $$w_t = w_{t-1} - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t$$

   where $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected estimates of the first and second moment respectively. Using bias terms β:

   $$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \text{ and } \hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

6. **Adamax**

   Adam uses the l_2 norm for calculating the v_t term. Adamax replaces that with the l_∞ norm, so its term v_t becomes

   $$v_t = \beta^\infty u_{t-1} + (1 - \beta^\infty)|g_t|^\infty = max(\beta * v_{t-1}, |g|_t)$$

   $$w_t = w_{t-1} - \frac{\eta}{\hat{v}_t + \epsilon}\hat{m}_t$$

   where $\hat{m}$ is again an estimation of the moment.

7. **Nadam**

   Includes the Nesterov-accelerated moment in Adam.

4.1.5 Current approach

Although the possibilities of network architectures are ever increasing, the quality of the training data is important for the application to be successful. In order to systematically explore the influences of different parameters and their contribution to the final, optimal network, the following approach is selected:

One parameter is varied with different levels or possibilities, while the others are kept constant. Then the optimal parameter value is selected, and the next parameter is examined in the same way. The selection is based on the underlying theory of the parameter and its suitability to the problem, as well as the performance in the experiment. The order of examination of parameters is the following:

1. Loss
2. Optimization
3. Layer number and size

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.1.6 Ablation study

Once the optimal architecture is chosen, the influence of different parameters in the training data is checked. This is traditionally done with an ablation study, leaving out the entire dimension of a parameter one at a time, and examining the effect on model performance. Unlike in most cases however, here the input data and its characteristics can be controlled entirely. This allows for a comparison of different levels of the parameters like vibrato and noise, ranging from high, physiologically realistic, to absent.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.2 Data synthesis

4.2.1 Datasets

The data with which the network was trained, was entirely synthesized, as well as one half of the testsets. The other half of the testsets was a synthesized sine wave with harmonics, including noise and vibrato, filtered by a real tube and recorded. The synthesized utterances are the middle of a vowel, assuming a steady-state with an unchanging and not interacting conformation of the sound source and vocal tract. An overview of the used synthesized datasets is given in table 4.2.1. They are designed as follows: Testset 0 has the same structure and levels of breathiness and noise as the training data, but is randomized with a different seed. Testset 1 also has the same parameter levels, but a higher f_o range to examine extrapolation possibilities to higher frequencies. Testsets 2 and 3 have intermediate and absent levels of breathiness respectively, while keeping vibrato amplitude at the standard level. Testsets 4a, 4b and 5 have progressively lower vibrato amplitude, while keeping breathiness at the standard level. This is to examine the influence of the changed parameter separately from all others. The vibrato frequency is set constant to 5.0 Hz throughout this entire work. Finally, testset 6 leaves out both breathiness and vibrato and testset 7 has both at higher levels than in the training data.

Entirely Synthesized Datasets

ID	Number of Sounds	Fo range [Hz]	Noise [%]	Vibrato amplitude [cents]
Train	10 000	65.41 - 1046.50	10	75
0	200	65.41 - 1046.50	10	75
1	200	554.37 - 1567.98	10	75
2	200	65.41 - 1046.50	5	75
3	200	65.41 - 1046.50	0	75
4a	200	65.41 - 1046.50	10	50
4b	200	65.41 - 1046.50	10	25
5	200	65.41 - 1046.50	10	0
6	200	65.41 - 1046.50	0	0
7	200	65.41 - 1046.50	20	100

4.2.2 Synthesis

The synthesis was performed by Bodo Maass using the sound generator of his software VoceVista [32]. The sampling rate is 44100 Hz. The sounds are normalized at different points in the pipeline to avoid clipping. The method is the following:

1 - Sine wave

A standard simple sine wave is synthesized with the desired fundamental frequency and an amplitude of 1.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

2 - Vibrato

Vibrato is added as a periodic modulation of the fundamental frequency. The vibrato frequency is set constant at 5.0 Hz, the amplitude is varied according to the experiment. A realistic vibrato amplitude in human singing is up to a maximum of 100 cents or 1 semitone. In the training set, the amplitude is set to 75 cents, with intermediate levels of 50 cents and 25 cents. For the high level, the vibrato is 100 cents.

3 - Harmonics

Harmonics are then added as multiples of the fundamental frequency at each point in time. They therefore automatically copy the vibrato. The number of harmonics is set to the maximum to still remain under half the sampling rate.

4 - Noise

Noise at the desired percentage α is added. The sound including harmonics is rescaled to an amplitude $1 - \alpha$ and summed up with α noise, so the total amplitude remains 1. The standard noise percentage in the training set is 10%, which is audible with the human ear and clearly visible in the spectrogram. The intermediate level is 5%, which is hard to discern for humans but still visible in the spectrogram. For the high level, the noise is 20%.

In order to express these percent values as signal to noise ratio (SNR), they are converted to RMS. If the peak amplitude of the rescaled sound, all harmonics included, is $1 - \alpha$, the RMS is $\frac{1-\alpha}{\sqrt{2}}$, independently of the number of harmonics. The RMS of the noise is $\frac{\alpha}{\sqrt{2}}$. The SNR is then

$$\mathrm{SNR} = \frac{\frac{1-\alpha}{\sqrt{2}}}{\alpha/\sqrt{2}} = \frac{1-\alpha}{\alpha}$$

For 10% noise, that gives an SNR of 9. For 5% noise, the SNR is 19 and for 20% noise, the SNR is 4.

5 - Filtering

The resonance filtering is applied using an fft-filter. A Fourier transform is performed on the source sounds, the result is multiplied with the vocal tract transfer function and then transformed back with the inverse Fourier transform. The transfer function is constructed by randomizing resonances with the following constraints: f_{R1} falls between 300 Hz and 1000 Hz, f_{R2} between 700 Hz and 2300 Hz, f_{R3} between 1700 Hz and 3000 Hz, f_{R4} between 3000 Hz and 4000 Hz, f_{R5} between 4000 Hz and 5000 Hz, and f_{R6} between 5000 Hz and 6000 Hz. These values are based on [28], with the added rule that the minimal distance between two resonances should be at least 150 Hz. It is a simplified approach to approximate physiological values, but sufficient for the scope of this exploratory study.

6 - Spectrogram

Finally the spectrogram is calculated and a slice comprising exactly one vibrato cycle is cut out. The spectrogram parameters are as follows:

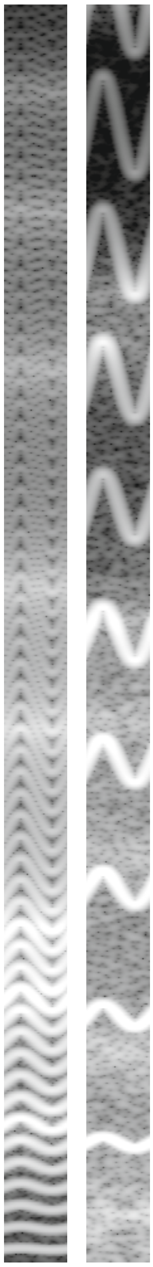
- Sampling rate: 44100 Hz
- FFT Size: 8192 samples
- Window function: Blackman-Nuttal
- Hoplength: 176 frames per step

The chosen spectrogram size is 50 pixels on the time axis and 1000 pixels on the frequency axis. After rescaling the spectrogram to fit these dimensions, the resulting time resolution, for 50 pixels on the time axis, covering a range of 0.2 seconds, is 3680 frames per pixel. The frequency resolution, for 1000 pixels for a frequency range of 50 Hz - 8000 Hz, is 7.95 Hz per pixel. These calculations and spectrograms are created by Bodo Maass. The dynamic range of 88 dB is rescaled into values from 0 to 1. An example of a resulting spectrogram is shown in figure 4.1.

Figure 4.1: *Two examples of spectrograms of sounds from the standard testset, with 10% noise and 75 ct vibrato amplitude.*

a: Example spectrogram for a sound with fundamental frequency 131.27 Hz. The first six resonances are at: F_{R1} *= 916.80 Hz,* F_{R2} *= 1720.42 Hz,* F_{R3} *= 2153.02 Hz,* F_{R4} *= 3419.03 Hz,* F_{R5} *= 4311.65 Hz and* F_{R6} *= 5697.20 Hz.*

b: Example spectrogram for a sound with fundamental frequency 803.12 Hz. The first six resonances are at: F_{R1} *= 351.81 Hz,* F_{R2} *= 1385.88 Hz,* F_{R3} *= 2602.35 Hz,* F_{R4} *= 3459.52 Hz,* F_{R5} *= 4093.05 Hz, and* F_{R6} *= 5864.16 Hz.*



(a) (b)

Figure 4.1

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.3 Lab experiments

In order to examine the influence of recorded input sounds, as will be the case in the final application of the model, experiments in the sound lab were performed. A testset with the same structure as the entirely synthesized one was created without adding resonances. They were played through the driver and filtered by a range of tubes of different lengths and with different resonances. The result was recorded, adjusted for the sound driver transfer function, and pre-processed to spectrograms as input to the neural network model. The resonances were predicted by the model and the same evaluation was performed as with the synthetic sounds, to analyze the model's performance in more realistic recording circumstances.

4.3.1 Materials

The experiments took place in the sound-treated room of IWK (the Department of Music Acoustics Wiener Klangstil) of the University of Music and Performing Arts (mdw) in Vienna, Austria. The following material was used:

- Laptop: Sony Vaio SVP132A1CL
- Sound driver: MONACOR KU-516
- Pre-amplifier: Crown XLS 1002 DRIVECORE
- Interface: Behringer U-PHORIA UMC404HD
- Microphone: Beyerdynamic MM1 Germany nr. 30202
- Plastic tubes set 1: 32 mm diameter, 2 mm wall thickness, lengths 5, 9, 13, 17, 21, 25 cm
- Adapters (accessory to 32 mm tube)
- Plastic tubes set 2: 20 mm diameter, cut to the needed lengths from 13 to 19 cm in steps of 0.5 cm
- 3D printed model of a vocal tract
- Carton
- Play doh
- tesa Masking tape 50mm
- Pattex power tape 50mm

The material is displayed in figure 4.2.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

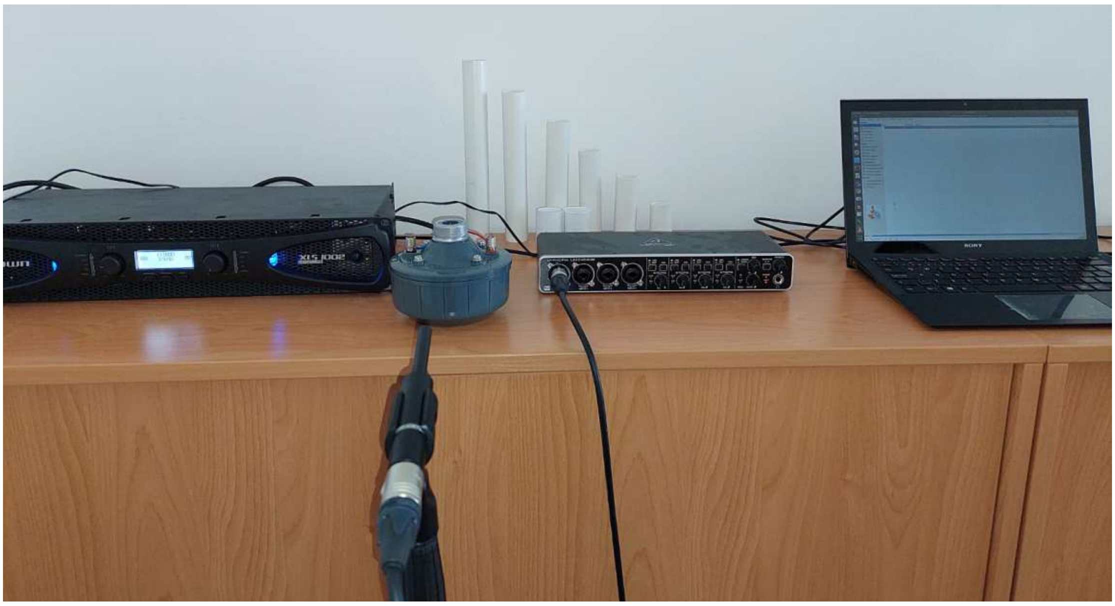
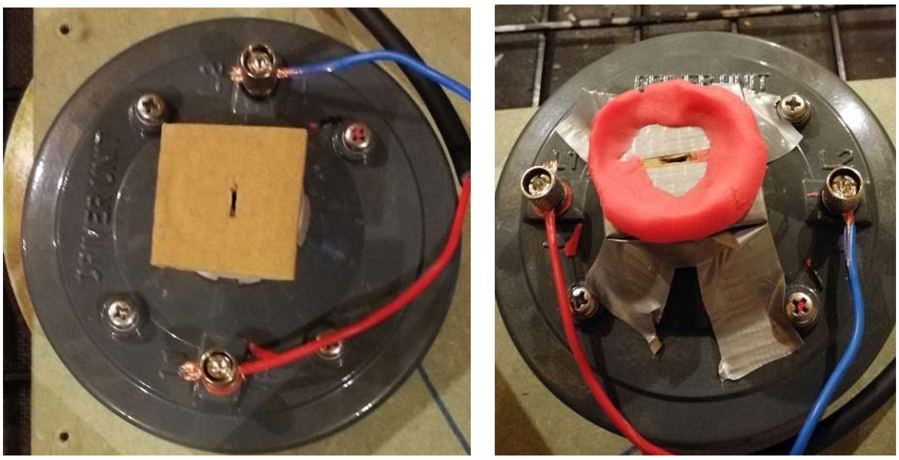


Figure 4.2: *Used playback and recording devices. Front: Microphone. Back: The set of tubes with diameter 32 mm. Middle from left to right: Pre-amplifier, sound driver, microphone interface and laptop.*



(a) *The carton with slit used to close off the sound driver.*

(b) *The carton fastened and insulated with power tape and prepared play doh for insulation of the tube.*

Figure 4.3

Synthesized Datasets used for Recording				
ID	Number of Sounds	Fo range [Hz]	Noise [%]	Vibrato amplitude [cents]
0	42	98.0 - 1046.5	10	75
1	20	523.3 - 1568.0	10	75
2	42	98.0 - 1046.5	5	75
3	42	98.0 - 1046.5	0	75
4a	42	98.0 - 1046.5	10	50
4b	42	98.0 - 1046.5	10	25
5	42	98.0 - 1046.5	10	0
6	42	98.0 - 1046.5	0	0
7	42	98.0 - 1046.5	20	100

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

4.3.2 Methods

The sounds were generated using VoceVista to obtain testsets 0 – 7 as described in table 4.3.1. The frequencies corresponded to pitches G2 (98 Hz) to C6 (1046.5 Hz) in semitone steps for sets 0 and 2 - 7. Set 1 had a different range; C5 (523.3 Hz) to G6 (1568.0 Hz). The 42 (20 for set 1) semitones per set were played for 1.5 s each and concatenated to one, 9:53 minute long soundfile.

They were played using Windows Media Player with laptop loudness set to 69 and the pre-amp to 3 p.m.. The driver was placed in the sound-treated room and its flange was covered with carton and power tape to leave a 1.5 x 5 mm slit, shown in figure 4.3. The tube was mounted and isolated with play doh. The microphone was placed 30 cm away from the sound source with gain set to 12 p.m. Figure 4.4 shows a 32 mm wide tube with an adapter taped to it.

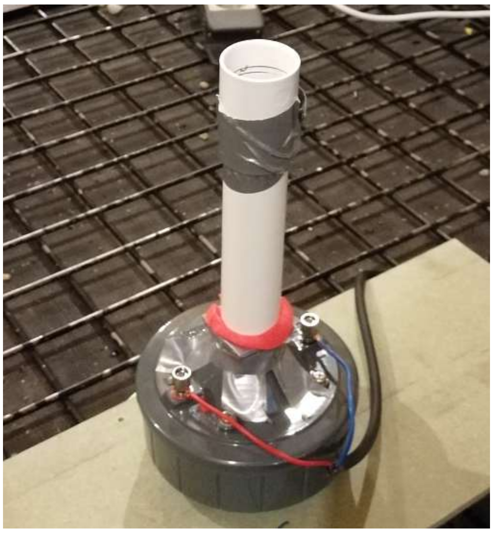


Figure 4.4: *The tube placed on top of the driver and carton with the slit, isolated with play doh.*

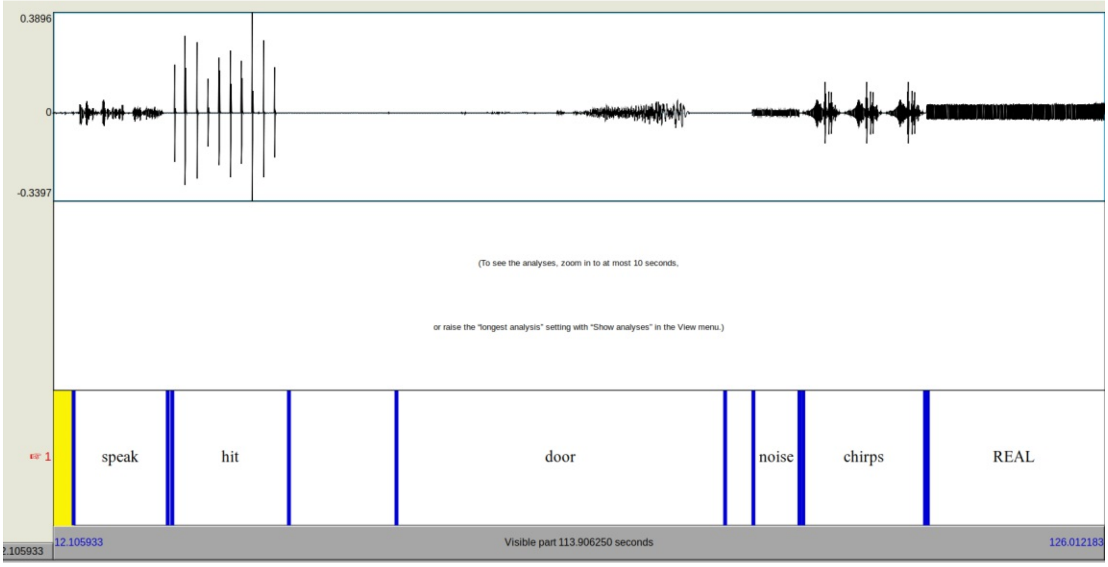
The background noise and a calibrating 1000 Hz sine wave were played and recorded for reference. The transfer function of the set-up was recorded using white noise and 4 second chirps. For each tube length, an approximately 12 minute recording was obtained, containing the following parts:

1. spoken date, time and tube length
2. hitting the tube several times in front of the microphone, closing off one end entirely with the palm of the hand
3. closing the door of the sound treated room
4. 5s of white noise and 3 4s chirps

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

5. the full soundfile with the testdata
6. 5s of white noise and 3 4s chirps
7. opening the door of the room
8. spoken date, time and tube length
9. hitting the tube again

An example of the first 2.5 minutes of an annotated recorded soundfile is shown in figure 4.5.


(To see the analyses, zoom in to at most 10 seconds,

or raise the "longest analysis" setting with "Show analyses" in the View menu.)

Figure 4.5: *An annotated soundfile, showing the experiment protocol. First speaking of the time and tube length, then hitting of the tube. Then the room is left and the door is closed, so the noise, chirps and test sounds can be played. Screenshot from Praat [3].*

The full experiment was recorded with both 32 mm and 20 mm diameter tubes, but only the 20 mm data was post-processed and used for analysis and further testing. A slimmer tube has narrower and clearer resonance peaks, resulting in a higher data quality for the current purposes.

Finally, one recording was made with a 3D printed vocal tract, provided by Christian Herbst. It was printed at 3dee.at using powder, to avoid fragments of support structures remaining in the hollow model, as they would significantly alter the acoustic characteristics of the model. The model was supported by a thin tube standing on carton, to avoid structures in the vicinity of the model and the microphone that could act as acoustic reflectors. Like the tubes, it was set and insulated on the slit using play doh. The model is pictured in figure 4.6 and the complete set-up in figure 4.7.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

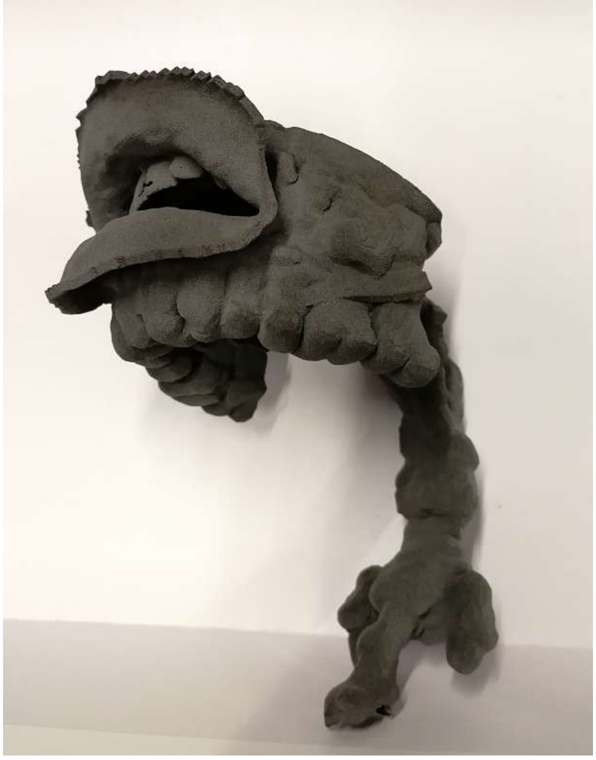


Figure 4.6: *The 3D-printed vocal tract model, used in the final phase of the lab experiments.*

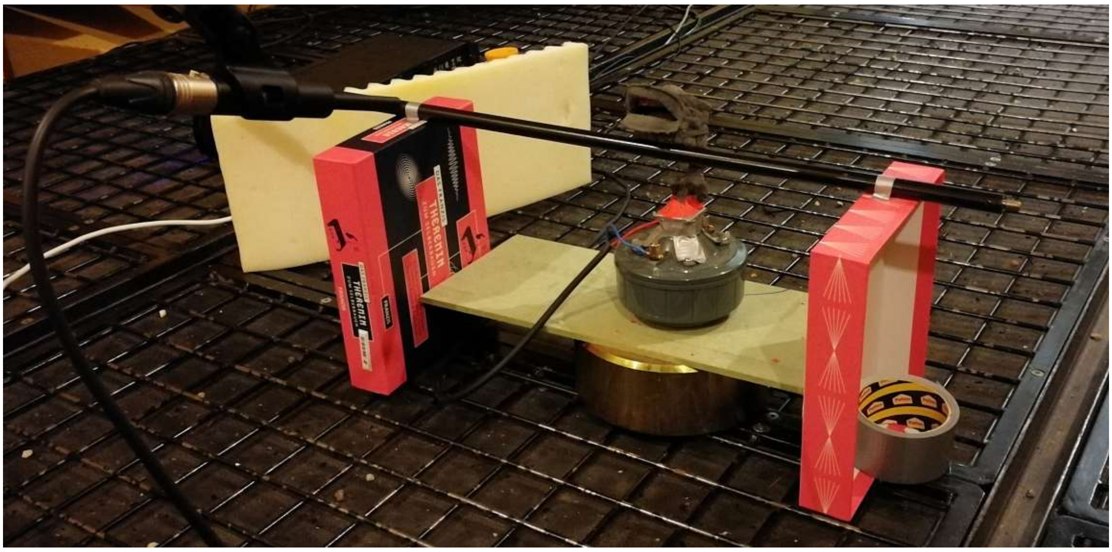


Figure 4.7: *The set-up of the vocal tract model on top of the sound driver. The model was supported by a rod and placed at a distance of 30 cm from the microphone.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

4.3.3 Analysis

All recorded files were annotated using Praat [3], [4]. The result of this annotation is shown in figure 4.5.

The transfer function of the system including the sound driver, room and microphone was extracted from separate noise and chirps, recorded without a tube. All recorded sounds were filtered with the inverse of this transfer function to avoid its influence. The sounds were cut into separate files of 0.5 seconds length per pitch, organized per test set, and converted to spectrograms as input for the network. The spectrogram parameters are described in section 4.2.2. The ground truth was extracted from the noise and chirps played through the tubes, also after adjusting for the transfer function of the sound driver and room.

4.4 Evaluation

The performance of the network was assessed using several error metrics. Although the final application target are recorded vowels sung by humans, testing on them is impossible because in that case the ground truth is not known, and there is no sufficiently accurate standard method to compare to. The network was tested on both synthesized data and on recorded data where the ground truth could be exactly determined.

The exact structure of the test sets is shown in 4.2.1 and 4.3.1.

Because the ground truth values are know exactly, the error metrics are based on [44], where Y_{pred} is the predicted resonance frequency and Y_{true} is the true resonance frequency.

$$\text{Absolute Error} = |Y_{pred} - Y_{true}|$$

$$\text{Directional Percent Error} = 100 * \frac{Y_{pred} - Y_{true}}{Y_{true}}$$

$$\text{Absolute Percent Error} = 100 * \frac{|Y_{pred} - Y_{true}|}{Y_{true}}$$

The absolute error should be as low as possible. Directional percent error shows whether there is systematic over- or underestimation. Finally, absolute percent error relates the error to the frequency of the resonance itself, and not of the sample sound. For higher resonances, this metric would then be less strict than for lower, which is not the goal in this application. However, comparing absolute error and absolute percent error uncovers favoring of lower over higher resonances, or vice-versa.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The 6 resonances of the same test set were also estimated using Praat, using the Burg's LPC method. [3] A formant ceiling of 6050 Hz was chosen because the upper limit for the sixth resonance is 6000 Hz, and the Praat algorithm removes all formants higher than 50 Hz below the formant ceiling [4]. The performances of both were compared, also in relation to the fundamental frequency of the sound. In order for the current method to be successful, it should outperform Praat in the frequency range 300 Hz - 1000 Hz, and for all 6 considered resonances.

4.5 Implementation

The data processing pipeline was written in Python 3.9.6 with the use of the library librosa 0.8.1 for audio preprocessing. Keras 2.6.0, a widely used machine learning library with TensorFlow 2.4.1 as a backend, was used for the training and application of the neural network models. Keras was run on NVIDIA-SMI 440.33.01 GPU with CUDA Version 10.2. The data visualisation was performed with Matplotlib 3.4.2.

Other used software is Praat 6.0.46 and VoceVista Video Pro 5.4.1. The entire process was run on Ubuntu 18.04.

The entire list of dependencies of the Conda environment is shown in table 2 in the appendix.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 5

Results and Discussion

In this chapter, the individual steps used to obtain the final model are described in sections 6.1 - 6.3. In section 6.4, the performance of the model on different datasets is described, and the influence of the different parameters is analyzed. Also the comparison to the baseline method is handled in section 6.4. Finally section 6.5 considers the recorded datasets with both the final model and a model with the same structure, trained on recorded data.

5.1 Loss function

Models were trained with the following parameters for each of the losses: 1000 epochs, Optimizer SGD, 2 convolutional layers with respectively 32 and 64 nodes and 3 dense layers with respectively 128, 64, and 64 nodes. This network size was chosen with wider and smaller layers, to be a standard, middle-sized model. The network architecture in this case is not as important, as the goal is to compare only the influence of the considered hyperparameter.

The loss, validation loss, RMSE and validation RMSE of each of these 7 models were recorded during training. Assuming that the model is stabilized in the last 10 epochs, the mean of the metrics in those last 10 epochs is calculated to account for small differences. In table 5.1, these values are shown.

As the formula for the loss is different for each model, they cannot be compared. Therefore, only the RMSE and validation RMSE are pictured in figure 5.1. The model trained to minimise MAPE returned no loss during training and NaN as predictions, so it is not represented in these figures and table. The performance of MSLE and CosSim are a lot worse than MSE, MAE, Huber, and LogCosh, so only these last four are shown with an appropriate zoom.

Loss	**Loss**	**Val Loss**	**Δ Loss**	**RMSE**	**Val RMSE**	**Δ RMSE**
MSE	1.148e-05	1.699e-05	5.514e-06	3.388e-03	4.122e-03	7.339e-04
MAE	1.425e-03	2.664e-03	1.239e-03	1.632e-03	3.440e-03	1.808e-03
Huber	8.772e-06	1.027e-05	1.502e-06	4.188e-03	4.533e-03	3.444e-04
LogCosh	9.765e-06	1.140e-05	1.638e-06	4.420e-03	4.776e-03	3.560e-04
MSLE	1.009e-01	1.010e-01	1.437e-04	5.349e-01	5.349e-01	-5.168e-05
CosSim	-1.000e+00	-1.000e+00	1.248e-05	8.787e-02	8.695e-02	-9.238e-04

Table 5.1: *The results of loss, validation loss, RMSE and validation RMSE of after training. The shown values are the mean of epochs 991 to 1000.*

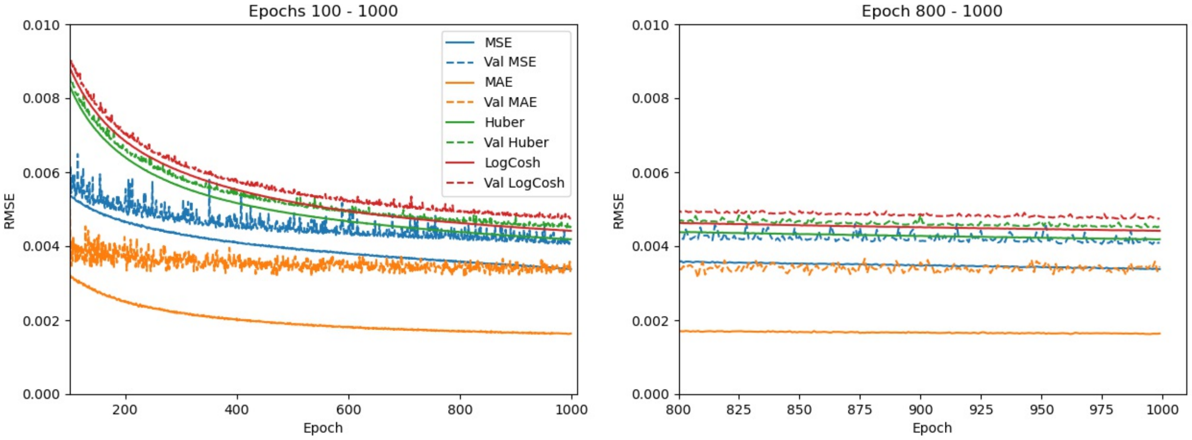


Figure 5.1: *The RMSE during training for the four best performing loss functions. The network trained with loss function MAE is almost stabilized after 1000 epochs of training, whereas MSE, Huber and LogCosh still show a slight decline. MAE shows strong overfitting, as the validation RMSE is double of the training RMSE.*

In figure 5.2, the overall error of all models is shown on a testset of 200 sounds, generated with the same settings as the training set, but a different random seed. The same for the best 4 losses separately is shown in figure 5.3. In figures 5.4 and 5.5, the error related to the f_o of the sound is pictured for all 6 and the best 4 models. The formula for absolute error is $|Y_{pred} - Y_{true}|$, for the directional percent error $100\% * \frac{Y_{pred}-Y_{true}}{Y_{true}}$ and for the absolute percent error $100\% * \frac{|Y_{pred}-Y_{true}|}{Y_{true}}$. The percent error is relative to the frequency of the resonance itself, not the fundamental frequency of the sound.

MAE overfits strongly, as seen by the big difference between training and validation loss and RMSE in figure 5.1 and table 5.1. This difference should be as small as possible. The large overfitting in MAE is a reason to not continue with it. Overfitting is less of an issue in the current problem, as well-constructed training data should be able to cover the parameter space that can be encountered in real-life applications. However, a

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

large difference between the training and validation RMSE from the same dataset is not desirable. Of the others, MSE is most stabilized after 1000 epochs training and performs best in figure 5.3. Its validation RMSE is lower that of the two other, it is most stabilized and the overfitting, although more than the others, is limited. For those reasons, the next experiments continue with this loss function.

Its compatibility with the data can be explained by the random factor in generating the data. Although there are constraints on the intervals in which resonances can be found, the data is sufficiently Gaussian.

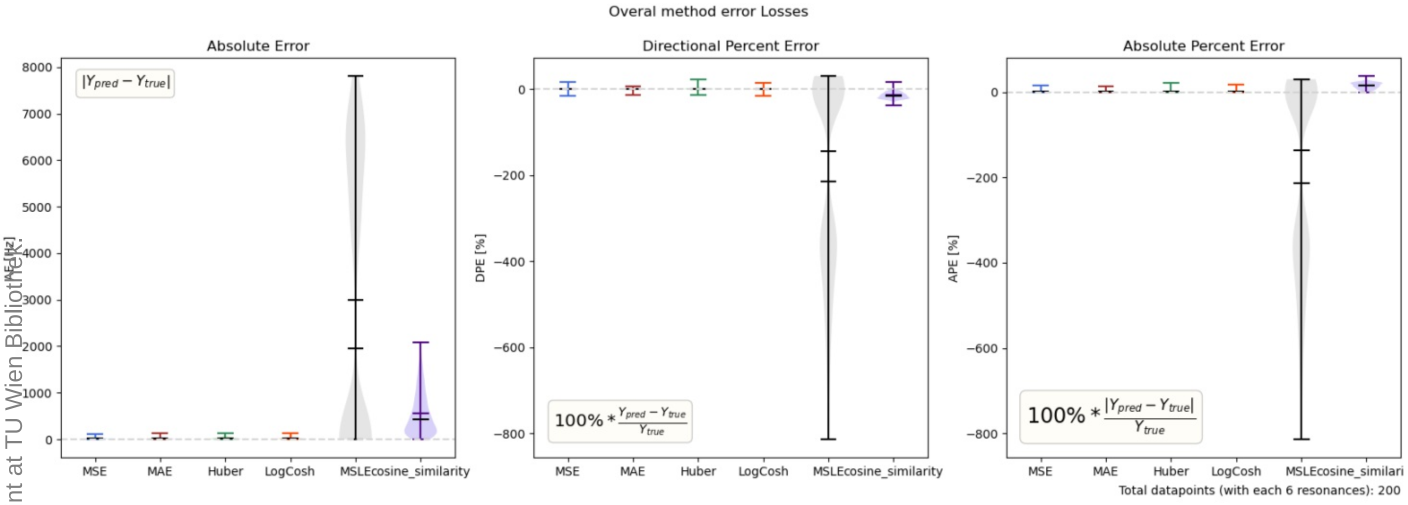


Figure 5.2: *Violinplots of the performance of networks trained with different loss functions on predicting the resonances in a testset of 200 sounds. As seen before from the training data, the performance of MSLE and cosine_similarity is significantly worse than that of the other four loss functions. Violinplots show the distribution of the data in the thickness of the violin. The black horizontal line marker is the median value, the coloured one is the mean value of the data.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

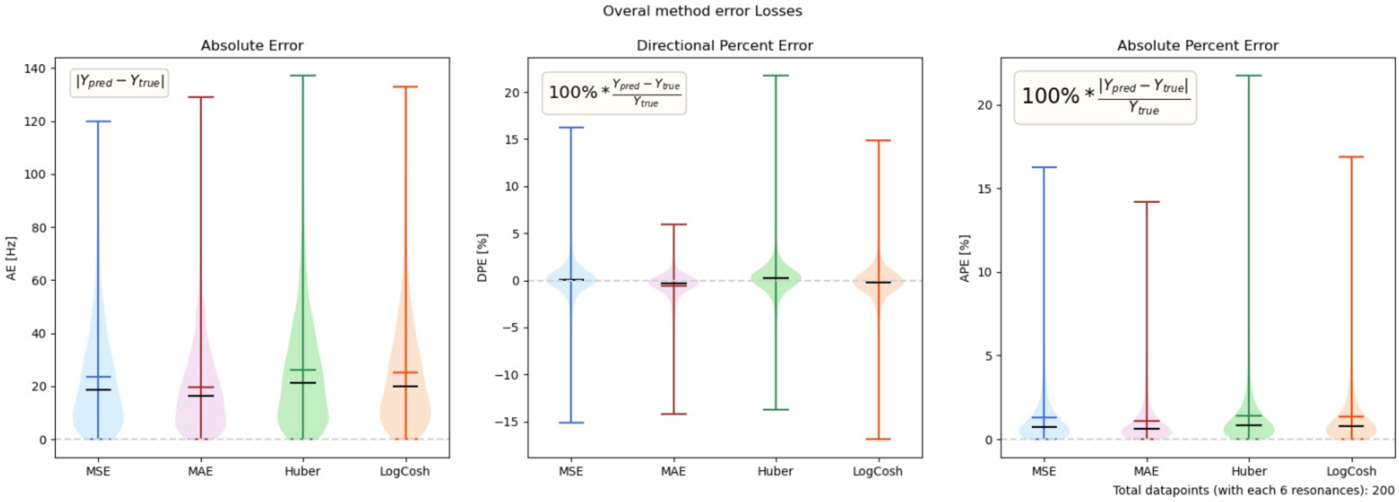


Figure 5.3: *Violinplots of the performance of networks trained with the four best performing loss functions on the testset. The performance is comparable with the mean and median of the absolute error all lying around 20 Hz. The performance of MAE is best, as expected from the RMSE during training, but at the price of strong overfitting. The second best is MSE.*

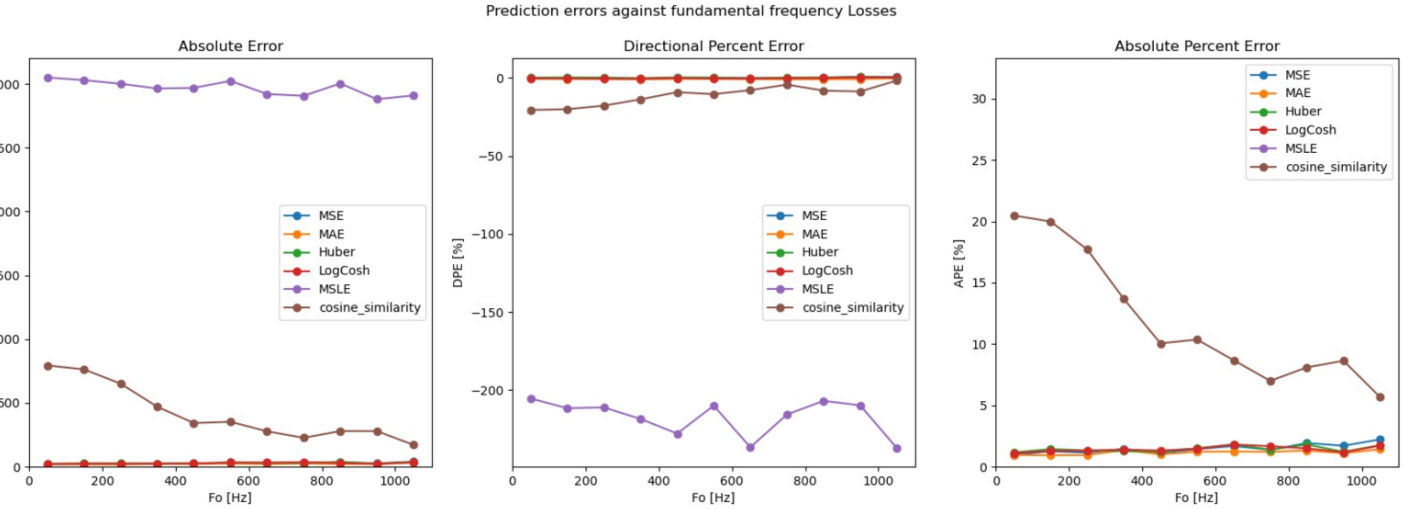


Figure 5.4: *The performance on the testset of all networks visualized with regard to the fundamental frequency of the sound. MSLE and cosine_similarity again perform worse than the others, causing them to be indiscernible in this graph.*

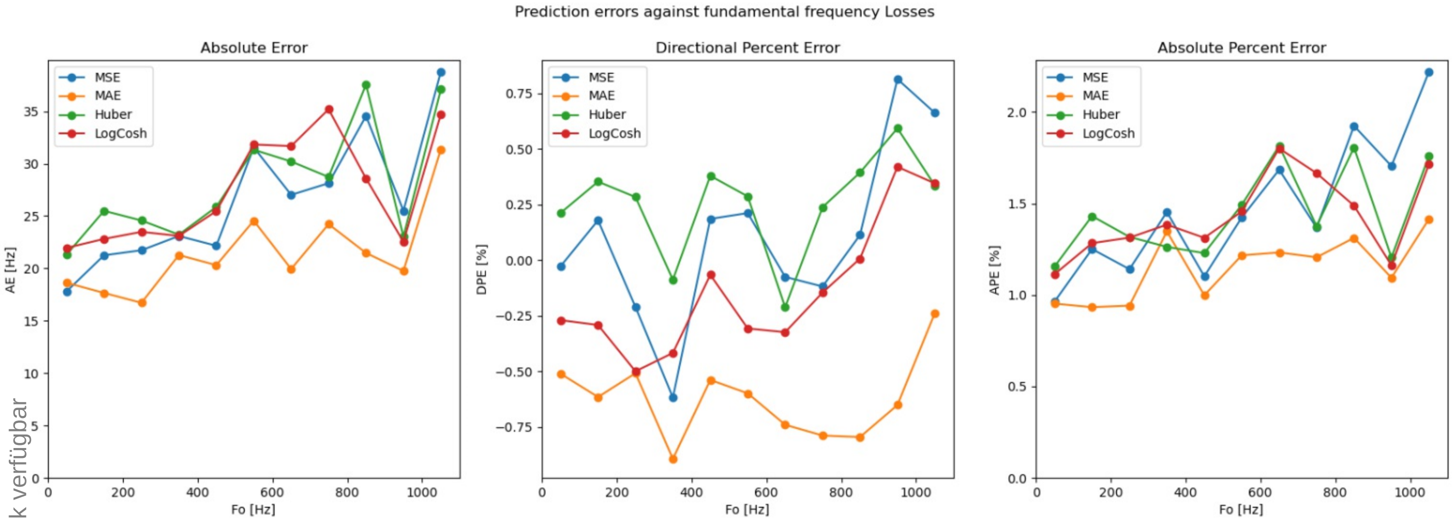


Figure 5.5: *The performance on the testset of the best four networks visualized with regard to the fundamental frequency of the sound. The performance is again comparable, with MAE slightly lower errors than the others.*

Optimizer	**Loss**	**Val Loss**	**Δ Loss**	**RMSE**	**Val RMSE**	**Δ RMSE**
SGD	1.012e-05	1.585e-05	5.725e-06	3.181e-03	3.978e-03	7.966e-04
Adagrad	1.538e-05	2.184e-05	6.469e-06	3.921e-03	4.672e-03	7.512e-04
Adam	7.958e-07	1.966e-05	1.887e-05	8.908e-04	4.434e-03	3.543e-03
Adamax	3.292e-07	2.351e-05	2.319e-05	5.735e-04	4.849e-03	4.276e-03
Nadam	7.584e-07	2.739e-05	2.663e-05	8.638e-04	5.233e-03	4.370e-03
Adadelta	2.835e-05	3.411e-05	5.760e-06	5.325e-03	5.841e-03	5.158e-04
RMSprop	5.233e-06	2.691e-05	2.167e-05	2.287e-03	5.170e-03	2.883e-03

Table 5.2: *Optimizer results: The loss, validation loss, RMSE and validation RMSE in the last ten epochs of training. The shown values are the mean over the last 10 epochs of training.*

5.2 Optimizer

The same network architecture as in the previous experiment was used to train models with loss MSE, and each optimizer. The loss and validation loss during training are recorded again. As the loss function is the same for each model, they can be compared and RMSE is not necessary to consider.

The models were trained with 1000 epochs each and each of them stabilized. In figure 5.6, the loss (MSE) during training is shown for all optimizers. On the left, epochs 100 - 1000 are shown and on the right only epochs 800 - 1000. In table 5.2, the values are

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

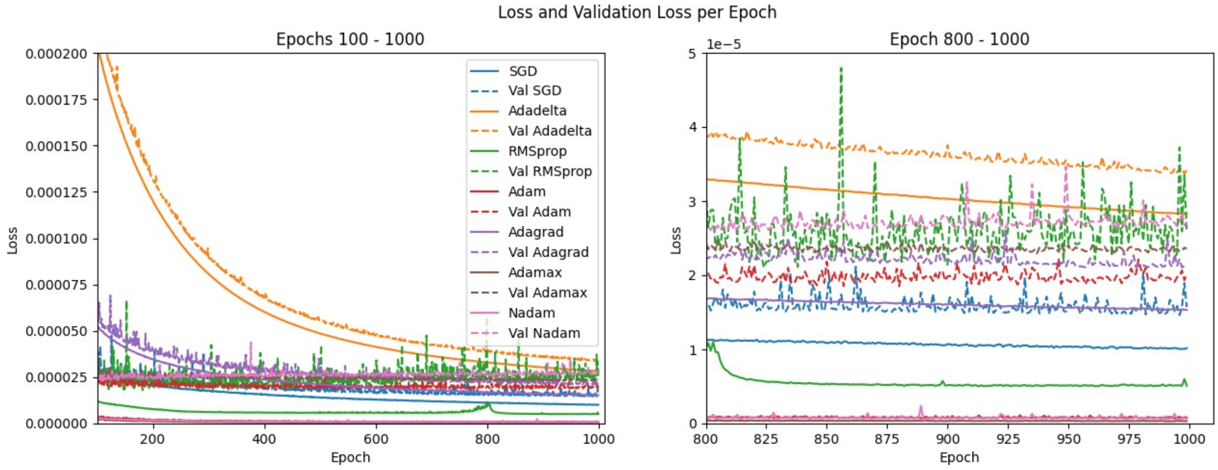


Figure 5.6: *The loss during training of the networks trained with different optimizers. Adam and RMSprop have very low training loss, but the validation loss is a lot higher. The network with Adadelta has not entirely stabilized in 1000 epoch of training. The lowest loss with a reasonable distance between training and validation data is using SGD.*

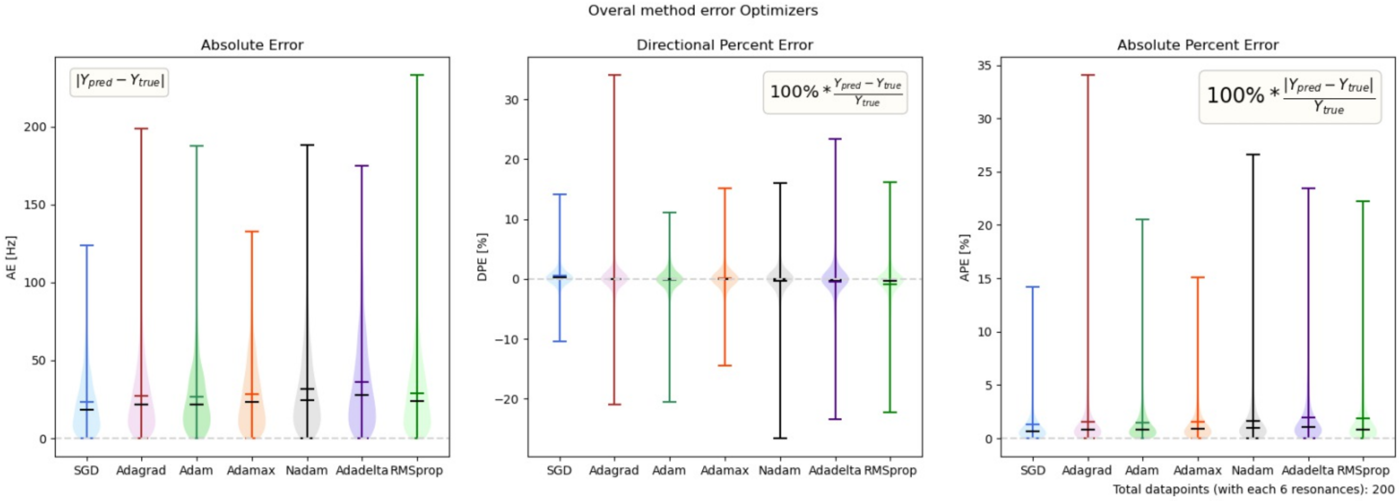


Figure 5.7: *The overall errors on predicting the testsets of the networks trained with different optimizers. The mean errors are comparable. SGD has the lowest AE and the unstabilized Adadelta the highest.*

given, again as the mean of the last 10 epochs. Most networks stabilized after 1000 epochs, except the one using Adadelta. For the others, the training and validation losses differ notably. Although Adam, Adamax and Nadam have very low training losses, their validation losses are higher than the lowest validation loss, SGD. The difference between the training and validation loss is biggest for these three, and smallest for SGD and

Adadelta. As Adadelta is not fully stabilized after 1000 epochs, and SGD has half the training and validation loss, SGD is preferred after this step.

The performance on the test set is generally good, with overall mean errors ranging from 20 Hz to 60 Hz. SGD has the lowest error here as well, as well as the smallest range, as seen in figure 5.7. Also in this regard, the performance of SGD is the best. The next steps will therefore continue with loss MSE and optimizer SGD.

5.3 Network architecture

5.3.1 Large network

The first step in searching the optimal network architecture was to find the performance of a large network with many parameters and long training. Then the network would be simplified while maintaining performance, as long as that would be possible. A network with 4 convolutional layers and 5 dense layers with 64 nodes each was trained with 2500 epochs.

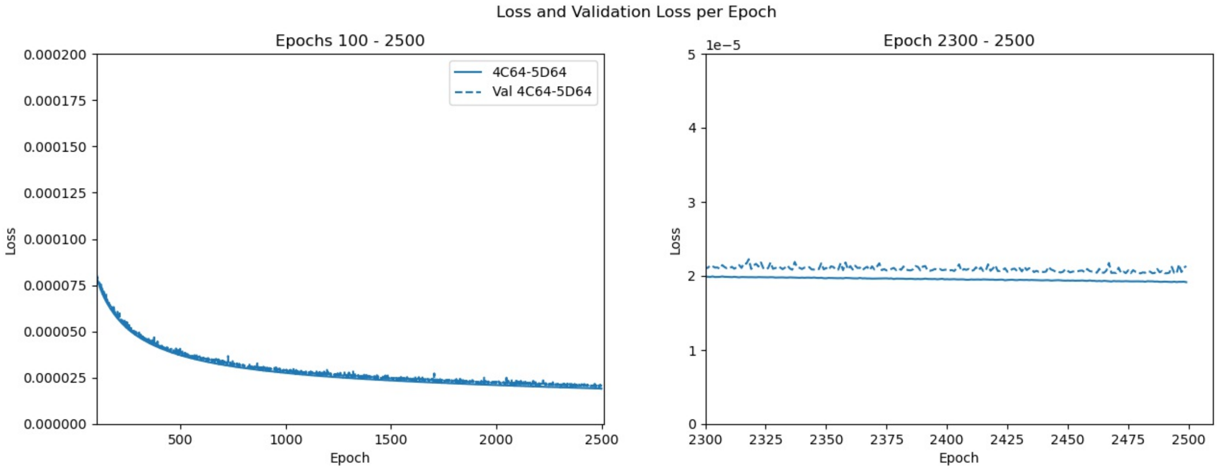


Figure 5.8: *The loss during training of a network with 4 convolutional layers with max-pooling and 5 dense layers. Each layer has 64 nodes. In the last 200 epochs of training, the training loss decreased from* $1.98857 * 10^{-5}$ *(mean of epochs 2300 - 2309) to* $1.92012 * 10^{-5}$ *(mean of epochs 2491-2500) and the validation loss from* $2.11837 * 10^{-5}$ *to* $2.07640 * 10^{-5}$. *This decrease indicates that further training could decrease loss even more.*

The network did not stabilise entirely after 2500 epochs, as visible in figure 5.8. Its mean training loss in the last 10 epochs was $1.92 * 10^{-5}$ and the mean validation loss $2.11 * 10^{-5}$, so the difference is $.56 * 10^{-6}$. The overall errors on the testset are AE 27.4895 Hz, DPE 0.3084% and APE 1.6314%. These results are the goal to also be achieved with a simpler network.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Convolutional layers	1C		4C	
Number of nodes	16	64	16	64
Dense layers	4D64 5D64	4D16 5D16	1D64 5D64	1D16 5D16

Table 5.3: *The trained networks to explore the limits of the parameter space and the influence of the parameters. Less nodes in convolutional layers were combined with more in the dense layers and vice-versa. All networks were trained with 1000 epochs. 4D64 means 4 dense layers with 64 nodes each.*

5.3.2 Extremes of the parameter space

The considered parameter space encompasses 1 to 4 convolutional layers, 1 to 5 dense layers and 16, 32 or 64 nodes per layer. To consider the extremes of the parameter space, the networks in table 5.3 were trained with 1000 epochs each. The code for the architecture is for example in 4C16-1D64: 4 convolutional layers with 16 nodes each, and 1 dense layer with 64 nodes. The number of nodes for all layers of the same type is kept equal to limit the parameter space.

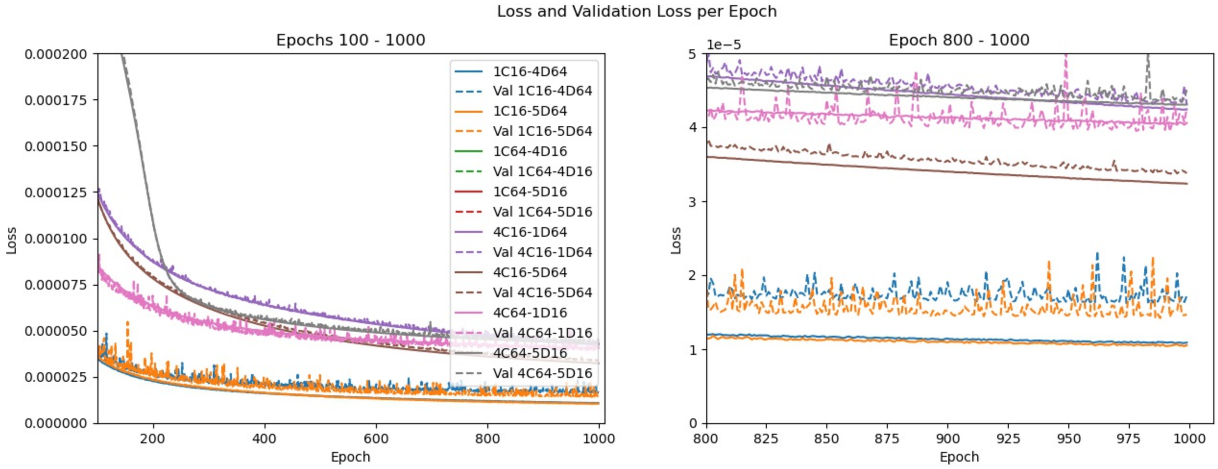


Figure 5.9: *The loss and validation loss during training of the networks from table 5.3. The smallest networks have the lowest loss, the networks with 1 large convolutional layer and respectively 4 and 5 small dense layers have higher losses than the depicted range.*

1C64-4D16 and 1C64-5D16 are not even visible on figure 5.9, with loss during training values of $3.3 * 10^{-4}$ and $3.2 * 10^{-4}$ respectively at the end of training. They also have correspondingly larger overall errors on the testset in figure 5.10. Excluding these two, the visible trends in figure 5.10 are:

- More dense layers mean a lower AE, except in 1C64-4D16 versus 1C64-5D16.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

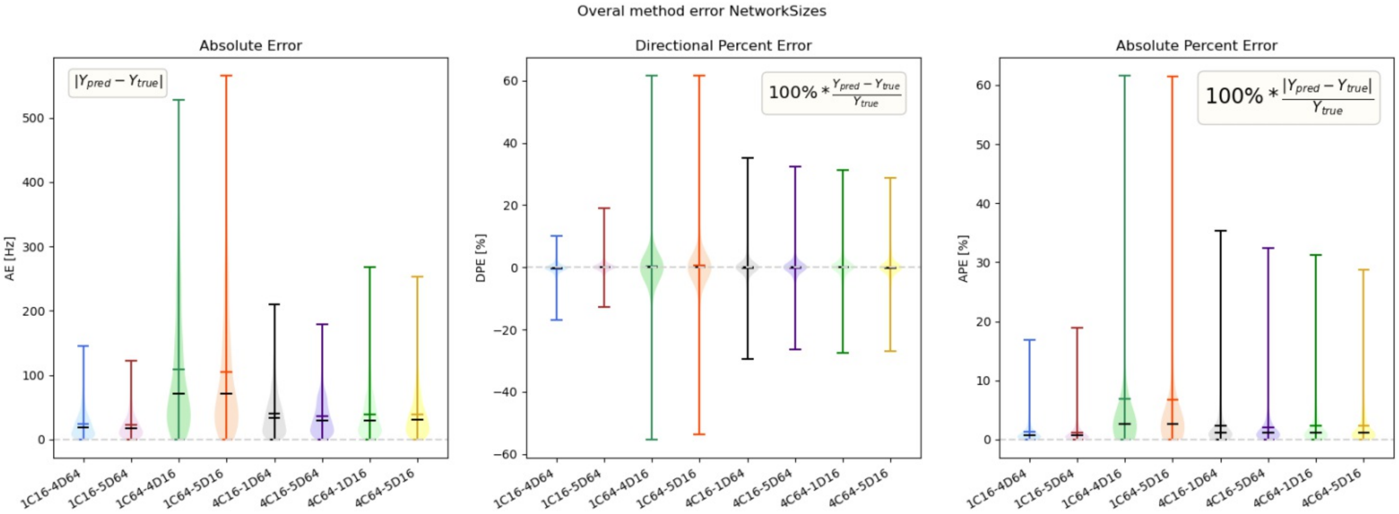


Figure 5.10: *The absolute error, directional percent error and absolute percent error for predicting the testset, for the networks defined in table 5.3.*

That could by caused by the incomplete training, which is even more pronounced in a slightly larger network.

- More convolutional layers mean a higher AE, probably due to incomplete training
- Larger convolutional layers and smaller dense layers cause higher AE. A possible reason is that large convolutional layers need more training, and their shortcomings cannot be compensated for by the smaller dense layers
- The directional error is symmetric around zero, indicating no trend of over- or underestimation

The simplest networks go below the pre-set goal in section 1, although with a larger validation loss. It may therefore be reasonable to even surpass that large network, if the validation loss can come closer too. For the larger networks, 1000 epochs of training do not suffice.

5.3.3 Sampling the parameter space

In the next step, a sampling of the parameter space is done with intermediate numbers of convolutional layers, 2 and 3, and numbers of dense layers ranging from 2 to 5. Again, small convolutional layers are combined with large dense layers and vice-versa.

Because of the large number of networks, the loss and validation loss during training is shown in figures 5.11 and 5.12 for the eight networks with 2 and 3 convolutional layers, respectively. In figure 5.11, the effect of the size of the layers is clearly visible. The four with small convolutional layers and large dense layers perform a lot better than the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Convolutional layers	2C		3C	
Number of nodes	16	64	16	64
Dense layers	2D64	2D16	2D64	2D16
	3D64	3D16	3D64	3D16
	4D64	4D16	4D64	4D16
	5D64	5D16	5D64	5D16

Table 5.4: *The trained networks to explore the influence of the parameters in the center of the parameter space. Less nodes in convolutional layers were combined with more in the dense layers and vice-versa. All networks were trained with 1000 epochs. 4D64 means 4 dense layers with 64 nodes each.*

others, and in between them, the effect of the number of dense layers is limited. In the other four, however, the differences are substantial. 2C64-2D16 and 2C64-5D16 approach the others by the end of training, thanks to a sudden drop in loss around epochs 200 and 400 respectively. The other two, with 3 and 4 dense layers, do not show these drops and keep a higher loss throughout. Interestingly, no unidirectional influence of the number of dense layers is deducible from these results.

In the networks with 3 convolutional layers in figure 5.11, only two layers have larger losses than the others: 3C64-4D16 and 3C64-5D16. The former also shows sudden drops, and the latter, the largest network from this set, may have needed longer training. The better performing group with 3 convolutional layers performs generally worse than the one with 2, with losses at the end of training around $2.5 * 10^{-5}$ to $3 * 10^{-5}$ versus around $2 * 10^{-5}$. The difference between loss and validation loss is smaller for the networks with 3 convolutional layers than those with 2, but overall sufficiently small. Finally, some networks with 3 convolutional layers may benefit from longer training, as they show a slight slope over the last 200 epochs of training.

In figures 5.13 and 5.14, the networks with larger dense layers show consistently lower errors than those with larger convolutional layers. The influence of the number of dense layers is not clear from these results.

5.3.4 Longer training

In order to successfully implement more convolutional layers, more training epochs should be provided to ensure stabilization. Larger and more convolutional layers are included and the networks are trained with 1500 epochs. The following 4 larger networks are trained: 3C32-4D64, 3C32-5D64, 4C32-4D64 and 4C32-5D64.

The loss and validation loss during training of the network with 3 convolutional layers and respectively 4 and 5 dense layers, is very similar. For the networks with 4 convolutional layers, adding a fifth dense layer even worsens performance. For that reason, 4 dense

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

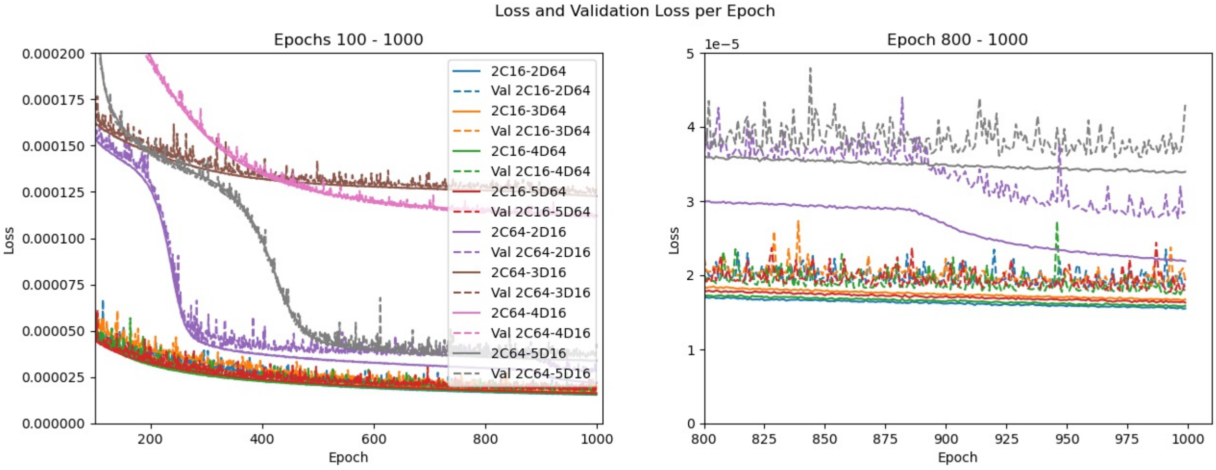


Figure 5.11: *The loss and validation loss during training of the networks with 2 convolutional layers, as defined in the left column of table 5.4.*

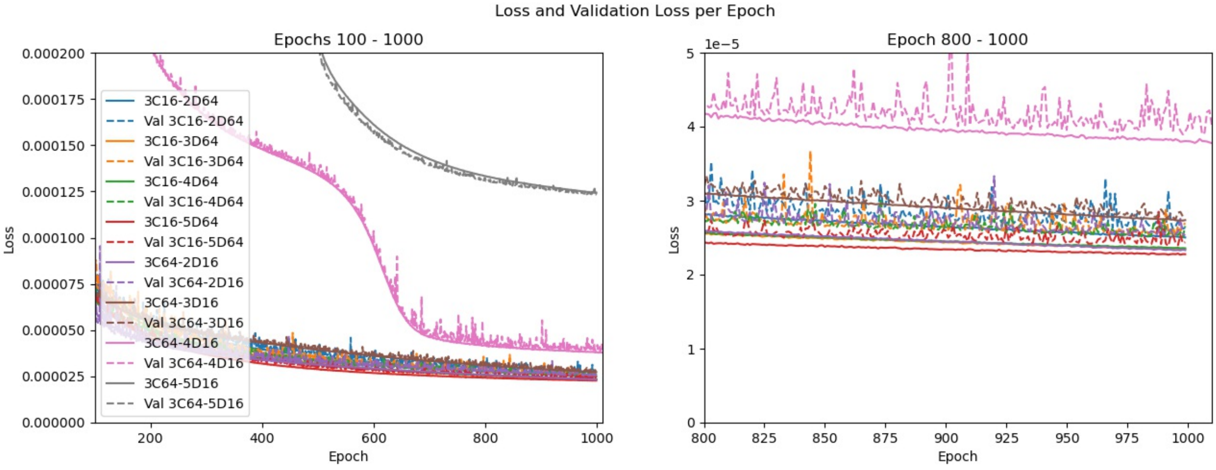


Figure 5.12: *The loss and validation loss during training of the networks with 3 convolutional layers, as defined in the right column of table 5.4.*

layers are kept to continue with, as well as 3 convolutional layers. This combination equals the performance of the first large layer with a mean AE of 27.3344 Hz.

5.3.5 Final model

In order to fine-tune the ideal number of nodes for both the convolutional and dense layers, 8 possible combinations are compared. The considered numbers of nodes are 32 and 64 for the convolutional layers and 16, 32, and 64 for the dense layers. In figure 5.17 the losses and validation losses during training are shown. All networks stabilize after

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

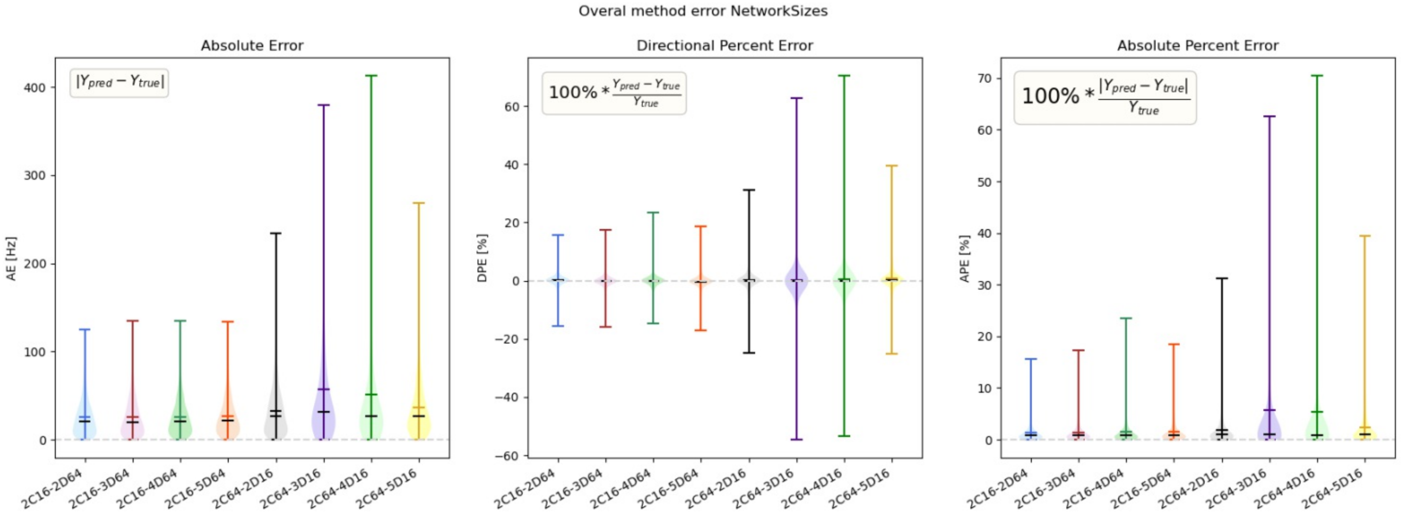


Figure 5.13: *The absolute error, directional percent error and absolute percent error for predicting the testset for the networks with 2 convolutional layers, as defined in the left column of table 5.4.*

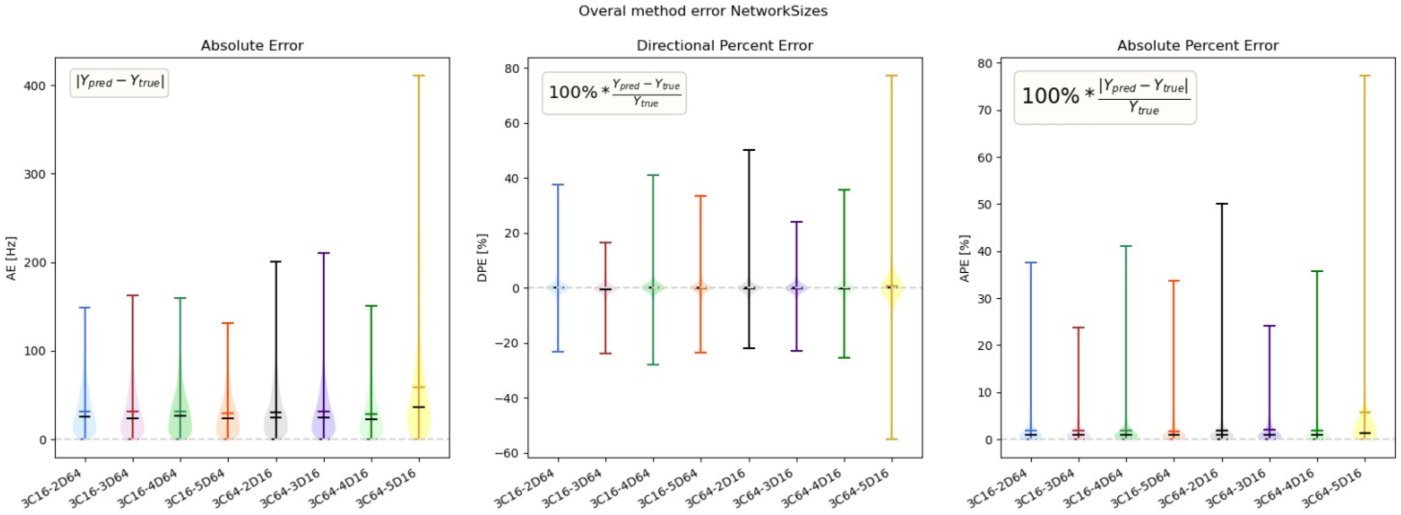


Figure 5.14: *The absolute error, directional percent error and absolute percent error for predicting the testset for the networks with 3 convolutional layers, as defined in the right column of table 5.4.*

1500 epochs and larger layers do clearly have lower losses. The loss of the largest of the considered networks, 3C64-4D64, even surpasses the loss of the first large network in step 1. The network shows a smooth training and early stabilization, a small difference between loss and validation loss, as well as the lowest overall errors in figure 5.18. Its values of AE 23.6235 Hz, DPE 0.0248% and APE 1.2163% are better than the expected

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

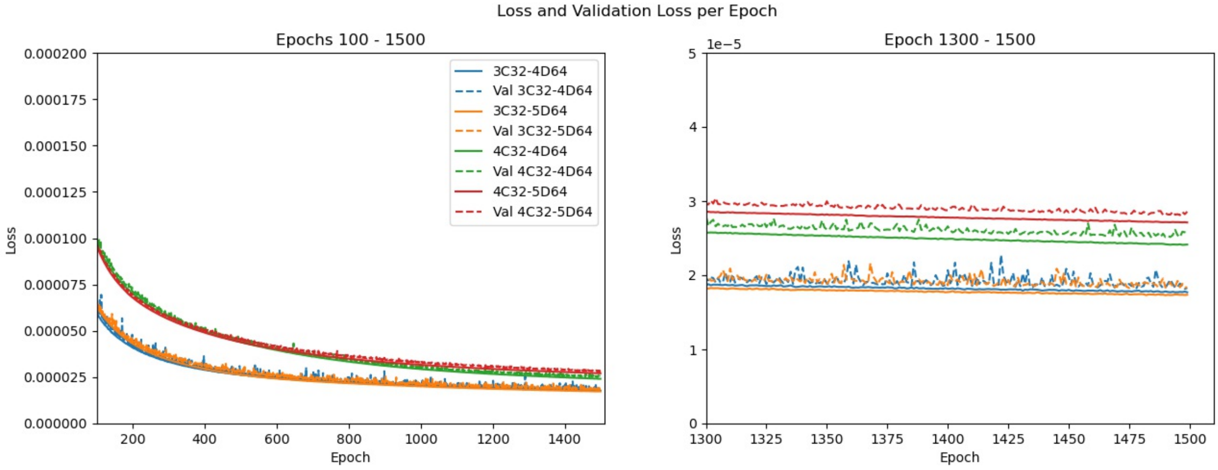


Figure 5.15: *The loss and validation loss during training for networks with 3 or 4 convolutional layers with 32 nodes each and 4 or 5 dense layers with 64 nodes each, trained with 1500 epochs. The networks with 3 convolutional layers are fully stabilized at the end of training and have a loss and validation loss corresponding to the target values, with a sufficiently small difference. The networks with 4 convolutional layers are not yet stabilized after 1500 epochs of training and have a higher loss and validation loss.*

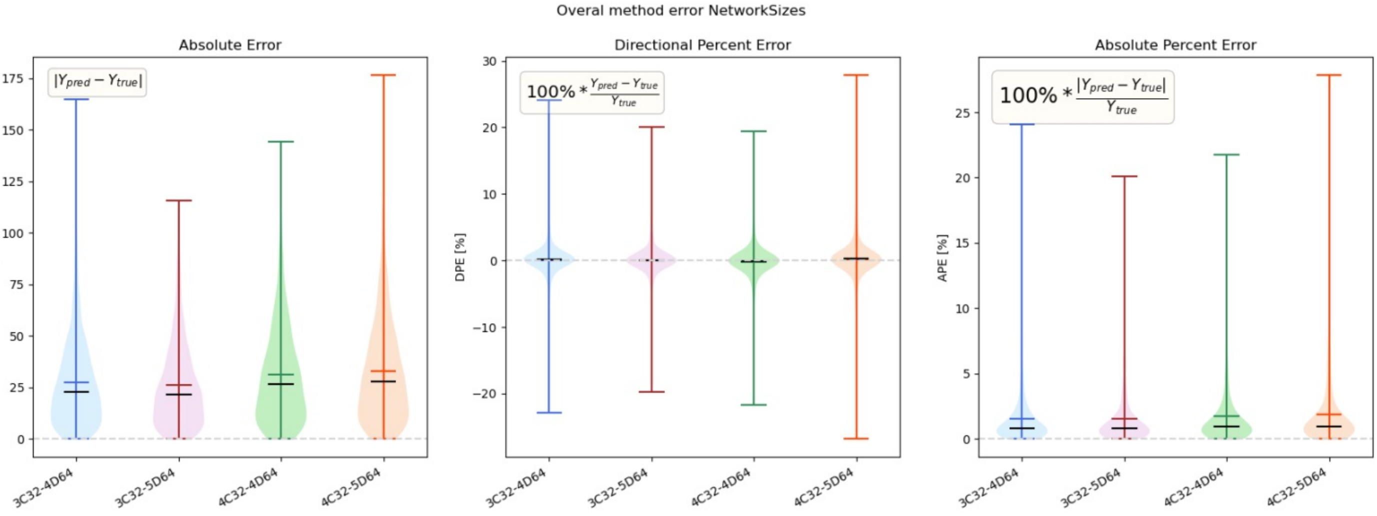


Figure 5.16: *The overall error on predicting the testset of networks with 3 or 4 convolutional layers with 32 nodes each and 4 or 5 dense layers with 64 nodes each, trained with 1500 epochs. The range of absolute and absolute percent error is lowest for network 3C32-5D64, but the mean error is comparable to network 3C32-4D64. An added convolutional layer worsens performance.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

ones in the large network, which are AE 27.4895 Hz, DPE 0.3084% and APE 1.6314% on the same testset. Better values could not be achieved with smaller networks in this experiment. For all these reasons, the chosen network architecture is 3 convolutional layers with 64 nodes each and maxpooling and 4 dense layers with 64 nodes each, trained with 1500 epochs with optimizer SGD to minimize MSE.

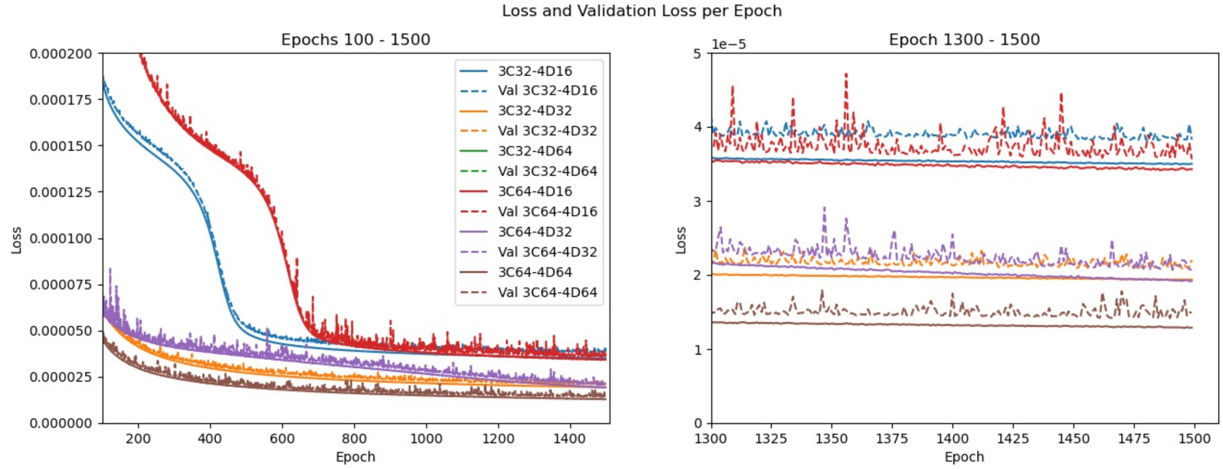


Figure 5.17: *The loss and validation loss during training of networks with 3 convolutional and 4 dense layers and different numbers of nodes. The largest network, where all layers have 64 nodes, outperforms all others and the pre-set goal.*

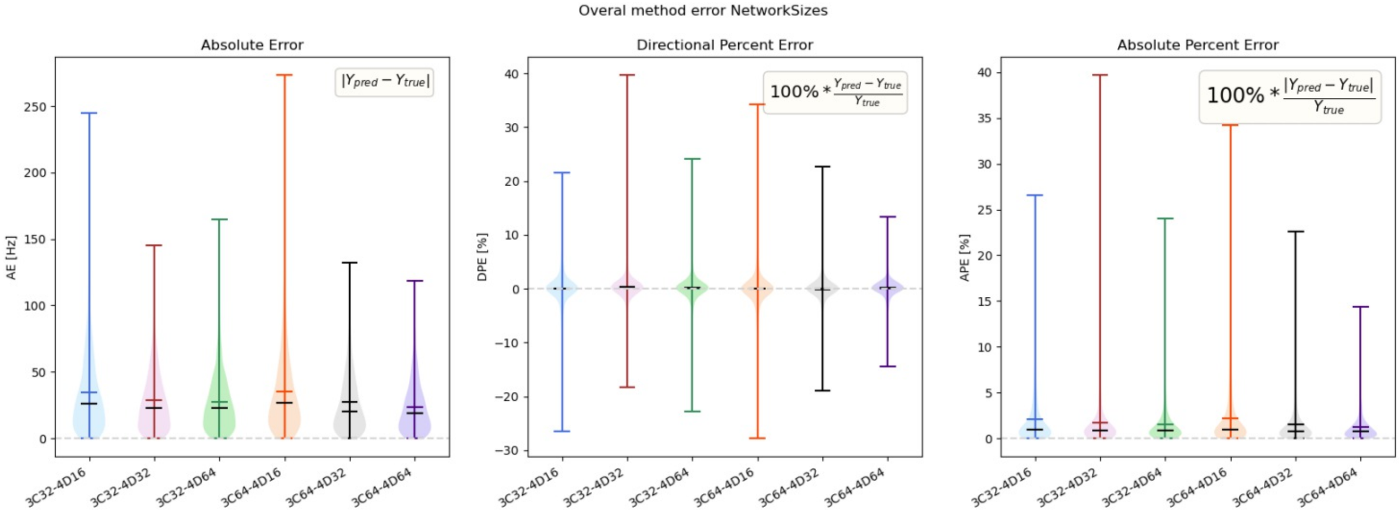


Figure 5.18: *The overall error on predicting the testset of the networks with 3 convolutional and 4 dense layers and different numbers of nodes. Small dense layers perform a lot worse than medium and big dense layers. The best performance on all error metrics is by the largest network with 64 nodes in each layer.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

5.4 Performance of the network on different datasets

The performance of the selected network is examined on different datasets. Dataset 0, with the same vibrato amplitude and noise level as the training data, has already been used for the violinplots in the last section. A detailed overview of the predicted and true resonances is shown in figure 5.19.

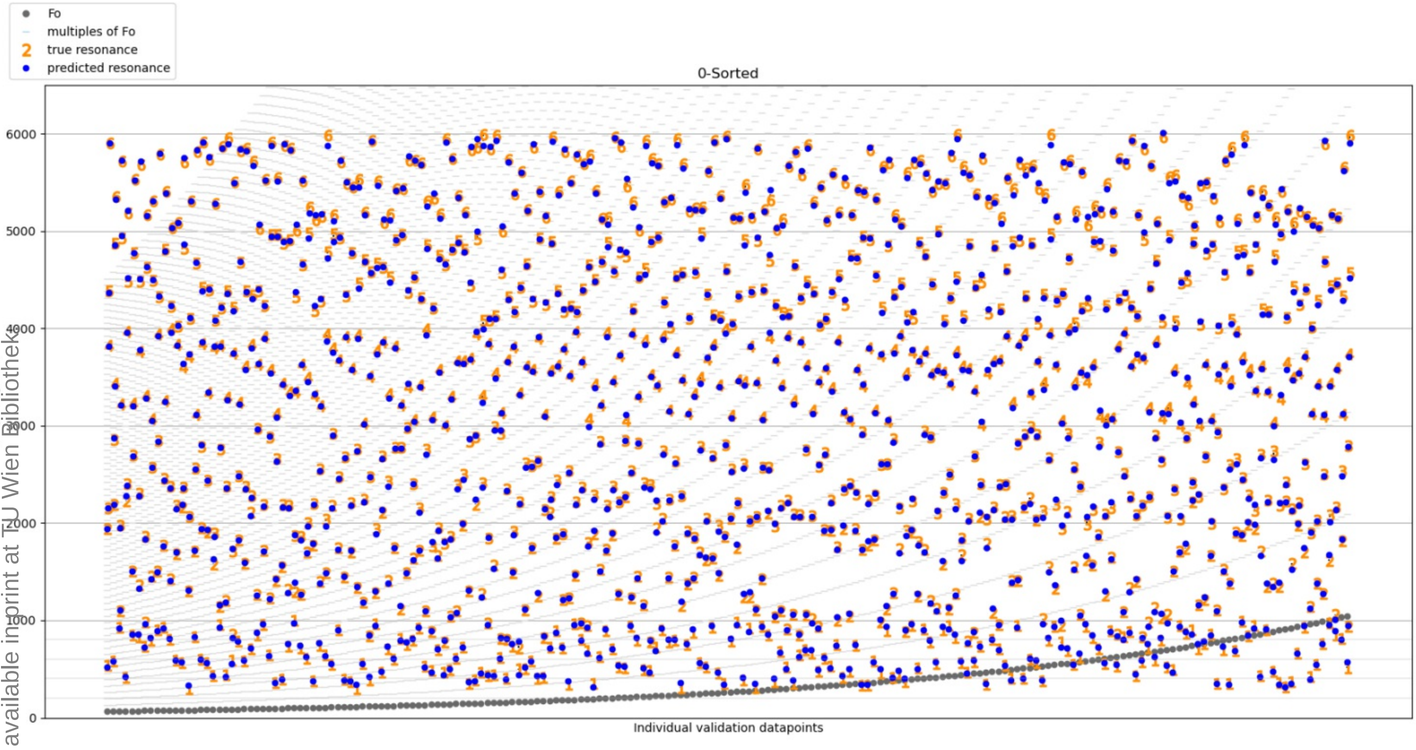


Figure 5.19: *An overview of the entire testset 0 with true and predicted resonance for each of the 200 sounds. The x-axis shows the individual sounds - samples of the testset - and the y-axis shows frequency. Each sound has 6 true resonances, marked with orange numbers, symbolizing the number of the resonance. The values predicted by the network are marked with dark blue dots. The closer the dot is to the corresponding number, the lower the error. If the corresponding true and predicted resonance are spaced farther apart, a dotted line connects them. Additionally, the fundamental frequency of each individual sound is marked with a grey dot at the bottom of the plot, and the harmonics are marked with thin grey lines. In the case of this network and dataset, the true and predicted resonances are close together. There is no region where performance would be visibly worse or better.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Entirely Synthesized Datasets				
ID	Number of Sounds	Fo range [Hz]	Noise [%]	Vibrato amplitude [cents]
Train	10 000	65.41 - 1046.50	10	75
0	200	65.41 - 1046.50	10	75
1	200	554.25 - 1567.98	10	75
2	200	65.41 - 1046.50	5	75
3	200	65.41 - 1046.50	0	75
4a	200	65.41 - 1046.50	10	50
4b	200	65.41 - 1046.50	10	25
5	200	65.41 - 1046.50	10	0
6	200	65.41 - 1046.50	0	0
7	200	65.41 - 1046.50	20	100

Table 5.5: *An overview of the parameter settings in the used testsets.*

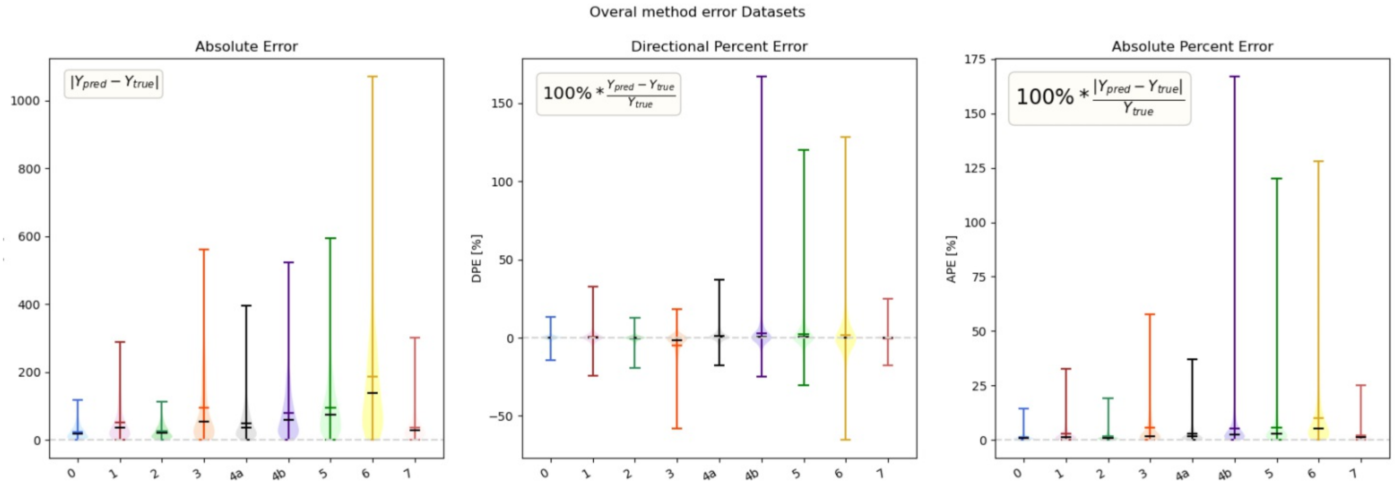


Figure 5.20: *The overall error on predicting each of the testing datasets, as defined in table 5.5. The effect on changing different parameters in the test data is treated in detail in the following paragraphs.*

The testing datasets are once more defined in table 5.5 An overview of the performance predicting these testsets is shown in figure 5.20 as overall error and in figure 5.21 as error related to the f_o of the sound. The mean error values, shown in figure 5.20 as coloured horizontal markers,are also given in table 5.6.

The lowest error, as expected, is the testset corresponding to the training data, as that is the same data structure as the network is used to. Importantly, this testset is not a subset of the training set, but randomized with the same settings and a different seed.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Mean errors on predicting the datasets									
Error	0	1	2	3	4a	4b	5	6	7
AE (Hz)	23.62	51.62	25.40	94.88	49.56	79.74	94.67	185.65	35.69
APE(%)	0.02	0.32	-0.55	-5.03	1.31	2.93	2.20	1.60	-0.25
DPE(%)	1.22	2.71	1.42	5.60	2.67	5.15	5.67	9.97	1.83

Table 5.6: *The mean absolute error, directional percent error and absolute percent error of the prediction of different datasets by the selected network.*

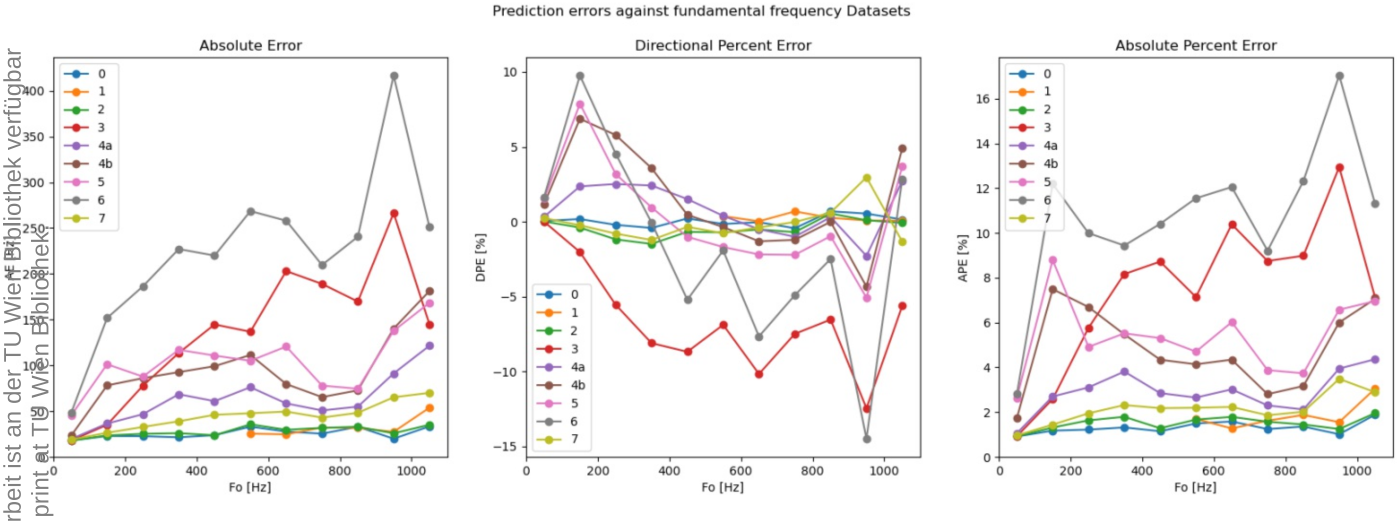


Figure 5.21: *The error for predicting the training testsets related to the f_o of the sound.*

On a testset with higher f_o, the error increases. Halving the noise does not significantly change performance, but removing noise entirely does. Lowering the vibrato amplitude does progressively worsen performance. The testset without noise or vibrato does have the worst performance, confirming that the network has learned to use these parameters to predict the resonances.

5.4.1 Influence of f_o

As shown by the standard testset, the network performs steadily over the entire f_o-range of the training data. The complexity and duration of training were sufficient to cover this range. At low fundamental frequencies below 100 Hz, all datasets can be predicted well. This is approximately the range for male speech, where other approaches are also successful. On datasets with at least 50 ct vibrato, this good performance of less than 50 Hz AE continues up to 200 Hz. Dataset 3 with no noise is hardest to predict about 600 Hz.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

When asked to extrapolate the frequency range, to examine whether the network has learned any underlying structure, the network performs worse. The performance on testsets 0 and 1 is shown separately in figure 5.22, displaying the entire considered frequency range. A steep increase in error is visible right at the upper end of the range of the training data.

The difference between the orange and blue lines on the last frequency interval where they are both represented, can be caused by the construction of the figure: Frequency bands of 100 Hz are taken and the mean error for all included samples is calculated. In the band of 1000 - 1100 Hz, only three samples are included in testset 0, but many more in testset 1, surpassing the training frequency range in the middle of this interval. The lack of training is therefore immediately visible. For optimized performance of the final network, the entire frequency range that could be encountered in the application should therefore already be present in the training data.

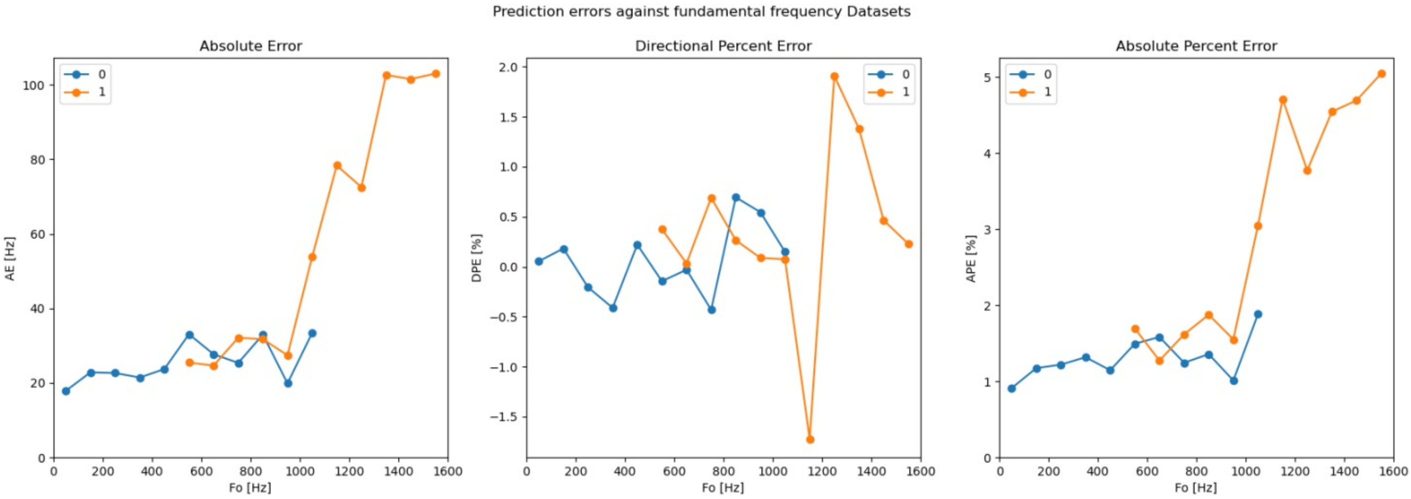


Figure 5.22: *The error for predicting the training testsets 0 and 1 related to the* f_o *of the sound. Testset 0 has the same structure, parameter values and frequency range as the training data, testset 1 also has the same structure and parameter values, but a higher frequency range. The network performs a lot worse at fundamental frequencies that it has not been trained on.*

5.4.2 Breathiness

In testsets 0, 2, and 3, the breathiness level decreases from 10% via 5% to 0%. The scatterplot overviews of the entire testsets are attached in the appendix. The overall absolute error is very similar in testsets 0 and 2, concretely AE 23.6235 Hz and 25.4046 Hz. Completely removing noise, however, increases overall error to 94.8784 Hz. The low amplitude shape of the vocal tract transfer function with local peaks at the location of the resonances is therefore sufficient and does not need to be strong to help the network,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek WIEN Your knowledge hub

as long as it is present. The influence of breathiness increases as fundamental frequency increases, as shown by the red lines on figure 5.21. At low fundamental frequencies, the red line is at a similar level as the original testset, but around 300 Hz fundamental frequency it crosses 100 Hz of mean absolute error and stays worse than most other testsets. The only worse testset at f_o above 300 Hz is testset 6, which also has 0% noise. The presence of noise is therefore crucial for network performance. Luckily, in the human voice, a little noise is almost always present, but it is also difficult to discern form background noise. It is important not to filter it out. Varying the breathiness in the input data will also help to optimize performance.

5.4.3 Vibrato

Testsets 0, 4a, 4b and 5 have progressively lower vibrato amplitude; 75 ct, 50 ct, 25 ct, and 0 ct. The increase in overall error in these same testsets indicates that breathiness is used by the network to predict the resonances. The AE increases steadily, from 23.6235 Hz to 49.5573 Hz, 79.7419 Hz and 94.6746 Hz respectively. A wider vibrato range helps predicting the resonance frequencies more accurately, as a sweep through more frequencies has a larger probability of passing through a peak, or allows to more accurately extrapolate the location of the peak from the slope of the intensity in its neighbourhood.

5.4.4 Combination of parameters

The absence of both vibrato and breathiness from the testdata yields the worst prediction results, of AE 185.6497 Hz and APE 9.9676 Hz. As the error on the datasets with 0% noise and 0 ct vibrato are 94.8784 Hz and 94.6746 Hz respectively, the current error could be considered almost a summation of those two. That would indicate that the information extracted from noise is different from the information from vibrato, and one cannot be used to replace the lack of the other.

The comparable shape of the grey and red lines in the AE and APE plots indicates no trend of favoring certain resonances, which is also visible in the scatterplots in the appendix. The predictions change and move farther from the true resonances in the entire frequency range and for all resonances, but more for higher fundamental frequencies.

Higher values of noise and vibrato amplitude do worsen performance slightly, but less than lower values. A sufficient representation of low vibrato amplitudes and noise levels will be important in the further development of the network.

5.4.5 Lab recorded sounds

The resonances of the different tubes, extracted from the noise and chirps played through them, as well as the theoretically predicted values, are shown in table 1 in the appendix. The resonances from noise and chirp recordings are often close together with differences of a few Hz, and more different from the theoretical values. A more detailed statistical

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Mean errors on predicting the datasets									
Error	0	1	2	3	4a	4b	5	6	7
AE (Hz)	348.14	358.70	356.87	381.70	359.34	357.95	356.07	372.98	349.93
APE(%)	3.68	-2.41	2.06	-0.59	0.51	1.35	2.99	10.53	1.73
DPE(%)	12.94	11.28	13.05	13.69	12.76	13.06	13.75	19.09	12.44

Table 5.7: *The mean absolute error, directional percent error and absolute percent error of the prediction of different datasets by the selected network.*

analysis of these values and their relationship to the formula is outside the scope of this thesis. The ground truth for the resonance estimation was based on the values extracted from the noise.

In table 5.7 and figures 5.23 and 5.24, the results are presented on the recorded datasets, predicted by the previously selected network, trained on only synthesized datasets. As the structure and characteristics of these datasets are substantially different, the network is not able to accurately predict the resonances of these data. The detailed results in the form of scatterplots sorted according to either f_o or tube lengths, can be found in the appendix. With mean absolute errors in the range from 348 Hz to 382 Hz, the performance is significantly worse, and does not change much for different datasets or depending on fundamental frequency.

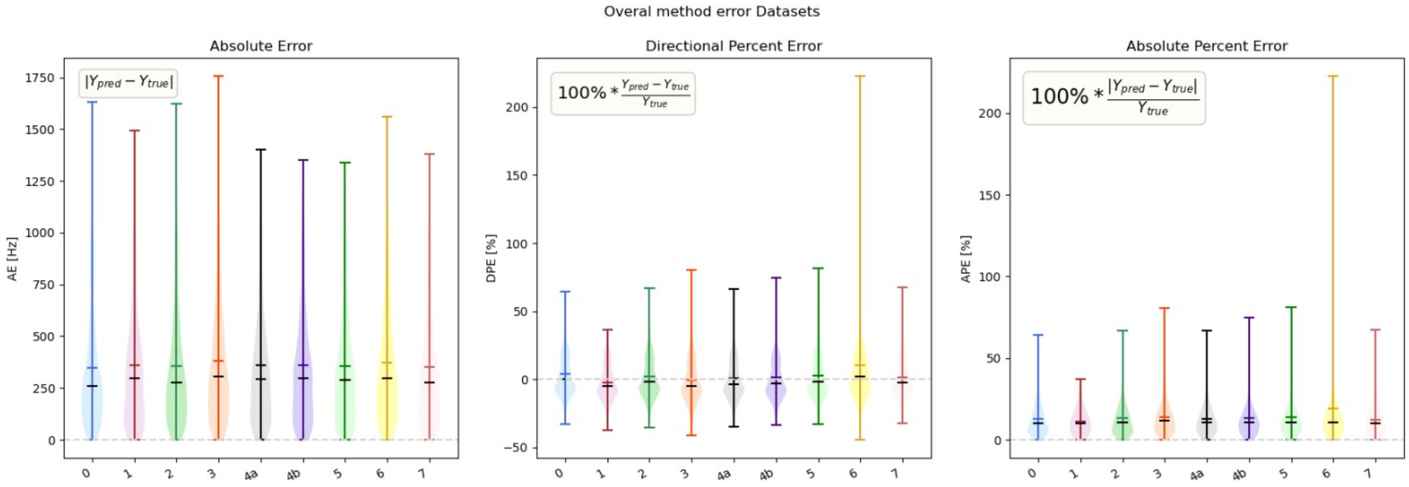


Figure 5.23: *The overall errors for predicting the different recorded datasets by the network, trained purely on synthesized data. The performance is poor for all datasets.*

A closer look at the scatterplot for dataset 0 in figure 5.25 allows a better interpretation of these results:

Overall, the predictions of the network stay constant independently of the changing

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
Bibliothek, Your knowledge hub
TU WIEN

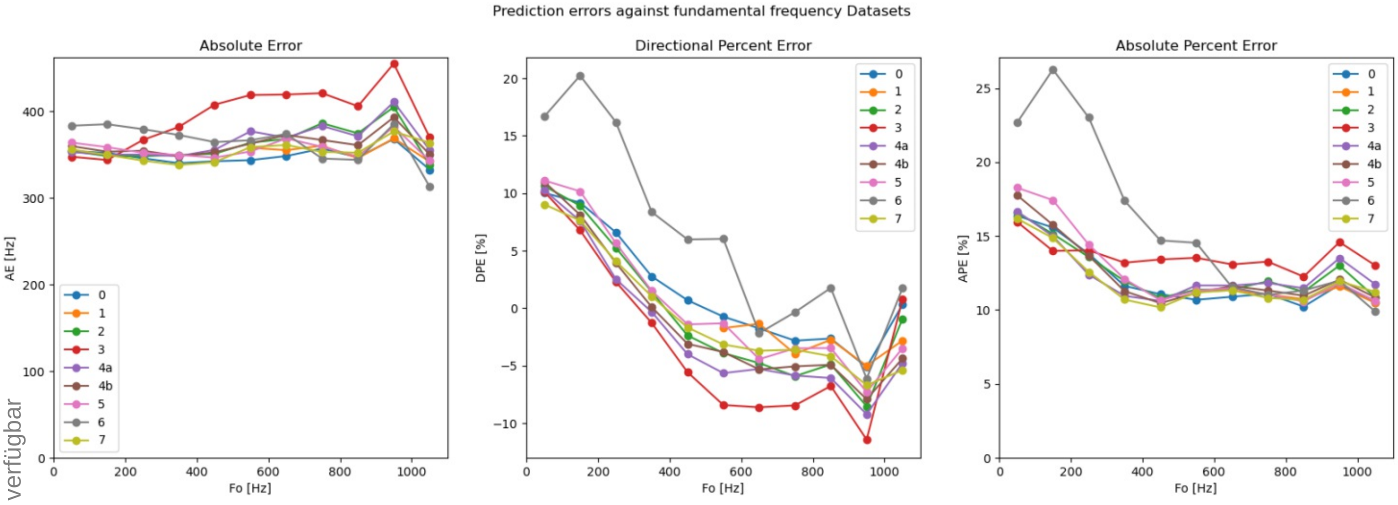


Figure 5.24: *The error for different fundamental frequencies. The performance is poor regardless of fundamental frequency.*

true resonance frequencies. The resonances predicted for sounds of higher fundamental frequencies are slightly higher, although the true resonances stay constant.

An apparent reason for the low performance is the discrepancy in resonance spacing between the training data and the tubes. The locations of the resonances in the training data were defined by clear rules: f_{R1} between 300 Hz and 1000 Hz, f_{R2} between 700 Hz and 2300 Hz, f_{R3} between 1700 Hz and 3000 Hz, f_{R4} between 3000 Hz and 4000 Hz, f_{R5} between 4000 Hz and 5000 Hz, and f_{R6} between 5000 Hz and 6000 Hz. The highest possible resonance was therefore 6000 Hz, whereas in the tubes, it exceeds 7000 Hz.

Importantly, at middle tube lengths and central resonance frequencies, the predicted and true resonances are closer together, as the horizontal line of the predicted resonances crosses the decreasing line of the true resonances. The resonance spacing of the corresponding tubelengths, 15 cm - 17 cm, is equivalent to the physiological resonance spacing, which was the basis for the locations of the resonances in the training data. The network tries to apply the structure it has learned, to the data. At lower tubelengths, it predicts resonances on the location it has learned they are, putting the highest at around 6000 Hz. Remarkably, the tubelengths above 16 cm still do all follow the resonance spacing rules, except the last one, but the network predicts similar values to the 15 cm - 17 cm tubes. For the shortest and longest tubes in figure 5.25, a wrong resonance is predicted accurately on the location of another.

Additionally, for the right column, corresponding to tubelength 19 cm, resonance 4 does not fit the line of the previous resonances, but the prediction of resonance 3 corresponds with it perfectly. There could be an artifact in the sounds, causing both the peak-picking algorithm for extracting the resonances from noise, and the network to choose it.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

In the used set-up, the recorded data differ from the training data in two aspects: the resonance spacing and the presence of the recording pipeline. It is not evident to interpret which of these two has which influence on the performance, but the results on the intermediate tubelengths could be an indication, concretely 15 cm, 15.5 cm, 16 cm, 16.5 cm and 17 cm. These 5 tubes, with each 42 sounds, yield together a dataset of 210 sounds, comparable to the size of the other testsets. The AE in this case is 175 Hz. That could mean that changing resonance spacing in the data increases the error from 175 Hz to 350 Hz, and using recorded instead of synthesized sounds increases the error from 25 Hz to 175 Hz, if no other parameter dependencies occur. In either way, these parameter variations should be accounted for in the training data, if the network should perform well in its application environment.

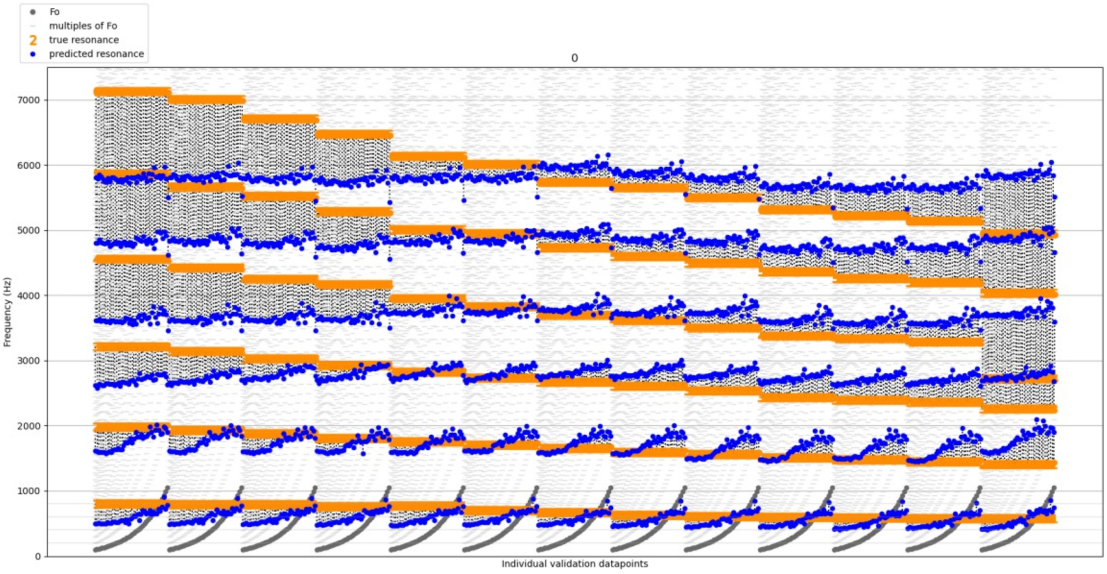


Figure 5.25: *Scatterplot of results for predicting recorded testset 0 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 75 ct vibrato.*

5.4.6 Comparison to the baseline method

The predictions on the standard testset by the baseline method, Burg's LPC method as implemented in Praat with the default settings and formant ceiling 6050 Hz, are shown in the scatterplot in figure 5.26. The formant ceiling of 6050 Hz was chosen because the upper limit for the sixth resonance is 6000 Hz, and the Praat algorithm removes all formants higher than 50 Hz below the formant ceiling [4]. The performance of Praat in the speech-relevant area is very good, for the first five resonances in sounds with fundamental frequencies up to 400 Hz. The sixth resonance is often missed entirely and the first resonance is consistently overestimated. Above 400 Hz, pitch locking is evident, as the predicted resonances align with the harmonics.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

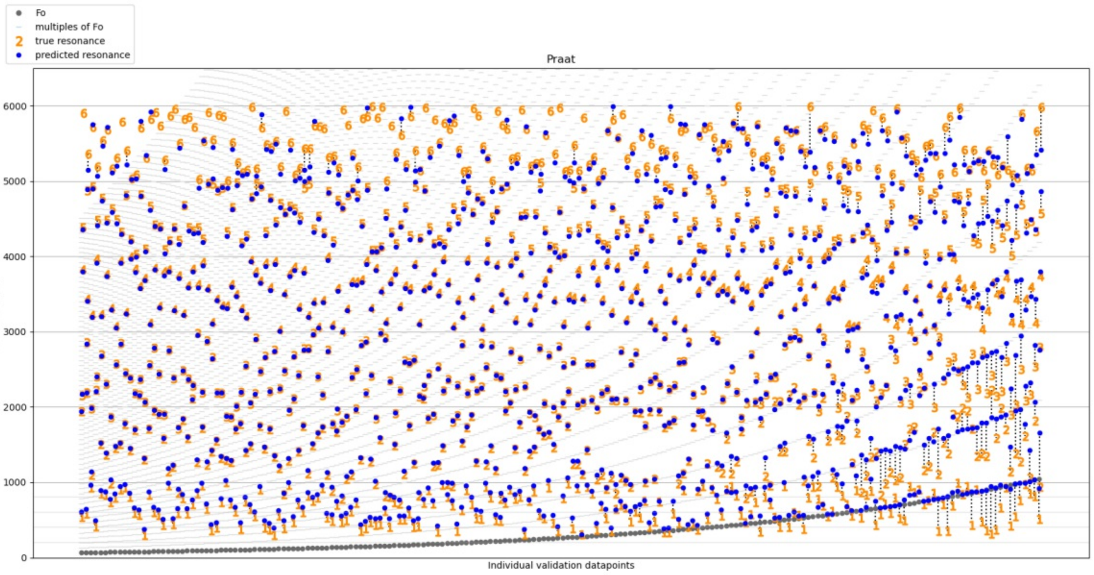


Figure 5.26: *Scatterplot of results for predicting synthesized testset 0 with Praat, using Burg's LPC algorithm with formant ceiling 6050 Hz.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 75 ct vibrato.*

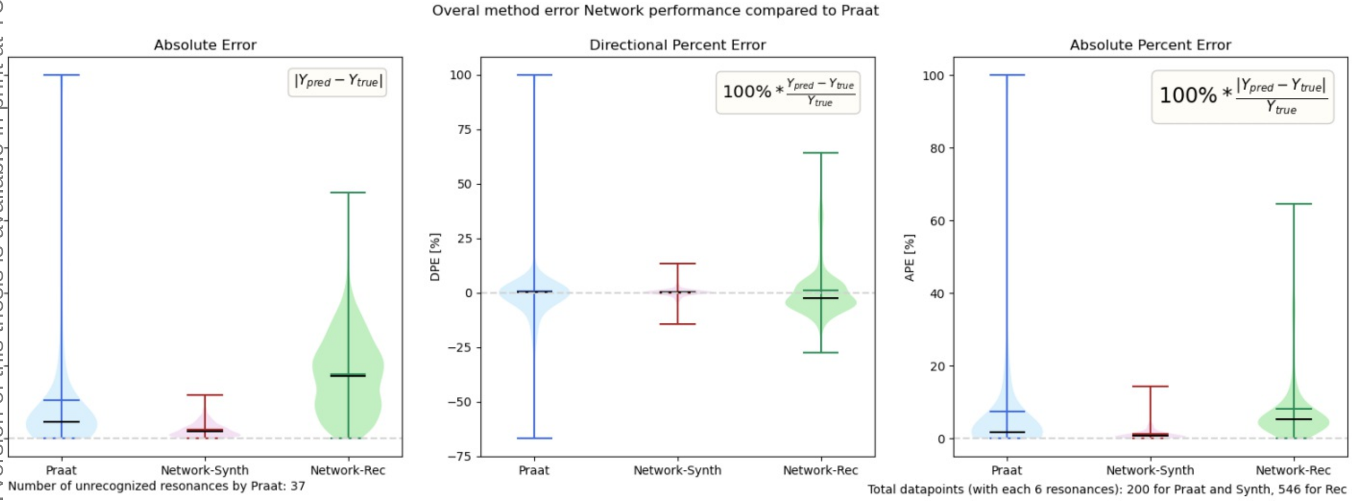


Figure 5.27: *The overall absolute error, directional percent error and absolute percent error for the baseline method (Burg's algorithm implemented by Praat) and the network on both the standard, synthesized dataset as on the standard recorded dataset, filtered by tubes of 15 to 17 cm.*

The total errors and errors relative to f_o are displayed in figures 5.27 and 5.28. The performance of the network stays stable over the full frequency range, with errors of about

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

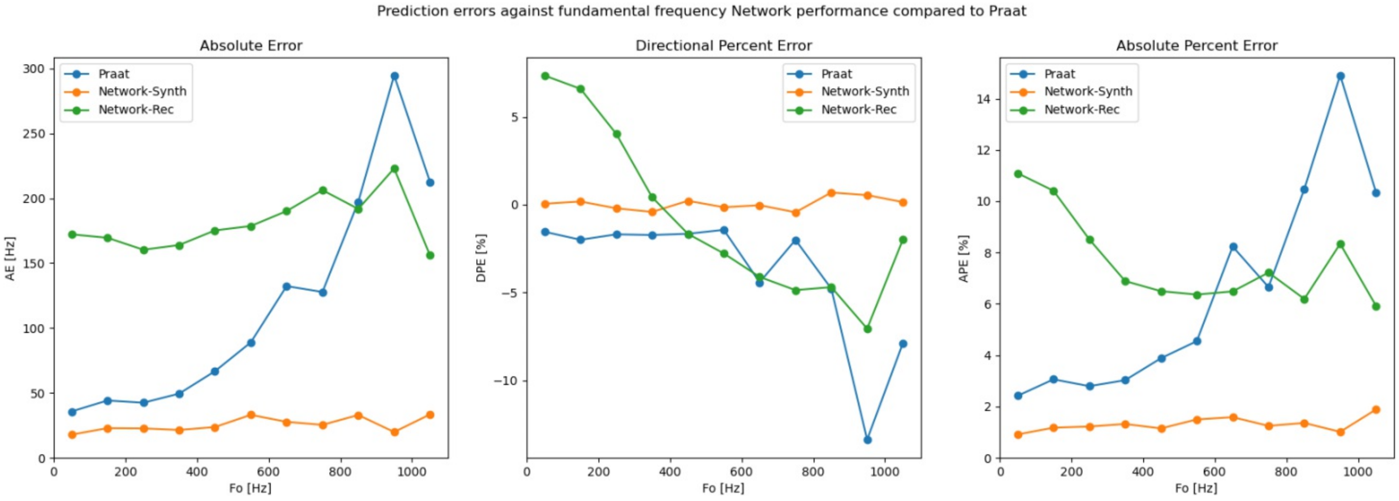


Figure 5.28: *The error for different fundamental frequencies for the baseline method (Burg's algorithm implemented by Praat) and the network on both the standard, synthesized dataset as on the standard recorded dataset, filtered by tubes of 15 to 17 cm. The network with synthesized data outperforms Praat on the entire frequency range. The performance of the network on both datasets is stable throughout the frequency range, but the performance of Praat worsens steeply above the frequency range of speech, below 400 Hz.*

Error	Praat	Network-Synth	Network-Recorded
AE (Hz)	76.710	23.626	175.105
APE(%)	-2.587	2.248	1.295
DPE(%)	4.474	1.216	8.190

Table 5.8: *Mean errors on the predicting the resonances of the standard dataset by Praat and the network and on the recorded dataset by the network.*

25 Hz, as seen before. Even at lower fundamental frequencies, the network performs better than Praat, which has errors below 50 Hz. Above f_o 400 Hz, Praat's error worsens a lot, peaking at 300 Hz mean errors when approaching f_o 1000 Hz.

Additionally, the performance of the network on some of the recorded data is part of the comparison. As discussed in the previous section, only the best performing, physiologically realistic middle tubelengths are considered. The performance on this data is also stable for the entire fundamental frequency range, but with absolute error values of about 175 Hz. For the highest fundamental frequencies above 800 Hz, this is better than the values of Praat.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

5.5 Network trained with lab recorded sounds and tested on the vocal tract model

The resonances of the 3D printed vocal tract model, extracted from the noise recordings in the same way as previously with the tubes, are shown in table 5.9. These resonances do obey the resonance spacing rules that were used for the construction of the training data, as they fit into the defined intervals. Additionally, the third and fourth resonance are close together with a distance of 176 Hz, which is more than the defined minimum. Their proximity to 3000 Hz corresponds to the singer's formant.

f_{R1}	f_{R2}	f_{R3}	f_{R4}	f_{R5}	f_{R6}
862.810 Hz	1226.679 Hz	2924.734 Hz	3101.155 Hz	4258.919 Hz	5907.355 Hz

Table 5.9: *Resonances of the 3D-printed vocal tract model, extracted by playing noise through it.*

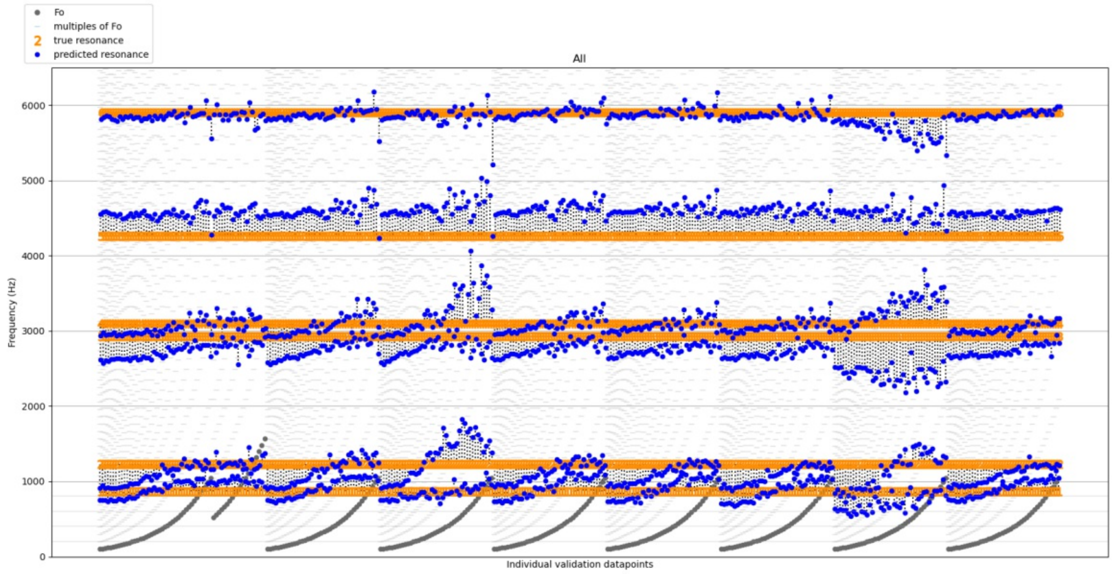


Figure 5.29: *Scatterplot of results for predicting the resonances of the 3D-printed vocal tract model by the network trained on synthesized data.*

The data in scatterplots 5.29 and 5.30 are organized by testset. The data sorted for fundamental frequency can be found in the appendix. The main network, trained on synthesized data, approaches the resonances better than the network trained on recorded data. The different overall errors are shown in table 5.10. The AE of 175 Hz for the synthetic network is equal to the AE for predicting the recorded sounds filtered by the 5 middle tubes in section 6.4.5. As both those tubes and the vocal tract model have the same resonance spacing as the training data, this error can be attributed to the recording.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The resonance predictions are consistently higher for higher fundamental frequencies. Just like for the recorded data filtered by the tubes, the influence of testset seems very limited here, except for testsets 3 and 6, both with 0% noise.

The network trained on recorded data has been trained on all testsets, but all training sounds have been filtered by cylindrical tubes with regularly spaced resonances. The network has learned to predict this regularity, so it is not able to predict the irregularly spaced resonances of the vocal tract model. It predicts a regular pattern, with the fourth predicted resonance on the location where the third and fourth real resonances come closely together. The AE of 491 Hz indicates a very poor performance. The predictions of the third and sixth resonance are 1000 Hz away. This proves once more that the resonance spacing in the training data is crucial for a successful network. The influence of fundamental frequency is limited and, like before, testsets 3 and 6 differ most from the others. Both higher frequency and lower noise tend to make prediction less consistent.

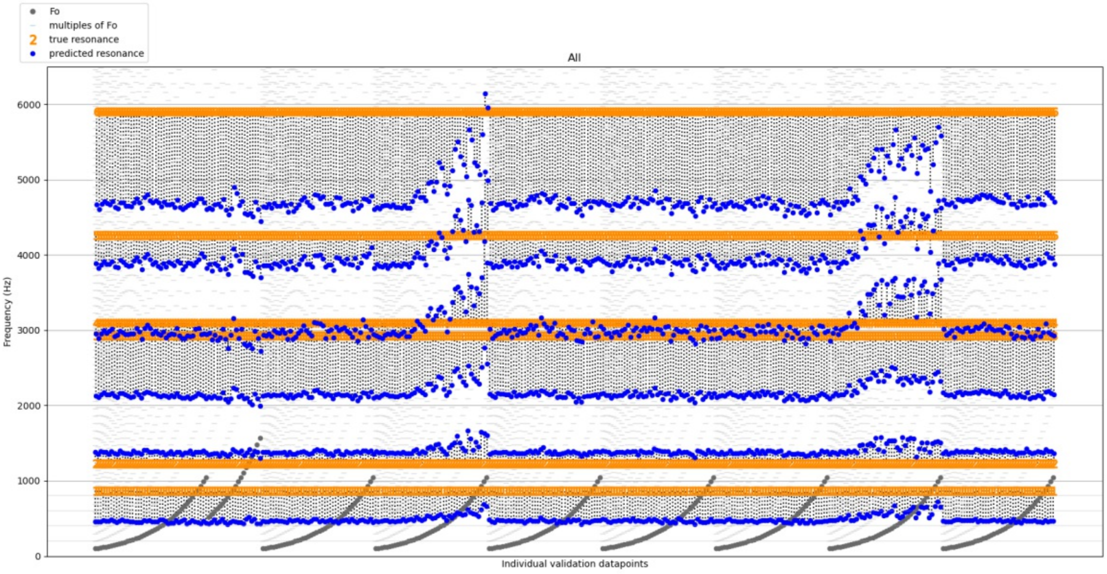


Figure 5.30: *Scatterplot of results for predicting the resonances of the 3D-printed vocal tract model by the network trained on recorded data.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Error	Network trained on synthetic data	Network trained on recorded data
AE (Hz)	175.24	490.68
APE (%)	2.06	22.78
DPE (%)	8.45	27.47

Table 5.10: *Mean errors on the predicting the resonances of sounds filtered by the 3D printed vocal tract by the network trained on synthesized sounds and by the network trained on recorded sounds.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 6

Summary and Future Work

In this work, a neural network was sought that would be able to estimate the first six resonance frequencies of a vocal tract filter using information that is not present in the spectral envelope. This approach overcomes shortcomings of the traditional envelope-based formant estimation methods, LPC and cepstral analysis, and can be applied at a much higher frequency range, at which the traditional methods fail due to wider spacing of the harmonics.

The network found to perform best on synthesized vowels has the following architecture: The input are raw spectrograms with 50x1000 pixels. The network has 3 convolutional layers with each 64 nodes, kernel size 3, activation relu, and each followed by 2-by-2 max-pooling to limit size. Then a flatten layer, 5 dense layers with 64 nodes each and a final output layer with 6 nodes, corresponding to the 6 predicted vocal tract resonances. The architecture and length of training has an important influence on the performance. With the selected network, the performance is stable over the frequency and resonance range of the training data, unlike with envelope-based formant estimation. Outside of this range, accuracy decreases. Both vibrato and noise are important for the resonance estimation and removing them significantly lowers performance. Higher noise does not improve performance compared to lower noise, but its presence is necessary for successful resonance estimation. In the case of vibrato, the accuracy improvement is related to its amplitude. As this approach is able to extract information from the data that is not in the envelope, it successfully avoids the shortcomings of envelope-based standard methods used in formant estimation in speech.

In conclusion, the network can outperform Praat (using Burg's LPC algorithm) over the entire frequency range for the synthetic data. As that was the main focus of this thesis, it reached its goal. The influence of data parameters outside the scope of the training data was systematically tested. The most important step in subsequent research will be careful construction of the testing data, completely covering the parameter space.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

Following recommendations for future work are formulated based on the results of this study:

- The fundamental frequencies should completely cover the range that could be encountered in the application, as the network is not able to extrapolate.
- Special attention should be given to low levels of noise and its absence, as that is a weak spot of the current network.
- Vibrato should vary in amplitude, but also in frequency and phase.
- Irregularities characteristic for human speech in pitch, vibrato and other aspects should be included.
- An additional parameter, not considered here, is spectral slope, which should also be sufficiently varied.
- A significant part of the training data should be recorded and not only synthesized
- Resonance spacing should cover all physiological possibilities for different vocal tracts lengths, e.g. also small children. The regular resonance spacing in cylindrical tubes is not suitable for the entire training set, as the network will then predict only regularly spaced resonances. Ideally, also the recorded data should be filtered by realistic filters, like the 3D vocal tract model.

Further work should also explore the implementation of the network for the applications mentioned before - language learning, singing training and speech therapy for timbre modification, such as in gender conversion. The network and training data may need to be adjusted in function of the exact application needs.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Appendices

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Scatterplots synthesized data

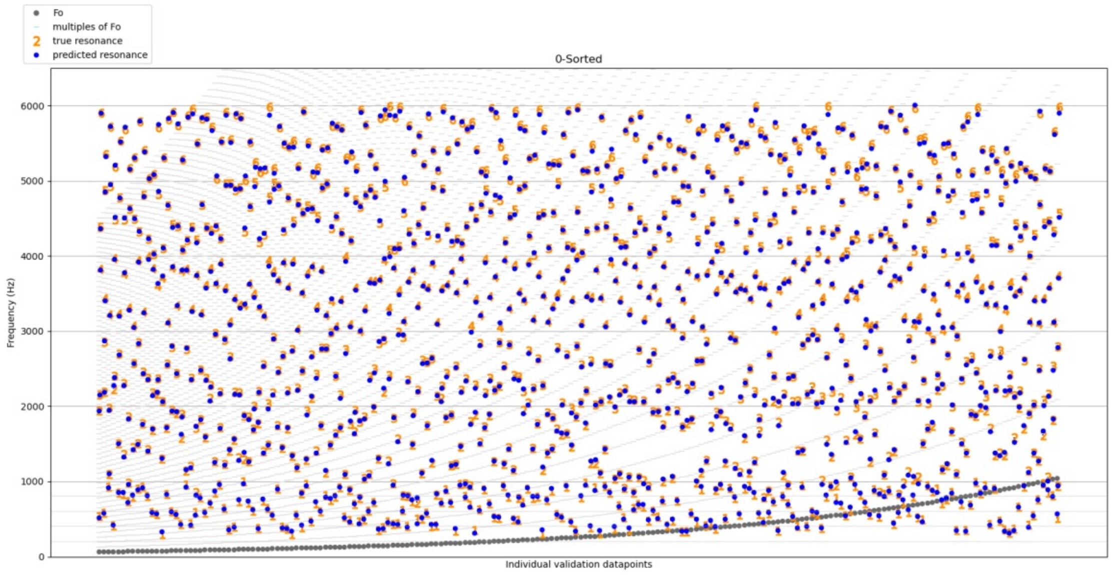


Figure 1: *Scatterplot of results for predicting fully synthesized testset 0 with a network trained on exclusively synthesized data.* f_o *ranging from 65.41 Hz to 1046.50 Hz, 10% noise and 75 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

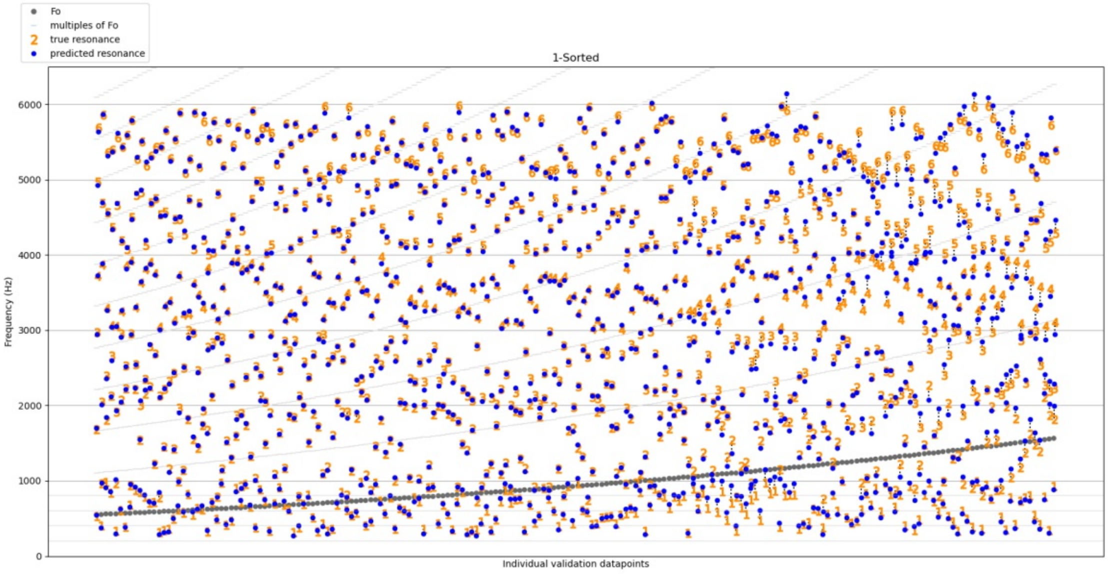


Figure 2: *Scatterplot of results for predicting fully synthesized testset 1 with a network trained on exclusively synthesized data.* f_o *ranging from 554.37 Hz to 1567.98 Hz, 10% noise and 75 ct vibrato.*

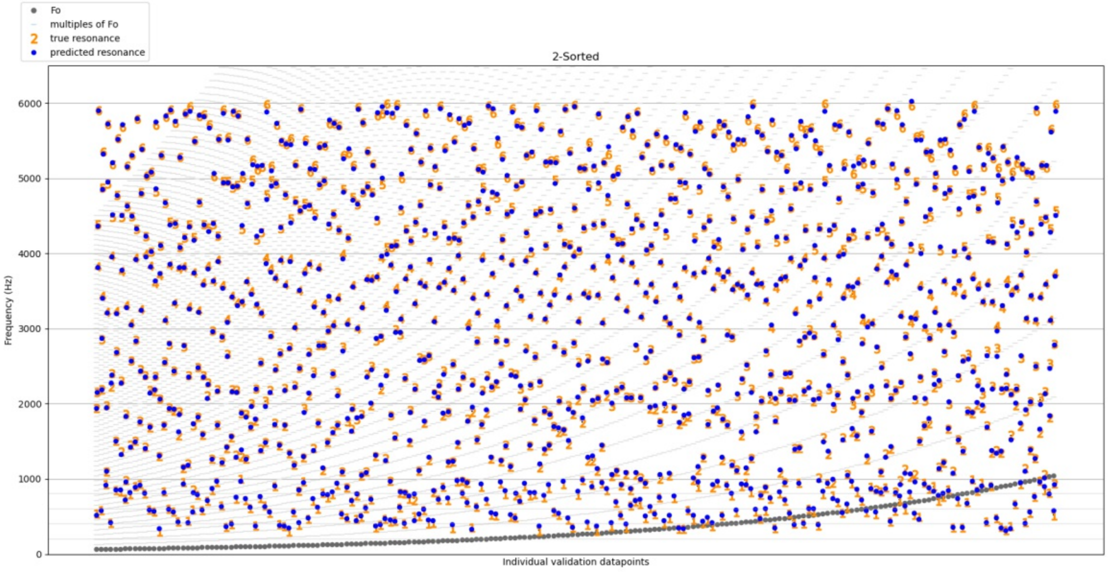


Figure 3: *Scatterplot of results for predicting fully synthesized testset 2 with a network trained on exclusively synthesized data.* f_o *ranging from 65.41 Hz to 1046.50 Hz, 5% noise and 75 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

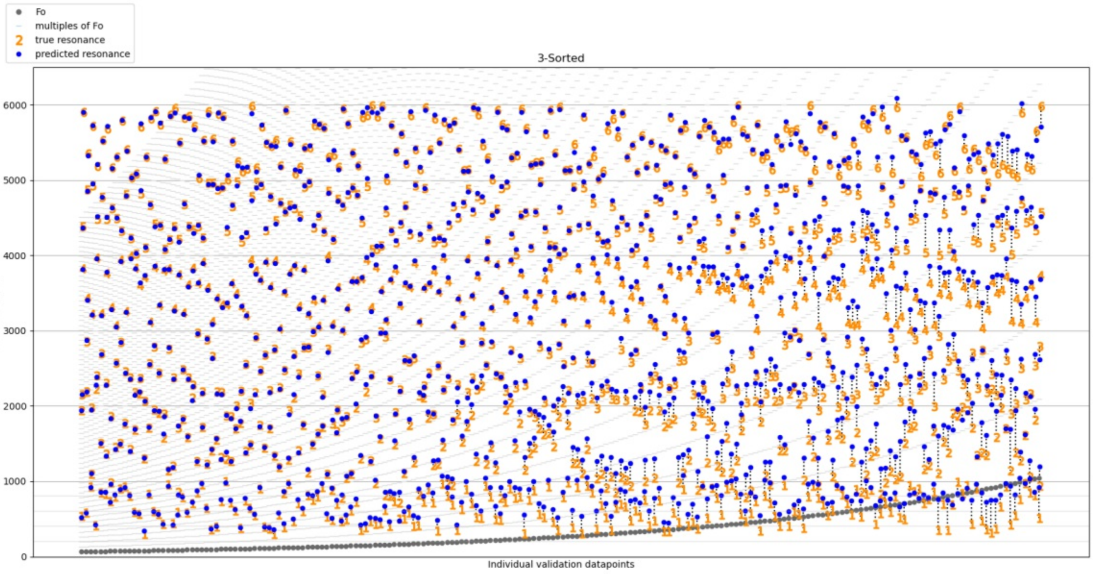


Figure 4: *Scatterplot of results for predicting fully synthesized testset 3 with a network trained on exclusively synthesized data.* f_o *ranging from 65.41 Hz to 1046.50 Hz, 0% noise and 75 ct vibrato.*

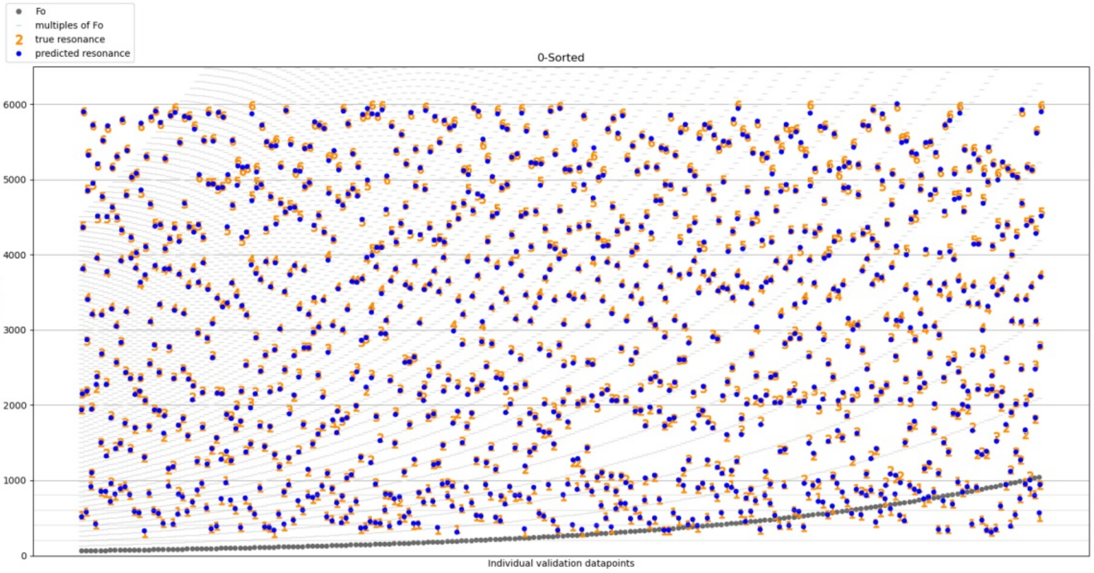


Figure 5: *Scatterplot of results for predicting fully synthesized testset 4a with a network trained on exclusively synthesized data.* f_o *ranging from 65.41 Hz to 1046.50 Hz, 10% noise and 50 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

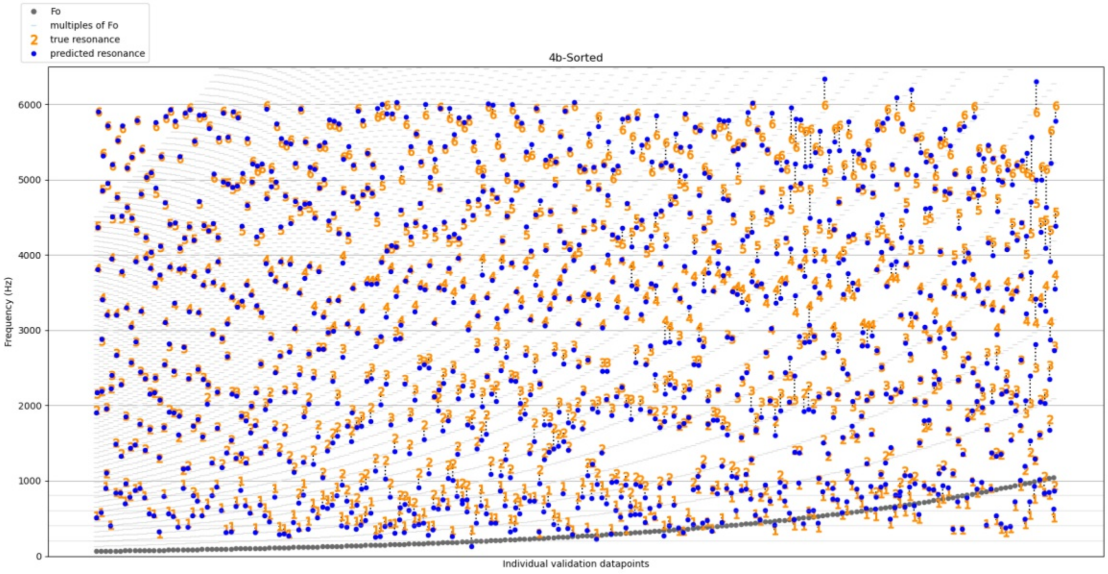


Figure 6: *Scatterplot of results for predicting fully synthesized testset 4b with a network trained on exclusively synthesized data.* f_o *ranging from 65.41 Hz to 1046.50 Hz, 10% noise and 25 ct vibrato.*

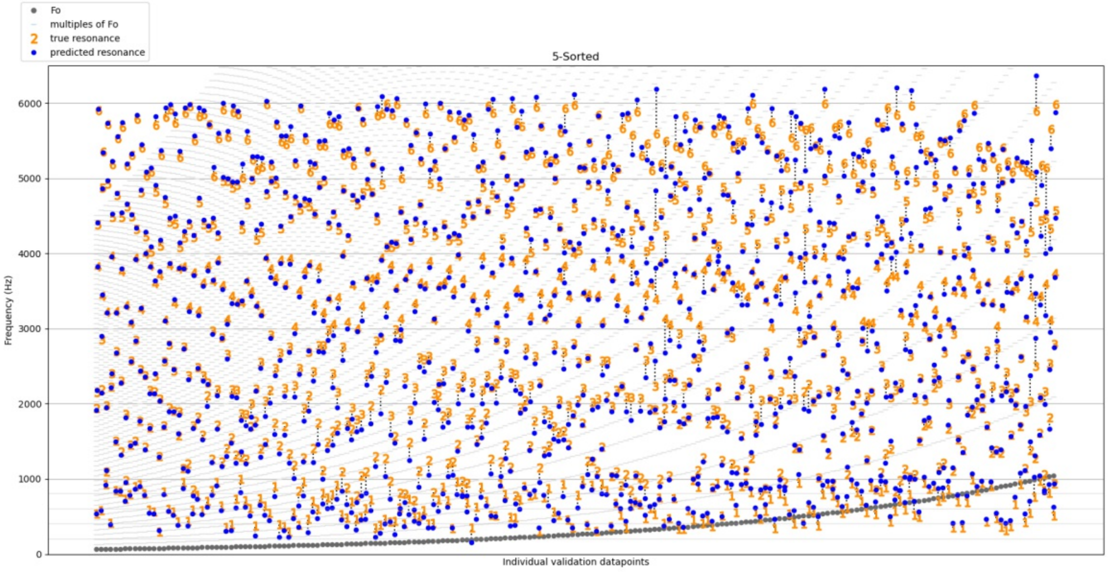


Figure 7: *Scatterplot of results for predicting fully synthesized testset 5 with a network trained on exclusively synthesized data.* f_o *ranging from 65.41 Hz to 1046.50 Hz, 10% noise and 0 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

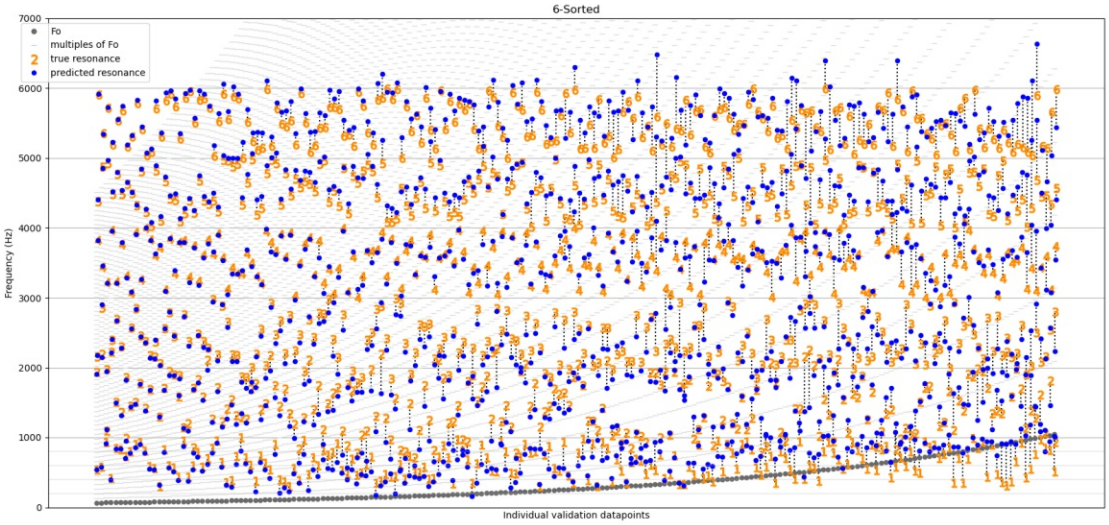


Figure 8: *Scatterplot of results for predicting fully synthesized testset 6 with a network trained on exclusively synthesized data.* f_o *ranging from 65.41 Hz to 1046.50 Hz, 0% noise and 0 ct vibrato.*

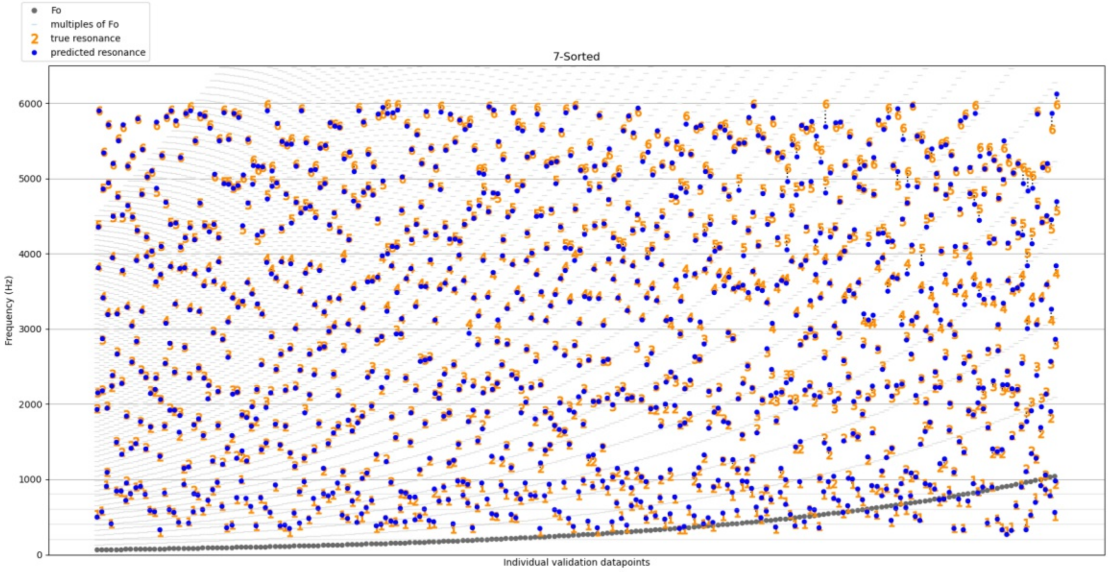


Figure 9: *Scatterplot of results for predicting fully synthesized testset 7 with a network trained on exclusively synthesized data.* f_o *ranging from 65.41 Hz to 1046.50 Hz, 20% noise and 100 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Scatterplot recorded data

In this section, the results of predicting the recorded testsets by the network trained on synthesized data are shown. They are visualised in two ways: sorted for increasing tube length (and therefore decreasing resonance frequencies) and for increasing fundamental frequencies. This allows trends to be visible more easily.

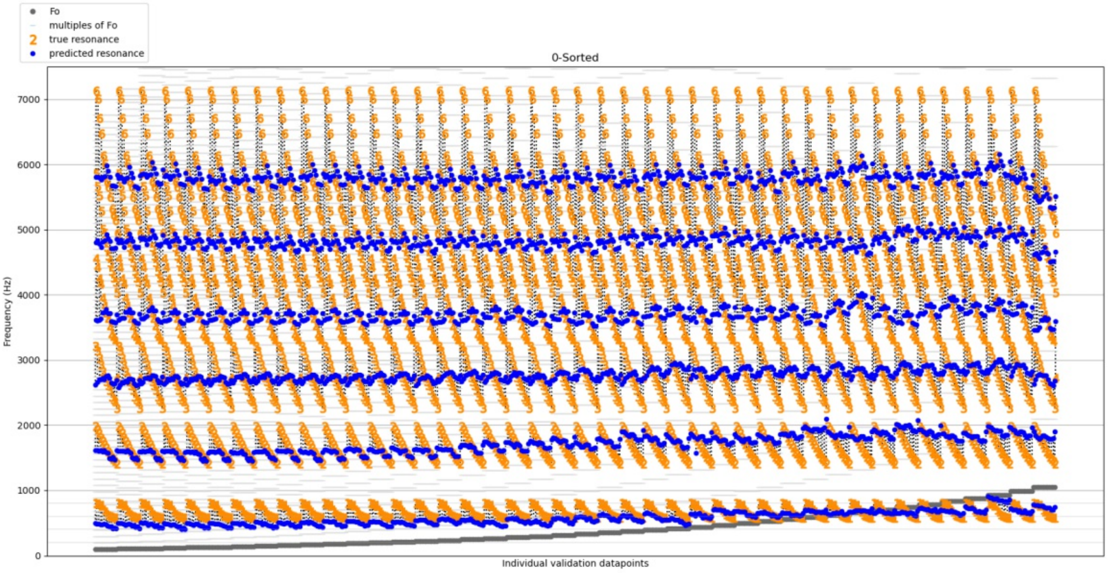


Figure 10: *Scatterplot of results for predicting recorded testset 0 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 75 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

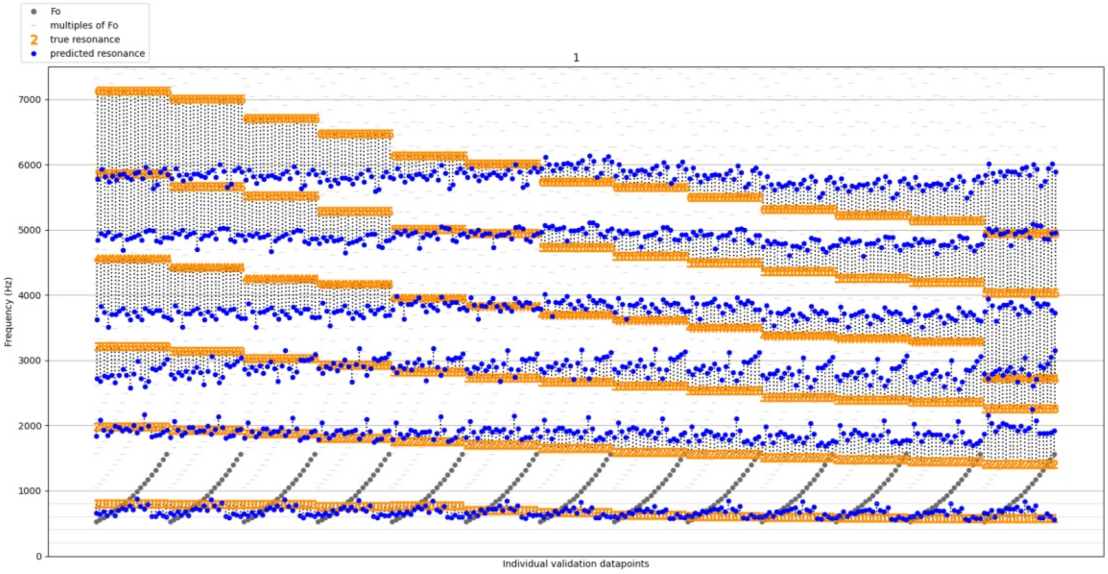


Figure 11: *Scatterplot of results for predicting fully synthesized testset 1 with a network trained on exclusively synthesized data.* f_o *ranging from 523.25 Hz to 1567.98 Hz, 10% noise and 75 ct vibrato.*

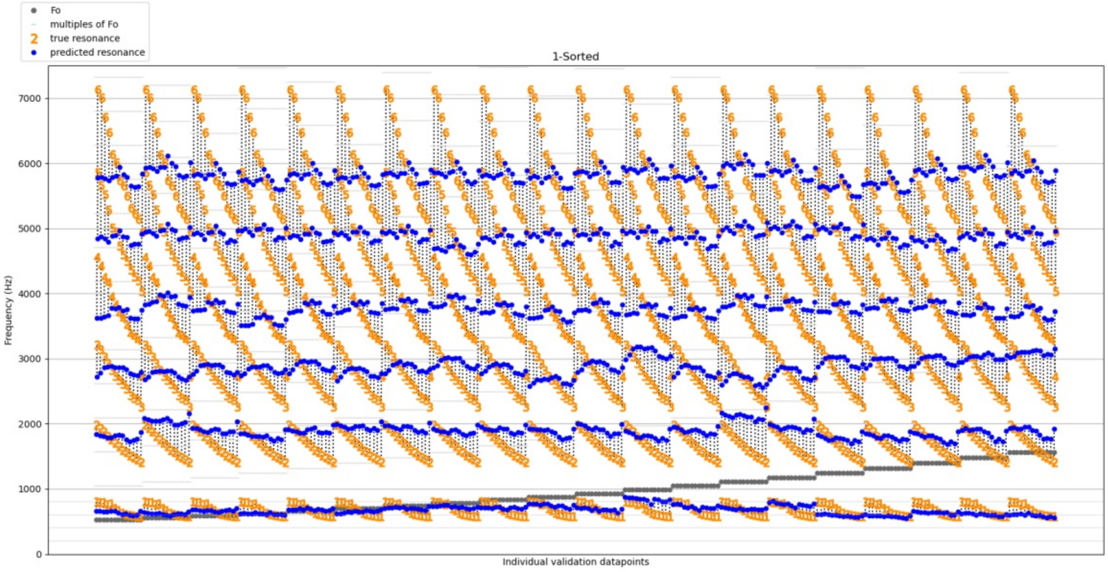


Figure 12: *Scatterplot of results for predicting fully synthesized testset 1 with a network trained on exclusively synthesized data.* f_o *ranging from 523.25 Hz to 1567.98 Hz, 10% noise and 75 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

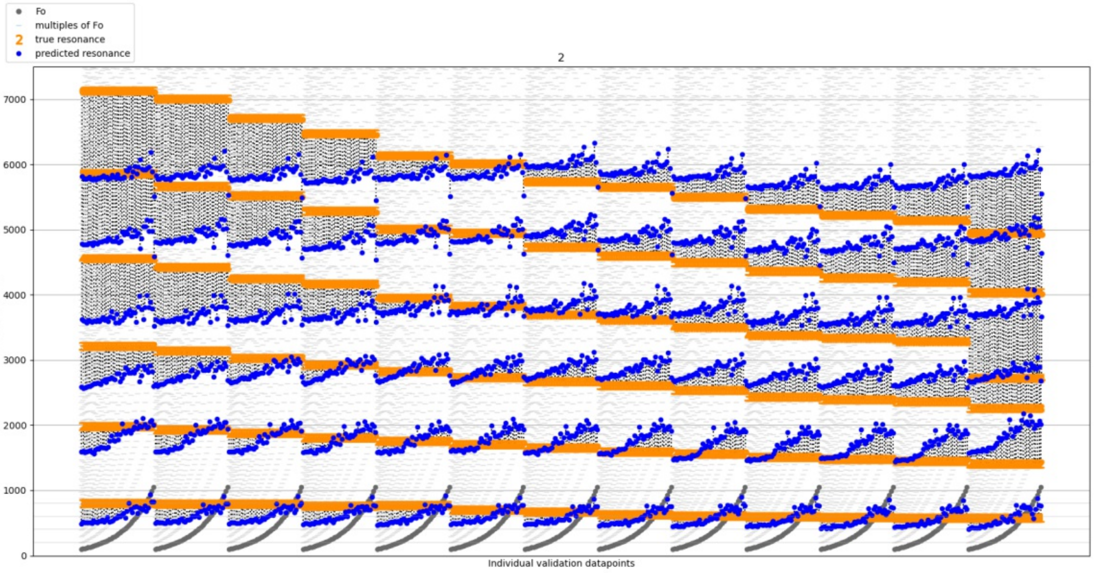


Figure 13: *Scatterplot of results for predicting fully synthesized testset 2 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 5% noise and 75 ct vibrato.*

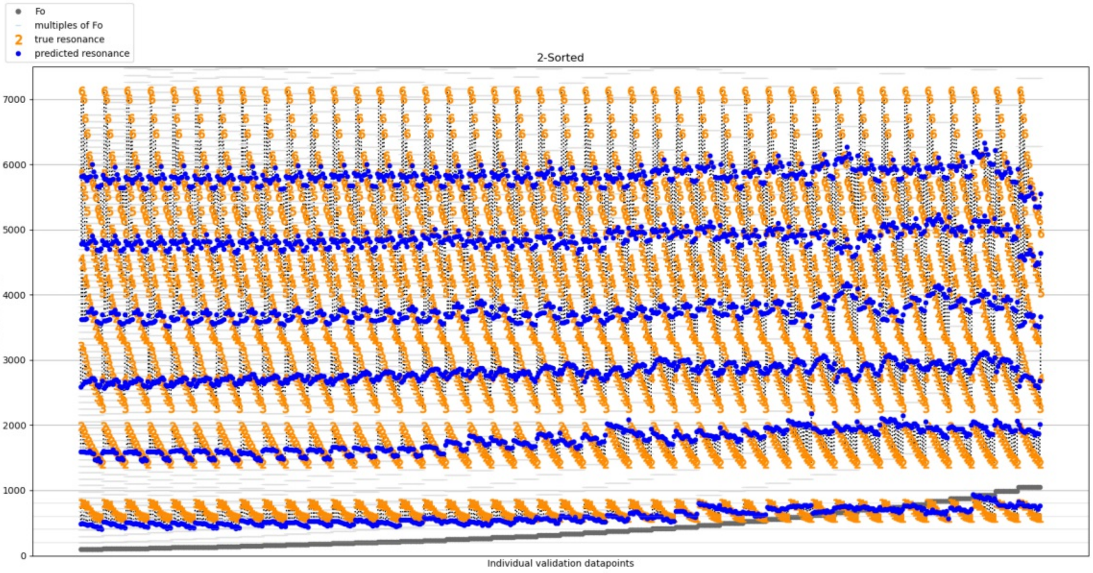


Figure 14: *Scatterplot of results for predicting fully synthesized testset 2 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 5% noise and 75 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

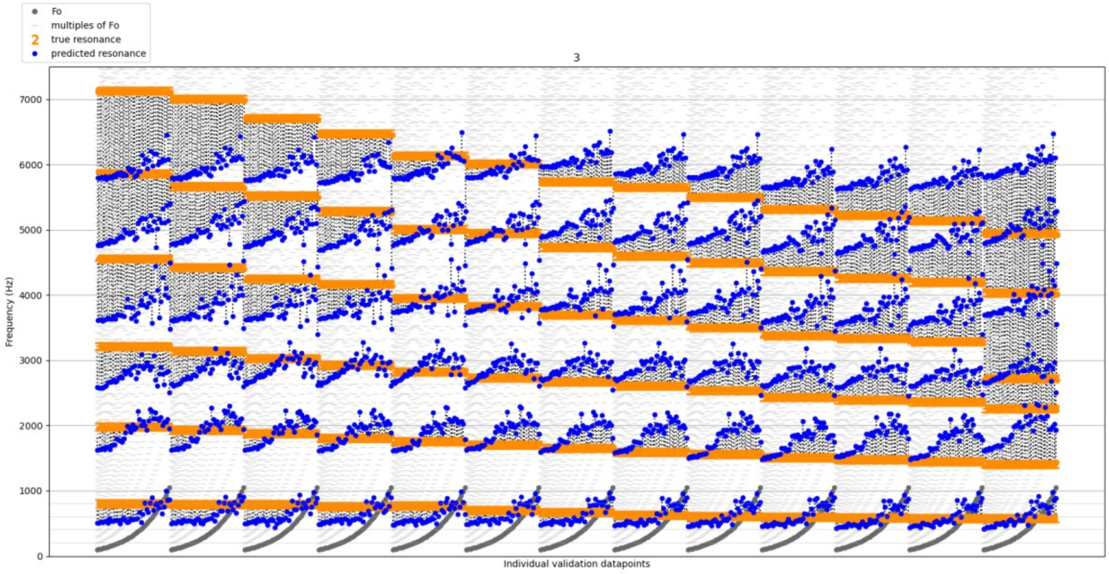


Figure 15: *Scatterplot of results for predicting fully synthesized testset 3 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 0% noise and 75 ct vibrato.*

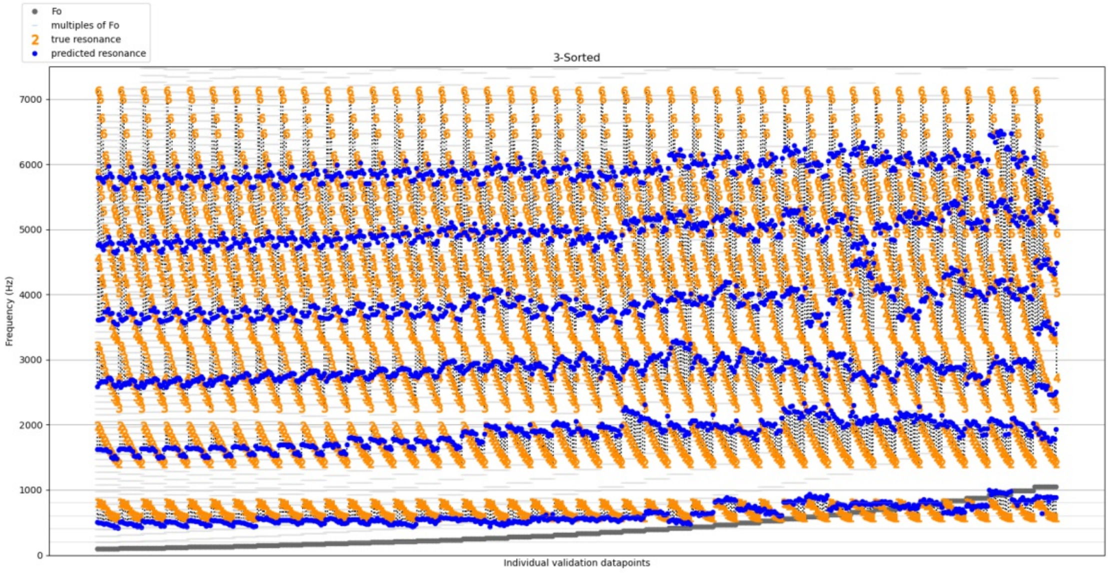


Figure 16: *Scatterplot of results for predicting fully synthesized testset 3 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 0% noise and 75 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

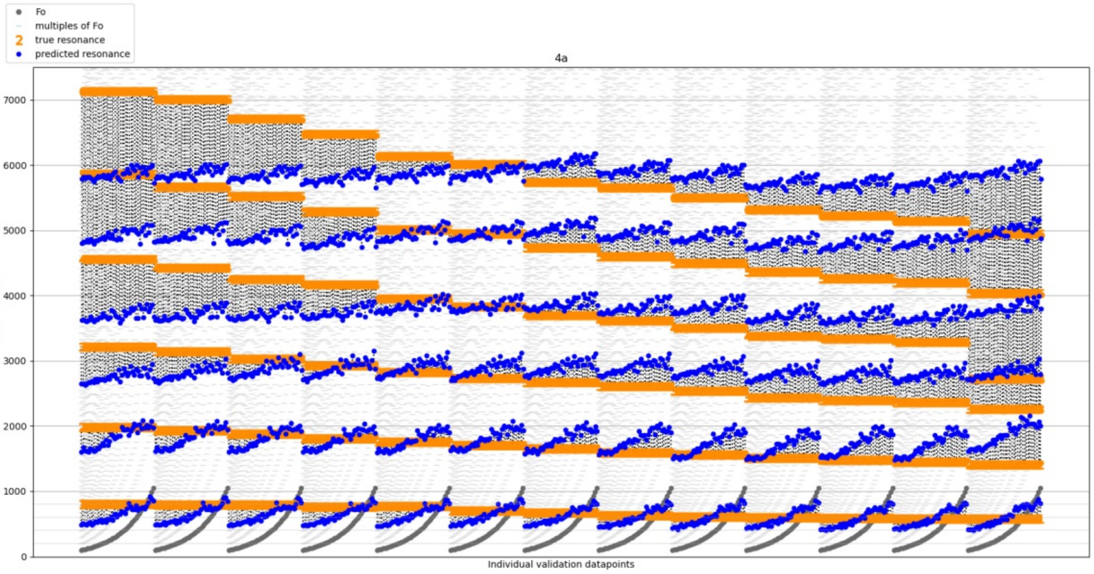


Figure 17: *Scatterplot of results for predicting fully synthesized testset 4a with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 50 ct vibrato.*

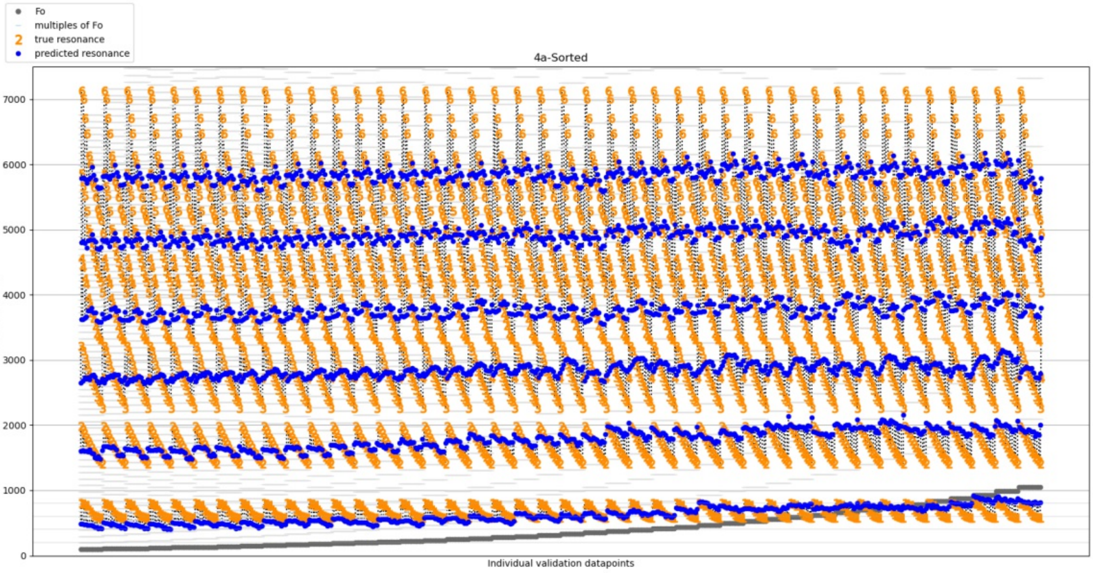


Figure 18: *Scatterplot of results for predicting fully synthesized testset 4a with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 50 ct vibrato.*

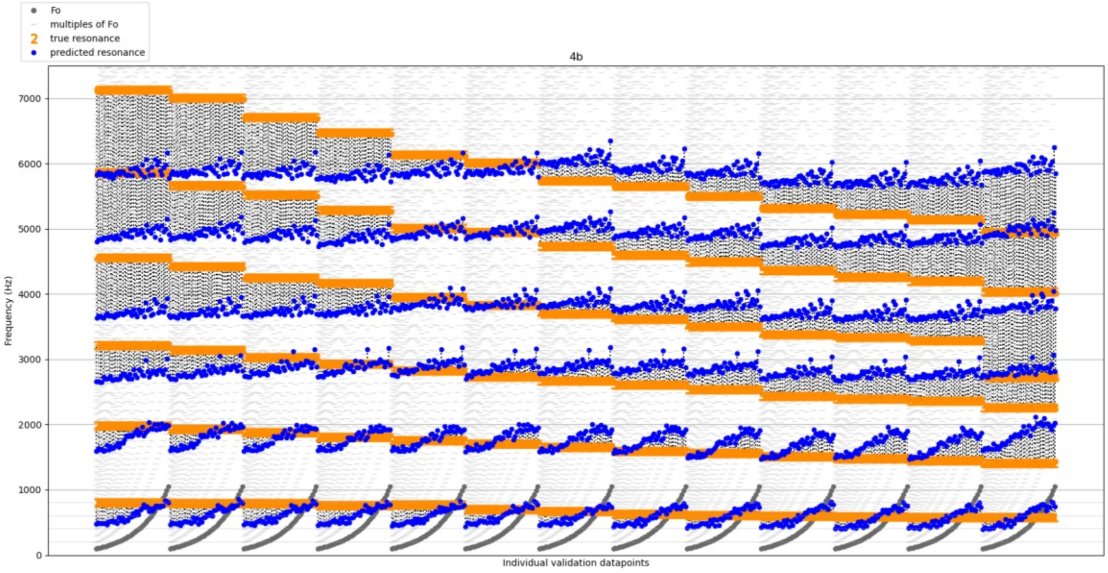


Figure 19: *Scatterplot of results for predicting fully synthesized testset 4b with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 25 ct vibrato.*

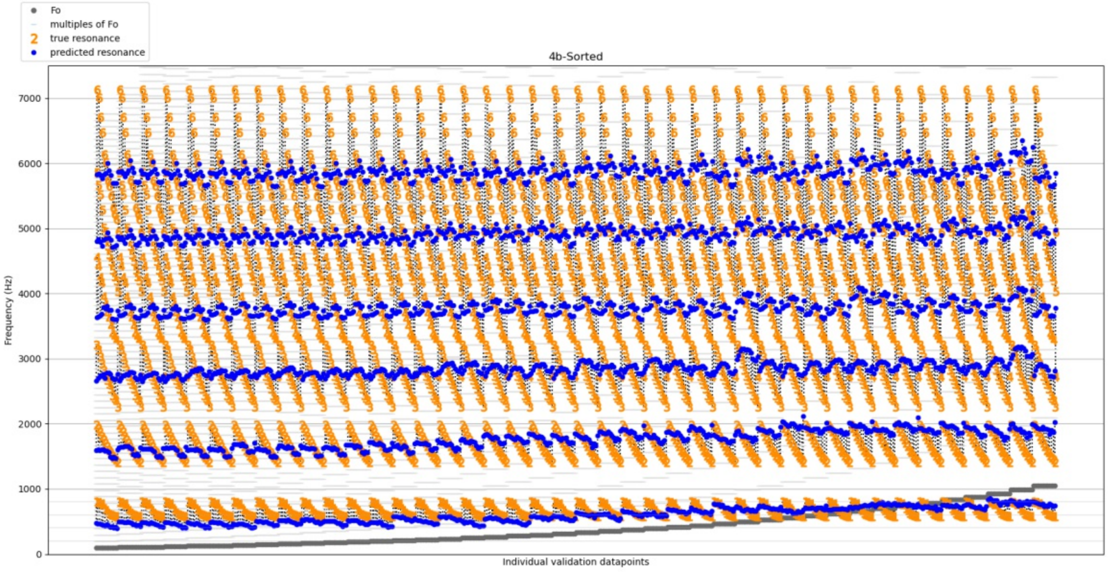


Figure 20: *Scatterplot of results for predicting fully synthesized testset 4b with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 25 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

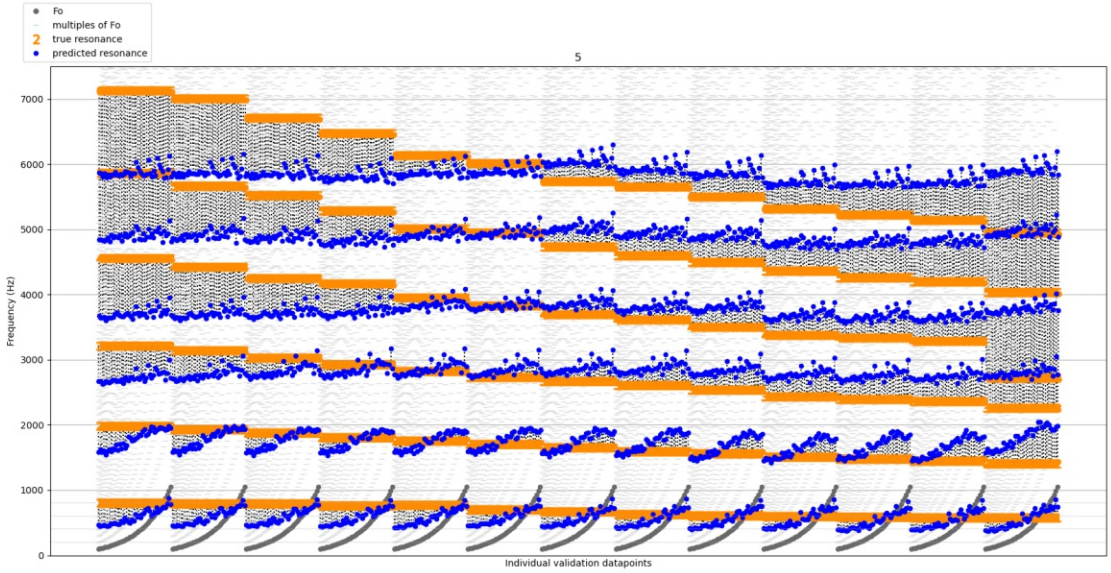


Figure 21: *Scatterplot of results for predicting fully synthesized testset 5 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 0 ct vibrato.*

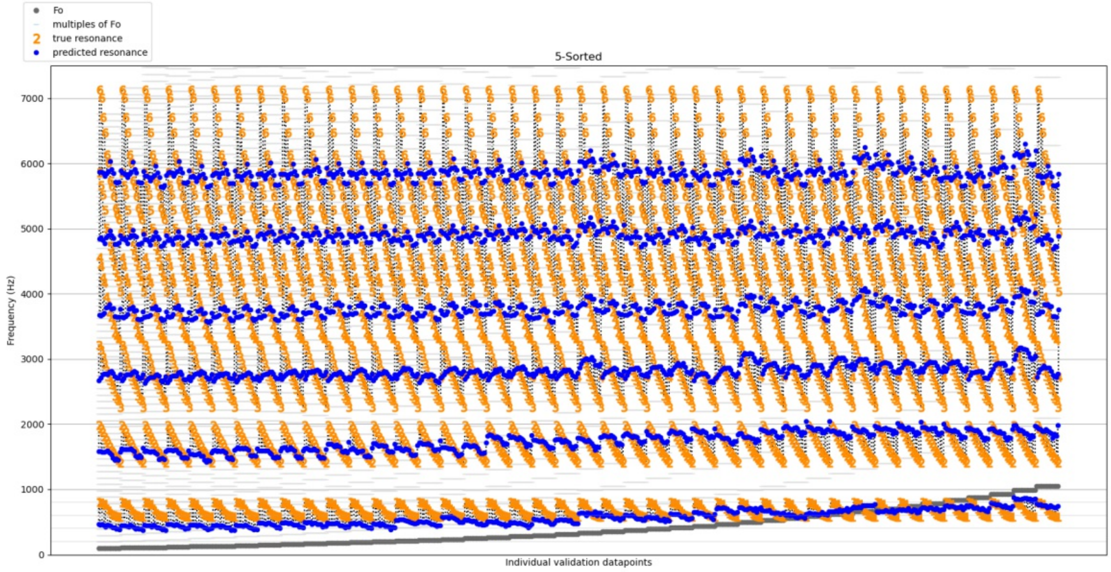


Figure 22: *Scatterplot of results for predicting fully synthesized testset 5 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 0 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

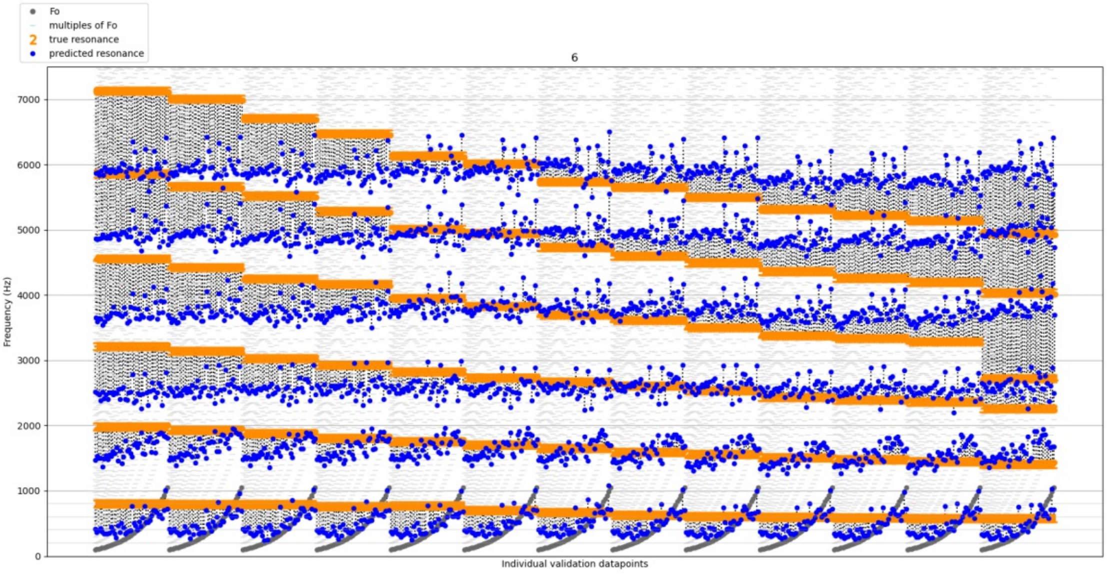


Figure 23: *Scatterplot of results for predicting fully synthesized testset 6 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 0% noise and 0 ct vibrato.*

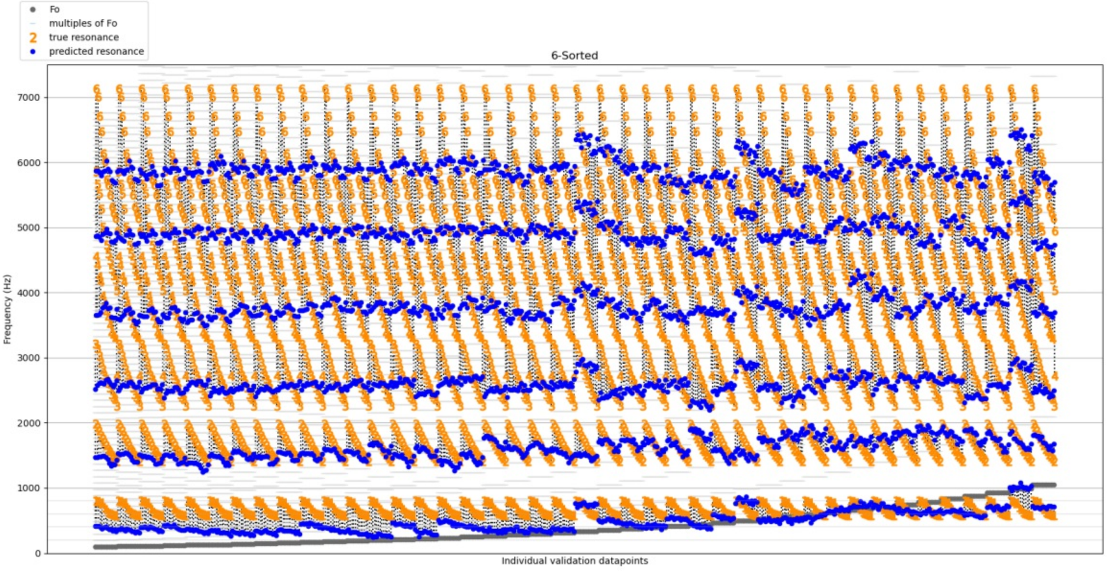


Figure 24: *Scatterplot of results for predicting fully synthesized testset 6 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 0% noise and 0 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

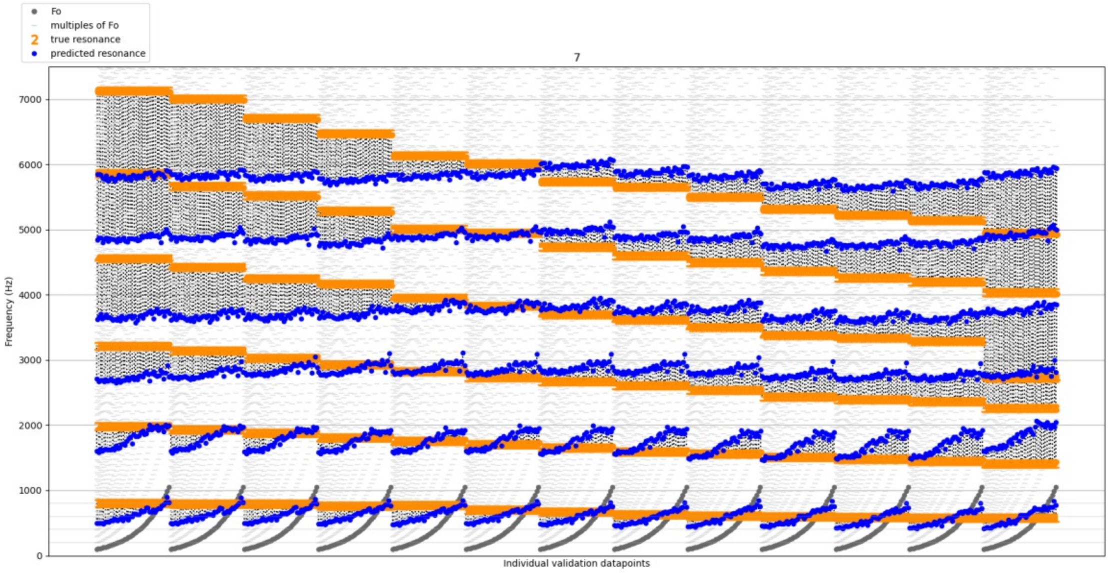


Figure 25: *Scatterplot of results for predicting fully synthesized testset 7 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 20% noise and 100 ct vibrato.*

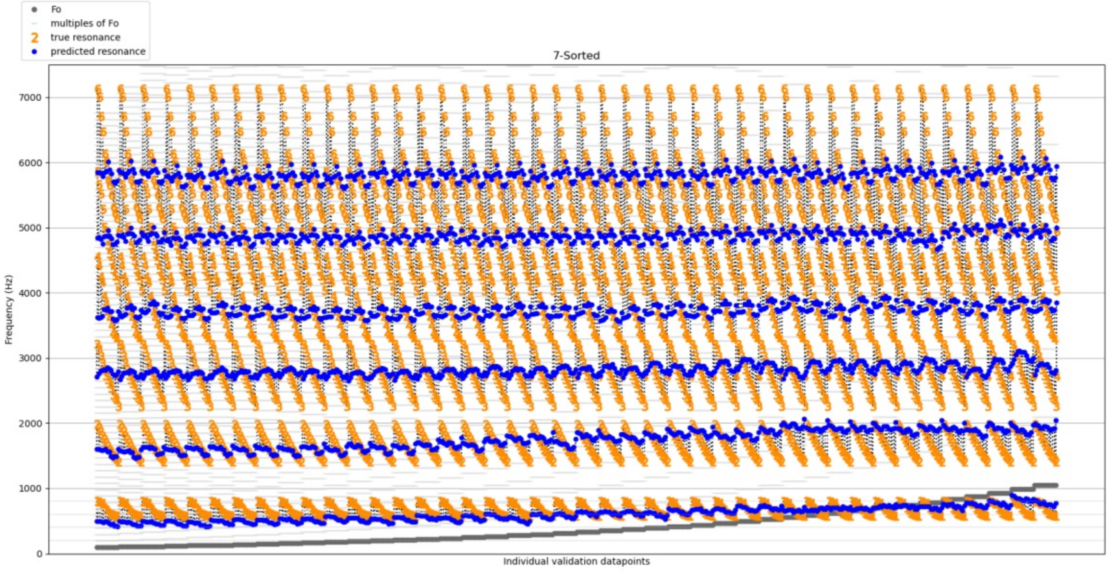


Figure 26: *Scatterplot of results for predicting fully synthesized testset 7 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 20% noise and 100 ct vibrato.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

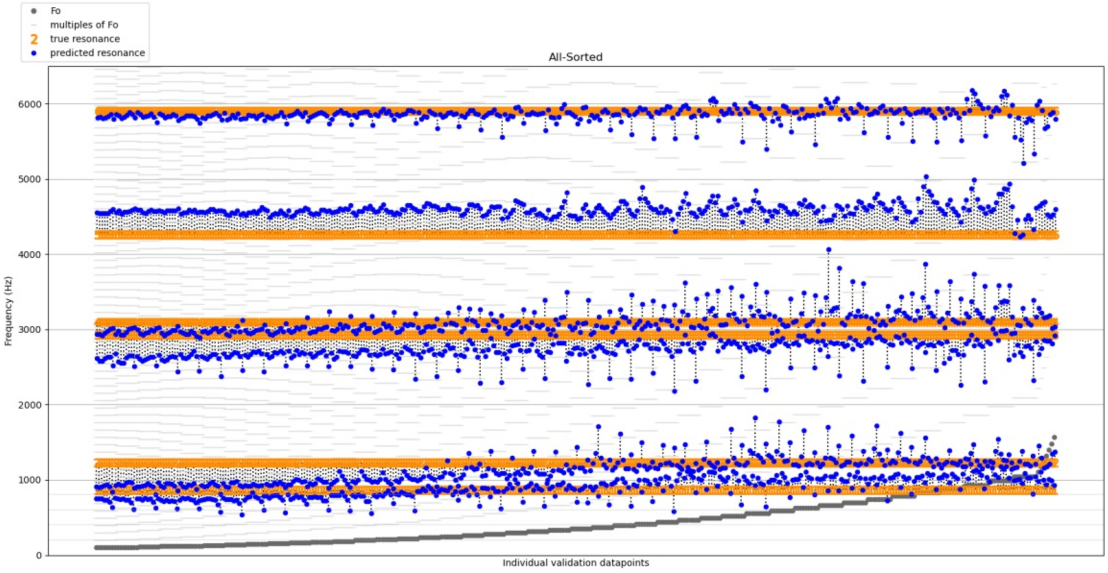


Figure 27: *Scatterplot of results for predicting the resonances of the 3D-printed vocal tract model by the network trained on synthesized data.*

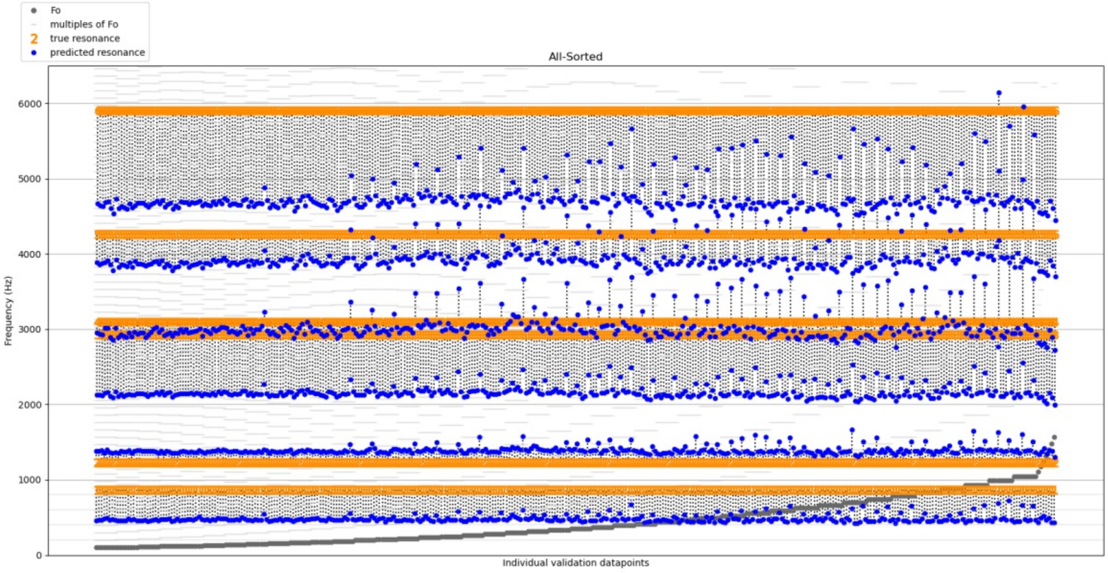


Figure 28: *Scatterplot of results for predicting the resonances of the 3D-printed vocal tract model by the network trained on recorded data.*

Resonances of lab recorded sounds

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Bibliothek Your knowledge hub

Tubelength	Source	1	2	3	4	5	6
0.130	Formula	659.615	1978.846	3298.077	4617.308	5936.538	7255.769
	Noise	807.679	1987.496	3216.931	4562.143	5857.737	7128.521
	Chirps	802.169	1976.478	3216.946	4556.651	5863.277	7136.823
0.135	Formula	635.185	1905.556	3175.926	4446.296	5716.667	6987.037
	Chirps	802.168	1932.370	3139.756	4424.327	5675.819	6938.338
	Noise	796.656	1937.886	3139.761	4424.334	5670.315	7010.020
0.140	Formula	612.500	1837.500	3062.500	4287.500	5512.500	6737.500
	Noise	791.143	1877.241	3029.497	4247.912	5526.972	6706.795
	Chirps	802.168	1877.238	3029.492	4253.418	5521.450	6695.757
0.145	Formula	591.379	1774.138	2956.897	4139.655	5322.414	6505.172
	Chirps	647.800	1800.056	2935.773	4165.214	5289.905	6464.214
	Noise	763.577	1805.569	2924.747	4165.214	5289.905	6475.241
0.150	Formula	571.667	1715.000	2858.333	4001.667	5145.000	6288.333
	Chirps	620.234	1755.951	2819.996	3950.200	5019.759	6144.449
	Noise	769.086	1755.943	2825.497	3950.182	5008.709	6138.908
0.155	Formula	553.226	1659.677	2766.129	3872.581	4979.032	6085.484
	Noise	702.928	1706.324	2737.286	3828.892	4948.065	6017.619
	Chirps	609.207	1700.819	2737.298	3839.936	4926.035	6023.159
0.160	Formula	535.938	1607.812	2679.688	3751.562	4823.438	5895.312
	Chirps	592.667	1656.711	2671.136	3696.587	4752.361	5841.214
	Noise	675.366	1656.713	2671.140	3696.593	4733.073	5736.473
0.165	Formula	519.697	1559.091	2598.485	3637.879	4677.273	5716.667
	Chirps	565.101	1585.039	2607.734	3627.672	4589.722	5664.792
	Noise	631.260	1590.555	2604.982	3619.409	4595.243	5653.775
0.170	Formula	504.412	1513.235	2522.059	3530.882	4539.706	5548.529
	Chirps	559.588	1540.934	2549.846	3520.165	4498.755	5524.206
	Noise	614.721	1562.989	2538.824	3503.632	4501.519	5504.919
0.175	Formula	490.000	1470.000	2450.000	3430.000	4410.000	5390.000
	Noise	598.181	1513.370	2434.073	3376.828	4358.176	5314.714
	Chirps	537.535	1507.855	2434.069	3382.336	4358.169	5306.436
0.180	Formula	476.389	1429.167	2381.944	3334.722	4287.500	5240.278
	Noise	592.665	1480.285	2389.957	3332.708	4258.919	5223.723
	Chirps	532.023	1474.778	2395.481	3321.696	4264.452	5234.773
0.185	Formula	463.514	1390.541	2317.568	3244.595	4171.622	5098.649
	Chirps	520.996	1452.725	2356.888	3277.591	4203.807	5146.562
	Noise	581.639	1452.719	2362.391	3283.089	4203.788	5141.026
0.190	Formula	451.316	1353.947	2256.579	3159.211	4061.842	4964.474
	Chirps	498.943	1403.104	2263.160	2726.267	4038.404	4942.566
	Noise	576.126	1408.613	2263.154	2726.260	4038.393	4948.065

Table 1: *The resonances of the different tubes, as predicted by the formula and as calculated from the noise and chirps.*

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Dependencies of the conda environment

Package	Version
_libgcc_mutex	0.1
_openmp_mutex	4.5
_tflow_select	2.1.0
absl-py	0.13.0
aiohttp	3.7.4
astor	0.8.1
astunparse	1.6.3
async-timeout	3.0.1
attrs	21.2.0
audioread	2.1.9
blas	1
blinker	1.4
brotli	1.0.9
brotlipy	0.7.0
c-ares	1.17.1
ca-certificates	2021.7.5
cachetools	4.2.2
certifi	2021.5.30
cffi	1.14.6
chardet	3.0.4
click	8.0.1
coverage	5.5
cryptography	3.4.7
cudatoolkit	10.1.243
cudnn	7.6.5
cupti	10.1.168
cycler	0.10.0
cython	0.29.24
dbus	1.13.18
decorator	5.0.9

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

expat	2.4.1
fontconfig	2.13.1
fonttools	4.25.0
freetype	2.10.4
future	0.18.2
gast	0.4.0
glib	2.69.0
google-auth	1.33.0
google-auth-oauthlib	0.4.4
google-pasta	0.2.0
grpcio	1.36.1
gst-plugins-base	1.14.0
gstreamer	1.14.0
h5py	2.10.0
hdf5	1.10.6
icu	58.2
idna	2.1
importlib-metadata	3.10.0
intel-openmp	2021.3.0
joblib	1.0.1
jpeg	9b
keras	2.4.3
keras-base	2.4.3
keras-preprocessing	1.1.2
kiwisolver	1.3.1
lcms2	2.12
ld_impl_linux-64	2.35.1
libffi	3.3
libgcc-ng	9.3.0
libgfortran-ng	7.5.0
libgfortran4	7.5.0
libgomp	9.3.0
libpng	1.6.37
libprotobuf	3.17.2
librosa	0.8.1
libstdcxx-ng	9.3.0
libtiff	4.2.0
libuuid	1.0.3
libwebp-base	1.2.0
libxcb	1.14
libxml2	2.9.12
llvmlite	0.36.0
lz4-c	1.9.3

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

markdown	3.3.4
matplotlib	3.4.2
matplotlib-base	3.4.2
mkl	2021.3.0
mklservice	2.4.0
mkl_fft	1.3.0
mkl_random	1.2.2
multidict	5.1.0
munkres	1.1.4
ncurses	6.2
numba	0.53.1
numpy	1.20.3
numpy-base	1.20.3
oauthlib	3.1.1
olefile	0.46
openjpeg	2.3.0
openssl	1.1.1k
opt_einsum	3.3.0
packaging	21
pcre	8.45
pillow	8.3.1
pip	21.2.4
pooch	1.4.0
protobuf	3.17.2
pyasn1	0.4.8
pyasn1-modules	0.2.8
pycparser	2.2
pyjwt	2.1.0
pyopenssl	20.0.1
pyparsing	2.4.7
pyqt	5.9.2
pysocks	1.7.1
python	3.9.6
python-dateutil	2.8.2
python-flatbuffers	1.12
pyyaml	5.4.1
qt	5.9.7
readline	8.1
requests	2.25.1
requests-oauthlib	1.3.0
resampy	0.2.2
rsa	4.7.2
scikit-learn	0.24.2

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

scipy	1.6.2
setuptools	52.0.0
sip	4.19.13
six	1.16.0
soundfile	0.10.3.post1
sqlite	3.36.0
tensorboard	2.4.0
tensorboard-plugin-wit	1.6.0
tensorflow	2.4.1
tensorflow-base	2.4.1
tensorflow-estimator	2.5.0
tensorflow-gpu	2.4.1
termcolor	1.1.0
threadpoolctl	2.2.0
tk	8.6.10
tornado	6.1
typing-extensions	3.10.0.0
typing_extensions	3.10.0.0
tzdata	2021a
urllib3	1.26.6
werkzeug	1.0.1
wheel	0.37.0
wrapt	1.12.1
xz	5.2.5
yaml	0.2.5
yarl	1.6.3
zipp	3.5.0
zlib	1.2.11
zstd	1.4.9

Table 2: *The dependencies of the conda environment used for training and using the network models and plotting the results.*

List of Figures

1.1 *Spectrogram of a vowel A at a frequency of 98 Hz (G2). The right part of the image shows the spectrum at the time selected by the green cursor. The vocal tract resonances are shown by the rulers and correspond with the peaks of the output spectrum.* 2
1.2 *Spectrogram of a vowel A at a frequency of 784 Hz (G5). The resulting spectrum consists only of the fundamental frequency and its harmonics. The vocal tract resonances are shown by the rulers and influence the amplitude of the harmonics, as the third harmonic has a higher amplitude than the second.* 2
1.3 *Spectrogram of a vowel A at a frequency of 784 Hz (G5) with 1% breathiness. The white noise filtered by the vocal tract follows the shape of the transfer function at a low amplitude. The peaks of this low amplitude spectrum correspond with the vocal tract resonances.* 4
1.4 *Spectrogram of a vowel A at a frequency of 784 Hz (G5) with 10 cent, 5 Hz vibrato. The amplitudes of the frequency fluctuations change according to their distance to nearby resonances. Left: crest and right: trough of the oscillation.* 4

2.1 *A schematic overview of the human body parts, contributing to voice production. Based on [36], modified by [22].* 10
2.2 *A schematic overview of the articulatory organs. From [42].* 12
2.3 *Illustration of glottal flow, spectral slope and filtering by the vocal tract by [42]. Upper left are two cycles of glottal oscillation in time, with a slower opening phase and abrupt closing phase, and zero closed phase. Upper middle is the schematic shape of a neutral vocal tract. Upper right is the output waveform of a glottal flow filtered by such a vocal tract in the time domain. Lower left, the glottal flow is presented in the frequency domain, where all harmonics, multiples of the fundamental frequencies, are represented with their intensity decreasing according to the spectral slope. Lower middle is the transfer function of the neutral vocal tract, with peaks at 500 Hz, 1500 Hz, 2500 Hz, 3500 Hz and 4500 Hz, as is the case for a theoretical 17 cm cylindrical tube. Lower right is the spectrum of the resulting output sound, shaped by both the spectral slope and the vocal tract transfer function. The peaks of the transfer function can be deduced from the output spectrum, as the harmonics are sufficiently closely spaced.* 15

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

2.4 *Vowel charts extracted from VoceVista [32]. The IPA symbols are shown in function of the frequencies of the first and second vocal tract resonance. It should be noted that the frequencies are not defined as exactly and vary with speaker sex and age [28].* 16

4.1 38
4.2 *Used playback and recording devices. Front: Microphone. Back: The set of tubes with diameter 32 mm. Middle from left to right: Pre-amplifier, sound driver, microphone interface and laptop.* 41
4.3 41
4.4 *The tube placed on top of the driver and carton with the slit, isolated with play doh.* 42
4.5 *An annotated soundfile, showing the experiment protocol. First speaking of the time and tube length, then hitting of the tube. Then the room is left and the door is closed, so the noise, chirps and test sounds can be played. Screenshot from Praat [3].* 43
4.6 *The 3D-printed vocal tract model, used in the final phase of the lab experiments.* 44
4.7 *The set-up of the vocal tract model on top of the sound driver. The model was supported by a rod and placed at a distance of 30 cm from the microphone.* 44

5.1 *The RMSE during training for the four best performing loss functions. The network trained with loss function MAE is almost stabilized after 1000 epochs of training, whereas MSE, Huber and LogCosh still show a slight decline. MAE shows strong overfitting, as the validation RMSE is double of the training RMSE.* 48
5.2 *Violinplots of the performance of networks trained with different loss functions on predicting the resonances in a testset of 200 sounds. As seen before from the training data, the performance of MSLE and cosine_similarity is significantly worse than that of the other four loss functions. Violinplots show the distribution of the data in the thickness of the violin. The black horizontal line marker is the median value, the coloured one is the mean value of the data.* 49
5.3 *Violinplots of the performance of networks trained with the four best performing loss functions on the testset. The performance is comparable with the mean and median of the absolute error all lying around 20 Hz. The performance of MAE is best, as expected from the RMSE during training, but at the price of strong overfitting. The second best is MSE.* 50
5.4 *The performance on the testset of all networks visualized with regard to the fundamental frequency of the sound. MSLE and cosine_similarity again perform worse than the others, causing them to be indiscernible in this graph.* 50
5.5 *The performance on the testset of the best four networks visualized with regard to the fundamental frequency of the sound. The performance is again comparable, with MAE slightly lower errors than the others.* 51

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

5.6 *The loss during training of the networks trained with different optimizers. Adam and RMSprop have very low training loss, but the validation loss is a lot higher. The network with Adadelta has not entirely stabilized in 1000 epoch of training. The lowest loss with a reasonable distance between training and validation data is using SGD.* 52

5.7 *The overall errors on predicting the testsets of the networks trained with different optimizers. The mean errors are comparable. SGD has the lowest AE and the unstabilized Adadelta the highest.* 52

5.8 *The loss during training of a network with 4 convolutional layers with max-pooling and 5 dense layers. Each layer has 64 nodes. In the last 200 epochs of training, the training loss decreased from* $1.98857 * 10^{-5}$ *(mean of epochs 2300 - 2309) to* $1.92012 * 10^{-5}$ *(mean of epochs 2491-2500) and the validation loss from* $2.11837 * 10^{-5}$ *to* $2.07640 * 10^{-5}$*. This decrease indicates that further training could decrease loss even more.* 53

5.9 *The loss and validation loss during training of the networks from table 5.3. The smallest networks have the lowest loss, the networks with 1 large convolutional layer and respectively 4 and 5 small dense layers have higher losses than the depicted range.* 54

5.10 *The absolute error, directional percent error and absolute percent error for predicting the testset, for the networks defined in table 5.3.* 55

5.11 *The loss and validation loss during training of the networks with 2 convolutional layers, as defined in the left column of table 5.4.* 57

5.12 *The loss and validation loss during training of the networks with 3 convolutional layers, as defined in the right column of table 5.4.* 57

5.13 *The absolute error, directional percent error and absolute percent error for predicting the testset for the networks with 2 convolutional layers, as defined in the left column of table 5.4.* 58

5.14 *The absolute error, directional percent error and absolute percent error for predicting the testset for the networks with 3 convolutional layers, as defined in the right column of table 5.4.* 58

5.15 *The loss and validation loss during training for networks with 3 or 4 convolutional layers with 32 nodes each and 4 or 5 dense layers with 64 nodes each, trained with 1500 epochs. The networks with 3 convolutional layers are fully stabilized at the end of training and have a loss and validation loss corresponding to the target values, with a sufficiently small difference. The networks with 4 convolutional layers are not yet stabilized after 1500 epochs of training and have a higher loss and validation loss.* 59

5.16 *The overall error on predicting the testset of networks with 3 or 4 convolutional layers with 32 nodes each and 4 or 5 dense layers with 64 nodes each, trained with 1500 epochs. The range of absolute and absolute percent error is lowest for network 3C32-5D64, but the mean error is comparable to network 3C32-4D64. An added convolutional layer worsens performance.* 59

5.17 *The loss and validation loss during training of networks with 3 convolutional and 4 dense layers and different numbers of nodes. The largest network, where all layers have 64 nodes, outperforms all others and the pre-set goal.* . . . 60

5.18 *The overall error on predicting the testset of the networks with 3 convolutional and 4 dense layers and different numbers of nodes. Small dense layers perform a lot worse than medium and big dense layers. The best performance on all error metrics is by the largest network with 64 nodes in each layer.* 60

5.19 *An overview of the entire testset 0 with true and predicted resonance for each of the 200 sounds. The x-axis shows the individual sounds - samples of the testset - and the y-axis shows frequency. Each sound has 6 true resonances, marked with orange numbers, symbolizing the number of the resonance. The values predicted by the network are marked with dark blue dots. The closer the dot is to the corresponding number, the lower the error. If the corresponding true and predicted resonance are spaced farther apart, a dotted line connects them. Additionally, the fundamental frequency of each individual sound is marked with a grey dot at the bottom of the plot, and the harmonics are marked with thin grey lines. In the case of this network and dataset, the true and predicted resonances are close together. There is no region where performance would be visibly worse or better.* 61

5.20 *The overall error on predicting each of the testing datasets, as defined in table 5.5. The effect on changing different parameters in the test data is treated in detail in the following paragraphs.* 62

5.21 *The error for predicting the training testsets related to the f_o of the sound.* 63

5.22 *The error for predicting the training testsets 0 and 1 related to the f_o of the sound. Testset 0 has the same structure, parameter values and frequency range as the training data, testset 1 also has the same structure and parameter values, but a higher frequency range. The network performs a lot worse at fundamental frequencies that it has not been trained on.* 64

5.23 *The overall errors for predicting the different recorded datasets by the network, trained purely on synthesized data. The performance is poor for all datasets.* 66

5.24 *The error for different fundamental frequencies. The performance is poor regardless of fundamental frequency.* 67

5.25 *Scatterplot of results for predicting recorded testset 0 with a network trained on exclusively synthesized data. f_o ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 75 ct vibrato.* 68

5.26 *Scatterplot of results for predicting synthesized testset 0 with Praat, using Burg's LPC algorithm with formant ceiling 6050 Hz. f_o ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 75 ct vibrato.* 69

5.27 *The overall absolute error, directional percent error and absolute percent error for the baseline method (Burg's algorithm implemented by Praat) and the network on both the standard, synthesized dataset as on the standard recorded dataset, filtered by tubes of 15 to 17 cm.* 69

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

5.28 *The error for different fundamental frequencies for the baseline method (Burg's algorithm implemented by Praat) and the network on both the standard, synthesized dataset as on the standard recorded dataset, filtered by tubes of 15 to 17 cm. The network with synthesized data outperforms Praat on the entire frequency range. The performance of the network on both datasets is stable throughout the frequency range, but the performance of Praat worsens steeply above the frequency range of speech, below 400 Hz.* 70

5.29 *Scatterplot of results for predicting the resonances of the 3D-printed vocal tract model by the network trained on synthesized data.* 71

5.30 *Scatterplot of results for predicting the resonances of the 3D-printed vocal tract model by the network trained on recorded data.* 72

1 *Scatterplot of results for predicting fully synthesized testset 0 with a network trained on exclusively synthesized data. f_o ranging from 65.41 Hz to 1046.50 Hz, 10% noise and 75 ct vibrato.* 79

2 *Scatterplot of results for predicting fully synthesized testset 1 with a network trained on exclusively synthesized data. f_o ranging from 554.37 Hz to 1567.98 Hz, 10% noise and 75 ct vibrato.* 80

3 *Scatterplot of results for predicting fully synthesized testset 2 with a network trained on exclusively synthesized data. f_o ranging from 65.41 Hz to 1046.50 Hz, 5% noise and 75 ct vibrato.* 80

4 *Scatterplot of results for predicting fully synthesized testset 3 with a network trained on exclusively synthesized data. f_o ranging from 65.41 Hz to 1046.50 Hz, 0% noise and 75 ct vibrato.* 81

5 *Scatterplot of results for predicting fully synthesized testset 4a with a network trained on exclusively synthesized data. f_o ranging from 65.41 Hz to 1046.50 Hz, 10% noise and 50 ct vibrato.* 81

6 *Scatterplot of results for predicting fully synthesized testset 4b with a network trained on exclusively synthesized data. f_o ranging from 65.41 Hz to 1046.50 Hz, 10% noise and 25 ct vibrato.* 82

7 *Scatterplot of results for predicting fully synthesized testset 5 with a network trained on exclusively synthesized data. f_o ranging from 65.41 Hz to 1046.50 Hz, 10% noise and 0 ct vibrato.* 82

8 *Scatterplot of results for predicting fully synthesized testset 6 with a network trained on exclusively synthesized data. f_o ranging from 65.41 Hz to 1046.50 Hz, 0% noise and 0 ct vibrato.* 83

9 *Scatterplot of results for predicting fully synthesized testset 7 with a network trained on exclusively synthesized data. f_o ranging from 65.41 Hz to 1046.50 Hz, 20% noise and 100 ct vibrato.* 83

10 *Scatterplot of results for predicting recorded testset 0 with a network trained on exclusively synthesized data. f_o ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 75 ct vibrato.* 85

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

11 *Scatterplot of results for predicting fully synthesized testset 1 with a network trained on exclusively synthesized data.* f_o *ranging from 523.25 Hz to 1567.98 Hz, 10% noise and 75 ct vibrato.* 86
12 *Scatterplot of results for predicting fully synthesized testset 1 with a network trained on exclusively synthesized data.* f_o *ranging from 523.25 Hz to 1567.98 Hz, 10% noise and 75 ct vibrato.* 86
13 *Scatterplot of results for predicting fully synthesized testset 2 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 5% noise and 75 ct vibrato.* 87
14 *Scatterplot of results for predicting fully synthesized testset 2 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 5% noise and 75 ct vibrato.* 87
15 *Scatterplot of results for predicting fully synthesized testset 3 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 0% noise and 75 ct vibrato.* 88
16 *Scatterplot of results for predicting fully synthesized testset 3 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 0% noise and 75 ct vibrato.* 88
17 *Scatterplot of results for predicting fully synthesized testset 4a with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 50 ct vibrato.* 89
18 *Scatterplot of results for predicting fully synthesized testset 4a with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 50 ct vibrato.* 89
19 *Scatterplot of results for predicting fully synthesized testset 4b with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 25 ct vibrato.* 90
20 *Scatterplot of results for predicting fully synthesized testset 4b with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 25 ct vibrato.* 90
21 *Scatterplot of results for predicting fully synthesized testset 5 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 0 ct vibrato.* 91
22 *Scatterplot of results for predicting fully synthesized testset 5 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 10% noise and 0 ct vibrato.* 91
23 *Scatterplot of results for predicting fully synthesized testset 6 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 0% noise and 0 ct vibrato.* 92
24 *Scatterplot of results for predicting fully synthesized testset 6 with a network trained on exclusively synthesized data.* f_o *ranging from 98.0 Hz to 1046.50 Hz, 0% noise and 0 ct vibrato.* 92

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

25 *Scatterplot of results for predicting fully synthesized testset 7 with a network trained on exclusively synthesized data. f_o ranging from 98.0 Hz to 1046.50 Hz, 20% noise and 100 ct vibrato.* 93
26 *Scatterplot of results for predicting fully synthesized testset 7 with a network trained on exclusively synthesized data. f_o ranging from 98.0 Hz to 1046.50 Hz, 20% noise and 100 ct vibrato.* 93
27 *Scatterplot of results for predicting the resonances of the 3D-printed vocal tract model by the network trained on synthesized data.* 94
28 *Scatterplot of results for predicting the resonances of the 3D-printed vocal tract model by the network trained on recorded data.* 94

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Tables

5.1 *The results of loss, validation loss, RMSE and validation RMSE of after training. The shown values are the mean of epochs 991 to 1000.* 48
5.2 *Optimizer results: The loss, validation loss, RMSE and validation RMSE in the last ten epochs of training. The shown values are the mean over the last 10 epochs of training.* 51
5.3 *The trained networks to explore the limits of the parameter space and the influence of the parameters. Less nodes in convolutional layers were combined with more in the dense layers and vice-versa. All networks were trained with 1000 epochs. 4D64 means 4 dense layers with 64 nodes each.* 54
5.4 *The trained networks to explore the influence of the parameters in the center of the parameter space. Less nodes in convolutional layers were combined with more in the dense layers and vice-versa. All networks were trained with 1000 epochs. 4D64 means 4 dense layers with 64 nodes each.* 56
5.5 *An overview of the parameter settings in the used testsets.* 62
5.6 *The mean absolute error, directional percent error and absolute percent error of the prediction of different datasets by the selected network.* 63
5.7 *The mean absolute error, directional percent error and absolute percent error of the prediction of different datasets by the selected network.* 66
5.8 *Mean errors on the predicting the resonances of the standard dataset by Praat and the network and on the recorded dataset by the network.* 70
5.9 *Resonances of the 3D-printed vocal tract model, extracted by playing noise through it.* 71
5.10 *Mean errors on the predicting the resonances of sounds filtered by the 3D printed vocal tract by the network trained on synthesized sounds and by the network trained on recorded sounds.* 73

1 *The resonances of the different tubes, as predicted by the formula and as calculated from the noise and chirps.* 96

2 *The dependencies of the conda environment used for training and using the network models and plotting the results.* 100

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliography

[1] Paavo Alku et al. “Formant frequency estimation of high-pitched vowels using weighted linear prediction”. In: *The Journal of the Acoustical Society of America* 134.2 (2013), pp. 1295–1313. DOI: `10.1121/1.4812756`. URL: `https://asa.scitation.org/doi/10.1121/1.4812756`.

[2] Janwillem van den Berg. “Myoelastic-Aerodynamic Theory of Voice Production”. In: *Journal of Speech and Hearing Research* 1.3 (1958), pp. 227–244. DOI: `10.1044/jshr.0103.227`. URL: `https://pubs.asha.org/doi/abs/10.1044/jshr.0103.227`.

[3] Paul Boersma and Weenink David. *Praat: doing phonetics by computer*. Version Version 6.0.46. 2021. URL: `/www.praat.org/`.

[4] Paul Boersma and David Weenink. *Praat Manual*. June 2020. URL: `https://www.fon.hum.uva.nl/praat/manual/Sound__To_Formant__burg____.html`.

[5] Donald G Childers. *Modern spectrum analysis*. IEEE Pr., 1978.

[6] Francois Chollet et al. *Keras Manual*. 2015. URL: `https://keras.io/api/`.

[7] C. G. Clopper and D. B. Pisoni. *The Nationwide Speech Project CorpusThe Nationwide Speech Project Corpus*. 2002. URL: `https://www.asc.ohio-state.edu/clopper.1/nsp/index.html`.

[8] Wang Dai et al. *Formant Tracking Using Dilated Convolutional Networks Through Dense Connection with Gating Mechanism*. 2020. arXiv: `2005.10803 [eess.AS]`.

[9] Wang Dai et al. “Gated Bilinear Networks for Vowel Formant Estimation”. In: *2020 International Conference on Asian Language Processing (IALP)*. 2020, pp. 205–209. DOI: `10.1109/IALP51396.2020.9310481`.

[10] Li Deng et al. “Adaptive Kalman Filtering and Smoothing for Tracking Vocal Tract Resonances Using a Continuous-Valued Hidden Dynamic Model”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.1 (2007), pp. 13–23. DOI: `10.1109/TASL.2006.876724`.

[11] Y. Dissen and J. Keshet. *DeepFormants*. 2016. URL: `https://github.com/MLSpeech/DeepFormants`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

[12] Yehoshua Dissen, Jacob Goldberger, and Joseph Keshet. "Formant estimation and tracking: A deep learning approach". In: *The Journal of the Acoustical Society of America* 145 (Feb. 2019), pp. 642–653. DOI: `10.1121/1.5088048`.

[13] Yehoshua Dissen and Joseph Keshet. "Formant Estimation and Tracking Using Deep Learning". In: *Interspeech 2016*. 2016, pp. 958–962. DOI: `10.21437/Interspeech.2016-490`. URL: `http://dx.doi.org/10.21437/Interspeech.2016-490`.

[14] E. M. Dogo et al. "A Comparative Analysis of Gradient Descent-Based Optimization Algorithms on Convolutional Neural Networks". In: *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*. 2018, pp. 92–99. DOI: `10.1109/CTEMS.2018.8769211`.

[15] Thomas Drugman and Yannis Stylianou. "Fast Inter-Harmonic Reconstruction for Spectral Envelope Estimation in High-Pitched Voices". In: *IEEE Signal Processing Letters* 21.11 (2014), pp. 1418–1422. DOI: `10.1109/LSP.2014.2338399`.

[16] J. S. Garofolo et al. *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1*. NASA STI/Recon Technical Report N. Feb. 1993.

[17] J. S. Garofolo et al. *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. 1993. DOI: `https://doi.org/10.35111/17gk-bn40`.

[18] Douglas C. Giancoli. *Physics for Scientists and Engineers with Modern Physics*. Pearson Education Limited, 2014. ISBN: 978-1-78434-293-7.

[19] A. Gray and D. Wong. "The Burg algorithm for LPC speech analysis and synthesis". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.6 (1980), pp. 609–615. DOI: `10.1109/TASSP.1980.1163489`.

[20] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: `1512.03385 [cs.CV]`.

[21] Christian Herbst and Jan Svec. "Chapter-06 Basics of Voice Acoustics–A Tutorial". In: Jan. 2016, pp. 63–80. ISBN: 9789351524571. DOI: `10.5005/jp/books/12711_7`.

[22] Christian T. Herbst. "A Review of Singing Voice Subsystem Interactions—Toward an Extended Physiological Model of "Support"". In: *Journal of Voice* 31.2 (2017), 249.e13–249.e19. ISSN: 0892-1997. DOI: `https://doi.org/10.1016/j.jvoice.2016.07.019`. URL: `https://www.sciencedirect.com/science/article/pii/S0892199716302922`.

[23] Christian T. Herbst. "Registers - The Snake Pit of Voice Pedagogy". In: *Journal of Singing* 77.2 (Dec. 2020), pp. 175–190. URL: `https://www.nats.org/_Library/JOS_On_Point/JOS-077-02-2020-175.pdf`.

[24] Christian T. Herbst, David M. Howard, and Jan G. Švec. "The Sound Source in Singing". In: *The Oxford Handbook of Singing* (2015), pp. 108–144. DOI: `10.1093/oxfordhb/9780199660773.013.011`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[25] Christian T. Herbst et al. "Membranous and cartilaginous vocal fold adduction in singing". In: *The Journal of the Acoustical Society of America* 129.4 (2011), pp. 2253–2262. DOI: `10.1121/1.3552874`.

[26] J. Hillenbrand et al. "Acoustic characteristics of American English vowels." In: *The Journal of the Acoustical Society of America* 97 5 Pt 1 (1995), pp. 3099–3111.

[27] Geoffrey Hinton et al. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 82–97. DOI: `10.1109/MSP.2012.2205597`.

[28] Raymond D. Kent and Houri K. Vorperian. "Static measurements of vowel formant frequencies and bandwidths: A review". In: *Journal of Communication Disorders* 74 (2018), pp. 74–97. DOI: `10.1016/j.jcomdis.2018.05.004`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6002811/`.

[29] Liu Liqing and Tetsuya Shimamura. "Pitch-Synchronous Linear Prediction Analysis of High-Pitched Speech Using Weighted Short-Time Energy Function". In: *Journal of Signal Processing* 19 (Mar. 2015), pp. 55–66. DOI: `10.2299/jsp.19.55`.

[30] Qing Liu et al. "A Review of Image Recognition with Deep Convolutional Neural Network". In: *Intelligent Computing Theories and Application*. Ed. by De-Shuang Huang et al. Cham: Springer International Publishing, 2017, pp. 69–80. ISBN: 978-3-319-63309-1.

[31] Deepali Y. Loni and Shaila Subbaraman. "Formant estimation of speech and singing voice by combining wavelet with LPC and Cepstrum techniques". In: *2014 9th International Conference on Industrial and Information Systems (ICIIS)*. 2014, pp. 1–7. DOI: `10.1109/ICIINFS.2014.7036530`.

[32] Bodo Maass. *VoceVista Video / Overtone Analyzer. User Manual and Reference Guide*. Sygyt Software. Bochum, Germany. URL: `https://www.sygyt.com/en/documentation/`.

[33] Paul H. Milenkovic et al. "Effects of sampling rate and type of anti-aliasing filter on linear-predictive estimates of formant frequencies in men, women, and children". In: *The Journal of the Acoustical Society of America* 147.3 (2020), EL221–EL227. DOI: `10.1121/10.0000824`. URL: `https://asa.scitation.org/doi/full/10.1121/10.0000824`.

[34] Meinard Müller. *Fundamentals of Music Processing*. Springer International Publishing, 2015. ISBN: 978-3-319-21945-5. DOI: `10.1007/978-3-319-21945-5`.

[35] Douglas O'Shaughnessy. "Formant Estimation and Tracking". In: *Springer Handbook of Speech Processing*. Ed. by Jacob Benesty, M. Mohan Sondhi, and Yiteng Arden Huang. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 213–228. ISBN: 978-3-540-49127-9. DOI: `10.1007/978-3-540-49127-9_11`. URL: `https://doi.org/10.1007/978-3-540-49127-9_11`.

[36] Thomas D Rossing, Paul Wheeler, and F. Richard Moore. *The science of sound*. 1st ed. Addison Wesley, 2002.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[37] Sebastian Ruder. *An overview of gradient descent optimization algorithms.* 2017. arXiv: `1609.04747 [cs.LG]`.

[38] Vasantha Sama Sai et al. “Estimation of Vocal Tract Resonances Using Spectral Prominent Regions and Artificial Neural Networks”. In: *Circuits, Systems, and Signal Processing* 37.11 (2018), pp. 5087–5100. DOI: `10.1007/s00034-018-0808-6`. URL: `https://link.springer.com/article/10.1007/s00034-018-0808-6`.

[39] Belle A Shenoi. *Introduction to digital signal processing and filter design.* John Wiley & Sons, Inc., 2006. ISBN: 13 978-0-471-46482-2.

[40] Dongpeng Song et al. “FPGA implementation of covariance lattice LPC method using burg algorithm”. In: Nov. 2017, pp. 308–312. DOI: `10.1109/ICAIT.2017.8388936`.

[41] Brad Story. “Technique for "tuning" vocal tract area functions based on acoustic sensitivity functions (L)”. In: *The Journal of the Acoustical Society of America* 119 (Mar. 2006), pp. 715–8. DOI: `10.1121/1.2151802`.

[42] Brad Story. “The Vocal Tract in Singing”. In: Jan. 2016. DOI: `10.1093/oxfordhb/9780199660773.013.012`.

[43] Brad Story, Anne-Maria Laukkanen, and Ingo Titze. “Acoustic impedance of an artificially lengthened and constricted vocal tract”. In: *Journal of voice : official journal of the Voice Foundation* 14 (Dec. 2000), pp. 455–69. DOI: `10.1016/S0892-1997(00)80003-X`.

[44] Brad H. Story and Kate Bunton. “Formant measurement in children’s speech based on spectral filtering”. In: *Speech Communication* 76 (2016), pp. 93–111. DOI: `10.1016/j.specom.2015.11.001`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4743040/`.

[45] R. Sudharshan and C.S. Ramalingam. “A Data-Driven Weighted LP Method for Formant Estimation”. In: *2020 IEEE 4th Conference on Information Communication Technology (CICT)*. 2020, pp. 1–6. DOI: `10.1109/CICT51604.2020.9312110`.

[46] Johan Sundberg. “Level and Center Frequency of the Singer’s Formant”. In: *Journal of Voice* 15.2 (2001), pp. 176–186. ISSN: 0892-1997. DOI: `https://doi.org/10.1016/S0892-1997(01)00019-4`. URL: `https://www.sciencedirect.com/science/article/pii/S0892199701000194`.

[47] Jan G. Švec et al. “Integrative Insights into the Myoelastic-Aerodynamic Theory and Acoustics of Phonation. Scientific Tribute to Donald G. Miller”. In: *Journal of Voice* (2021). ISSN: 0892-1997. DOI: `https://doi.org/10.1016/j.jvoice.2021.01.023`. URL: `https://www.sciencedirect.com/science/article/pii/S0892199721000552`.

[48] Ingo Titze. “Nonlinear source-filter coupling in phonation: Theory”. In: *The Journal of the Acoustical Society of America* 123 (June 2008), pp. 2733–49. DOI: `10.1121/1.2832337`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub

[49] Ingo Titze et al. “Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization”. In: *The Journal of the Acoustical Society of America* 137 (May 2015), p. 3005. DOI: `10.1121/1.4919349`.

[50] Ingo R. Titze. *Principles of Voice Production*. 2nd ed. National Center of Voice and Speech, 2000.

[51] Boukaye Boubacar Traore, Bernard Kamsu-Foguem, and Fana Tangara. “Deep convolution neural network for image recognition”. In: *Ecological Informatics* 48 (2018), pp. 257–268. ISSN: 1574-9541. DOI: `https://doi.org/10.1016/j.ecoinf.2018.10.002`. URL: `https://www.sciencedirect.com/science/article/pii/S1574954118302140`.

[52] Tianyu T. Wang and Thomas F. Quatieri. “High-Pitch Formant Estimation by Exploiting Temporal Change of Pitch”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.1 (2010), pp. 171–186. DOI: `10.1109/TASL.2009.2024732`.

[53] B. Yegnanarayana, Anand Joseph, and Vishala Pannala. “Enhancing Formant Information in Spectrographic Display of Speech”. In: *Proc. Interspeech 2020*. 2020, pp. 165–169. DOI: `10.21437/Interspeech.2020-2653`. URL: `http://dx.doi.org/10.21437/Interspeech.2020-2653`.

[54] Dong Yu and Li Deng. *Automatic speech recognition*. 1st ed. Springer, 2015. ISBN: 978-1-4471-5779-3. DOI: `https://doi.org/10.1007/978-1-4471-5779-3`.

[55] Christos Zarras et al. “Cepstrum-based estimation of resonance frequencies (formants) in high-pitch singing signals”. In: *Fortschritte der Akustik - DAGA 2010* (Berlin, Germany). Ed. by Michael Möser, Brigitte Schulte-Fortkamp, and Martin Ochmann. Deutsche Gesellschaft für Akustik e.V. (DEGA), Berlin, 2010, 2010, pp. 661–662. ISBN: ISBN: 978-3-9808659-8-2. URL: `https://pub.dega-akustik.de/DAGA%5C_2010/data/articles/000360.pdf`.

[56] Emilia Zawadzka-Gosk, Krzysztof Wołk, and Wojciech Czarnowski. “Deep Learning in State-of-the-Art Image Classification Exceeding 99% Accuracy”. In: *New Knowledge in Information Systems and Technologies*. Ed. by Álvaro Rocha et al. Cham: Springer International Publishing, 2019, pp. 946–957. ISBN: 978-3-030-16181-1.

[57] Jiajia Zhang, Kun Shao, and Xing Luo. “Small sample image recognition using improved Convolutional Neural Network”. In: *Journal of Visual Communication and Image Representation* 55 (2018), pp. 640–647. ISSN: 1047-3203. DOI: `https://doi.org/10.1016/j.jvcir.2018.07.011`. URL: `https://www.sciencedirect.com/science/article/pii/S1047320318301810`.

[58] Zhao Zhang, Kiyoshi Honda, and Jianguo Wei. “Retrieving Vocal-Tract Resonance and anti-Resonance From High-Pitched Vowels Using a Rahmonic Subtraction Technique”. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 7359–7363. DOI: `10.1109/ICASSP40776.2020.9054741`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU WIEN Bibliothek Your knowledge hub