Perceived Breathiness, Pressedness, and Vocal Fry in Synthetic Voice Stimuli

Bachelor Thesis

Simon Windtner

Supervisor: o.Univ.Prof. Mag.art DI Dr.techn. Robert Höldrich
Supervisor: Univ.Ass. DI Dr.techn. Philipp Aichinger, Medical University of Vienna

Graz, November 4, 2021

institut für elektronische musik und akustik

for dad
∗ 1967 ∼ † 2020

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Acknowledgements

First of all, I would like to thank my professor Robert Höldrich for his patience. Finally, the time has come to finish my thesis after nearly four terms during the corona pandemic including a change of the topic as well. Many thanks for all the instructive conversations and all the questions you asked me in order to scrutinise my own thinking.

I would also like to thank Philipp Aichinger for all the conversations we had, the input for the listening experiments, all the basics and details about speech in physiological terms, in perceptional terms: Thanks for sharing your knowledge in this field with me.

My gratitude also goes to Dr. Jean Schoentgen, who always had an open ear for questions regarding the synthesizer as well as the tutorials he gave to me.

Many thanks also to Dr. Florian Wendt for the support regarding the statistical analysis of the listening experiments. In addition, I would like to thank DI Johannes Zmölnig for his support in hosting the listening experiments.

To my family, my mum Birgit, my brothers Matthias and Elias: Thank you for your love, support, and unwavering belief in me and also for having you in the good and even in the hard times we had the past year. Without you, I would not be the person I am today.

Abstract

In the present thesis, different voice types, i.e., breathy and pressed voice as well as vocal fry are generated by means of an existing model based synthesizer. The determination of the voice type is important for clinical health care, because it affects the indication, selection, evaluation, and optimisation of clinical treatment techniques. The voice samples are generated by using different parameters and values. In two listening experiments, the samples are assessed by experts and students. The listening experiments provide information on whether the chosen parameters can be considered representative and whether they were perceived as natural for the respective voice type. The thesis shows which parameter can be representative for pressed voice, breathy voice as well as vocal fry. Furthermore, parameters have been achieved in order to generate the aforementioned voice types. Results from the listening experiments showed that the opening quotient is highly significant for the perception of the respective voice type. Furthermore, an increase of the fundamental frequency leads to a weak perceived strength of the vocal fry.

In der vorliegenden Arbeit werden verschiedene Stimmtypen, behauchte und gepresste Stimme sowie Vocal Fry, mit Hilfe eines bestehenden modellbasierten Synthesizers erzeugt. Die Klassifizierung eines individuellen Stimmtyps ist für die klinische Gesundheitsversorgung sehr wichtig. Der Stimmtyp, in Verbindung mit möglichen Stimmkrankheiten, beeinflusst die Indikation, Auswahl, Bewertung und Optimierung klinischer Behandlungstechniken. Die Stimmmuster werden unter Verwendung verschiedener Parameter und Werte erzeugt. In zwei Hörversuchen wurden die Sprachmuster von Experten:innen und Studenten:innen bewertet. Die Hörversuche gaben Aufschluss darüber, ob die verwendeten Parameter als repräsentativ angesehen werden können und für den jeweiligen Stimmtyp natürlich wahrgenommen werden. Die Arbeit zeigt, welche Parameter für gepresste Stimme, behauchte Stimme sowie Vocal Fry repräsentativ sind. Darüber hinaus wurden Parameter gefunden mithilfe dessen die zuvor genannten Stimmtypen generiert werden können. Die Ergebnisse der Hörversuche zeigten, dass der Öffnungsquotient höchst signifikant für die Wahrnehmung des jeweiligen Stimmtyps ist. Darüber hinaus führt eine Erhöhung der Grundfrequenz zu einer schwächer wahrgenommenen Stärke des Vocal frys.

Contents

List of Figures

List of Tables

1 Introduction

The human voice is undoubtedly one of the most important means for human communication and interaction. Voice can not only transmit information from one human to another human. Furthermore voice is used to express emotions, feelings as well as meanings.

Another aspect regarding the human voice is in fact, that every person has it's own sound. Thus, voice indeed characterises every single human. Through voice it is even possible to recognise with whom we are talking, although we do not see the face of the person we are talking with.

Within the technical development of the past 30 years even electronic devices (e.g. mobile phone, notebook, tablets) have the ability to communicate via speech with humans. While the first synthetically generated voices sounded very artificial, nowadays they sound quite similar to a normal human voice, and even the gender of the voice can be selected. Actual scientific approaches use neural networks, e.g. WaveNet [vdODZ+16]

The determination of the voice type is important for clinical health care. Depending on the voice type, in association with possible voice disorders, it affects the indication, selection, evaluation, and optimisation of clinical treatment techniques. [WRD+21] [AV96] Under those circumstances models and tools are developed to get a better understanding of physical and physiological reasons of such disorders.

In the present thesis, different voice types, breathy and pressed voice as well as vocal fry are generated by means of an existing model based synthesizer. The voice samples are generated by using different parameters and values. In a listening experiment, the samples are assessed by experts and students. The listening test should provide information on whether the used parameters can be considered representative and perceive naturally for the respective voice type.

In order to have a common understanding of the voice types the current thesis focuses on, chapter 2 gives detailed definition and explanation for the used voice types. In chapter 3, the used Synthesizer is expounded in his general structure as well as the used parameters. The parameters are introduced in their technical - physical meaning. Furthermore the physiological aspects are pointed out. For the listening experiments and their evaluation the used statistical routines are shown in chapter 4. The precise setup of the listening experiments are documented in chapter 5. In detail, the used listening experiment environment as well as the used parameters and values for the generated voice samples are shown. Furthermore the methodological approach to evaluate the listening experiments is explained in detail. All observations from the listening experiments are also summarised in chapter 5. The last chapter 6 gives a conclusion about the whole thesis and points out further research topics based on to the lessons learned from this thesis.

2 Voice types

Voice types and voice quality are termini which have a wide range of possible meanings. [Abe67] Thus, this thesis will focus on pressed voice, breathy voice, modal voice as well as vocal fry. In order to get a better understanding of before mentioned voice types, the following section will explain the difference between that voice types in detail.

2.1 Pressed voice

Pressed voice is characterised by a long period where the vocal folds are closed and just a short open period. Thus, the open quotient (Q_o), which is explained in detail in chapter 3.2.4, is small ($Q_o \to 0$). From the perceptual point of view pressed voice is characterised by its amount of the adduction of the vocal folds. Furthermore, pressed voice is rich in harmonics from a spectral point of view (compare figure 9). [BTS04] Figure 1 shows a high-speed digital imaging (HSDI) of the vocal folds vibrations from a female speaker with recording speed of 2000 frames per second (fps). By having a closer look on what is shown in figure 1 several properties of pressed voice can be determined. First, one entire vocal fold cycle consists of 9 image frames. Therefore, the cycle time $T = 4.5ms$ and the fundamental frequency with $f0 = 222Hz$. When focusing whether the vocal folds are opened or closed, it can be observed, that in 2 frames the vocal folds are slightly opened and closed in the remaining 7 frames. These observations lead to an Q_o of 0.26. According to Bergan et al. the Q_o range for pressed voice is < 0.4. [BTS04]

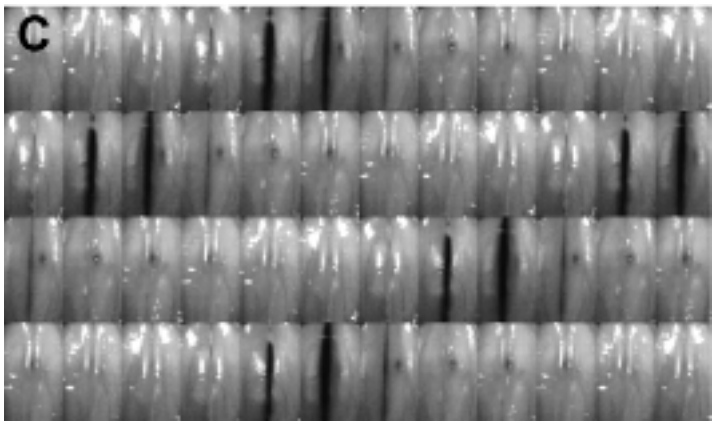


Figure 1 – High speed digital imaging of female pressed voice in sustained vowel /i/ [AYB12].

2.2 Modal voice

Modal voice is the transition between pressed voice and breathy voice and colloquially known as normal voice. This voice type can be characterised by a nearly equal duration of closed and opened vocal folds. Again, figure 2 shows a high-speed digital imaging

of the vocal folds vibrations from a female speaker with recording speed of 2000 fps. Comparison of figure 1 and figure 2 shows that for a modal voice type the duration of the opening phase of the vocal folds is much longer. In 4 frames (frame 9,10,16,17) the vocal folds are nearly closed. From frame 11 it can be observed that the vocal folds start to open and reach the maximum opening position in frame 13. Afterwards, frame 14 to 17 shows the closing phase of the vocal folds. Different experiments in the literature ended up with results for the Q_o for modal voice from $0.6 \sim 0.75$. [AYB12], [CL91] In fact that one entire cycle can be recognised with 9 frames leads to the same cycle time $T = 4.5ms$ and a fundamental frequency $f0 = 222Hz$ as for pressed voice. According to Bergan et al. the Q_o range for modal voice is from 0.4 to 0.7. [BTS04]

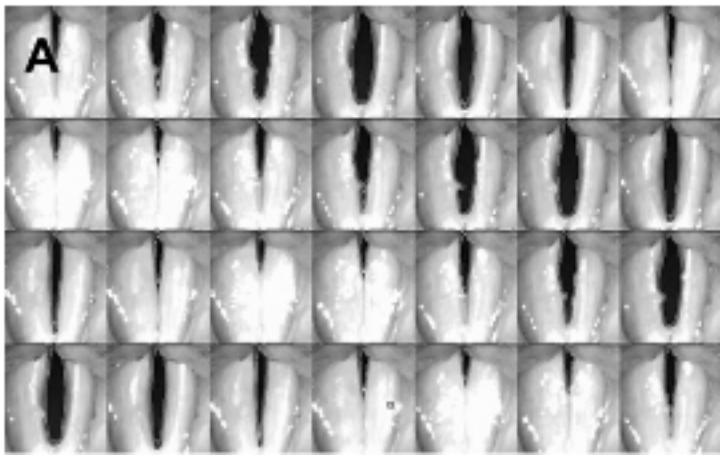


Figure 2 – High speed digital imaging of female modal voice in sustained vowel /i/ [AYB12].

2.3 Breathy voice

Breathy voice is characterised by a long open period of the vocal folds compared to the entire cycle in a physiological meaning. Consequently, the open Quotient tends towards 1 ($Q_o \rightarrow 1$). Breathiness can also be perceived as a superimposed noise of the air flow, which is the difference between modal voice and breathy voice especially when they have similar Q_o. Figure 3 shows a high-speed digital imaging of the vocal folds vibrations from a female speaker with recording speed of 2000 fps. By detailed analysis of figure 3 a lower fundamental frequency can be observed. Counting the frames of one entire cycle the cycle time $T = 6ms$, which leads to a fundamental frequency of $f0 = 166Hz$. The single frames illustrate the long open period of the vocal folds (frame 15 to 22). These observations lead to an Q_o of 0.83. It can also be observed that the vocal folds are widely opened compared to pressed voice and modal voice. According to Bergan et al. the Q_o range for breathy voice is > 0.7. [BTS04]

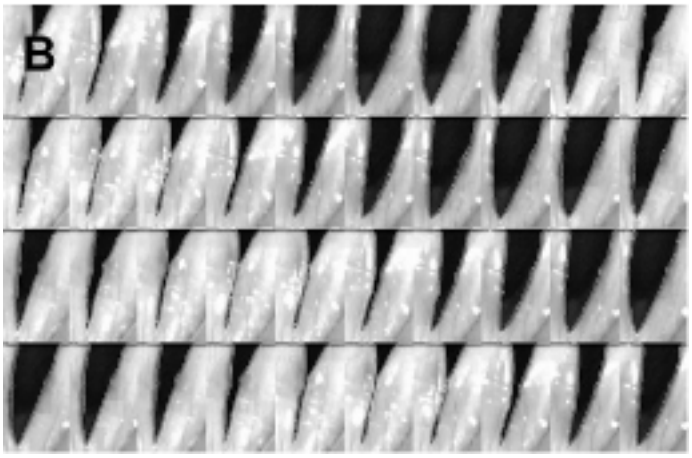


Figure 3 – High speed digital imaging of female breathy voice in sustained vowel /i/ [AYB12].

2.4 Vocal fry

Vocal fry is characterised by a low fundamental frequency and a small open Quotient. Furthermore, the mode of vibrations can be described as regular rather than aperiodic. The perception of the vocal fry is a combination of the low fundamental frequency $f0$ and the periodic continued short glottal pulse. Thus, the vocal folds are closed most of the cycle time and only open for a short period. Vocal fry is also described by synonyms like creaky voice, pulse register, glottal fry or creak, but terminology remains controversial. [BCNG98] Blomgren et al. compared various studies from literature where the fundamental frequency of vocal fry was evaluated. As a result, the range of the fundamental can be indicated from $20Hz$ to $70Hz$ with a mean of $50Hz$. Figure 4 shows a high-speed digital imaging of the vocal folds vibrations from a female speaker with a recording speed of 2000 fps. The HSDI shows the low fundamental frequency of $f0 = 80Hz$ which leads to a cycle time of $T = 12.5ms$. Furthermore, it can be observed that the vocal folds are closed in 20 frames by an entire cycle period of 25 frames. These observations lead to an Q_o of 0.18.

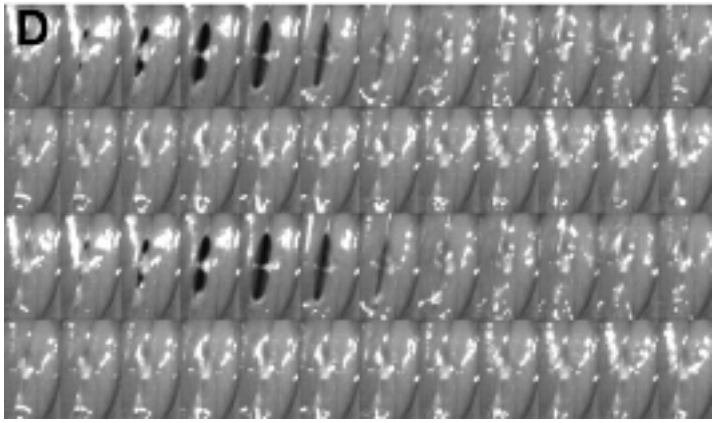
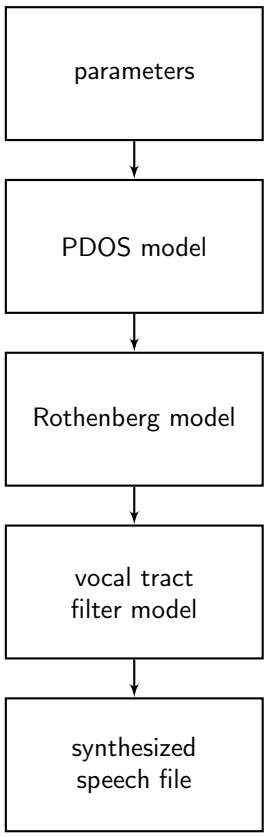


Figure 4 – High speed digital imaging of female vocal fry in sustained vowel /i/ [AYB12].

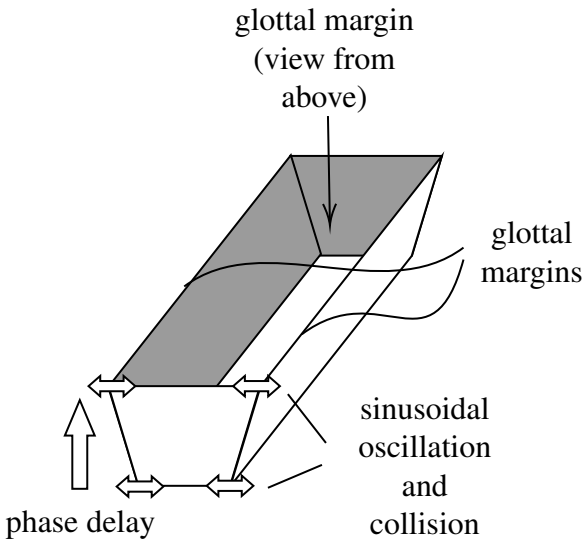
3 Synthesizer

As mentioned in the introduction, the voice stimuli for the listening experiments were generated by a synthesizer. The synthesizer which is used in this thesis is a formant synthesizer. The first version was developed and published in 2012 by Fraj et al. The aim of the authors was to develop a synthesizer that enables a computer based analysis - synthesis of human waveforms. In addition, the target was to develop a synthesizer based on models of the glottal source which can be controlled by parameters. [FSG12] The used updated version of the synthesizer comprises almost 100 parameters. In the following sections, the general structure of the synthesizer is explained. In chapter 3.2 the parameters itself and their physiological meaning, which were used and varied, are determined.

3.1 General structure



General structure of the synthesizer.



3D model of the glottis referred to the PDOS model of [TF06].

Figure 5 – General structure and 3D model.

Basically, the synthesis of the voice samples can be divided in three main parts. First, the glottal area function is generated by using a Phase Delayed Overlapping Sinusoid

(PDOS) model. The PDOS model takes account of the kinematics of the vocal fold vibrations, hence several voice types can be simulated and generated. The model comprises some physiological parameters. Hence, the glottal area function can be controlled with the open quotient Q_0, the fundamental frequency $f0$ and the pulse skewness as well as the pulse kurtosis.

In the second part, the glottal area function is inserted in the Rothenberg model to obtain the glottal flow rate based on a differential equation. [TF06]

In the last stage of the synthesizer, the glottal flow rate is processed by a cascade of five second order filters that represent the vocal tract resonances.

3.2 Parameters

The synthesizer comprises almost 100 parameters. In a first step it was mandatory to figure out which parameters are the ones that enable varying the voice quality from pressed voice to breathy voice and vocal fry. After literature research in addition with the experience of the supervisors of this thesis some major parameters were selected, which are described in the following sections. First, vowels were used for the target voice sample. Second, in order to control the pitch of the voice sample, the fundamental frequency of the vocal folds was selected. Third, for the purpose of achieving pressed voice, breathy voice as well as vocal fry it was mandatory to use the open quotient. To take account of the air flow through the vocal tract, the lung pressure enables varying the flow. The leakage describes a remaining open position of the vocal folds. Furthermore, another important aspect is to achieve a natural timbre for the voice samples. For this purpose, the parameter spectral slope of the aspiration noise can be used.

3.2.1 Glottal area function

As in chapter 3.1 mentioned the glottal area function is generated by using a Phase Delayed Overlapping Sinusoid model. Figure 6 shows the main components of the glottal area function.

In the first step 3, important parameters are determined once for a voice sample. The vocal fold length L_g which is a function of the reference vocal fold length L_o and the fundamental frequency $f0$. The amplitude of the vocal folds vibrations depend on the reference vocal fold length L_o the lung pressure Pl and again the fundamental frequency $f0$. Due to the fact, that the vocal folds do not have an initial position where they start oscillating, an offset - abduction - is calculated. The abduction is a function depending on the Amplitude A, the pulse skewness $skew$, the pulse convexity $conv$, the pulse kurtosis $kurt$ as well as the target open quotient. The skewness, convexity, kurtosis and the open quotient can be regarded as control parameters.
The following iteration is applied for every sample (voiceduration $\cdot$ sampling frequency). First, the actual sample phase ϕ is computed as a function of the sampling frequency f_s

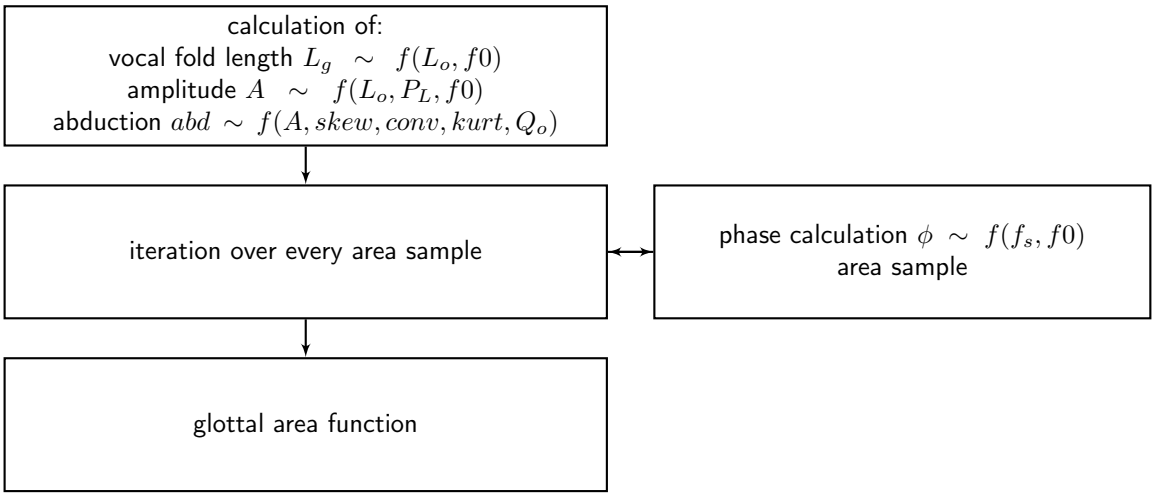


Figure 6 – Implemented code structure of the glottal area function.

and the fundamental frequency $f0$. In the next step the area sample is determined. The phase is modified to take account of the pulse sknewness, convexity and kurtosis as shown in equation 1 and 2.

$$t_1 = \phi + skew \cdot sin(\phi + conv \cdot sin(\phi)) \tag{1}$$

$$t_2 = kurt \cdot cos(\phi) \tag{2}$$

The vibration distance d of the vocal folds is computed as a sum of the abduction and the product of the amplitude times the sinus of the current phase as shown in equation 3.

$$d = abd + A \cdot sin(t_1 + t_2) \tag{3}$$

The width w of both vocal folds is achieved as two times the distance.

$$w = 2 \cdot d \tag{4}$$

To prevent a negative width after collision of the left and right folds a maximum operator ensures that the glottal opening is either zero or positive. Finally, the glottal area, in cm^2 is obtained as the product of the width times the vocal fold length L_g, assuming a rectangular glottis.

3.2.2 Vowels

From the linguistic point of view, sounds can either be classified as vowels or consonants. The word vowel has its origin from Latin word *vocalis* which means 'vocal' as related to the voice. Vowels are characterised by its formants which are ascribed by the acoustic resonances of the vocal tract. The vocal tract acts as a resonance tube which is determined by the position of the jaw, lips, and tongue. The aforementioned positions affect

the different formant values. In common understanding there are 5 different formant frequencies (F1 - F5). For instance, the first formant frequency of the vocal tract describes the openness. F2 represents the position of the tongue from the back to front and F3 the curvature of the lips. Furthermore, it can be distinguished between closed, mid and open vowels. Tabel 1 shows the used formant frequencies in the synthesizer. Due to the fact, that the vocal tract of male and female speakers have different geometric dimensions, the formant frequencies are also different. In this thesis, the male gender was chosen for the reason of experimental convenience only. No implications on superiority of any gender may be derived from this decision. [LJ15]

vowel	F1	F2	F3	F4	F5
/a/	640	1212	2254	3500	4500
/e/	343	2054	2609	3500	4500
/i/	287	2547	3469	3150	4050
/o/	400	868	2410	3500	4500
/u/	298	730	2172	3500	4500

Table 1 – Formant frequencies for each vowel [Hz] for male speaker.

3.2.3 Fundamental frequency $f0$

The fundamental frequency of the vocal folds is an important parameter in order to control the pitch of the vowels. Furthermore, the harmonics depend on the fundamental frequency.

3.2.4 Open quotient Q_o

The open quotient describes the ratio between the open phase of the vocal folds compared to the entire cycle. Figure 7 shows the time sections for the open phase (t_{op}) and the entire duty cycle (t_c). Furthermore, the open quotient has to be within the interval $[0, 1]$. From a physiological point of view an open quotient of 0 means permanently closed vocal folds consequently no voice is perceptible.

$$Q_o = \frac{t_{op}}{t_c} \tag{5}$$

Figure 8 shows the glottal area function of different Q_o, that are used later in the listening experiments.

From the spectrograms shown in figure 9 following aspects can be observed. Independent of the vowel, stimuli with smaller open quotients lead to more harmonics than stimuli with higher open quotient.

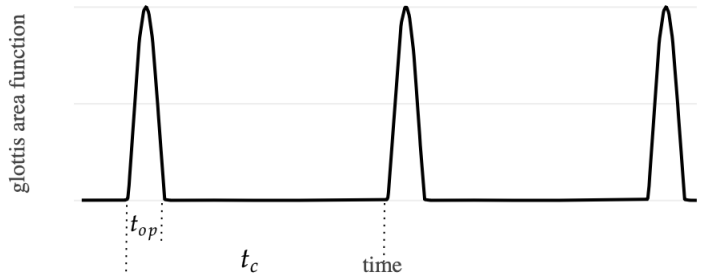
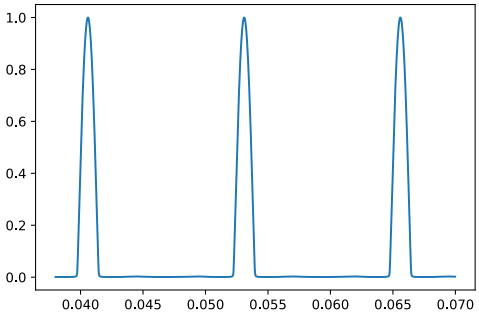
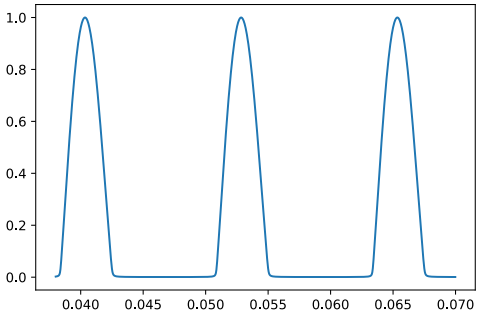


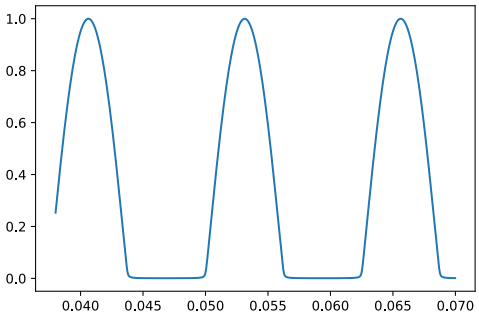
Figure 7 – Glottal area function with marked t_{op} and t_c.



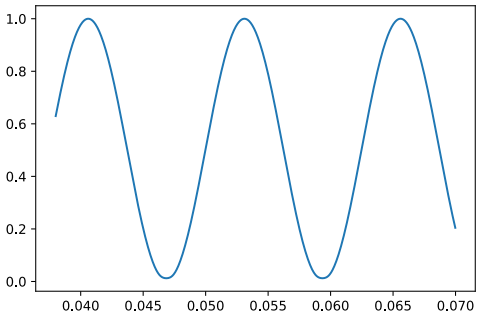
Target $Q_o = 0.1$, actual $Q_o = 0.13$.



Target $Q_o = 0.3$, actual $Q_o = 0.33$.

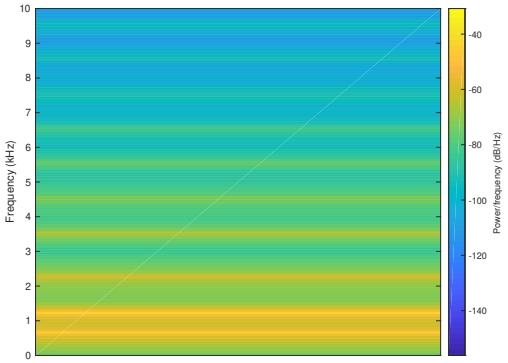


Target $Q_o = 0.5$, actual $Q_o = 0.48$.

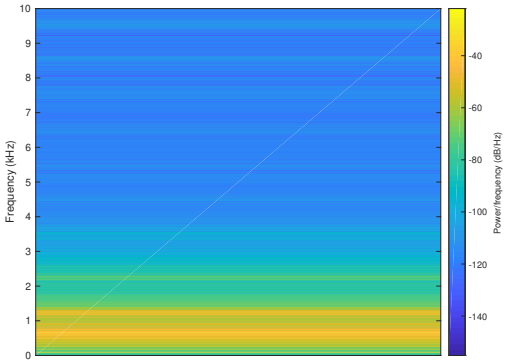


Target $Q_o = 1.0$, actual $Q_o = 0.98$.

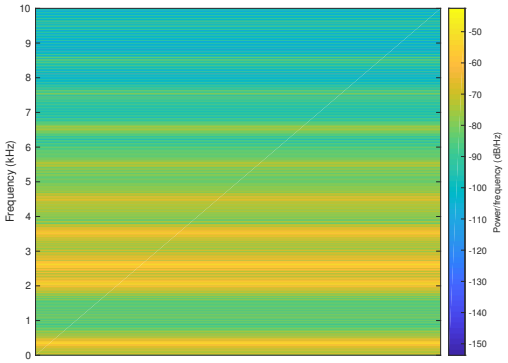
Figure 8 – Normalised glottal area function over time for vowel /a/ and $f0 = 80Hz$.



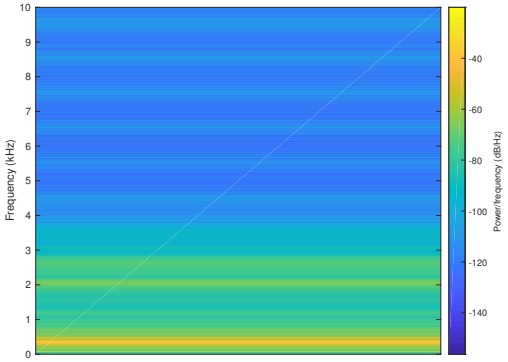
Vowel /a/, $f0 = 80Hz$, $Q_o = 0.1$.



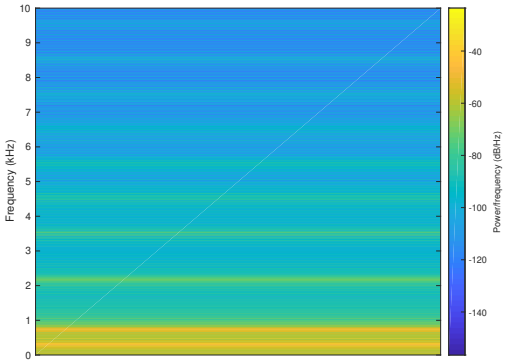
Vowel /a/, $f0 = 80Hz$, $Q_o = 1.0$.



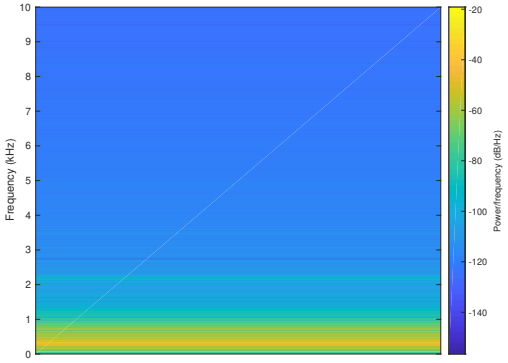
Vowel /e/, $f0 = 80Hz$, $Q_o = 0.1$.



Vowel /e/, $f0 = 80Hz$, $Q_o = 1.0$.



Vowel /u/, $f0 = 80Hz$, $Q_o = 0.1$.



Vowel /u/, $f0 = 80Hz$, $Q_o = 1.0$.

Figure 9 – Spectrograms for different vowels and different open quotients averaged over time.

3.2.5 Lung pressure P_L

The lung pressure provides information about the pressure of the air streaming from the lungs through the vocal tract. A certain pressure is mandatory in order to start the oscillation of the vocal folds previous studies pointed out that there is a relation between the fundamental frequency and the before mentioned phonation threshold pressure. [Tit88] By increasing the fundamental frequency the pressure threshold is also increasing. [Zha16] [TS92] For loud speech the lung pressure is typically between 10 and 15 cmH2O. [Sun14] In addition, lung pressure affects the amplitude of vocal fold vibration.

3.2.6 Leakage area of vocal folds

During incomplete closure of the vocal folds, a permanent leakage area remains. This leakage area leads to an airflow through the vocal folds and an audible noise. In purpose of controlling this phenomenon, the leakage area can be set by this parameter. Its unit is *cm*.

3.2.7 Spectral slope of aspiration noise

In order to control the timbre of the voice, the spectral slope of the aspiration noise is an important parameter. Inappropriate choice of the slope can lead to artificial audible behaviour, caused by the downstream vocal tract resonances. In the first version of the synthesizer there was a spectral slope, which followed the power law, in order to create blue, white, pink and red noise. After the first generated and reviewed voice sample the issue came up, that this implemented noise may not work as well as expected. Observations from the perception have led to the insight that the timbre of the generated voice samples did not sound naturally. In the latest version of the synthesizer, the power law has been changed to an exponential law, and with this implementation the timbre sounds quite well. The new spectral slope is based on previous studies. Hillman et al. investigated in their study the *glottal turbulent noise* by recording a whispered vowel of 5 male and 5 female speakers. As a result, comparing the mean of the spectral slope from male and female speakers with least squares approximation, a spectral slope of $-9.4db/kHz$ was determined. Figure 10 shows the resulting spectral slope.

In the synthesizer the spectral slope can be controlled as mentioned in equation 6 and 7

$$A_{spec} \sim 10^{\alpha \cdot f} \tag{6}$$

where α determines the parameter which can be accessed.

$$\alpha = \frac{-0.5}{1000} \qquad \left[\frac{1}{Hz}\right] \tag{7}$$

The given α in equation 7 corresponds to the spectral slope of Hillman et al..

One can argue, that noise emitted during whispering may be different from aspiration noise emitted during voiced phonation. In a previous study Aichinger investigated the

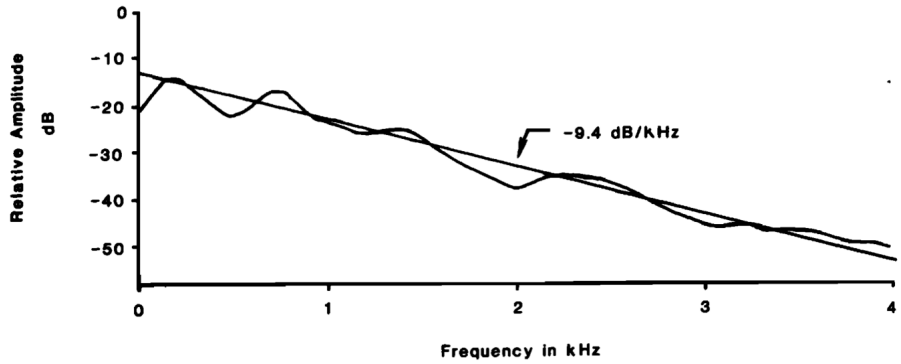


Figure 10 – Least squares linear approximation to the source spectrum of a whispered vowel [HOF83].

spectral slope of aspiration noise for breathy and normal voice. Figure 11 shows the power spectral density for breathy and normal voice. The spectral slope decreases with $\sim -1dB/kHz$.

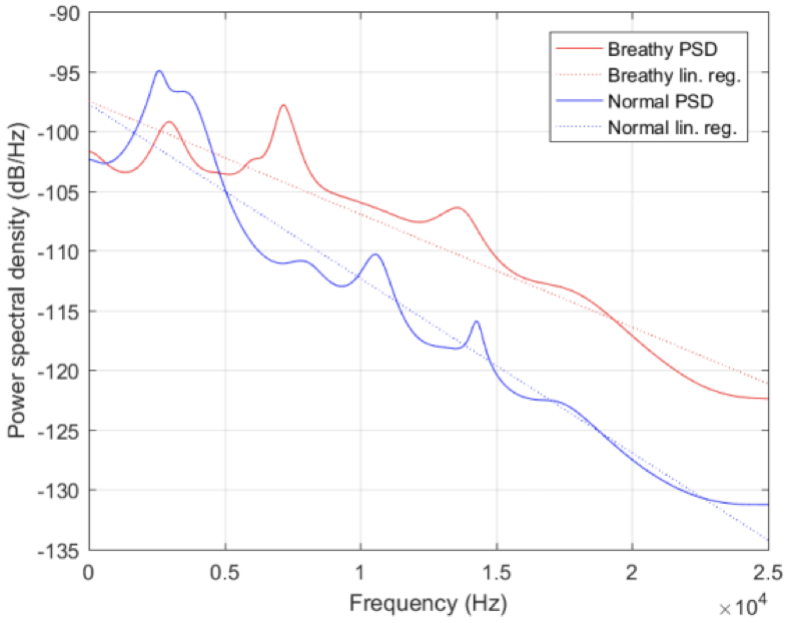


Figure 11 – Spectral slope for breathy voice [Aic19].

Different approaches in obtaining the spectral slope make a comparison between the two studies difficult. In the synthesizer the spectral slope is treated before the vocal tract and therefore the thesis follows the data from Hillmann et al..

4 Statistical test routines

For the evaluation of the results from the listening experiment, some statistical methods were used. The following sections will give an overview of the selected test routines. The Wilcoxon test is declared in section 4.1. The second section 4.2, explains the cluster analysis.

4.1 Wilcoxon test

The Wilcoxon test, also known as Wilcoxon signed rank test, provides information whether the trend behaviour of 2 dependent sample groups show different behaviours. Dependent sample groups can be described as below listed:

- rating iteration: the rating values descend from the same participant. (e.g. rating before and after a treatment, or different treatments on the same participant.)
- natural pairs: ratings from different participants which belong somehow together. (e.g. wife - husband, owner - renter, lawer - client)
- matching: ratings from different participants, which are matched together on a basis that is not evaluated

The Wilcoxon test can be performed in case following requirements are full-filled:

- The dependent variable is at least ordinally scaled
- There are two linked samples or groups, but the different pairs of measurements are independent of each other (e.g. pair A and pair B are independent of each other) [oZ08]

Thus, there is no need to have normal distributed datasets and the dependent variable need only to be ordinally scaled. The Wilcoxon test can also be performed with a small amount of samples and outliers.

4.2 Cluster analysis

Cluster analysis can be an efficient way for statistical analysis. The essence by clustering is to check how single datapoints are related to the others. Datapoints of one particular cluster are more similar referring to another cluster. Hence, cluster analysis can be used to reduce the amount of data or to formulate hypotheses that are not generally valid but applicable to a cluster. The methodical approach to cluster data depends on the type of data. Literature research has shown plenty of algorithms for cluster analysis. The choice, for the appropriate algorithm, depends on the one hand at the type of data that are processed and on the other hand on the experience in statistical signal processing. To be more precisely, the matlab function `linkage` was used to create a hierarchy tree. The distance between objects of 2 clusters is calculated with the nearest neighbour model. [lin] In order to visualise the clusters the matlab function `dendrogram` was used. [den]

5 Listening experiment

In order to rate the generated voice samples, two listening experiments were performed. The presented speech samples of vowels are generated by a synthesizer. The aim of the listening experiments were to investigate dependencies of perceived pressedness, breathiness and vocal fry on particular control parameters of the synthesizer.

Listening experiments are often used in audio research. Due to the fact that audio research is quite closely related to the human hearing and every single human has his or her unique ear in a physiological sense, general statements are quite difficult to phrase without a listening experiment. Thus, listening experiments are designed to get feedback of the participants after acoustical stimulation. For instance, participants should rate different stimuli according to before mentioned aspects. Another approach to use a listening experiment can be active adjustments of system-parameters. Hence, participants will adjust parameters in a way they would perceive an inquired acoustic event.

From a psychoacoustical point of view, different methods can be used to gather information from a test. For instance, a listening experiment with the forced choice method (AB) or the opposite unforced choice method (ABN) can be used. [PRA01] Also often used methods are the ITU-R BS.1534 (MUSHRA) and ITU-R BS.1116. [Uni15b] [Uni15a] In the present thesis the mushra method was chosen and described in the following section. Section 5.2 will explain the approach of the design and the evaluation of the listening experiments. The particular experiments are described in section 5.3 and 5.4.

5.1 General environment - WEBMushra

***MU**lti **S**timulus test with **H**idden **R**eference and **A**nchor - Mushra*
A mushra experiment contains several stimuli and a hidden reference as well as anchors. In 2018, Schoeffler et al. from the international audio laboratories Erlangen, developed a web based mushra version which has several advantages. There is no need to install additional software. Furthermore the listening experiment can be accessed from all over the world and is not restricted on a particular place. Also in the latest times, during the still ongoing corona pandemic, such tools as WEBMushra lead to even more possibilities to carry out listening experiments. The listening experiment can be setted up by changing the config file, which will be interpreted as an HTML Webpage. The audio is controlled and played back with Web Audio API. Therefore, sound files can be easily prepared before the listening experiment as a *.wav* file. Compared to the offline mushra there is no need anymore to use a DAW [1] where the soundfiles will be stored and controlled via OSC [2] from the mushra test system. [SBS+18] For the parts concerning pressedness and breathiness, participants were asked to rate the stimuli by means of a vertical slider (0-100%). Ratings towards 100% represent perceived breathiness and ratings towards 0% represent perceived pressedness. For the parts concerning vocal fry participants had the

1. DAW - Digital Audio Workstation
2. OSC - Open Sound Control

same slider which represented the strength of the vocal fry. Ratings towards to 100% were interpreted as strong vocal fry and ratings towards 0% were interpreted as slight or even no vocal fry. For all listening experiments and trials all stimuli were presented in random order. Figure 12 shows the WEBMushra appearance of one particular trial for pressed $\sim$ breathy.

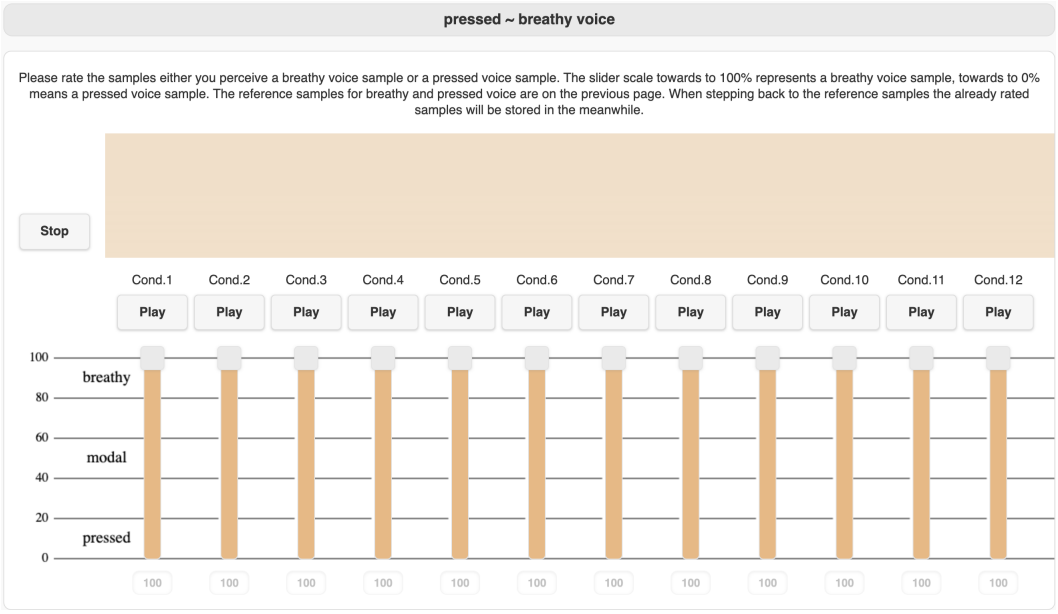


Figure 12 – Example of one WEBMushra trial.

5.2 Methodological approach

The methodological approach can be listed as follows:

1. Definition of the environment.
2. Definition of the listening experiment target(s) (hypotheses).
3. Generation of the required stimuli.
4. Performing of the listening experiment.
5. Analysis of the resulting data (visual analysis, statistical analysis)
6. Formulation of results

5.3 Listening experiment 1

The first listening experiment took place in the period from August 12 to August 23 2021. The listening experiment was divided into 2 parts.

— part I: pressed voice and breathy voice

— part II: vocal fry

The first part focused on pressed and breathy voice and comprised 9 different trials. The second part concerned vocal fry and comprised 3 different trials. Figure 13 shows the sequence of the trials, the particular stimuli for each trial are shown in table 2 and 3. After the introduction to the listening experiment the participants had the opportunity to adjust the volume to their convenience. In order to make the participants familiar with the presented voice samples, before each trial 2 reference samples for pressed voice and breathy voice were presented. For the trials regarding vocal fry 1, reference sample was presented before each trial. The first part was carried out with a Mushra test, thus each trial had a reference stimuli and a anchor stimuli. The second part of the experiment was realised without anchor, there was only a reference available.

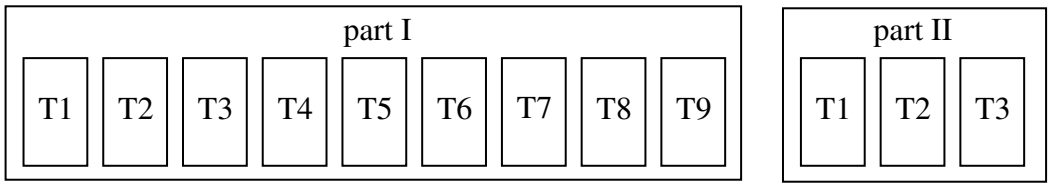


Figure 13 – Listening experiment 1 - overview.

5.3.1 Stimuli

This section gives an overview of the used parameters and values. To take account of the different formant frequencies the vowels /a/, /e/ and /u/ were chosen. The vowels /e/ and /i/ have similar formant frequencies and also the vowels /u/ and /o/. In order to increase the breathiness, a higher lung pressure and a greater leakage area, as previous studies figured out, were chosen. For the spectral slope of aspiration noise the slope from Hillman et al. was set as parameter value. For both parts of the experiment following parameters were fixed for all voice samples:

— gender of the voice samples: male
— lung pressure: $30.8cmH20$
— fixed spectral slope of aspiration noise: $-9.5dB/kHz$
— fixed leakage area: $0.3cm^2$

Following parameters were varied:

— vowel
— fundamental frequency $f0$
— open quotient Q_o

parameter / trial	T1	T2	T3	T4	T5	T6	T7	T8	T9
vowel	/a/	/a/	/a/	/e/	/e/	/e/	/u/	/u/	/u/
$f0[Hz]$	40	80	120	40	80	120	40	80	120
Q_o	for each trial $0.0 \sim 1.0$								
skewness	0	0	0	0	0	0	0	0	0
convexity	0	0	0	0	0	0	0	0	0
kurtosis	0	0	0	0	0	0	0	0	0
lung pressure $[cmH20]$	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8
spectral slope $[db/kHz]$	-9.5	-9.5	-9.5	-9.5	-9.5	-9.5	-9.5	-9.5	-9.5
leakage area $[cm^2]$	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
# of samples	11	11	11	11	11	11	11	11	11

Table 2 – Listening experiment 1 - part I stimuli parameters and values.

parameter/trial	T1				T2				T3			
vowel	/a/				/e/				/u/			
$f0[Hz]$	20	40	60	80	20	40	60	80	20	40	60	80
Q_o	for each trial 0.1 and 0.3											
skewness	0				0				0			
convexity	0				0				0			
kurtosis	0				0				0			
lung pressure $[cmH20]$	30.8				30.8				30.8			
spectral slope $[db/kHz]$	-9.5				-9.5				-9.5			
leakage area $[cm^2]$	0.3				0.3				0.3			
# of samples	12				12				12			

Table 3 – Listening experiment 1 - part II stimuli parameters and values.

Table 2 and 3 show the parameters and values in detail.

To complete the stimuli section, the reference and anchor stimuli are listed.

- part I
  - reference: particular fundamental frequency with a $Q_o = 0.0$
  - anchor: particular fundamental frequency with a $Q_o = 1.0$
- part II
  - reference: lowest fundamental frequency with a $Q_o = 0.1$

5.3.2 Participants

A total of $n = 8$ persons participated in the listening experiment for the first part, which can be differentiated in terms of gender, age and profession as listed below:

- gender:
  - male participants from 24 to 38 years with a mean of 29
- profession:

— audio students: 5
— audio experts: 3

For the second part $n = 7$ participants were attending. The list above has to be corrected in such a way that 4 audio students were participating and the mean of the age was 30 years.

5.3.3 Evaluation and Results

In a first step boxplot's of each trial were created. On the x-axis of the boxplot the target Q_o is shown, the y-axis represents the ratings of the participants. The division of the y-axis belongs to following aspects. The modal voice is placed at 60% of the scale due to the fact, that literature research pointed out a open quotient range for modal voice $Q_o = 0.6 \sim 0.75$. In Addition, the ratings of the second listening experiment for stimuli with $Q_o = 0.5$ have medians that are quite near to 50% perceived sample rating. Therefore a linear mapping of the open quotient to the perception as well as a interval scale is assumed. Figure 14, 15 and 16 show the ratings for the vowels /a/, /e/ and /u/ for the fundamental frequency $f0 = 80Hz$. Comparing the rating results for the different

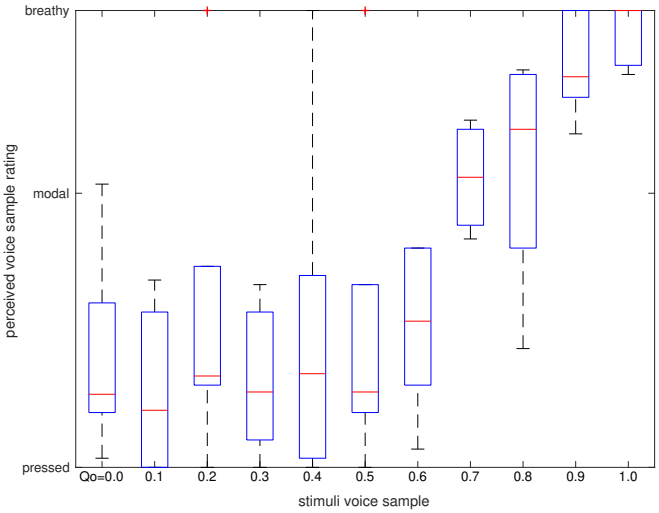


Figure 14 – Listening experiment 1 - pressedness, breathiness for vowel /a/ with $f0 = 80Hz$.

vowels at the same frequency one can assume they show similar trend behaviour. By increasing the Q_o from 0.6 to 1.0, participants will perceive breathy voice. Decreasing the Q_o, from 0.6 to 0.0 participants will perceive something between modal voice and pressed voice. Assuming similar trend behaviour after visually evaluation of the rating results a Wilcoxon signed rank test was performed. The Wilcoxon test should provide information whether the rating results of the previous shown boxplots belongs to the same distribution independent of the different vowels. In a statistical sense, a verification if the null-hypothesis (H0) cannot be rejected. The signed rank test will provide two details as a result. The p value indicates the significance level. The Wilcoxon test was performed

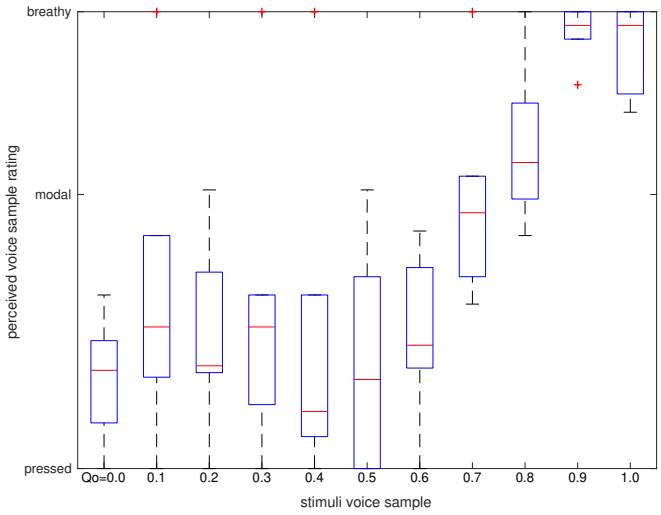


Figure 15 – Listening experiment 1 - pressedness, breathiness for vowel /e/ with $f0 = 80Hz$.

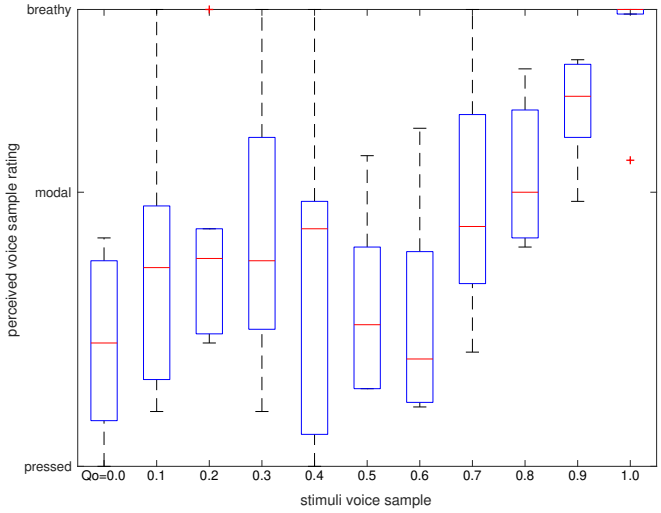


Figure 16 – Listening experiment 1 - pressedness, breathiness for vowel /u/ with $f0 = 80Hz$.

with the common $p < 0.05$ criterion. The h value indicates whether the null hypothesis is rejected ($h = 1$) or the failure to reject the null hypothesis ($h = 0$).

Q_o	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
p	0.84	0.30	0.73	0.46	0.79	0.81	0.84	0.41	0.96	0.21	0.54
h	0	0	0	0	0	0	0	0	0	0	0

Table 4 – Listening experiment 1 - wilcoxon test for vowel /a/ and /e/ with $f0 = 80Hz$.

Q_o	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
p	0.38	0.73	0.62	0.73	0.09	0.66	1.0	0.66	0.16	0.48	1.0
h	0	0	0	0	0	0	0	0	0	0	0

Table 5 – Listening experiment 1 - wilcoxon test for vowel /a/ and /u/ with $f0 = 80Hz$.

Q_o	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
p	0.95	0.84	0.26	0.31	0.59	0.56	0.84	0.75	0.43	0.12	0.75
h	0	0	0	0	0	0	0	0	0	0	0

Table 6 – Listening experiment 1 - wilcoxon test for vowel /e/ and /u/ with $f0 = 80Hz$.

Table 4, 5 and 6 shows that the null hypothesis is not rejected. Thus, I decided to merge the data. Due to the significance test, the data for one particular frequency was merged over all 3 vowels. The same approach was chosen for the second part of the listening experiment. The significance test was also performed for the vocal fry part. It came up that the data can be merged over the 3 vowels and also over the different open quotients.

Hence, the null hypothesis was applicable and the results of the first listening experiment can be described as follows. Figure 17, 18 and 19 show merged data for all 3 vowels for one particular frequency. Increasing the open Quotient from 0.5 to 1.0 leads to an increase of perceived breathiness. This trend can be observed for all 3 frequencies. Furthermore, it can be observed that the reference and anchor stimuli have been perceived by most of the participants. After detailed investigation of the results, due to the fact that the perceived voice type for an open quotient from 0.0 to 0.5 leads to quite similar ratings, a wrong parameter setting by the generation of the stimuli was detected. It turned out that a change of open quotient in the range of 0.0 to 0.5 had no effect on the actual open quotient of the voice samples. The actual open quotient remained at a value of 0.5. Therefore, the data for the open quotients from 0.0 to 0.5 were merged. A variation of the open quotient parameter Q_o from 0.6 towards to 1 leads to a variation of perceived breathiness, independent of the vowel and frequency. The current results do not allow to establish a hypothesis between the perceived pressedness and the open quotient, due to wrong parameter settings.
The before mentioned parameter issue affected also the part of vocal fry. Instead of having open quotients of 0.1 and 0.3 the actual open quotient was 0.5 for all samples. Nevertheless 2 hypothesis can be phrased. Samples with the lowest fundamental frequency where perceived as the highest strength of vocal fry. By increasing the fundamental frequency the perceived strength of the vocal fry decreases. Figure 20 shows the results.

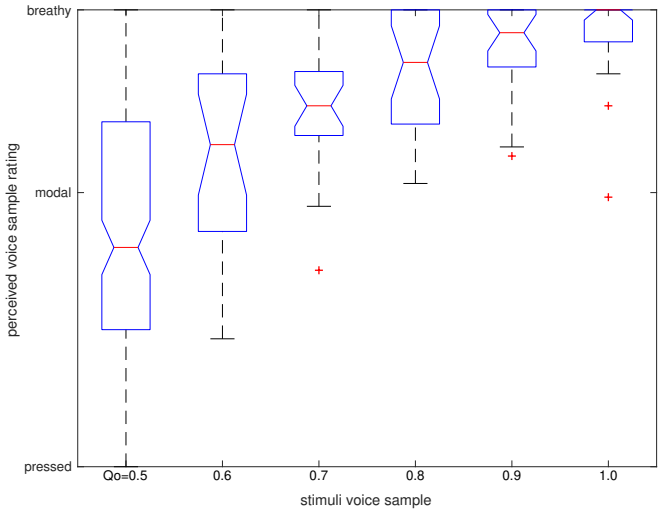


Figure 17 – Listening experiment 1 - perceived pressedness and breathiness with merged vowels for $f0 = 40Hz$.

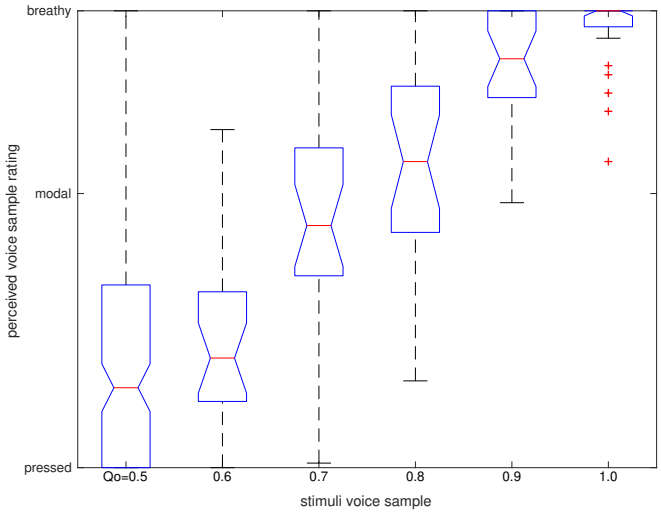


Figure 18 – Listening experiment 1 - perceived pressedness and breathiness with merged vowels for $f0 = 80Hz$.

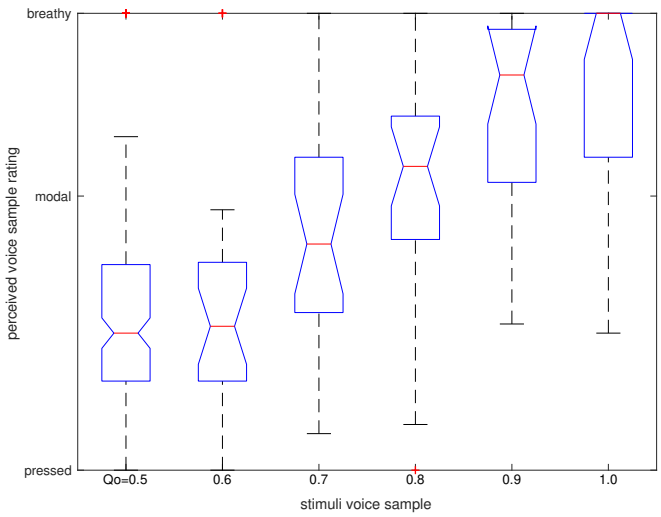


Figure 19 – Listening experiment 1 - perceived pressedness and breathiness with merged vowels for $f0 = 120Hz$.

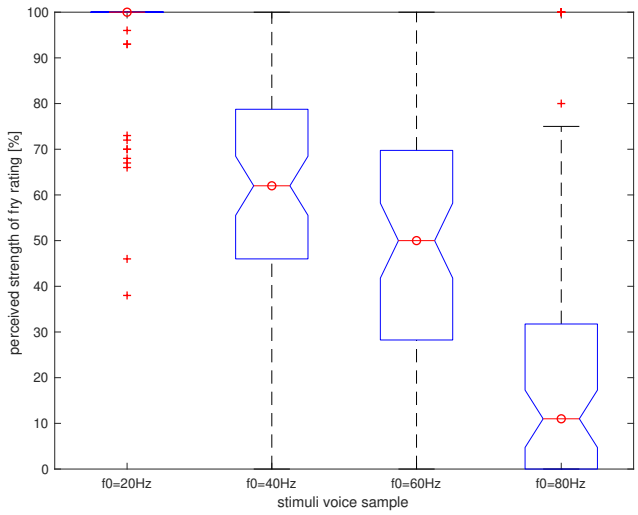


Figure 20 – Listening experiment 1 - perceived strength of vocal fry with merged vowels.

5.4 Listening experiment 2

The second listening test took place in the period from September 21 to October 21 2021. Compared to the first listening experiment the second one was improved in terms of technical issues as well as regarding the content. The first listening experiment pointed wrong parameter settings for an open quotient smaller than 0.5 out. Furthermore, the first listening experiment was limited due to the fact, that each trial was focusing on one vowel and one frequency. Thus, a absolute comparison between different vowels was not possible. Furthermore, the second listening experiment contains no reference and no anchor stimuli. Hence, the second listening experiment was adapted as listed below:

— trial with combined vowels
— reducing amount of trials
— reducing amount of fundamental frequencies

In detail, 3 different trials were conducted. The first and the second trial concerned pressed voice and breathy voice, the third one was related to vocal fry. To get rid of the limitations of the first experiment in the first and second trial, 2 vowels were presented. The aim was to get information how participants will rate pressedness or breathiness if two vowels are presented in the same trial.

5.4.1 Stimuli

To take account of the wrong parameter settings from the first listening experiment, the parameters for the second experiment were checked carefully. In the second listening experiment, the following parameters were fixed for all voice samples:

— gender of the voice samples: male
— lung pressure: $30.8cmH20$
— fixed spectral slope of aspiration noise: $-9.5dB/kHz$
— fixed leakage area: $0.3cm^2$

Following parameters were varied:

— vowel
— fundamental frequency $f0$
— open quotient Q_o

The detailed control parameters are shown in table 7, 8 and 9.

parameter/trial	T1			
vowel	/a/		/e/	
$f0[Hz]$	80	120	80	120
Q_o	0.1, 0.5, 1.0	0.1, 0.5, 1.0	0.1, 0.5, 1.0	0.1, 0.5, 1.0
lung pressure $[cmH20]$	30.8			
spectral slope $[db/kHz]$	-9.5			
leakage area $[cm^2]$	0.3			
# of samples	12			

Table 7 – Listening experiment 2 - trial 1 stimuli parameters and values.

parameter/trial	T2			
vowel	/a/		/u/	
$f0[Hz]$	80	120	80	120
Q_o	0.1, 0.5, 1.0	0.1, 0.5, 1.0	0.1, 0.5, 1.0	0.1, 0.5, 1.0
lung pressure $[cmH20]$	30.8			
spectral slope $[db/kHz]$	-9.5			
leakage area $[cm^2]$	0.3			
# of samples	12			

Table 8 – Listening experiment 2 - trial 2 stimuli parameters and values.

parameter/trial	T3					
vowel	/a/			/e/		
$f0[Hz]$	20	40	80	20	40	80
Q_o	0.1, 0.3	0.1, 0.3	0.1, 0.3	0.1, 0.3	0.1, 0.3	0.1, 0.3
lung pressure $[cmH20]$	30.8					
spectral slope $[db/kHz]$	-9.5					
leakage area $[cm^2]$	0.3					
# of samples	12					

Table 9 – Listening experiment 2 - trial 3 stimuli parameters and values.

5.4.2 Participants

The second listening test took place in the period from September 21 to October 8 2021. The aim was to increase the number of participants compared to the first listening experiment. Hence, people with professions such as audio students and experts as well as phoniatrics students and experts were invited. Thus, invitations were sent to the students of audio engineering at the Institute for Electronic Music and Acoustic, the Division of Phoniatrics at the Medical University of Graz, the Division of Speech and Language Therapy at the Medical University of Vienna, the Institute of Logopedics from the University of Applied Sciences Joanneum as well as to the research community of Dr. Aichinger and the McGill auditory list. A total of $n = 39$ persons participated in the listening test, which can be differentiated in terms of gender, age and profession as listed below:

- gender:
  - 14 female participants from 18 to 65 years with a mean of 32,8
  - 25 male participants from 22 to 85 years with a mean of 37,2
- profession:
  - audio students: 13
  - audio experts: 7
  - phoniatrics students: 3
  - phoniatrics experts: 2
  - other professions: 14

5.4.3 Evaluation and Results

In a first step, boxplots of each trial were created. On the x-axis of the boxplot, the particular stimuli are given. Table 10 describes in detail the particular stimuli regarding their parameters for the trials about pressedness and breathiness. Table 11 describes in detail the particular stimuli regarding its parameters for the trial about vocal fry. The y-axis represents the ratings of the participants.

	T1			T2		
stimuli/parameter	vowel	$f0$	Q_o	vowel	$f0$	Q_o
1	/a/	80	0.1	/a/	80	0.1
2	/a/	80	0.5	/a/	80	0.5
3	/a/	80	1.0	/a/	80	1.0
4	/a/	120	0.1	/a/	120	0.1
5	/a/	120	0.5	/a/	120	0.5
6	/a/	120	1.0	/a/	120	1.0
7	/e/	80	0.1	/u/	80	0.1
8	/e/	80	0.5	/u/	80	0.5
9	/e/	80	1.0	/u/	80	1.0
10	/e/	120	0.1	/u/	120	0.1
11	/e/	120	0.5	/u/	120	0.5
12	/e/	120	1.0	/u/	120	1.0

Table 10 – Listening experiment 2 - overview stimuli for trial 1 and 2 - pressedness and breathiness.

	T3		
stimuli/parameter	vowel	$f0$	Q_o
1	/a/	20	0.1
2	/a/	20	0.3
3	/a/	40	0.1
4	/a/	40	0.3
5	/a/	80	0.1
6	/a/	80	0.3
7	/u/	20	0.1
8	/u/	20	0.3
9	/u/	40	0.1
10	/u/	40	0.3
11	/u/	80	0.1
12	/u/	80	0.3

Table 11 – Listening experiment 2 - overview stimuli for trial 3 - vocal fry.

Figure 21 shows the ratings from the first trial. The vowels /a/ and /e/ are combined in one trial with two different frequencies. The cluster analysis was performed by a near-

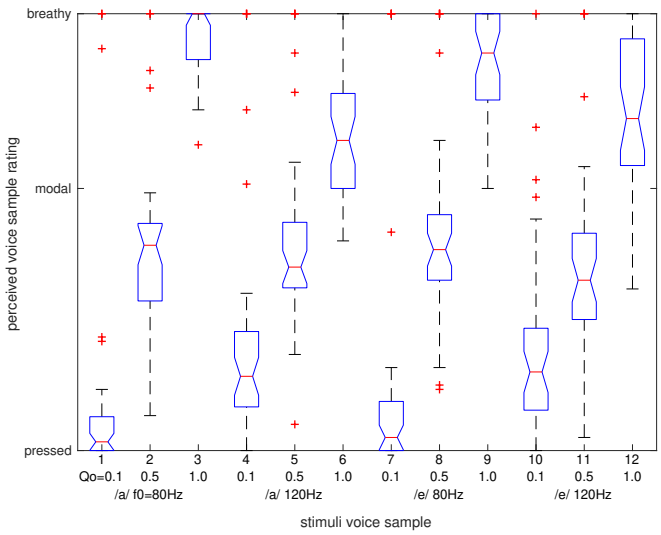


Figure 21 – Listening experiment 2 - pressedness, breathiness combined vowels /a/ and /e/ with $f0 = 80Hz$ and $f0 = 120Hz$.

est neighbour method. The nearest neighbour method uses the smallest distance between datapoints in two clusters. By focusing on the cluster analysis of the first trial, figure 22 indicates that participant 2, 11, 21 and 22 have a totally different rating than the other participants, therefore they will not be included in the results. For the second trial participant 2, 11 and 22, as shown in figure 24, are not included in the following results. The participants do not have the same profession, therefore the rating behaviour can not be argued based on their profession. Figure 23 shows the ratings from the second trail. The ratings for the vowel /a/ show similar trend behaviour. After performing a Wilcoxon signed rank test datapoints for vowel /a/ from trial 1 and trial 2 descend from the same distribution, as shown in table 12 therefore the data for vowel /a/ are merged in the subsequently figures 27 and 28.
After evaluation of the cluster analysis for the trial regarding vocal fry participant 10, 11 and 20 are not included in the following results, as given in figure 26. The rating behaviour is too far away from the other participants.

stimuli	1	2	3	4	5	6
p	0.20	0.93	0.85	0.10	0.24	0.10
h	0	0	0	0	0	0

Table 12 – Listening experiment 2 - wilcoxon test for vowel /a/ from first and second trial.

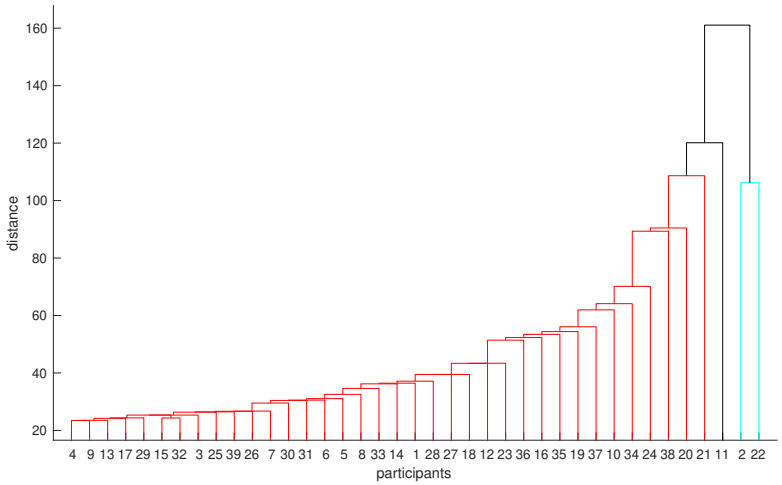


Figure 22 – Listening experiment 2 - pressedness, breathiness combined vowels /a/ and /e/ with $f0 = 80Hz$ and $f0 = 120Hz$ cluster analysis.

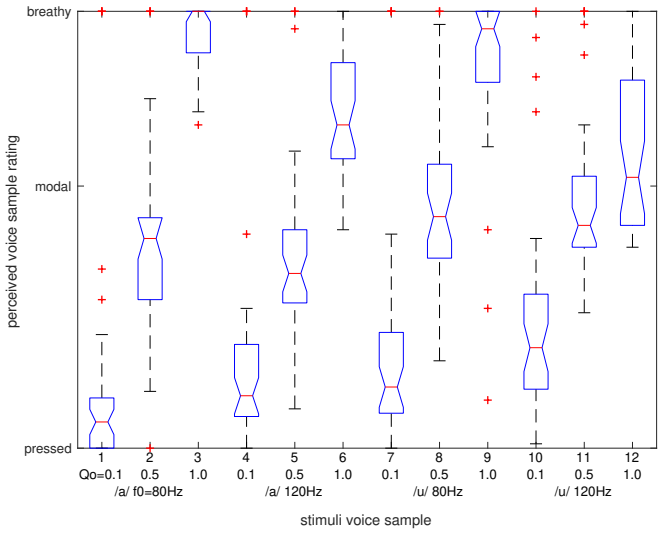


Figure 23 – Listening experiment 2 - pressedness, breathiness combined vowels /a/ and /u/ with $f0 = 80Hz$ and $f0 = 120Hz$.

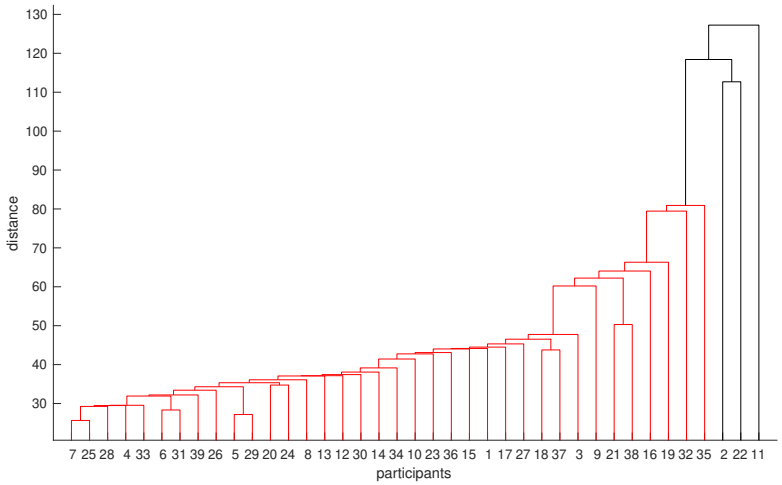


Figure 24 – Listening experiment 2 - pressedness, breathiness combined vowels /a/ and /u/ with $f0 = 80Hz$ and $f0 = 120Hz$ cluster analysis.

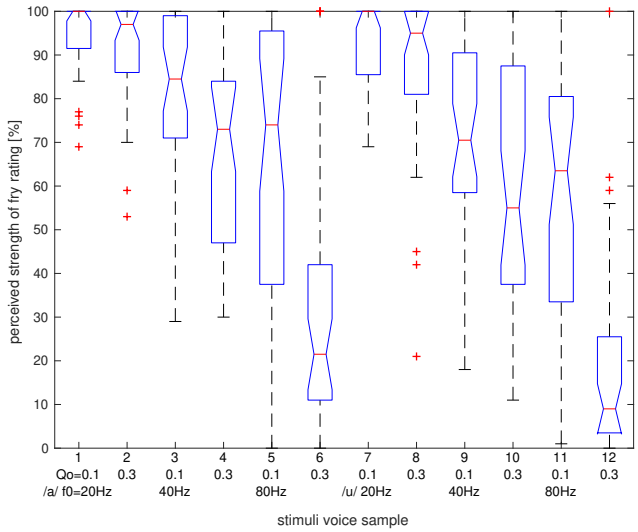


Figure 25 – Listening experiment 2 - vocal fry combined vowels /a/ and /u/ with $f0 = 20Hz$,$f0 = 40Hz$ and $f0 = 80Hz$.

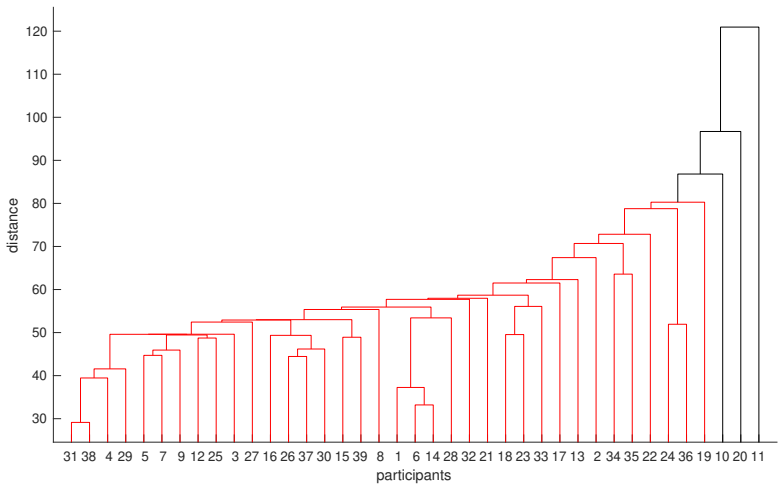


Figure 26 – Listening experiment 2 - vocal fry combined vowels /a/ and /u/ with $f0 = 20Hz$,$f0 = 40Hz$ and $f0 = 80Hz$ cluster analysis.

As mentioned before, some of the data were merged. Therefore, the stimuli have to be explained again. Figure 27 shows the results for a fundamental frequency of $f0 = 80Hz$, stimuli 1∼3 belong to the vowel /a/, stimuli 4∼6 belong to the vowel /e/ and stimuli 7∼9 belong to the vowel /u/. The open quotients for each stimuli can be described as follows:

— stimuli 1,4,7: open quotient $Q_o = 0.1$
— stimuli 2,5,8: open quotient $Q_o = 0.5$
— stimuli 3,6,9: open quotient $Q_o = 1.0$

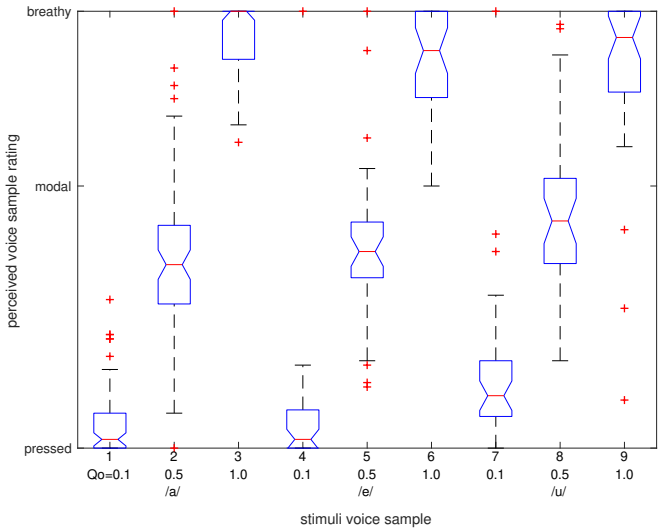


Figure 27 – Listening experiment 2 - perceived pressedness and breathiness for vowels /a/, /e/ and /u/ with $f0 = 80Hz$.

By detailed analysis of figure 27 some aspects can be observed. First, all vowels (/a/, /e/, /u/) show the same monotonously increasing trend behaviour by variation of the open quotient towards 1. Second, the medians from vowel /a/ and /e/ are almost identically. Third, the perception of the stimuli of vowel /u/ is not perceived as pressed or breathy as /a/ and /e/, which depends on the vocal tract formant frequencies of the vowel /u/.

Before mentioned stimuli explanation is also valid for figure 28 except the fundamental frequency because that was changed from $80Hz$ to $120Hz$. The vowels /a/ and /e/ show again monotonously increasing trend behaviour. For the vowel /u/ the trend is monotonously increasing from pressed voice to modal voice but from modal voice to breathy voice it shows sigmaoid behaviour. All 3 vowels were not perceived as pressed as for the lower fundamental frequency. The same fact can be observed for breathy voice in addition the vowel /u/ was again not perceived as much as vowel /a/ and /e/.

A short summarise of before mentioned aspects. The open quotient Q_o the is highly significant for the perception of the respective voice type. For this purpose the effect size of Cohen's d were determined and shown in table 13 for $f0 = 80Hz$ and $f0 = 120Hz$.

[Coh13] At high fundamental frequency, voice types are perceived less strongly in terms

vowel	$f0[Hz]$	Q_o versus Q_0	Cohen's d'
/a/	80	0.1 vs. 0.5	1.4
/e/	80	0.1 vs. 0.5	1.25
/u/	80	0.1 vs. 0.5	1.32
/a/	80	0.1 vs. 1.0	4.2
/e/	80	0.1 vs. 1.0	3.1
/u/	80	0.1 vs. 1.0	2.6
/a/	120	0.1 vs. 1.0	2.35
/e/	120	0.1 vs. 1.0	2.84
/u/	120	0.1 vs. 1.0	1.83

Table 13 – Listening experiment 2 - perceived pressedness and breathiness Cohen's d'.

of their extreme position than at low fundamental frequency. Therefore, the effect size Cohen's d' is lower at a fundamental frequency of $f0 = 120Hz$, but nevertheless still a strong effect size. The differences between the vowels are not significant. Exceptions are vowel /a/ (/e/) and vowel /u/ at $f0 = 80Hz$, $Q_o = 0.1$ with a low effect size Cohen's $d' = 0.39$.

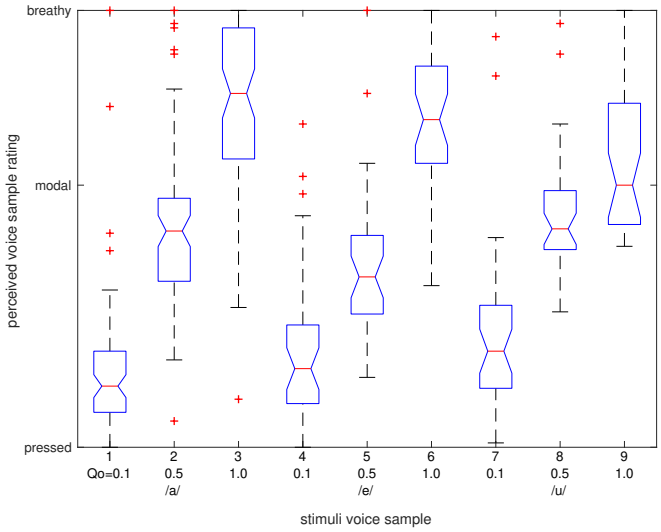


Figure 28 – Listening experiment 2 - perceived pressedness and breathiness for vowels /a/, /e/ and /u/ with $f0 = 120Hz$.

In order to analyse the results from the perceived strength of vocal fry, the used stimuli are listed.

- stimuli 1,2,7,8: fundamental frequency $f0 = 20Hz$
- stimuli 3,4,9,10: fundamental frequency $f0 = 40Hz$
- stimuli 5,6,11,12: fundamental frequency $f0 = 80Hz$

While, odd stimuli have an open quotient of $Q_o = 0.1$, even stimuli have an open quotient of $Q_o = 0.3$. First, stimuli with the lowest fundamental frequency are perceived as the strongest vocal fry. Variation of the open quotient do not lead to significant distinctions. By increasing the fundamental frequency the strength of vocal fry is also decreasing. The vowel /a/ for $f0 = 80Hz$ show more strength than for the vowel /u/. For the stimuli with an open quotient of $Q_o = 0.3$ both vowels point greater variances than for the stimuli with $Q_o = 0.1$. For the highest fundamental frequency both vowels have similar medians and variances for the stimuli with $Q_o = 0.3$. The stimuli with an open quotient of $Q_o = 0.1$ show great variances. Some participants have perceived them as the same strength as the stimuli with $f0 = 80Hz$. For the stimuli with a fundamental frequency of $f0 = 80Hz$ and a $Q_o = 0.1$ it has to be further proofed if this stimuli can be generated in terms of physiological meanings.

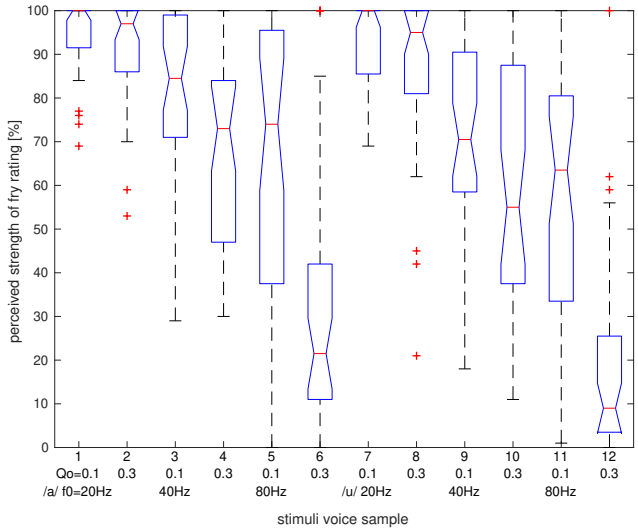


Figure 29 – Listening experiment 2 - perceived strength of vocal fry for for vowels /a/ and /u/ with $f0 = 20Hz$,$f0 = 40Hz$ and $f0 = 80Hz$.

A short summarise of before mentioned aspects. Increasing the fundamental frequency $f0$ leads to a weaker perceived vocal fry. At the lowest fundamental frequency the effect size between $Q_o = 0.1$ and $Q_o = 0.3$ is small. For the fundamental frequencies of $f0 = 40Hz$ and $f0 = 80Hz$ there are medium to strong effect sizes between $Q_o = 0.1$ and $Q_o = 0.3$ as shown in table 14.

vowel	$f0[Hz]$	Q_o versus Q_0	Cohen's d'
/a/	20	0.1 vs. 0.3	0.41
/u/	20	0.1 vs. 0.3	0.38
/a/	40	0.1 vs. 0.3	0.65
/u/	40	0.1 vs. 0.3	0.95
/a/	80	0.1 vs. 0.3	0.46
/u/	80	0.1 vs. 0.3	1.46

Table 14 – Listening experiment 2 - perceived vocal fry Cohen's d'.

6 Conclusion and outlook

In the present thesis, different voice types, breathy and pressed voice as well as vocal fry were generated by means of an existing model based synthesizer. In a listening experiment, the samples were assessed by experts and students.

Both listening experiments approved following hypothesis:

- a gradual increase of the open Quotient parameter Q_o from 0.5 towards to 1 leads to a gradual increase of perceived breathiness.
- a variation of the open Quotient parameter Q_o from 0.5 towards to 0 leads to more and more perceived pressed voice.
- samples with the lowest fundamental frequency were perceived as the ones with the strongest vocal fry.
- by increasing the fundamental frequency the perceived strength of vocal fry decreases.
- the influence of vowel formants on the perception of the voice types is small.

Furthermore, parameter values for the pulse skewness, convexity and kurtosis were found in order to generate pressed voice and breathy voice. These values were determined during the generation of the voice samples used in this work and are shown in figure 15.

Q_o	skewness	convexity	kurtosis
0.1	0.05	0.025	-2.75
0.3	0.2	0.1	-0.6
0.5	0	0	0
1.0	0	0	0

Table 15 – Skewness, convexity and kurtosis for different Q_o.

Another aspect that was observed, the perception of breathy voice and pressed voice depends on the vowel and also on the fundamental frequency. In the early stage of the thesis the question of the right timbre for the aspiration noise turned into a fundamental aspect for generating breathy voice samples.

As this thesis is not the end of scientific research, regarding synthesis of human speech, this work can be continued in serveral aspects.

- dependency of other parameters to the perception of pressed and breathy voice
- further listening experiments also in comparison with a clinical study to evaluate the differences between generated voice samples and natural voice samples from humans with either pressed voice or breathy voice as well as vocal fry
- improving of usability of the synthesizer as well as further documentation of the synthesizer functions and the general structure

References

[Abe67] D. Abercrombie, *Elements of general Phonetics*. Edinburgh University Press, 1967.

[Aic19] P. Aichinger, "Characterization of turbulence noise in breathy human phonation," *International Congress on Acoustics*, 2019.

[AV96] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers," *Folia Phoniatr Logop*, vol. 48, no. 5, pp. 240–254, 1996.

[AYB12] K. Ahmad, Y. Yan, and D. M. Bless, "Vocal fold vibratory characteristics in normal female speakers from high-speed digital imaging," *J Voice*, vol. 26, no. 2, pp. 239–253, Mar 2012.

[BCNG98] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *J Acoust Soc Am*, vol. 103, no. 5 Pt 1, pp. 2649–2658, May 1998.

[BTS04] C. Bergan, I. Titze, and B. Story, "The perception of two vocal qualities in a synthesized vocal utterance: Ring and pressed voice," *Journal of voice : official journal of the Voice Foundation*, vol. 18, pp. 305–17, 10 2004.

[CL91] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, 1991. [Online]. Available: https://doi.org/10.1121/1.402044

[Coh13] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* -. New York: Routledge, 2013.

[den] "Dendrogram plot - MATLAB dendrogram - MathWorks Deutschland — de.mathworks.com," https://de.mathworks.com/help/stats/dendrogram.html?s_tid=doc_ta, [Accessed 16-Oct-2021].

[FSG12] S. Fraj, J. Schoentgen, and F. Grenez, "Development and perceptual assessment of a synthesizer of disordered voices," *J Acoust Soc Am*, vol. 132, no. 4, pp. 2603–2615, Oct 2012.

[HOF83] R. E. Hillman, E. Oesterle, and L. L. Feth, "Characteristics of the glottal turbulent noise source," *The Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 691–694, 1983. [Online]. Available: https://doi.org/10.1121/1.389854

[lin] "Agglomerative hierarchical cluster tree - MATLAB linkage - MathWorks Deutschland — de.mathworks.com," https://de.mathworks.com/help/stats/linkage.html?searchHighlight=linkage&s_tid=srchtitle, [Accessed 16-Oct-2021].

[LJ15] P. Ladefoged and K. Johnson, *A course in phonetics*. Cengage Learning, 2015.

[oZ08] U. of Zürich, "Wilcoxon Test," https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/unterschiede/zentral/wilkoxon.html, 2008, [Online; accessed 20-September-2021].

[PRA01] J. L. Punch, B. Rakerd, and A. M. Amlani, "Paired-comparison hearing aid preferences: evaluation of an unforced-choice paradigm," *J Am Acad Audiol*, vol. 12, no. 4, pp. 190–201, Apr 2001.

[SBS+18] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra â a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, 02 2018.

[Sun14] J. Sundberg, "Re: What is the typical lung pressure for normal human phonation/speech?" https://www.researchgate.net/post/What_is_the_typical_lung_pressure_for_normal_human_phonation_speech/5463131acf57d7863d8b4600/citation/download, 2014, [Online; accessed 24-September-2021].

[TF06] I. R. Titze and A. Fariborz, *The myoelastic aerody- namic theory of phonation*. National Center for Voice and Speech, Iaowa, 2006.

[Tit88] I. R. Titze, "The physics of small-amplitude oscillation of the vocal folds," *J Acoust Soc Am*, vol. 83, no. 4, pp. 1536–1552, Apr 1988.

[TS92] I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *The Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2936–2946, 1992. [Online]. Available: https://doi.org/10.1121/1.402929

[Uni15a] I. T. Union, "Recommendation ITU-R BS.1116-3: Methods for the subjective assessment of small impairments in audio systems." https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1116-3-201502-I!!PDF-E.pdf, 2015, [Online; accessed 24-September-2021].

[Uni15b] ——, "Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems." https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-3-201510-I!!PDF-E.pdf, 2015, [Online; accessed 24-September-2021].

[vdODZ+16] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Arxiv*, 2016. [Online]. Available: https://arxiv.org/abs/1609.03499

[WRD+21] F. Wendt, I. Roesner, V. Devaraj, J. Schoentgen, and P. Aichinger, "Auditory perception of impulsiveness and tonality in vocal fry," *Acoustical Society of America*, 2021.

[Zha16] Z. Zhang, "Respiratory Laryngeal Coordination in Airflow Conservation and Reduction of Respiratory Effort of Phonation," *J Voice*, vol. 30, no. 6, pp. 7–760, Nov 2016.