Graz University of Technology

Master Thesis

Learning Effects for Electromyographically controlled Electrolarynx Speech

conducted at the
Signal Processing and Speech Communications Laboratory
Graz University of Technology, Austria

by
Hartwig Klammer, 0631694

Supervisors:
Dipl.-Ing. Dr. Martin Hagmüller
Dipl.-Ing. Anna Katharina Fuchs

Assessors/Examiners:
Dipl.-Ing. Dr. Martin Hagmüller

Graz, January 7, 2015

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 sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

date	(signature)

Abstract

Laryngeal cancer is a rare disease but the effects of medical treatment, often a total removal of the larynx, are an enormous disturbance for the affected persons. The larynx with its vocal chords is the speech producing organ and a loss leads to loss of the voice. Possibilities to be able to speak again can be tracheoesophageal speech, esophageal speech and electrolarynx speech. The electrolarynx speech is the easiest way to speak but with the major disadvantage of monotonic voice and the necessity of using a hand while speaking.

A new way to control the electrolarynx is using electromyography to turn on the device and vary the fundamental frequency via neck muscle tension. The main advantage of this device is that the user does not need his or her hand anymore to handle the electrolarynx. In an earlier thesis a unit was developed to control the on-/off-function of the electrolarynx through electromyography.

Until now there are no analyses about the learning effects of electromyographically controlled electrolarynx speech in German language so this thesis will approach the question how much the handling of an electrolarynx can be improved through practising with the electromyographically controlled electrolarynx. Therefore four participants had to undergo a defined training where recordings were made before and after training. For recordings a listening test should be developed to estimate the pleasantness of electromyographically controlled electrolarynx speech and to figure out the improvement through practising with the device. The results show an obvious handling improvement and also an increase of pleasantness.

Kurzfassung

Kehlkopfkrebs ist eine seltene Krankheit, aber die Auswirkungen der medizinischen Behandlung, oftmals eine totale Entfernung des Kehlkopfes, führen zu einer enormen Belastung für die betroffenen Personen. Der Kehlkopf (lat. larynx) mit den Stimmbändern ist das sprachproduzierende Organ und ein Verlust führt zum Verlust der Stimme. Mögliche Varianten, um die Sprache wiederzuerlangen, sind die die tracheoösophageale Ersatzstimme (Sprache über ein Shunt-Ventil), die ösophageale Ersatzstimme (Ruktusstimme) und die elektronische Sprechhilfe (Elektrolarynx). Die elektronische Sprechhilfe ist die einfachste Möglichkeit, hat jedoch den Nachteil der monotonen, technischen Stimme und benötigt eine Hand zur Bedienung.

Eine neue Möglichkeit den Elektrolarynx zu steuern, ist die Verwendung der Elektromyographie. Hier wird das Gerät über die Anspannung der Halsmuskulatur ein- und ausgeschaltet. Der Hauptvorteil liegt in der Benutzerfreundlichkeit, da keine Hand für die Bedienung benötigt wird. In einer früheren Masterarbeit wurde ein Steuergerät entwickelt, um den Elektrolarynx mit Hilfe der Elektromyographie ein- und auszuschalten.

Bis jetzt gibt es keine Untersuchungen über die Lerneffekte in der Elektrolarynx-Sprache mit Steuerung über die Elektromyographie mit deutscher Sprache, sodass diese Arbeit der Frage nachgehen wird, wie sehr die Steuerung des Elektrolarynx über Training mit dem Gerät verbessert werden kann. An der Studie nahmen vier Personen teil, die sich einem definierten Training unterzogen. Jeweils vor und nach dem Training wurden Aufnahmen für die Auswertung gemacht. Über einen Hörtest wurden die Verbesserungen der Annehmlichkeit der künstlichen Sprache ausgewertet. Die Ergebnisse zeigen eine erkennbare Verbesserung in Handhabung als auch in der Annehmlichkeit der Sprache.

Acknowledgments

First of all I am indebted to my supervisors Dipl.-Ing. Dr. Martin Hagmüller and Dipl.-Ing. Anna Katharina Fuchs who gave me patient guidance and advice to all challenges and questions in this work. A big thanks also to Dipl.-Ing. Clemens Amon and Dipl.-Ing. (FH) Klaus Jänsch for their help and to the participants of the training for their endurance.

My parents deserves a lot of gratefulness as well. Thank you for your generous support in all my decisions throughout my studies.

Contents

List of Tables

1 Introduction

Fortunately, laryngeal cancer is a rare disease in society. Nevertheless there are more than 600 000 people in the world who have to face the fact that they got laryngeal cancer [3]. The cancer is caused by heavy smoking and abuse of alcohol especially alcoholic spirits. One common form of treatment is total laryngectomy which means the removal of the larynx with the big drawback of the loss of the natural sound source. There are basically three types of artificial speech to maintain the possibility of conversation. The electrolarynx (EL), esophageal speech and tracheoesophageal speech.

1.1 Motivation

"Speech is an arrangement of notes
that will never be played again."
– F. Scott Fitzgerald

For laryngectomees the mutilating total removal of the larynx brings along not only the loss of the voice but also a social stigmatization is not the voice one of the most human way to communicate. An extensive voice rehabilitation is required to bring laryngectomees back to an adequate life. Using the EL is one possibility to be able to speak again after total laryngectomy. One major drawback of this device is that the patient has to use one of his or her hand while speaking to press the on-/off-button of the electrolarynx. The use of surface electromyography (sEMG) can solve this problem. An electromyographically controlled electrolarynx (EMG-EL) can be controlled by the neck muscle activity so that the hands are free while speaking and the patient gets back a bit more of freedom.

Constructing on a former thesis by Amon [4] where an EMG-EL was developed now the learning effects of the EMG-EL-speech will be examined. Like all muscles in the body also the neck muscles can be trained to improve their functionality in controlling the device so that the produced speech is clearly intelligible without any interruptions during speaking. Therefor a training protocol was created and participants had to do consecutive training of nine sessions to improve the handling of the EMG-EL. This thesis should reveal the influence of training effects on EMG-EL-speech and how much the EL-handling, the speech intelligibility and naturalness of speech can be improved by regular training.

1.2 Related publications

In 2007, Goldstein et al. [5] examined the training effects on speech production using an electromyographically controlled electrolarynx (EMG-EL). The participants were trained on seven stages: vowel initiation, vowel duration, vowel termination, words, sentences, a paragraph and intonation contrasts. The test participants were three total laryngectomees (three men), so people who underwent a total laryngectomy surgery and four people with normal voice (two men and two women).
The test setting consists of a condenser microphone, a video monitor to present the stimulus material. A photo cell on the video monitor was used for timing accuracy to measure the time delay between the stimulus and the voice initiation and termination. To evaluate the reaction time the test with vowel initiation, duration and termination was used. For the evaluation of words and sentences the Yorkston and Beukelman test was used. In each session 10 sentences were used. To test the intonation a pool of 20 short declarative sentences that could be spoken as a statement or as a question was used.
The participants had to undergo a training for each skill which consists of ten 10 - 60 min sessions for ten consecutive days. To evaluate the training effects the recordings of the first three sessions were compared to the last three ones. There was no listening test. The author found out that all participants could improve their skills with a few hours of training.

Goldstein et al. [6] compared in 2004 the user performance with an EMG-EL to other voice sources like normal voice, tracheoesophageal speech and manual EL. The focus was on the voice initiation time (VIT) and the voice termination time (VTT). The test setting was the same as described by the above paragraph. The VIT and VTT were defined as the time delay between the appearance of the visual cue and the voice initiation by using the vowel "a" as speech material. Four males and three females were used as normal subjects and one male underwent a laryngectomy. The result shows that VIT is nearly the same for all voice sources but that VTT for EMG-EL-speech is longer than for other sources.

In 2006, Van As-Brooks [7] did a perceptual evaluation with 39 laryngectomized patients where 29 were men and 10 were women. The speech material consisted of three sustained vowels "a" and a standard read-aloud text. The vowels were acoustically analysed by narrow-band spectrograms to detect harmonics. For acoustic analysis, seven voice-quality measures were chosen: to reflect the pitch (fundamental frequency) and quality (periodicity, harmonicity) of the voice, fundamental frequency, standard deviation of fundamental frequency, jitter, HNR, and %voiced speech. The quality of the read-aloud text was judged by four trained speech language pathologist. The used acoustic signal types are a good basis to perform more acoustic analyses.

The thesis written by Stepp [8] in 2008 shows a perceptual evaluation of seven different electrode positions on neck and face. Word and pause production was examined. The participants had to say several words each time with a pause in between. The test was the Yorkston and Beukelman test. The seven electrode positions were tested before where the participant had a screen to get a visual feedback of the sEMG envelope. After producing serial speech with pauses between each words the best two positions were selected for the Yorkston and Beukelman test. For evaluation the voicing performance and the pause performance was weighted. The test was judged by a speech language pathologist and by the author. With a right electrode position all participants could produce continuous speech. The radiation therapy was also discussed. The question was if the radiation therapy reduces the muscle integrity but the author could not determine a relationship. Prospective the author will study the effects of training.

In 2002, Rossum [9] evaluated whether a laryngectomized speaker can produce pitch accent. The test involved 10 tracheoesophageal, 9 esophageal and 10 laryngeal (control) speakers. As

stimulus material Rossum used 10 sentences, each read two times with different emphasize on one word. For the listening test 22 listeners evaluated the recorded material. For each sentence they could choose between two questions which asked for the emphasize. For example for the sentence "The ball flew over the fence" where "ball" and "fence" were accentuated the two questions were "What flew over the fence" and "What did the ball fly over". All participating listeners were unfamiliar with alaryngeal speech and inexperienced in speech evaluation to eliminate the effects of experience. The results show that all speakers were able to convey accent but some alaryngeal speakes had problems in controlling fundamental frequency.

In 1994, Vinton [10] made a study about the Tamil language to figure out the placement of stress. He distinguishes between the length, loudness, pitch and sound quality of syllables. As test material five sentences were used, each spoken two times with different emphasize on one noun, respectively. To analyze the recorded sentences a spectrogram was used. The result shows that for a correct prediction of emphasis frequency and amplitude is necessary.

1.3 Thesis organization

Chapter 2. In this chapter the theoretical aspects of the thesis will be discussed. The differences between normal speech and alaryngeal speech will be explained likewise the technical background for electrolarynx, surface electromyography, used devices, activity detection algorithms and electrodes.

Chapter 3. This chapter describes the training with the electromyographically controlled electrolarynx. There is information on the trained protocol with its different training steps, speech material and the test setting. Also a description about the criteria for evaluation can be found here.

Chapter 4. Here a detailed description about the listening test developed in MATLAB is presented. There is general information about a meaningful listening test and also details about the evaluation of words, sentences and a paragraph.

Chapter 5. This chapter contains the evaluation of the recorded speech material. It is divided into objective part of the evaluation of recorded vowels and subjective results of the listening test.

Chapter 6. In the last chapter a conclusion is presented. It provides a summary of the results and an outlook into the future.

In **Appendix A** the lists for words, sentences and the paragraph of the recorded speech material for the evaluation are attached. There are also plots of latency measurement.

In **Appendix B** more results of the listening test are presented. It gives an overview of the evaluation of each speaker and each listener.

2 Theoretical and technical background

2.1 Anatomy of speech

Natural speech. Speaking is a complex process where respiration, vocalization and articulation have to be complied. Figure 2.1 shows these parts in detail. The respiration needs an interaction between diaphragm, chest and abdominal muscles, lungs and trachea to produce an air stream. In the larynx the vocal chords modulate the airflow into speech by cutting it off periodically. The air stream opens the vocal fold between the vocal chords but due to elastic properties and Bernoulli forces the vocal chords are pulled back together immediately. Dependent on the lengths and tension of the vocal chords different tone pitches can be produced. For normal breathing and unvoiced sounds the fold is open so that the air stream can pass through nearly without any resistance.

The articulation occurs in the vocal tract. With tongue, mouth, teeth and lips all kinds of speech sounds can be produced. The pharyngeal, oral and nasal cavity act as a resonator and strengthen and filter the speech and give it an individual and personal sound [8, 11, 12].

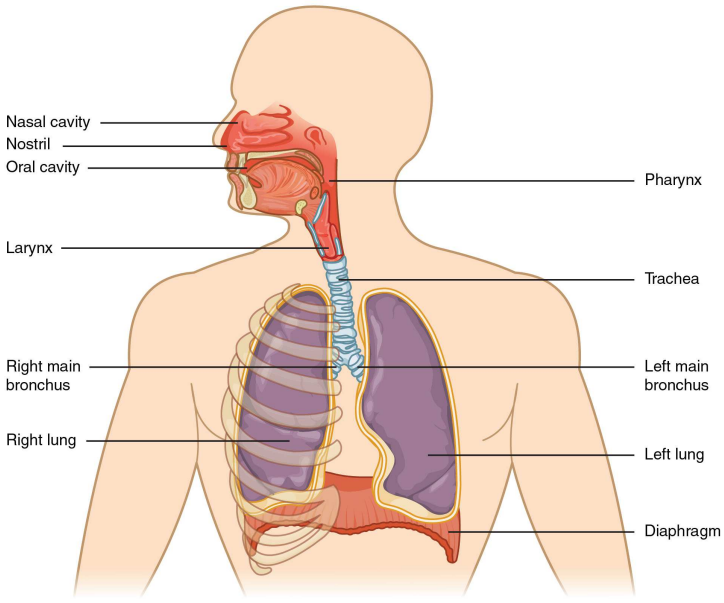
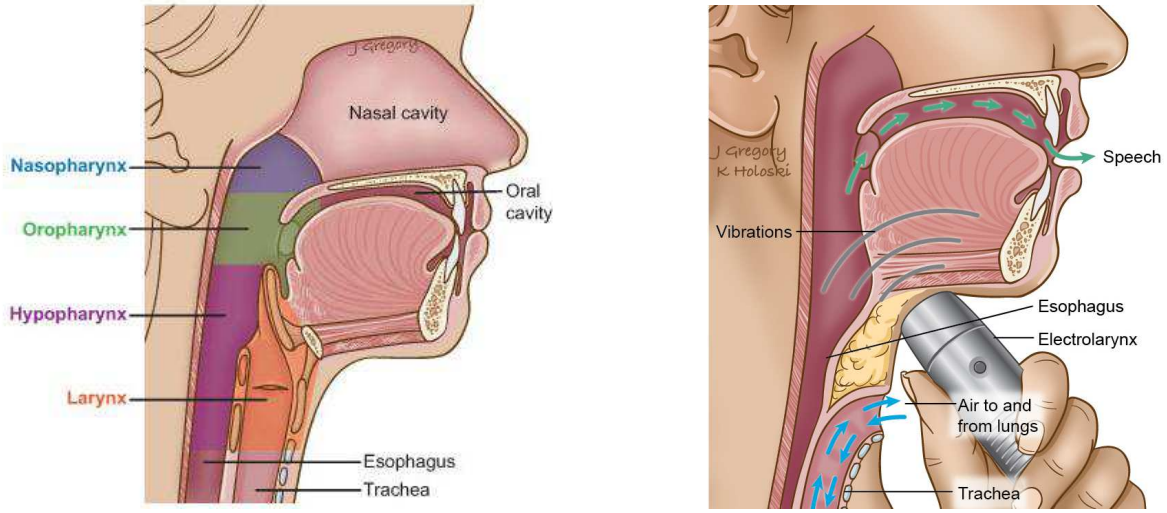


Figure 2.1: Anatomy of the human respiration system [13].

Alaryngeal Speech. In the case of a total laryngectomy the whole larynx has to be removed. The larynx can be seen as a valve which prevents foot, liquids and saliva from percolating into the lungs. After removal there is no longer a connection between mouth/nose and lungs and the trachea is attached to a hole, the so called laryngostome or stoma, placed in the front of the neck to enable breathing (compare figure 2.2(a) to figure 2.2(b)).

Thereby, it is now impossible to produce voiced speech since vocal chords are removed. Only the production of unvoiced consonants is still possible. For speaking using an electrolarynx (EL) the device is held against the neck. It induces the sound into the oral and pharyngeal cavities so that the user can now modulate the vibrations into speech by using the articulatory mechanisms of the vocal tract [3, 14, 15].



(a) Healthy vocal tract

(b) Laryngectomee using an EL

Figure 2.2: Anatomy of a healthy vocal tract and a vocal tract after ectomy of the larynx [16].

2.2 Technical devices

Electrolarynx. The idea of using an artificial larynx or electrolarynx (EL) arose in 1876. In 1930 Bell Telephone Laboratories started developing the first models which had various deficiencies and in the 1950's the first usable devices were developed [14]. Until presence the fundamental basic technology has not changed. Figure 2.3 shows some different types of common used EL. Some devices offer the opportunity to change the fundamental frequency dynamically but the major drawback is still a monotonic voice and the necessity to use a hand to hold the device against the neck and turn it on and off.

The electrolarynx is a battery driven hand-held device which induces vibrations into the oral and pharyngeal cavities to produce speech as mentioned in section 2.1 before. The vibrations are produced by an electromechanical vibrator with a fundamental frequency adapted to a male or female voice.

Surface Electromyography. To remove the drawback of utilizing a hand to control the EL the usability of the device can be improved by using two different methods of electromyography (EMG) to detect speech activity and provide hands-free interaction.

Surface electromyography (sEMG) is a common non-invasive method to measure motor unit action potentials (MUAPs). Motor units are the basic functional units for excitation and contraction in muscles and they can be activated and controlled voluntarily by the nervous system

Figure 2.3: Different electrolarynx devices [17].

and the brain. To record the external activity of a muscle, a so called EMG, the rhythmic series of action potentials of the motor units are detected by electrodes placed on the surface of the muscle. The size of contraction of the underlying muscle correlates directly to the measured EMG-activity. The activity mostly comes from tongue root musculatur and suprahyoid musculatur. These muscles are used by healthy speakers for articulation and laryngeal control [4, 8, 18–20].

The second form, the invasive needle electromyography is an accurate method and often used in medical diagnostics but intra-muscular recordings can be inconvenient for the patient. So in this thesis surface electromyography (sEMG) is used due to the fact of an easier handling.

Figure 2.4 shows the recorded EMG-signal (black) of the electrodes from a male person during the articulation of three words. During speech activity also EMG-activity increases. For a better display the speech signal (blue) is shifted by +1.5. An EMG-signal has its most energy in the frequency range of 10 to 500 Hz with amplitudes in the range of tenth of mV [6]. The envelope of the EMG-signal (red) helps to provide a continuous excitation signal for the EL. The idea behind sEMG is that values of the envelope above a defined threshold turn on the EL. If the envelope falls below the threshold the EL stops. Now the EL can be controlled without any interactions of hands.

Hard- and software. Figure 2.5 illustrate the test set-up which consists of two separate paths, the recording path and the EL-path. For pre-training and post-training recording the software SpeechRecorder [21] was used. The recording settings were set to 44.1 kHz and 16 bit. To ensure that all recordings can be done under the same condition and that the distance between microphone and mouth is the same for all participants a headset (condenser microphone AKG HC 577L) was used.

The electrodes measure the EMG-signal. The following bio-shield, a ARDUINO©DUE compatible bio-signal shield with an ARDUINO©DUE micro-controller board was developed in a former thesis by Amon [4]. It amplifies and filters the EMG-signal for the threshold detection in MATLAB [22]. For training and recording a root mean square (RMS) envelope calculation in combination with single threshold detection (STD) was used. The excitation signal of the shaker (electrolarynx), a gaussian pulse, was produced in MATLAB.

To ensure that training and recording sessions can be done in real-time it is necessary to figure out which of the two available systems, one is using Playrec and one is using Data Acquisition Toolbox, has the shortest latency period. This was done by sending a 1 kHz rectangular wave impulse with the signal generator Agilent 33120 A into the computer and measuring the time difference between the input and the output with the automatic delay-measurement of an Agilent 54622 D oscilloscope to get an approximate value. After this a 1 Hz rectangular wave was used to measure the exact latency. Figure A.1(a) in appendix A shows the oscilloscope screen with the rectangular input and the the delayed output using the Data Acquisition Toolbox. This system has a latency of 244 ms. The system using Playrec, which was used in this thesis is

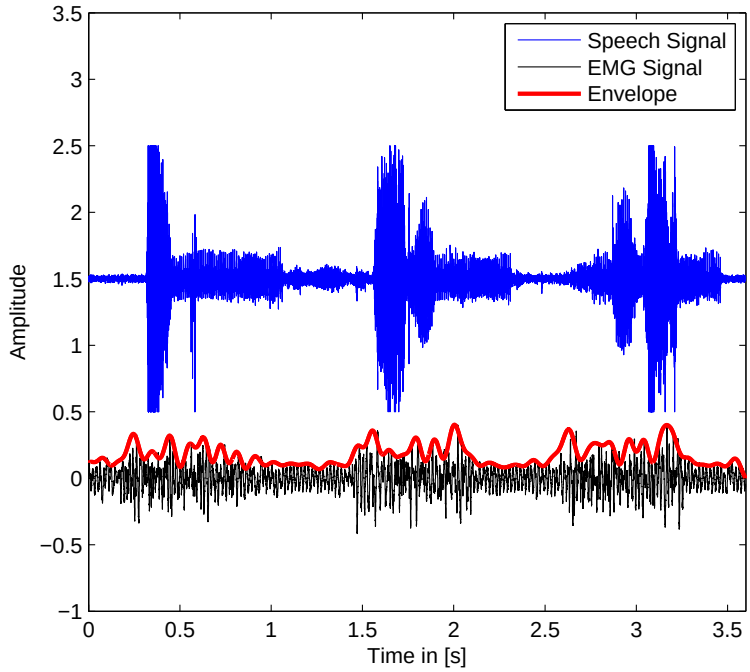


Figure 2.4: EMG-signal with its envelope and speech signal of the words "schwimmen", "schaffen" and "waschen".

shown in figure A.1(b) and has a latency of 76 ms.

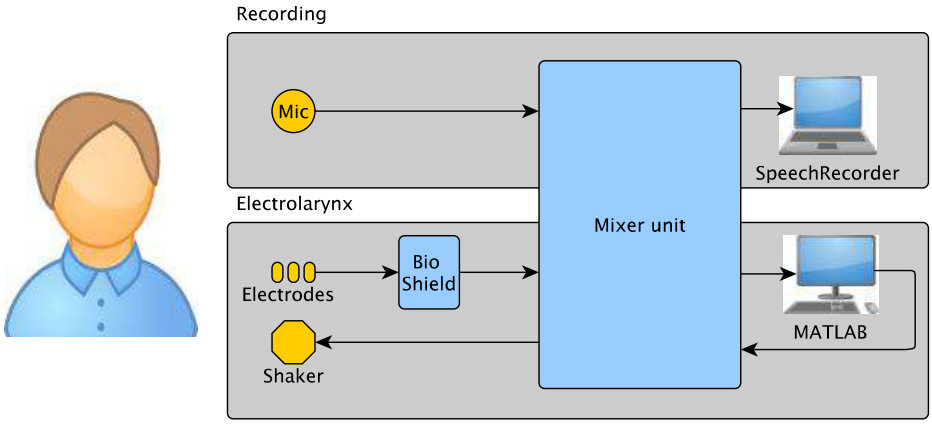
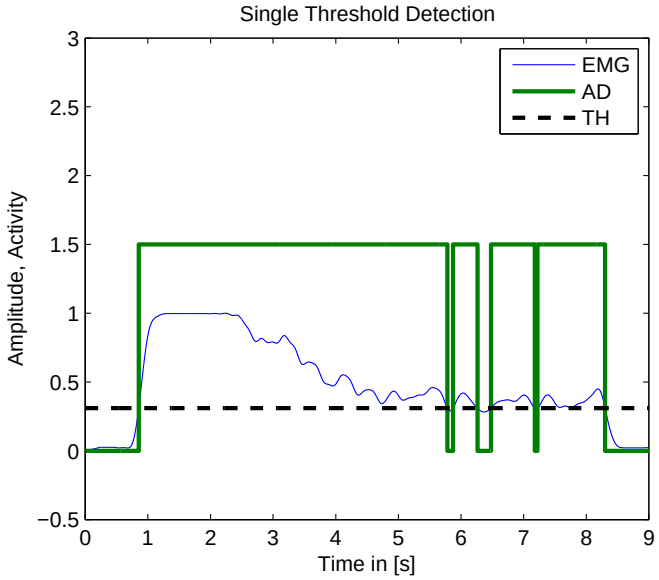


Figure 2.5: Block diagram of the test set-up.

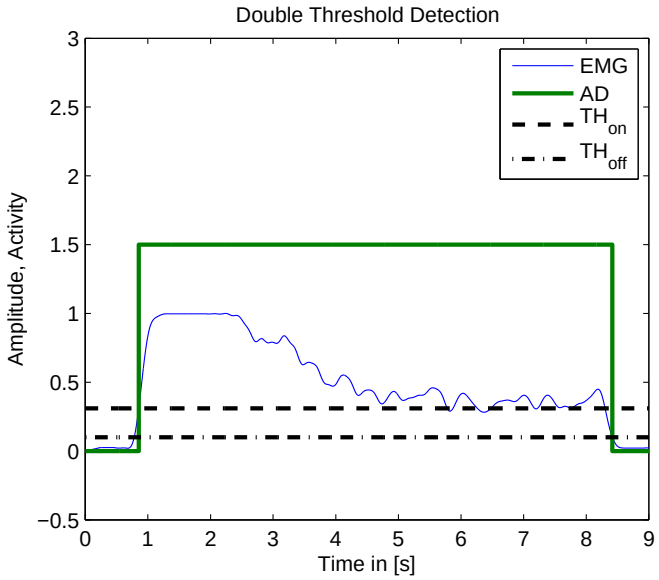
Single Threshold vs. Double Threshold Detection. The single threshold detection (STD) is the simplest activity detection algorithm in EMG-processing. Values of the EMG-signal above the threshold turn on the EL and values below the threshold turn off the device. Figure 2.6(a) shows a not optimal adjustment of the single threshold. The threshold is set too high, so a low EMG-signal can cause interruptions of the electrolarynx. To avoid these interruptions here the threshold has to be set lower but then disturbing noises brought about by e.g. swallowing can turn on the device accidentally.

To avoid this problem the double threshold detection (DTD), as can be seen in figure 2.6(b) uses a combination of two thresholds where the offset threshold is below the onset threshold so

the amplitude level of the EMG-signal has to fall under the lower offset threshold to turn off the device. The consequence should be a EL-speech with less interruptions. For the example in figure 2.6 the DTD is the better solution. According to Amon [4] the optimal threshold for STD is between 15 % and 25 % of the maximum amplitude.



(a) Single threshold



(b) Double threshold

Figure 2.6: Differences of activity detection (AD) between single and double threshold detection. Note that TH of single threshold is equal to TH_{on} of double threshold.

Electrodes. Using self-adhesive electrodes are an easy way to measure MUAPs. There are wet and dry sEMG-electrodes available. Wet electrodes need a conductive gel to ensure good conductivity.

All of the tested electrodes shown in figure 2.8 provide a very good signal quality. The main benefits of these electrodes are the uncomplicated and hygienic handling and the prevention of background noise. So the alternating current hum does not interfere the signal and also the

close placement of the electrolarynx (EL) to the electrodes does not cause any problems. There is almost no feedback between the EL and the electrodes.

In the thesis by Amon [4] a sensor strap with reusable Ag/AgCl (silver/silver-chloride) electrodes was used. Due to regularly usage the connection between electrodes and the wiring became fragile and caused a lot of noise. Also the close placement of the EL next to the electrodes caused some feedback problems. To get rid of these problems the Skintact®RT 34 electrode shown in figure 2.8(a) was used for the training described in chapter 3.1. These electrodes are the cheapest of the tested objects ($\sim 0.1\,€$/piece) and the connection can be easily made by alligator clamps.

To be sure to get comparable results a defined position of the self-adhesive electrodes is necessary. Figure 2.7 shows the ideal position according to Amon [4] and Stepp [8]. The two signal electrodes are placed with a distance of 5 mm from each other on the sternocleidomastoid muscle (lat. Musculus sternocleidomastoideus) right or left to the larynx. The third reference electrode is placed on the backside of the neck near the spinal column.

The EL is placed above the electrodes and could be adjusted by the subjects itself so that they can easily produce intelligible speech. An important part is the skin preparation before putting on the electrodes. The skin should be cleaned with a disinfectant to ensure good conductivity and adhesive bond and depending on the design of the electrodes the usage of electrode cream is recommended to receive a good conduction. For training the participant should relax to avoid unwanted muscle tensions. During training it turned out that abdominal breathing causes less neck muscle tensions than chest breathing and that also swallowing can produce enough EMG-signal to turn on the electrolarynx.

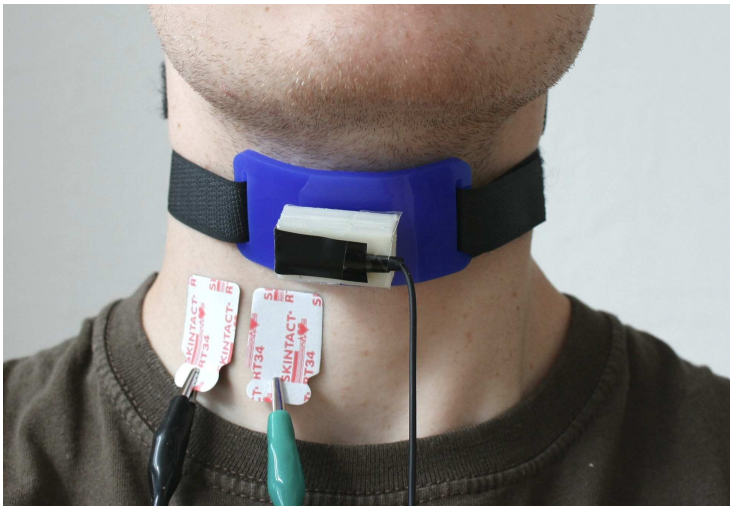
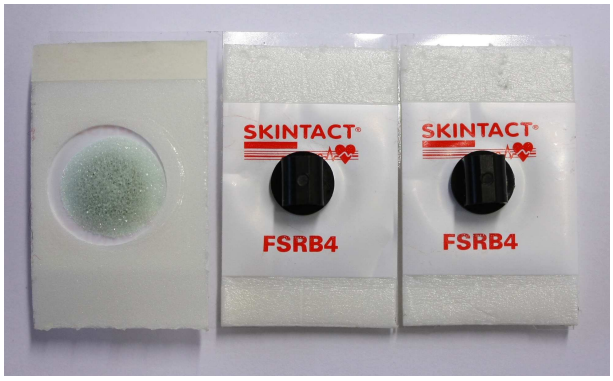
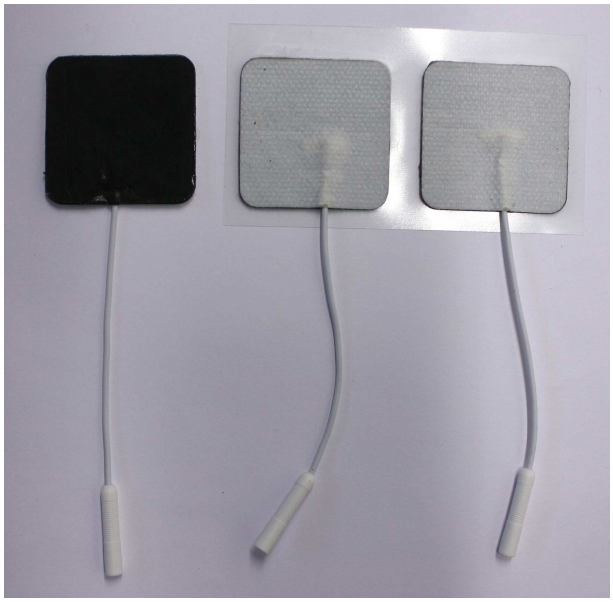


Figure 2.7: Position of self-adhesive electrodes and electrolarynx (white box). The third reference electrode is placed at the back of the neck near the cervical spine.

(a) Dry electrode: Skintact®RT 34, 32 x 41 mm



(b) Wet electrode: Skintact®FSRB 4, 44 x 28 mm



(c) Dry electrode: Textile Electrode tens/ems, 40 x 40 mm

Figure 2.8: Different wet and dry self-adhesive electrodes. For training and recording the Skintact® RT 34 was used.

3 Training Protocol

Since the electromyographically controlled electrolarynx (EMG-EL) is used by humans the handling with the device can be trained to improve the speech results. To figure out the detailed improvement a defined training protocol was used for all participants.

3.1 Procedure and speech material

The study involved four participants, three women (F1, F2 and F3, ages 28, 27 and 33 years) and one man (M1, age 40 years). All of them were students or graduated students with healthy vocal tracts and normal neck anatomy.

The training protocol for all four participants consists of nine training sessions which had to be done within two weeks. The first session of the training protocol was used to get in contact with the performance of the device and administration. The participants had time to try out how to produce intelligible speech for the following recordings. The second and last session was used for making pre- and post-recordings of the speech material. The recorded material was used for the evaluation in chapter 5. After the second session training started where participants were asked to repeat exercises for approximately 50 to 60 min while getting feedback about their performance and instructions for the improvement. Each participant starts with vowel initiation and moves on with vowel duration, vowel termination, words, sentences and paragraphs. For each skill one session was used so that the training was finished after nine sessions. Table 3.1 gives an overview of the training protocol.

The following training criteria are based on Goldstein et al. [5] but were adapted for this experiment:

Vowel initiation. To practice vowel initiation and to figure out the voice initiation time (VIT) the vowel "a" is used so that the participant is not distracted by articulation and can concentrate on muscle activities [7]. The participant should learn how to produce an adequate neck muscle EMG-activity to turn on the device quickly and consistently. While not speaking the participant should learn how to relax the neck muscles to prevent to turn on the electrolarynx unintentionally. For the evaluation 40 vowels were recorded once at the pre- and once at the post-recording.

Vowel duration. The participant should produce continuous vowels for different time intervals with 2 sec, 2.5 sec and 3 sec. Therefore the EMG-signal must be high enough during the whole stimulus period. For the evaluation in section 5.3 forty vowels "a" were pre- and

Table 3.1: Training Protocol with six consecutive training steps, two recording sessions and a try-out day. Sessions should be held off within two weeks.

Day	Exercise
Day 1	Time to try out the device
Day 2	Pre-Training recording
Day 3	Vowel initiation
Day 4	Vowel duration
Day 5	Vowel termination
Day 6	Words
Day 7	Sentences
Day 8	Paragraph
Day 9	Post-Training recording

post-recorded. For a successful stage a sustained voicing without any interruptions during the whole time interval is required. More details are explained in section 3.2.

Vowel termination. The participant should learn how to relax the neck muscles to stop the device and articulation to achieve interword pauses during an ongoing speech. For evaluation 40 vowels "a" were used like described.

Words. To combine the articulation with the skills trained during the last steps the participant should read words from a printed list. For recording the words in table A.1 in appendix A are used. These words are phonetically balanced with a focus on the consonants "m" and "n", because these cause more problems to produce a proper EMG-signal. For the evaluation and the listening test 40 words were pre- and post-recorded. For training a second list with 200 different words was used to avoid habituation. A token can count as successful if the device starts and ends at the same time as the first and last phoneme in the word, respectively and when an uninterrupted voicing of the articulation of the word is achieved so that the word is fully intelligible.

Sentences. For the evaluation of sentences the participant should read 30 sentences from a printed list shown in appendix A. The execution was the same as for words. For the training a second list with approximately 350 sentences was used.

Paragraph. For the evaluation of the paragraph the participant should read *Der Nordwind und die Sonne* from a printed list shown in appendix A. For training the paragraphs *Unser Garten* and *Die Buttergeschichte* were used. During a phrase there should be no interruptions in voicing and the participant should string several phrases together. Pauses between words and phrases are allowed.

3.2 Voice initiation time, voice duration and voice termination time measurement

To train vowel initiation, duration and vowel termination the software SpeechRecorder [23] was used. Figure 3.1 shows the useful traffic light of SpeechRecorder. It was used as an optical signal to define different pre-recording, recording and post-recording times. To avoid the

effect of anticipation and that participants get used to the traffic light phases during training and recording different time intervals where used so that is was necessary to be alert all time. Table 3.2 shows the different time intervals of the traffic light for voice initiation time (VIT), voice duration (VD) and voice termination time (VTT) measurement.

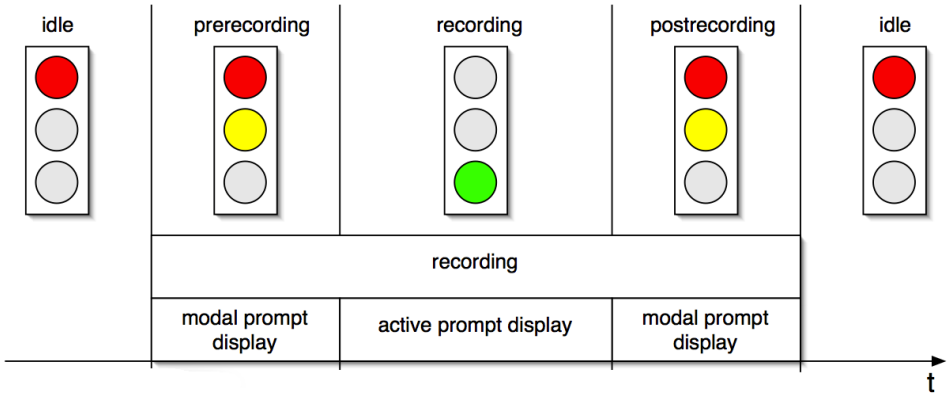


Figure 3.1: Recording phases of the traffic light of the recording software REAPER. *First yellow phase: pre-recording, green phase: recording, second yellow phase: post-recording [21].*

Table 3.2: Recording phase times for voice initiation time (VIT), voice duration (VD) and voice termination time (VTT).

Measurement	Pre-Rec [sec]	Rec [sec]	Post-Rec [sec]
VIT	1.0/1.5/2.0	2.5	1.0
VD	1.0	2.0/2.5/3.0	1.0
VTT	1.0	2.0/2.5/3.0	1.5

The measurement of VIT, parts of VD and VTT was done by hand with the software REAPER [24]. Therefore the time-cursor was set to the moments when the EL started and stopped. By subtracting pre-recording and recording-times of table 3.2 from the received time of the cursor VIT and VTT could be calculated. Figure 3.2 shows a recorded vowel with 1 sec pre-recording time, 2 sec recording time and 1.5 sec post-recording time. With these defined times a precise measurement of VIT and VTT was possible.

To get a rough overview of the VD the amount of errors was counted by hand. If there was a continuous EL-signal during the whole recording time or green phase it was counted as successful. If there were one or more interruptions as shown in figure 3.2 it was counted as an error.

But since this "passed/failed" evaluation just gives a small amount of information over the speech data a more detailed examination is needed. Therefore, in accord with Amon [4] and Beritelli [25] the error is divided into front end error (FEE), middle speech error (MSE), back end error (BEE) and noise detected as speech (NDS). Figure 3.3 presents the definition of the four errors. These errors are dependent on time appearance of the error compared with the ground truth.

The ground truth (GT) is a combination of the recording phase times, the reaction time of the speaker from the visual cognition to the begin of articulation and the delay of the system and it defines the starting point and endpoint of the articulation of the speaker. As shown as in figure 3.3 the front end error describes the time corridor 200 ms before and after the onset

event of the ground truth. The back end error describes the time corridor 200 ms before and after the offset event of the ground truth. The middle speech error describes the time corridor in between and the NDS describes the time corridor before the FEE and after the BEE and detects unwanted activations of the EMG-EL. According to equation 3.1 the total error is the averaged sum of all separated errors for number of N vowels i.

$$Total\ error = \frac{1}{N}\sum_{i=1}^{N} \mathrm{FEE_i} + \mathrm{BEE_i} + \mathrm{MSE_i} + \mathrm{NDS_i} \tag{3.1}$$

The block detection ratio is defined as the ratio between the number of active blocks in the detection vector and the desired number of blocks in the ground truth vector. A block detection ratio (BDR) of 1 would be a perfect result [4, 26].

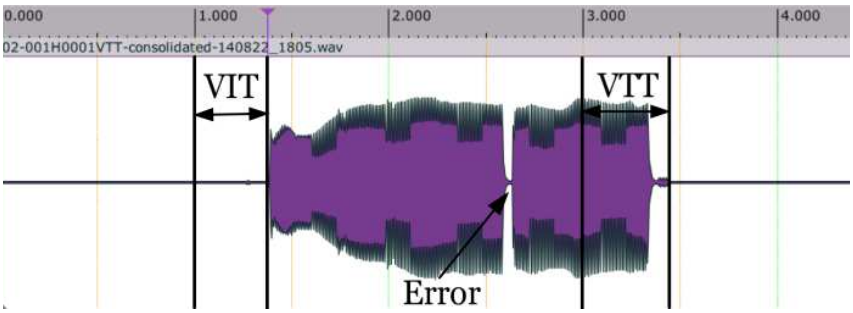


Figure 3.2: Recorded vowel "a" in REAPER *with the definition of VIT, VTT and error.*

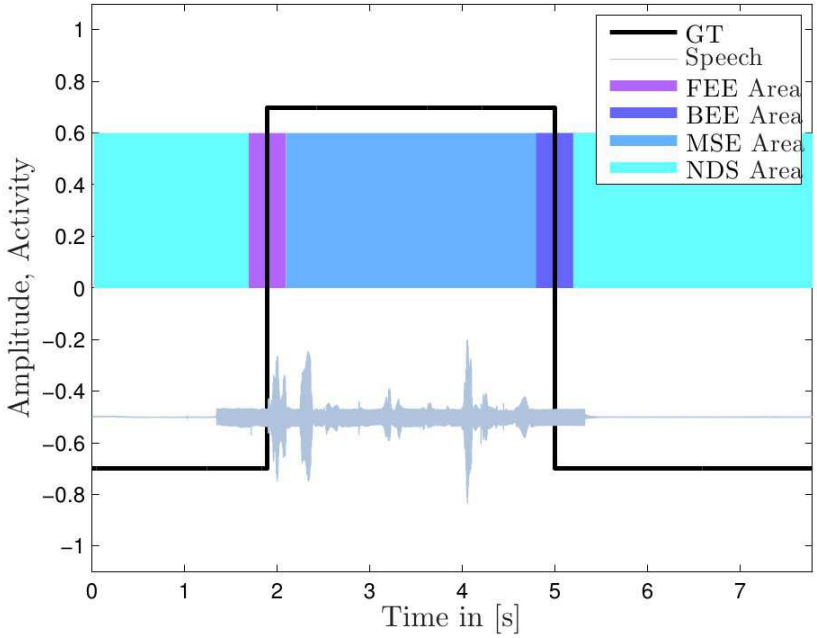


Figure 3.3: Regions of separated errors front end error (FEE), middle speech error (MSE), back end error (BEE) and noise detected as speech (NDS) of a EL-sentence [4].

4 Listening Test

This chapter will show the different steps to build up a meaningful listening test. Starting with a hypothesis a corresponding test should be developed to validate the hypothesis and to get the anticipated answers.

4.1 General

To get meaningful results it is necessary to know what should be achieved. In this work an improvement of pleasantness of electromyographically controlled electrolarynx-speech should reached due to training. This includes a decrease of interruptions of the EL, the speaking rate, the conscious stops of speech at commas between two clauses and the way of articulation. With this information it is possible to establish a hypothesis which will be proven by the test. The listening test should be developed in a way so that with the received information an answer can be given.

Alternative hypothesis H_1: $H_1 : \bar{x} > \mu$

Post-training electromyographically controlled electrolarynx (EMG-EL)-speech is more pleasant than pre-training EMG-EL-speech.

Null hypothesis H_0: $H_0 : \bar{x} = \mu$

Training does not affect the pleasantness of electromyographically controlled electrolarynx (EMG-EL)-speech.

According to Rossum [9] the attributes naturalness and intelligibility can be determined by a pairwise comparison of examples. Here, the pre- and post-recorded speech sample of the same word, sentence and paragraph should be compared to detect the influence of training. So a AB listening test in accord with the comparison mean opinion score (CMOS) method [2] with a comparison category rating (CCR) scale was used for the evaluation. Since there is not much experience available in evaluation of EL-speech this standard was used for this thesis despite the fact that this assessment method is used for measurement of listening quality for healthy speech.

Thirteen listeners attended the test, three women and ten men between the ages of 22 and 40 (mean 27 years). Two of them had already experience with EL-speech, the rest were naive listeners. The speech material consists of 160 words, 120 sentences and 24 longer sentences of the paragraph.

The pre- and post-recording times of the stimulus material were removed in REAPER. To ensure that all samples have the same volume a RMS-adjustment was done in MATLAB [22]. The listening test was held in an ordinary office were the stimulus material was played via a laptop with MATLAB and AKG K 271 headphones.

4.2 Implementation

For the execution of the listening test a graphical user interface (GUI) was developed in MATLAB. Figure 4.1 shows the GUI for the evaluation of words. The buttons "Play A" and "Play B" play the recordings of the same example from pre- and post-recording sessions described in chapter 3.1. With the seven evaluation buttons it is possible to make a personal choice how much sample A is better or worse than sample B. The recordings were randomly distributed for the playback. Each listener got his or her own sequence of speech samples and also the pre- and post-recordings of the same sample were played back randomly. The evaluation for sentences and the paragraph works in the same way with a separate evaluation-GUI. Before each evaluation an information window describes the following task and offers two examples of a good working EL and two examples of a bad working EL to hear the possible range of quality.

The criterion for evaluation is the pleasantness of the EL-speech. If a participant is able to produce high muscle tension during speaking, which produces a good EMG-signal, the electrolarynx produce a continuous excitation signal. As a result continuous speech is possible. If the EMG-signal is not high enough there will be unpredictable interruptions in the speech which cause difficulties in intelligibility and increase annoyance. So speech interruptions play a major role in the evaluation but in the case that both speech samples contains EL-dropouts also the intelligibility of pronunciation itself should be considered for the evaluation by the test participants. In the case of uncertainty the example with better intelligibility should get better assessment.

Because of the big amount of speech samples and to avoid that listeners became unconcentrated due to a too long evaluation the listening test is divided into three parts (HT1, HT2 and HT3 in the evaluation plots) so that the execution of each part lasts only around 30 min. Each part was evaluated by four men and one woman. One of the male speakers evaluated all three parts with longer pauses between the tests. Table 4.1 gives an overview of distribution of the speech samples.

Table 4.1: Distribution and quantity of speech samples for the three parts of the listening test. Each part consists of speech samples of all four speakers, e.g. part 1 consists of 13 words, 10 sentences and 2 samples of the paragraph of speaker F1, the same applies to F2, F3 and M1.

	Words	Sentences	Paragraph
Part 1 (HT1)	$13 * 4 = 52$	$10 * 4 = 40$	$2 * 4 = 8$
Part 2 (HT2)	$14 * 4 = 56$	$10 * 4 = 40$	$2 * 4 = 8$
Part 3 (HT3)	$13 * 4 = 52$	$10 * 4 = 40$	$2 * 4 = 8$
Total	160	120	24

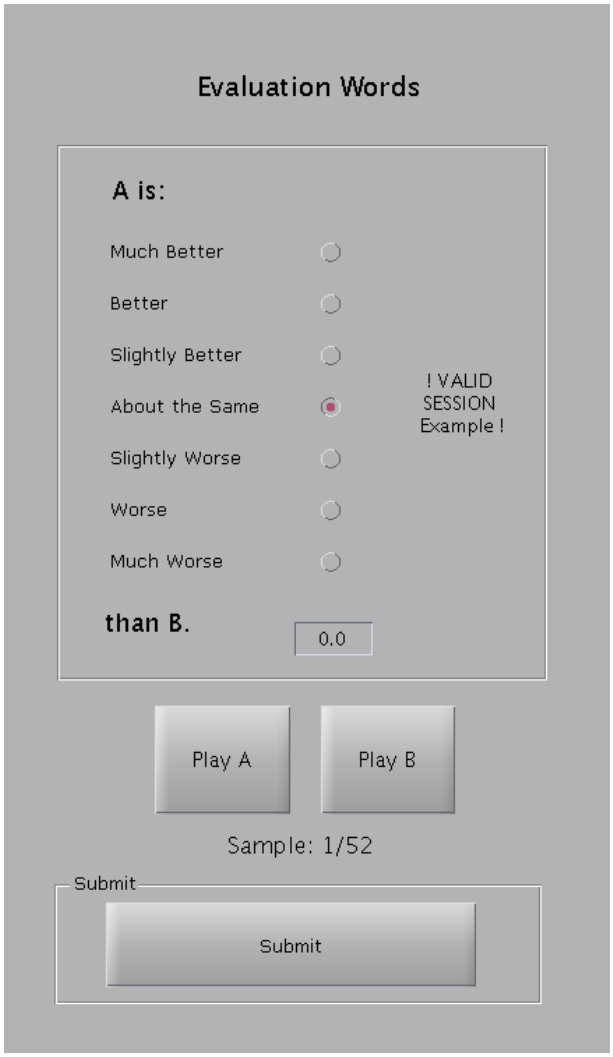


Figure 4.1: MATLAB-*GUI for evaluation of words with two playback-buttons, seven evaluation options, a sample counter, a display panel for the personal choice and a submit-button.*

4.3 Significance test

The results of the listening test in section 5.5 illustrate an obvious improvement of EMG-EL-speech due to training. To prove that these results are meaningful a hypothesis validation on the expected value according to Brell [27] has to be made.

Table 4.2 shows the calculation of hypothesis validation. After formulation of the null hypothesis H_0 and alternative hypothesis H_1 in section 4.1 the significance level was set to $\alpha = 0.05$. Conformable to the null hypothesis the asserted median μ was assumed with $\mu = 0$ since if there is no difference due to training all samples would be evaluated with "0" (About the same). The alternative hypothesis asserts that the median μ will be above "0". So post-training will be slightly better, better or much better than pre-training. Because of using a ordinal scale in the listening test the median and not the mean value is used for calculation.

After calculation of the standard deviation s of the evaluated samples, the standard deviation of the median σ_M and the test factor T_{pruef}, the critical values $z_{1-\alpha}$ could be read out of the standard deviation table for $P = 0.95$. Despite the fact that evaluation data does not have normal distribution (Lilliefors test in MATLAB) this is possible since the quantity of evaluations is $n > 30$. A comparison of test factor T_{pruef} with the critical values $z_{1-\alpha}$ shows that the null

hypothesis H_0 can be refused and the alternative hypothesis H_1 can be accepted in all cases. So the results in section 5.5 show a trustworthy improvement of pleasantness in EMG-EL-speech for words, sentences and the paragraph. Also figure 4.2 supports this statement. It shows boxplots of the distribution of evaluations from the listening test. For words, sentences and the paragraph the medians with 25th and 75th percentiles are above "0".

The author is aware that this significance test does not provide statistically meaningful results and just gives a rough estimate because the test's population is based on normal distribution. Due to the used CCR-scale in the listening test the received evaluation data distinguished from normal distribution.

To find out more about the validity of the listening test in appendix B figures B.8 to B.14 show the distribution of judgements of each recorded speech sample. Each speech sample was evaluated by five listeners. In almost all cases the listeners judged similar. Just for words in part 1 (figure B.8) and part 2 (figure B.9) of the listening test individual judgements differ a bit more.

Table 4.2: Hypothesis validation on the expected value with significance level $\alpha = 0.05$.

	All data ($n = 1520$)	Words ($n = 800$)	Sentences ($n = 600$)	Paragraph ($n = 120$)
Asserted median μ	0	0	0	0
Received median $\bar{x}$	1	1	1	1
Received mean value x	1.03	0.86	1.25	1.05
Standard deviation s	1.38	1.40	1.31	1.41
Significance level α	0.05	0.05	0.05	0.05
Probability $P = 1 - \alpha$	0.95	0.95	0.95	0.95
Normal distribution (Lilliefors test)	$\times$ $p < 0.001$	$\times$ $p < 0.001$	$\times$ $p < 0.001$	$\times$ $p < 0.001$
Standard deviation of median $\sigma_M = \frac{s}{\sqrt{n}}$	0.036	0.05	0.053	0.13
Test factor $T_{pruef} = \frac{\bar{x}-\mu}{\sigma_M}$	28.18	20.22	18.76	7.78
Critical value $z_{1-\alpha}$	1.645	1.645	1.645	1.645
Refuse H_0 and accept H_1 if: $T_{pruef} > z_{1-\alpha}$	✓	✓	✓	✓

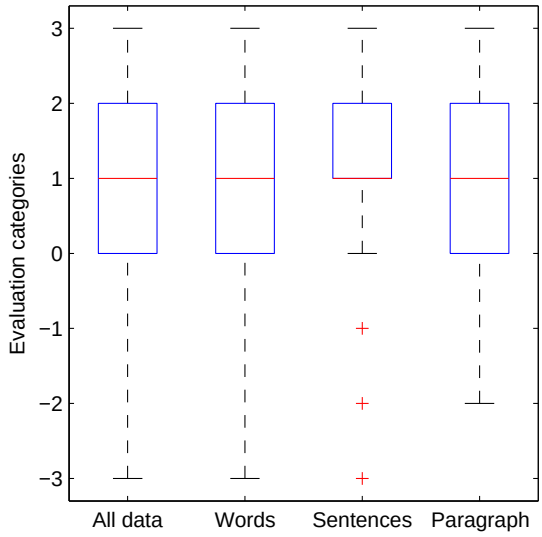


Figure 4.2: Medians with 25th and 75th percentiles of the listening test for all speech samples, words, sentences and the paragraph.

5 Evaluation

After the explanation of recording and evaluation of the speech material in the previous chapters here a detailed elaboration of the results will follow. At first there are general questions which arose during the thesis. In the next part the recorded vowels are used to examine the voice initiation time (VIT), voice duration (VD) and voice termination time (VTT) and to give a detailed overview on different kinds of errors. Also the results of the listening test will discussed here.

5.1 General questions

Does the EMG-signal change over time? The results in sections 5.2, 5.3, 5.4 and 5.5 show a noticeable improvement of speech quality and handling. Also the participants who did the training sessions in chapter 3 observed an improvement of the handling of the EL. To find an answer to this question the EMG-signal was recorded over a whole training session of one hour.

Table 5.1 exhibit the signal to noise ratio (SNR)-values of a female speaker, once for words and once for sentences and for a male speaker who spoke words. For the calculation a 30 sec interval of the EMG-signal from the begin and the end of a training session was used. It can be seen that the SNR raised by 1 dB for the female speaker and 2 dB for the male speaker during a one hour training. Unlike the thesis by Amon [4], he measured a higher EMG-SNR for male speakers, here the male SNR is worse. Although the male EMG-signal in general is higher than the female one a bad adhesion of the electrodes can cause unwanted noise which lower the SNR. To avoid this problem good skin preparation is recommended.

Table 5.1: Signal to noise ratio (SNR) of the EMG-signal before and after a training session for male and female speaker for spoken words and sentences.

	Before	After
Female speaker: Words	9 dB	10 dB
Female speaker: Sentences	10 dB	11 dB
Male speaker: Words	8 dB	10 dB

Does the electrolarynx have an influence on the electrodes and the EMG-signal? If the EL is directly placed on the self-adhesive electrodes an influence of the excitation signal within the area of 800 Hz to 1800 Hz can be seen in figure 5.1(a). This impact can interfere the intentional deactivation of the EL and the device works until it is removed from the electrodes by hand so that the feedback is interrupted. Since nearly 90 % of the energy of an EMG-signal is between 10 Hz and 500 Hz and the signal is filtered in the bio-signal shield by a low-pass filter with a cut-off frequency at 1 kHz the influence does not cause any problems with a correct position of the EL. If the EL is placed on the larynx as shown in figure 2.7 there is almost no influence noticeable as can be seen in figure 5.1(b). As a result an additional adaptive filtering of the excitation signal was not necessary.

Is there a difference between single threshold detection and double threshold detection? A look on figure 5.2(b) illustrates the answer to this question. Using double threshold detection (DTD) only 12 % of the utterances had a mistake whereas single threshold detection (STD) causes errors in 45 % of all utterances. The speech material consists of 40 vowels "a" recorded by a female speaker with each one of the two detection algorithms. For the participant it was much easier to produce a continuous electrolarynx-speech with DTD but it caused some problems in turning off the device as can be seen on the right side of figure 5.2(a). Here the variance of the voice termination time is higher which indicates worse handling. Since DTD supports the production of continuous speech it is more difficult to turn off the electrolarynx. The mean value and the variance of the VTT increased.

5.2 Voice initiation time

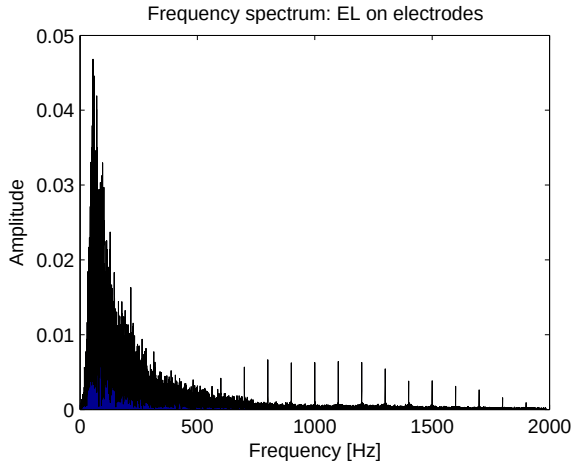
The voice initiation time (VIT), voice duration (VD) and voice termination time (VTT) can be seen as indicators for the handling of an EMG-EL. To turn on the device on purpose within a short time the VIT has to be as short as possible. In figure 5.3 there is a comparison of the VIT of all four speakers for pre-recording (left side) and post-recording (right side). The mean values of speaker F1 and F3 became a bit smaller, the mean values of speaker F2 and M1 became a bit bigger but all fluctuate between 400 ms and and 480 ms, depending on the alertness and condition of the speakers at the recording sessions. The VIT consists of the reaction time of the participant (time between visual recognition of the traffic light and muscle tension) and the delay of the system.

The main difference between pre- and post-recording is the variance. Before training the values varies between 100 ms and 750 ms so participants did not have an adequate influence on the device. After training the values fluctuate within an area of around 200 ms. This decrease of variance leads to an improvement of handling of the EMG-EL.

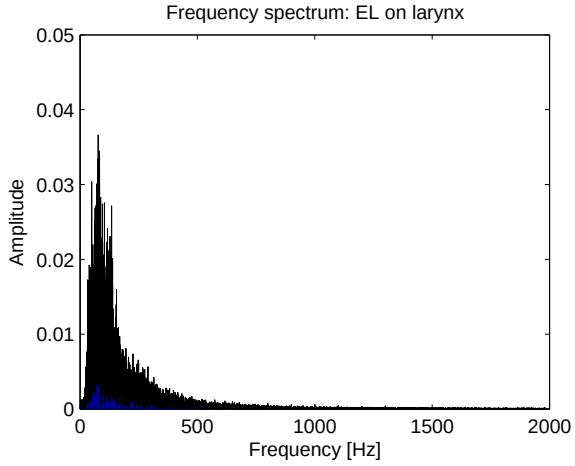
5.3 Voice duration

According to figure 3.2 a speech sample is counted as failed as long as there is at least one interruption or error in it. Figure 5.4 shows this "passed/failed" results for the 40 pre- and post-recorded samples of the vowel "a" for all four speakers. It is evident that training has a influence in maintaining a proper EMG-signal to produce a fluent speech. Before training between 38 % (F1) and 95 % (M1) of all samples had one or more interruptions. After the training sessions all participants could obvious reduce unwanted interruptions.

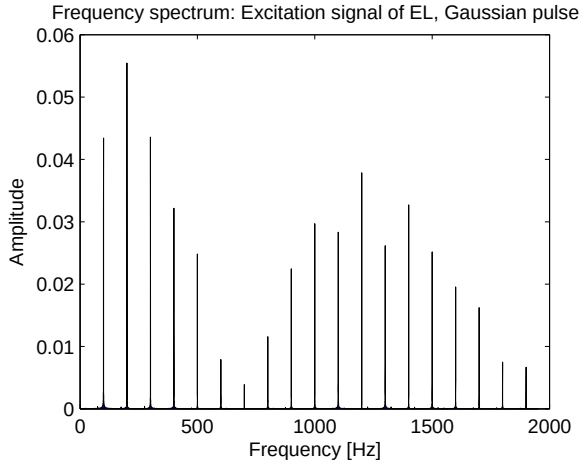
A more detailed examination of the errors in figure 5.5 points out some interesting correlations with the conclusion of sections 5.2. A comparison of the pre-training errors in figure 5.5(a) with



(a) Frequency spectrum of an EMG-signal with an electric larynx placed on electrodes.

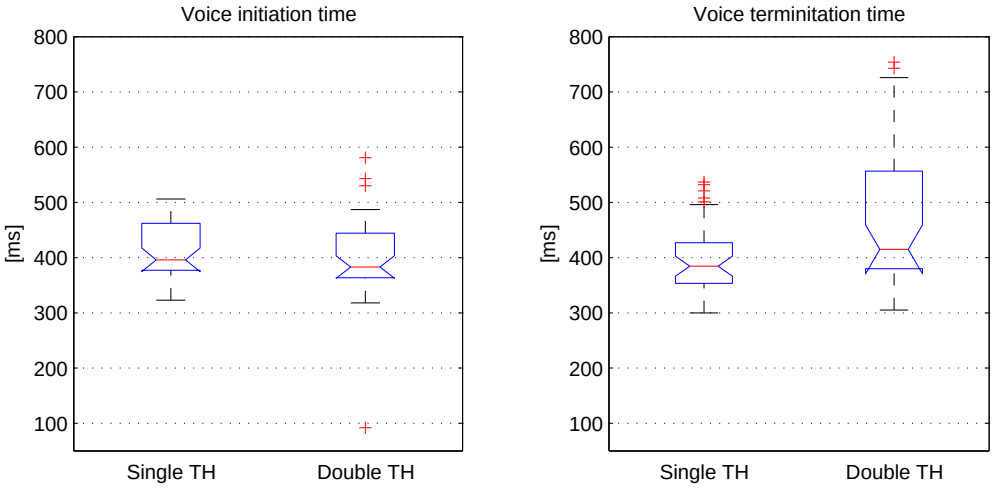


(b) Frequency spectrum of an EMG-signal with an electric larynx placed on larynx.

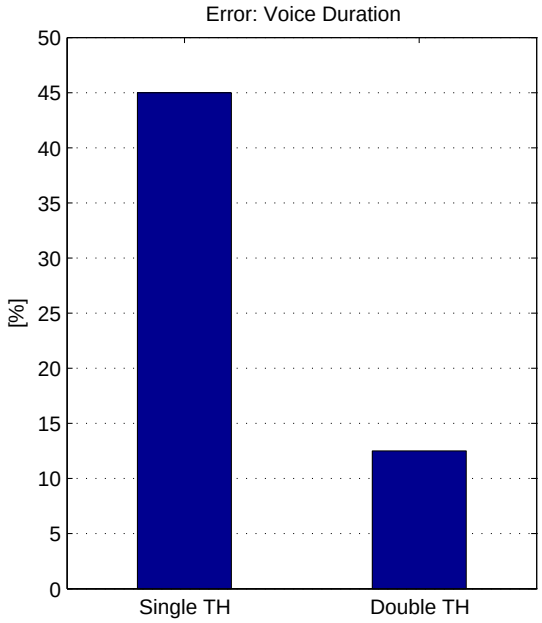


(c) Frequency spectrum of the excitation signal, a Gaussian pulse.

Figure 5.1: Frequency spectra of EMG-signals with different positions of the electric larynx and of the excitation signal.



(a) VIT (left figure) and VTT (right figure): Single vs. Double threshold detection.



(b) VD: Single vs. Double threshold detection.

Figure 5.2: Differences between single and double threshold detection for voice initiation time (VIT), voice duration (VD) and voice termination time (VTT).

the post-training errors in figure 5.5(b) shows at first a decrease of the total error for all four participants. The total error is the sum of the separated errors FEE, MSE, BEE and NDS Contrary to the error results before in figure 5.4 here the errors are normalized to the total length of the recorded vowels.

The decrease of the FEE correlates with the decrease of the variance of the VIT. The reduction of the NDS error of speaker F1 and F2 is caused by the increase of the threshold. Before the training the threshold had to be quite low due to a weak EMG-signal so the ground noise or some unwanted muscle activities (e.g. strong breathing, swallowing) can already activate the EL. After the training the threshold could be put on a higher level so that the disturbing noise did not have any influence but the participants were still able to produce a more continuous

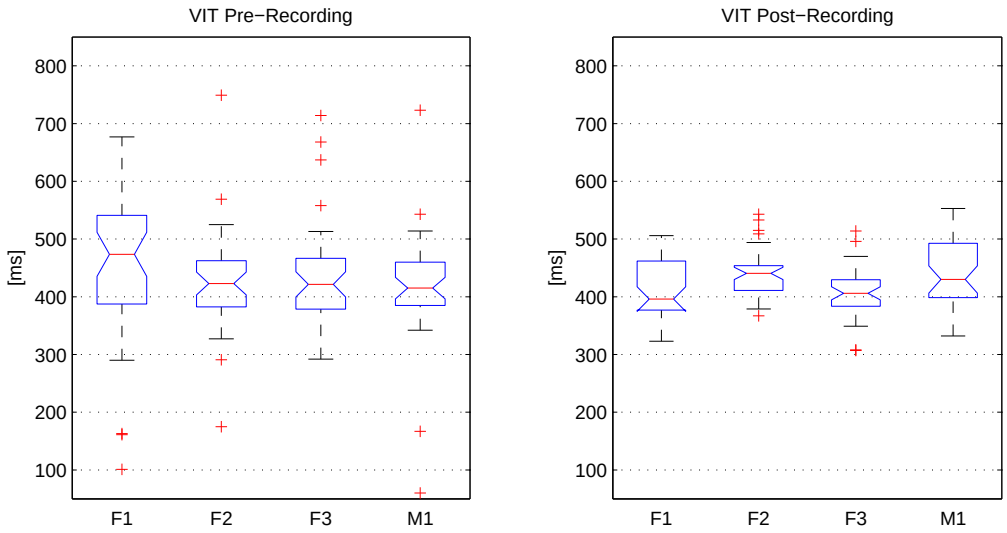


Figure 5.3: Voice initiation time (VIT) for pre- and post-recording for three female speakers (F1, F2, F3) and one male speaker (M1).

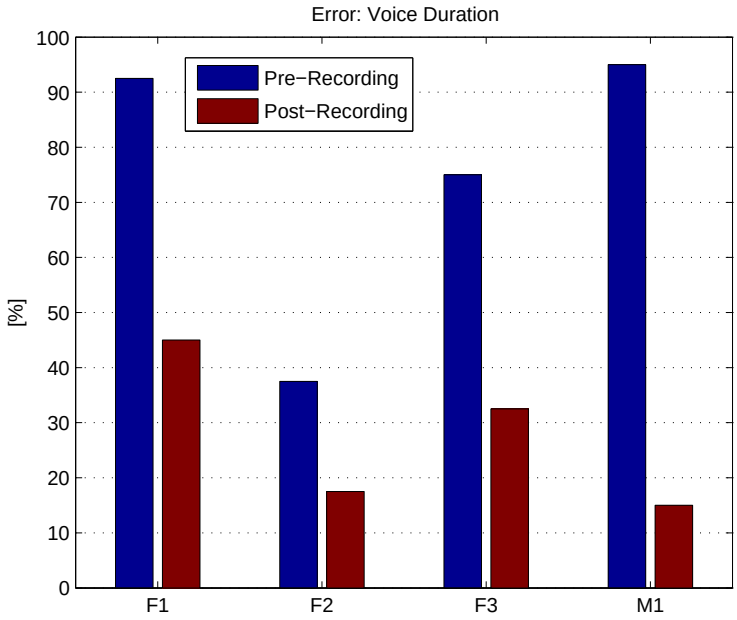


Figure 5.4: Errors of voice duration (VD) for pre- and post-recording for three female speakers (F1, F2, F3) and one male speaker (M1).

speech than before as the reduction of the MSE shows. The detection of interruptions during ground truth activity causes a BDR higher than the optimum of 1. It is connected to the middle speech error. So the received reduction of MSE leads to a decrease of BDR.

A comparison with Fuchs [26] and Amon [4] reveals some interesting improvements. Fuchs figured out a total error of 12.2 % for a female speaker and 12 % for a male speaker using the RMS envelope calculation method and STD. In this work female speakers had total errors between 4.1 % and 7.3 % at pre-training and these errors decrease to 0.8 % and 2.8 % at post-training.

The total error values of the male speaker also decrease from 9.6 % at pre-training to 2.1 % at post-training.

It should be mentioned that Fuchs and Amon used recorded sentences as speech samples for the evaluation of separated errors which in general are more difficult to articulate than the vowels used in this evaluation.

5.4 Voice termination time

The voice termination time (VTT) is an important factor to turn off the EMG-EL on purpose within a short time period. For all speakers it was quite difficult to produce a proper EMG-signal to get a speech signal without interruptions and to relax their neck muscles immediately after saying a word or sentence to turn off the device. Figure 5.6 shows a comparison of VTT from all four speakers for their pre- and post-recordings. The mean values drop by around 100 ms and also the variances decrease. These measured results indicate an enhancement of EL-handling and they correlate with the perceived perception of the participants. All of them could audibly improve their handling with the device within nine training sessions.

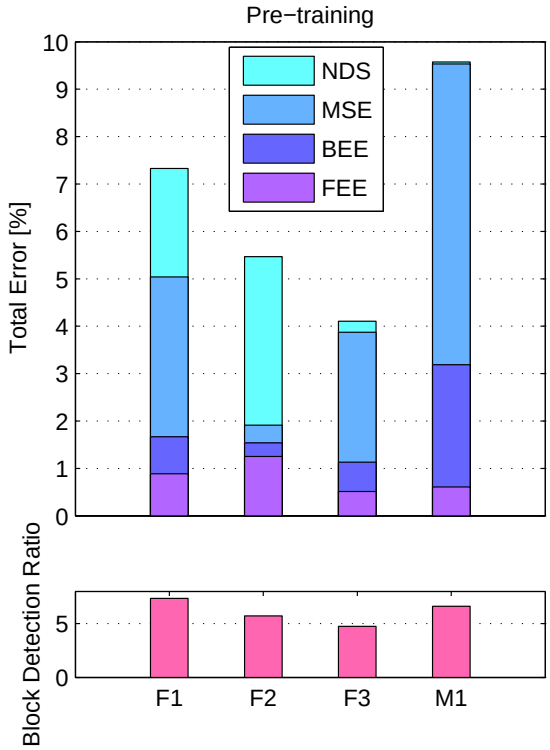
5.5 Results of the listening test

As already mentioned in chapter 4.1 thirteen people attended the evaluation of the listening test who evaluated collectively fifteen times. The used data consists of the evaluations of all three parts of the listening tests, fifteen evaluation units and four speakers and the whole recorded speech material. All in all there are 800 evaluations for words, 600 evaluations for sentences and 120 evaluations for the paragraph, altogether 1520 evaluations. For this examination a good result would be a choice of "slightly better", "better" or "much better", since the question was if sample A is better or worse than sample B. So sample A (post-recording) should be at least slightly better than sample B (pre-recording) to prove that training does have an effect on EL-speech.

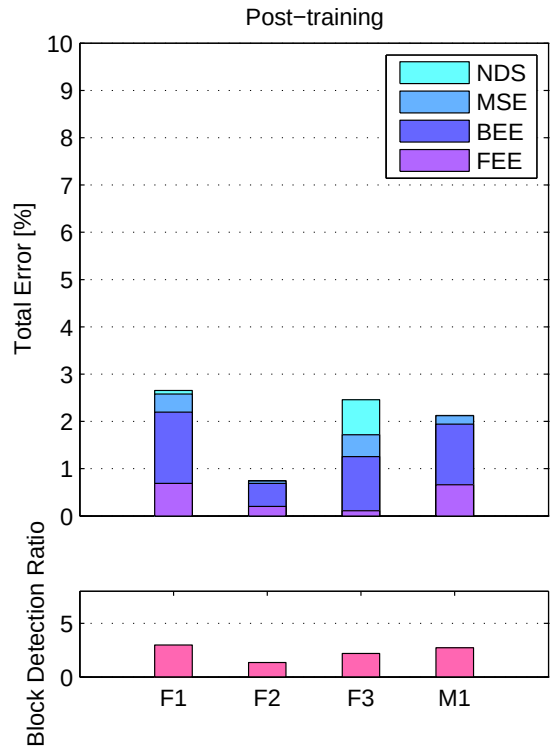
Figure 5.7 shows the main result of the test where test subjects determine that 71 % of all post-recordings are slightly better, better or much better than the recordings before training and only 13 % of the post-recordings were evaluated with slightly worse, worse or much worse. For the remaining 16 % no differences were recognizable.

To find out why some of the post-recorded samples appears to be worse than the pre-recorded ones a more detailed examination is necessary. Therefore in figure 5.8 the results are separated into words, sentences and the paragraph. Here, the results of words indicates a small tendency of a worse evaluation than for the rest. Due to the missing context words are more difficult to understand and pronounce even if the EL is working perfectly. Also one missing letter can already change the meaning of the word or make the word unintelligible, like a missing "g" in "Glas" or a missing "r" in "rot". This difficulties also can be seen in figures B.8 and B.9 in appendix B. Here, judgements varies a bit more than in comparison with judgements of sentences or the paragraph.

Figure B.3 in appendix B gives more answers about some unexpected evaluations. It shows the results for the evaluation of female speaker F3 separated into words, sentences, the paragraph and the three parts of the listening test. During training the speaker had big problems in producing understandable speech with the EMG-EL and holding her´s breath while speaking. So, especially on pre-recordings, sometimes a whisper was recognizable next to the EL-speech. Since people are more familiar with whispering than with EL-speech some of the listening test participants (VP5 in figure B.5 and VP10 in figure B.6 in appendix B) preferred the samples with the whisper as can be seen in the results of words and paragraph for listening test 1 (HT 1) in the figure.



(a) Errors pre-training.



(b) Errors post-training.

Figure 5.5: Separated errors: front end error (FEE), middle speech error (MSE), back end error (BEE), noise detected as speech (NDS) and block detection ratio (BDR) of voice duration (VD) before and after training for three female speakers (F1, F2, F3) and one male speaker (M1).

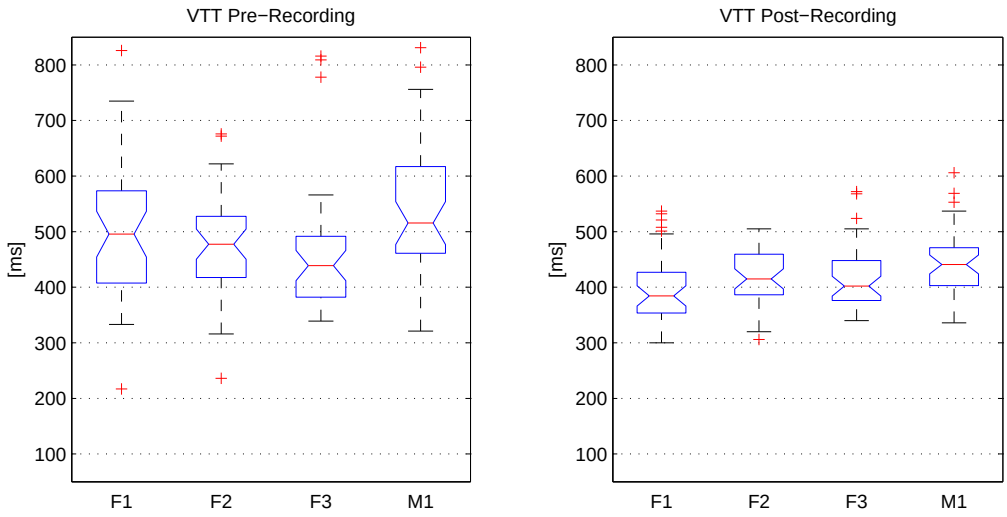


Figure 5.6: Voice termination time (VTT) for pre- and post-recording for three female speakers (F1, F2, F3) and one male speaker (M1).

Table 5.2 gives an overview of all "much worse" evaluations in the listening test with the impacted speech samples. A review done by the author revealed that only two of the sentences were evaluated unexpected. Here, the participants of the listening test maybe made a mistake in evaluation by mistake because post-recordings were obviously better than pre-recordings. The rest of the samples were evaluated reasonable where for female speaker F3 the influence of whispering is audible. Seven post-recording speech samples of speaker F3 were evaluated with "much worse" in comparison with pre-recording samples.

Female speaker F1 and male speaker M1 show a distinct improvement in speech quality as can be seen in figure B.1 and figure B.4 in appendix B were they got nearly perfect results in the evaluation of the paragraph and also sentences.

Table 5.2: Distribution and quantity of "much worse" evaluations of the listening test with impacted speech samples.

	Words	Sentences	Paragraph	Speech sample
Female speaker F1	0	1	0	Auf dem Brett leuchten bunte Tulpen.
Female speaker F2	1	0	0	furchtbar
Female speaker F3	7	2	0	2x Suppe, Arzt, Glas, furchtbar, Papagei, rot, Messer und Gabel liegen auf dem Teller. Jetzt suche ich das Weißbrot.
Male speaker M1	1	1	0	Nachrichten, Günther muss noch einkaufen gehen.

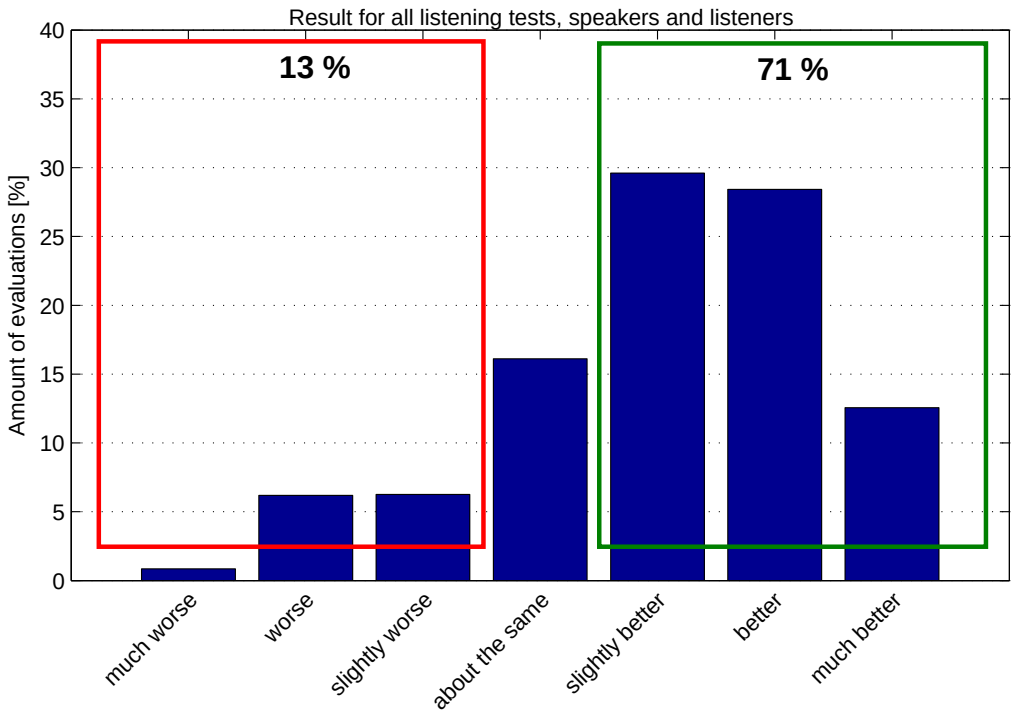


Figure 5.7: Results of all three parts of the listening tests, all speakers, the whole speech material and all listeners. (1520 evaluations)

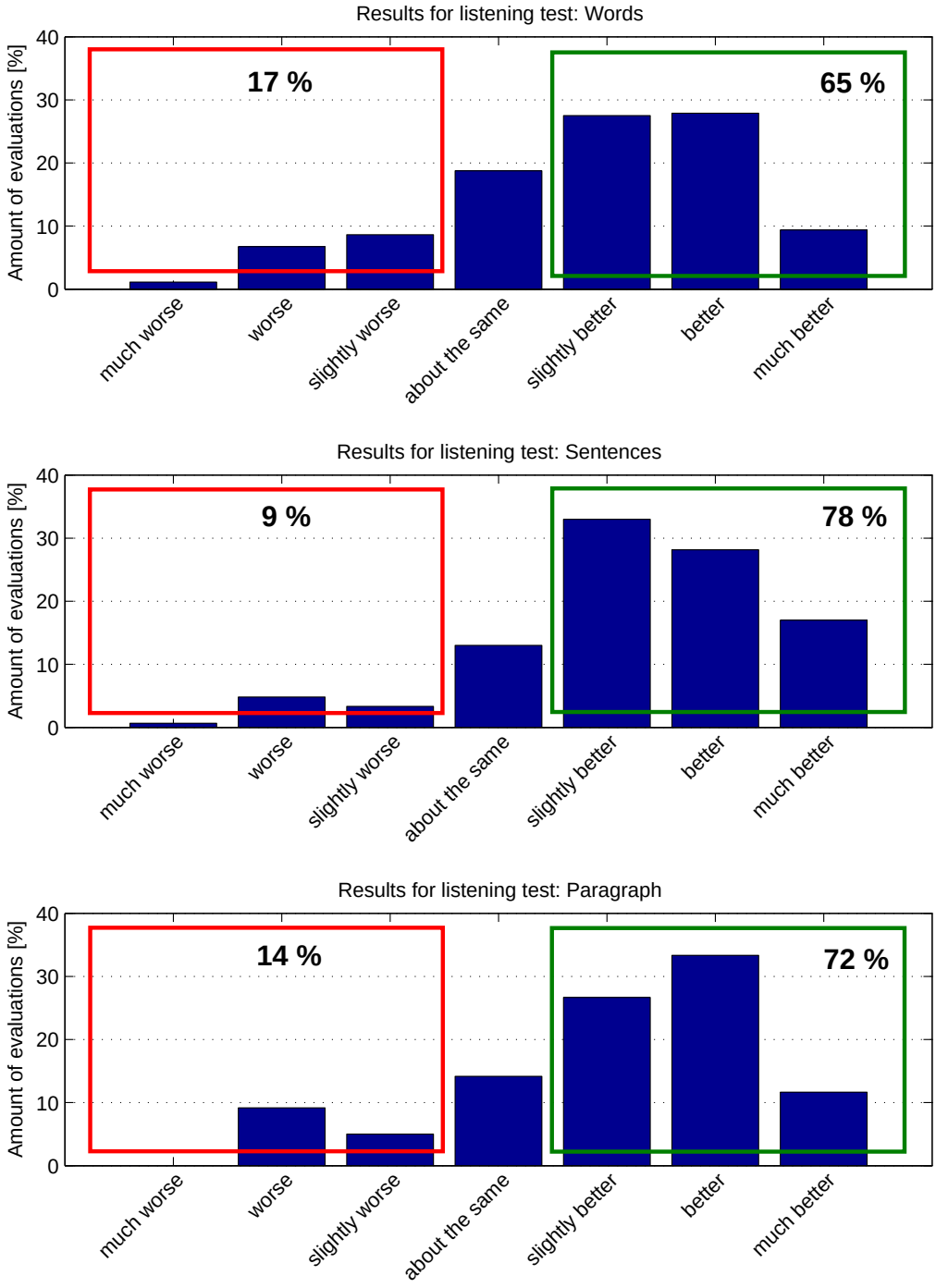


Figure 5.8: Results of the listening test for words, sentences and paragraph, averaged over all listeners and speakers. (800 evaluations for words, 600 evaluations for sentences and 120 evaluations for the paragraph)

6 Conclusion

6.1 Training

To examine the influence of training on EMG-EL speech a training protocol was developed. Three female and one male participant had to do nine training sessions á one hour with consecutive tasks. The program started with vowel initiation, followed by vowel duration and vowel termination. For that the vowel "a" was used so that the participants could concentrate on neck muscle tension and articulation to produce an audible sound. After that the training went on with the articulation of words, sentences and a paragraph. The participants should learn to produce an intelligible speech with defined turn on and off times of the electrolarynx. At the first and last day of training recordings were made for the evaluation.

With the recorded vowels the voice initiation time, voice duration and voice termination time was measured. These values give information about the handling of the electrolarynx. The detected reduction of variance of VIT, VD and VTT leads to the conclusion that all participants could improve the handling with the device within nine training sessions. Also the decrease of the mean value of the voice termination time indicates that the participants learned to relax their neck muscles to switch off the electrolarynx on time.

For recording of words, sentences and the paragraph phonetically balanced speech data was used with a focus on "m" and "n" as initial letters for words because these letters cause problems in producing a proper EMG-signal. The recorded speech material was evaluated in a listening test.

6.2 Listening Test

The listening test was designed to give information about pleasantness of EL-speech. Therefor each of the four participants of the training recorded 40 words, 30 sentences and 6 long sentences of a paragraph before and after six consecutive training sessions. An AB listening test with a comparison category rating scale was used for the evaluation were each pre- and post-recording of the same speech sample was compared. To get meaningful results 13 listeners attended the execution of the listening test. Due to the big amount of speech samples a single test would lead to fatigue of the listeners. So the test was divided into three parts to ensure that each part only last around 30 min.

The results show that 71 % of the post-recorded speech samples were evaluated slightly better, better or much better than the samples recorded before training. The differences between

words, sentences and the paragraph are marginal but for all an improvement of pleasantness is identifiable.

6.3 Outlook

In a former thesis Amon [4] developed a device to control the electrolarynx via neck muscle tension. Based on that this thesis revealed that there are measurable and audible learning effects in EMG-EL-speech and that it is possible to improve handling, intelligibility and pleasantness. So, one major drawback of occupation of a hand to turn on and off the electrolarynx could be eliminated. To improve the acceptance of the EMG-EL and eliminate the other major drawback of monotonic sound a more pleasant excitation signal with a varying fundamental frequency must be developed.

In this work the training protocol consists of nine training sessions which last around 50 to 60 min each. This long training was quite exhausting for the participants so an optimization of the training protocol could be useful for further tests. Depending on the participants a reduction of time of a single session to around 30 min would possibly yields in nearly the same results. Also a reduction of days could be possible in some cases.

Furthermore, an adaptive algorithm for threshold detection which regards the speaker would also help to improve EMG-EL and for daily usage a portable, independent and reliable working device must be designed which provides an easy handling and a comfortable wearing of the electrodes.

Acronyms

BDR block detection ratio 14

BEE back end error 13

CCR comparison category rating 15

CMOS comparison mean opinion score 15

DTD double threshold detection 7

EL electrolarynx 1

EMG electromyography 5

EMG-EL electromyographically controlled electrolarynx 1

FEE front end error 13

GUI graphical user interface 16

GT ground truth 13

sEMG surface electromyography 1

MSE middle speech error 13

MUAPs motor unit action potentials 5

NDS noise detected as speech 13

RMS root mean square 6

SNR signal to noise ratio 20

STD single threshold detection 6

VD voice duration vi

VIT voice initiation time vi

VTT voice termination time vi

A Appendix A

A.1 Speech Database

The phonetically balanced speech database for recording presented in table A.1, A.2 and table A.3 consists of 40 words, 30 sentences and the paragraph *Der Nordwind und die Sonne*. The words are chosen with focus on "m" and "n" as initial letters because these consonants cause more problems to produce a proper EMG-signal.

Table A.1: List of speech samples for the evaluation of words.

No.	Words
WD_1	Licht
WD_2	Besuch
WD_3	Nachrichten
WD_4	Durchsage
WD_5	Arzt
WD_6	Wohnung
WD_7	Glas
WD_8	Suppe
WD_9	Fischotter
WD_10	Mitternacht
WD_11	furchtbar
WD_12	müssen
WD_13	jagen
WD_14	munter

WD_15	schwärmen
WD_16	verletzen
WD_17	neun
WD_18	leuchten
WD_19	aussagen
WD_20	studieren
WD_21	Leuchtturm
WD_22	Zimmerleute
WD_23	rücksichtslos
WD_24	Menschenmenge
WD_25	warum
WD_26	Natur
WD_27	Eilzug
WD_28	Amsel
WD_29	innerhalb
WD_30	Reihe
WD_31	Telefon
WD_32	Jubel
WD_33	Idee
WD_34	Rad
WD_35	Papagei
WD_36	ebenso
WD_37	jemand
WD_38	rot
WD_39	echt
WD_40	ignorieren

Table A.2: List of speech samples for the evaluation of sentences.

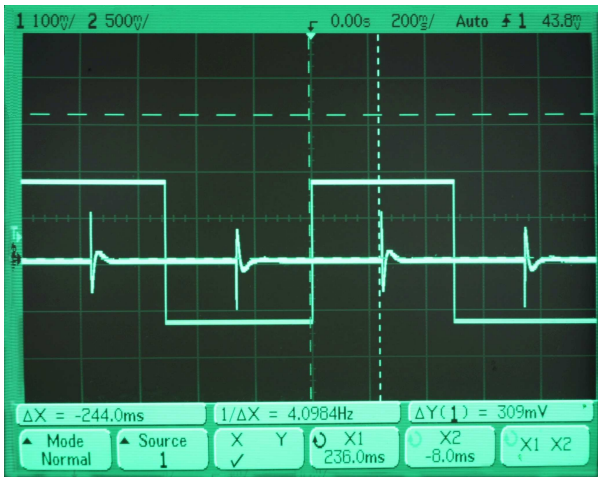
No.	Sentences
SE_1	Heute ist schönes Frühlingswetter.

SE_2	Am blauen Himmel ziehen die Wolken.
SE_3	Riecht ihr nicht die frische Luft?
SE_4	Messer und Gabel liegen neben dem Teller.
SE_5	Im Geschäft stehen viele Leute.
SE_6	Jetzt suche ich das Weißbrot.
SE_7	Ich müsste lesen und rechnen.
SE_8	Du solltest weniger rauchen.
SE_9	Die Kartoffeln gehören zum Mittagessen.
SE_10	Können wir nicht Tante Erna besuchen.
SE_11	Durch Wald und Feld führt unser Weg.
SE_12	Dahinter liegt der Rosengarten.
SE_13	Am Zaun steht eine Regentonne.
SE_14	Aus dem Radio klingt Musik.
SE_15	Der gelbe Küchenofen sorgt für Wärme.
SE_16	Zum Schnitzel gibt es Erbsen.
SE_17	Frische Gardinen hängen am Fenster.
SE_18	Wir wollen heute spazieren gehen.
SE_19	Überquere die Straße vorsichtig!
SE_20	Manche Obstbäume blühen prächtig.
SE_21	Dazu essen wir den Salat.
SE_22	Die Bremsen quietschen gräßlich.
SE_23	Er gewinnt sechs Spiele nacheinander.
SE_24	Der Pflug zieht tiefe Furchen.
SE_25	Ein Bauer arbeitet auf seinem Acker.
SE_26	Wir haben ein Abteil extra für uns.
SE_27	Schon bald sind wir zu Hause.
SE_28	Vater will sich eine Pfeife anzünden.
SE_29	Günther muss noch einkaufen gehen.
SE_30	Auf dem Brett leuchten bunte Tulpen.

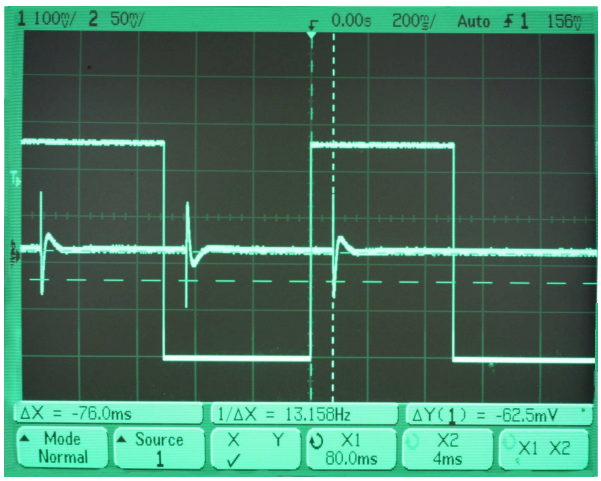
Table A.3: Paragraph: Der Nordwind und die Sonne. (Modified from J. G. Herder [1])

No.	Paragraph
NW_1	Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges kam.
NW_2	Sie wurden einig, dass derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzunehmen.
NW_3	Der Nordwind blies mit aller Macht, aber je mehr er blies, desto fester hüllte sich der Wanderer in seinen Mantel ein.
NW_4	Endlich gab der Nordwind den Kampf auf.
NW_5	Nun erwärmte die Sonne die Luft mit ihren freundlichen Strahlen, und schon nach wenigen Augenblicken zog der Wanderer seinen Mantel aus.
NW_6	Da musste der Nordwind zugeben, dass die Sonne von ihnen beiden der Stärkere war.

A.2 Measurement of system latency



(a) Latency of the Data Acquisition Toolbox.



(b) Latency of Playrec.

Figure A.1: Measurement of latency of the systems. Because of the shorter latency of 76 ms in this work the system running with Playrec was used.

B
Appendix B

B.1 Diagrams of results of the listening test

The figures on the following pages reveal more details of the listening test. Figures B.1, B.2, B.3 and B.4 show the evaluation of each of the four speakers F1, F2, F3 and M1 separated into the three parts of the listening test and into the three tasks of speaking (words, sentences and the paragraph).

Figures B.5, B.6 and B.7 provide more details about the evaluation of each of the 15 listeners (VP1 to VP15) who attended the listening test. Each evaluation of a listener is separated into words, sentences and the paragraph. A listener had to evaluate 52 (part 1)/56 (part 2)/52 (part 3) samples of words, 40 samples of sentences and 8 samples of the paragraph. It should be concerned about that the samples of one part of the listening test consist of speech samples of all four speakers (compare table 4.1). So a deviating evaluation of the three parts of the listening test, e.g. sentences of speaker F3 in figure B.3, is not surprising since one speech sample is already a tenth of all samples of one speaker.

Figures B.8 to B.14 give information about individual evaluations of each recorded speech sample of the four speakers. In the listening test each speech sample was judged by five listeners. It can be seen that in most cases the listeners evaluated similar.

For a better display the notation of the histogram abscissa and the boxplot ordinate respectively is changed to numbers. Table B.1 defines the notation in the figures.

Table B.1: Possible evaluation categories of the listening test for comparison category rating [2].

Category	Score
-3	much worse
-2	worse
-1	slightly worse
0	about the same
1	slightly better
2	better
3	much better

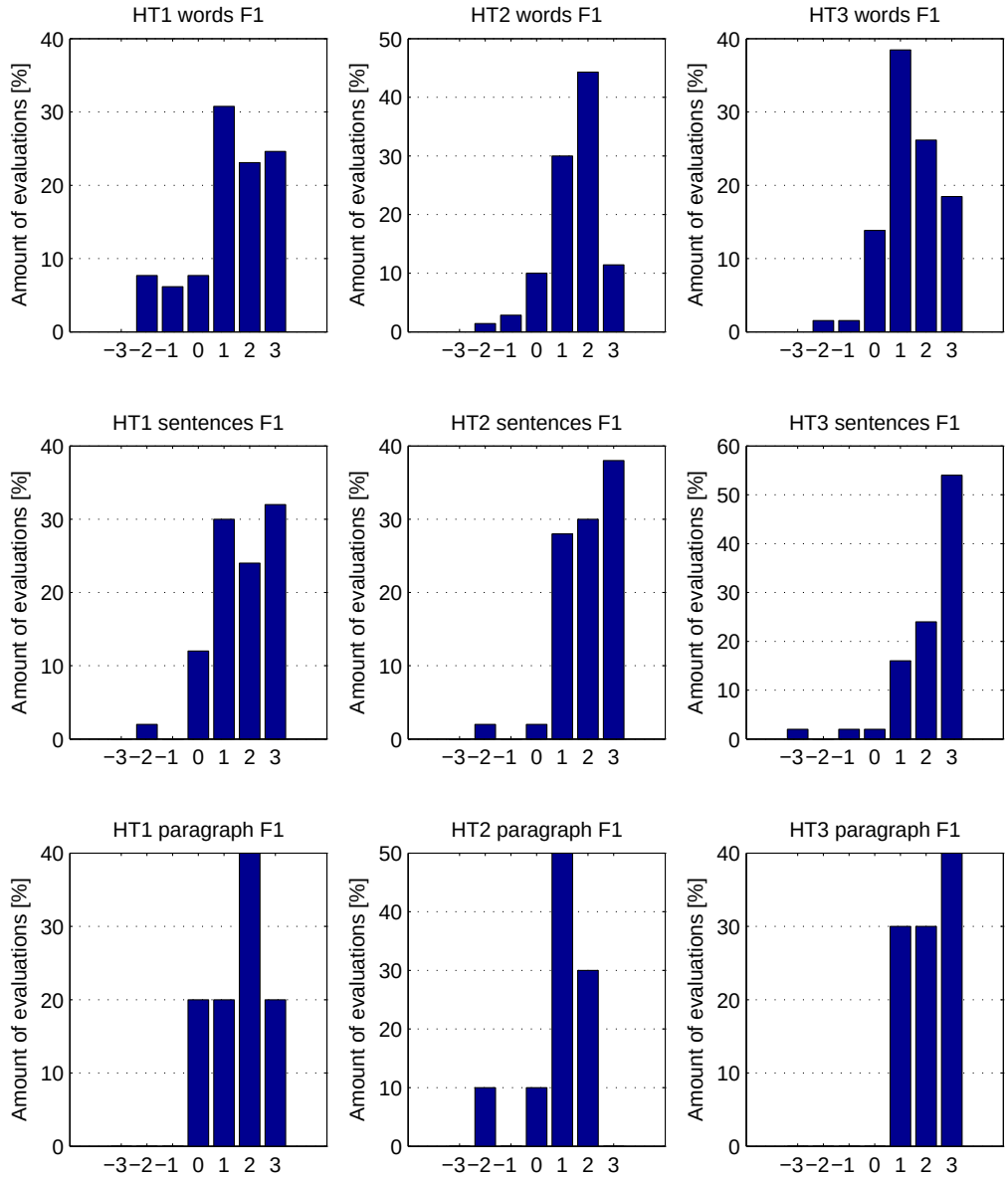


Figure B.1: Overview of the evaluation of female speaker F1. (200 evaluations for words, 150 evaluations for sentences and 30 evaluations for the paragraph separated into three parts of listening test)

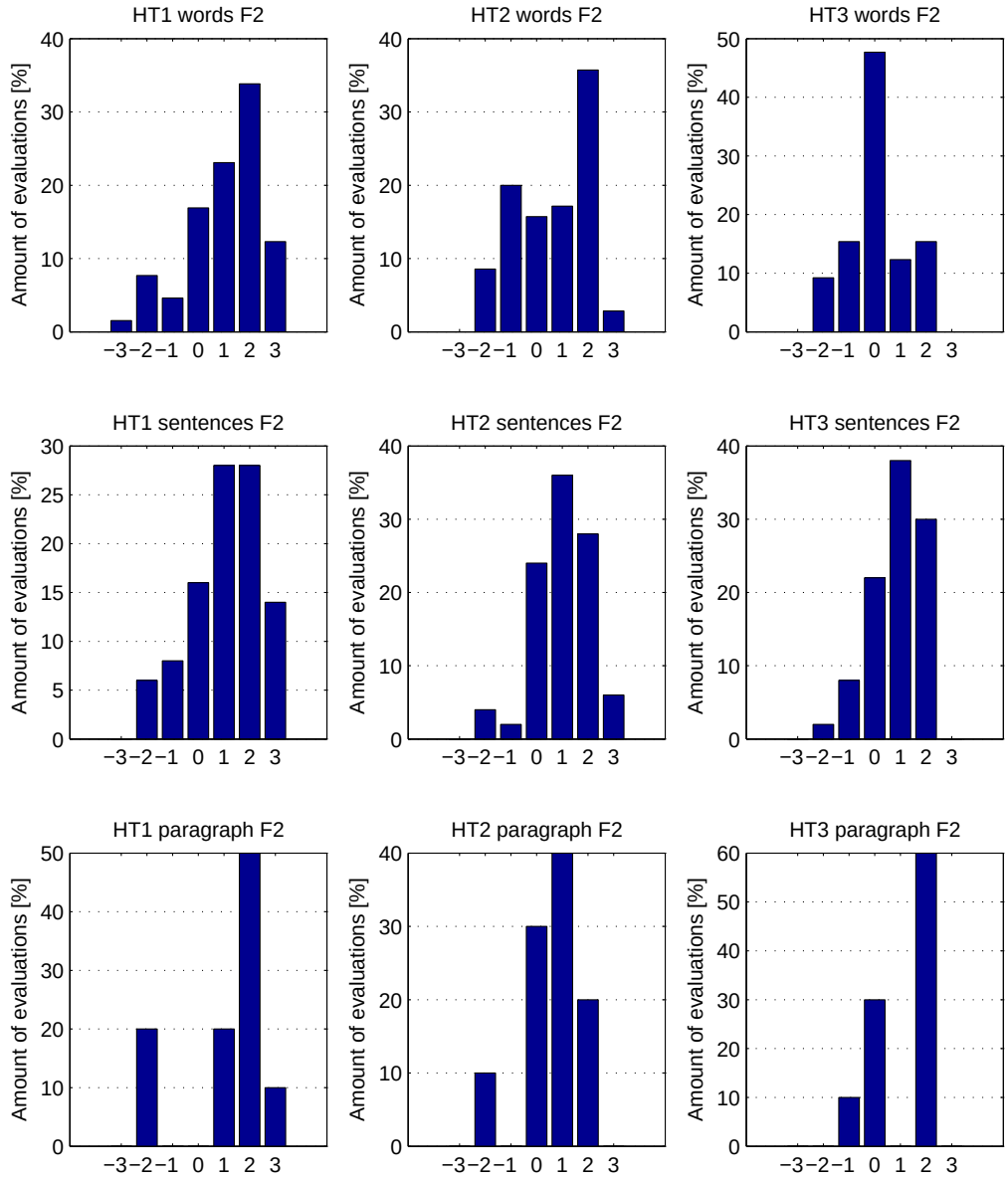


Figure B.2: Overview of the evaluation of female speaker F2. (200 evaluations for words, 150 evaluations for sentences and 30 evaluations for the paragraph separated into three parts of listening test)

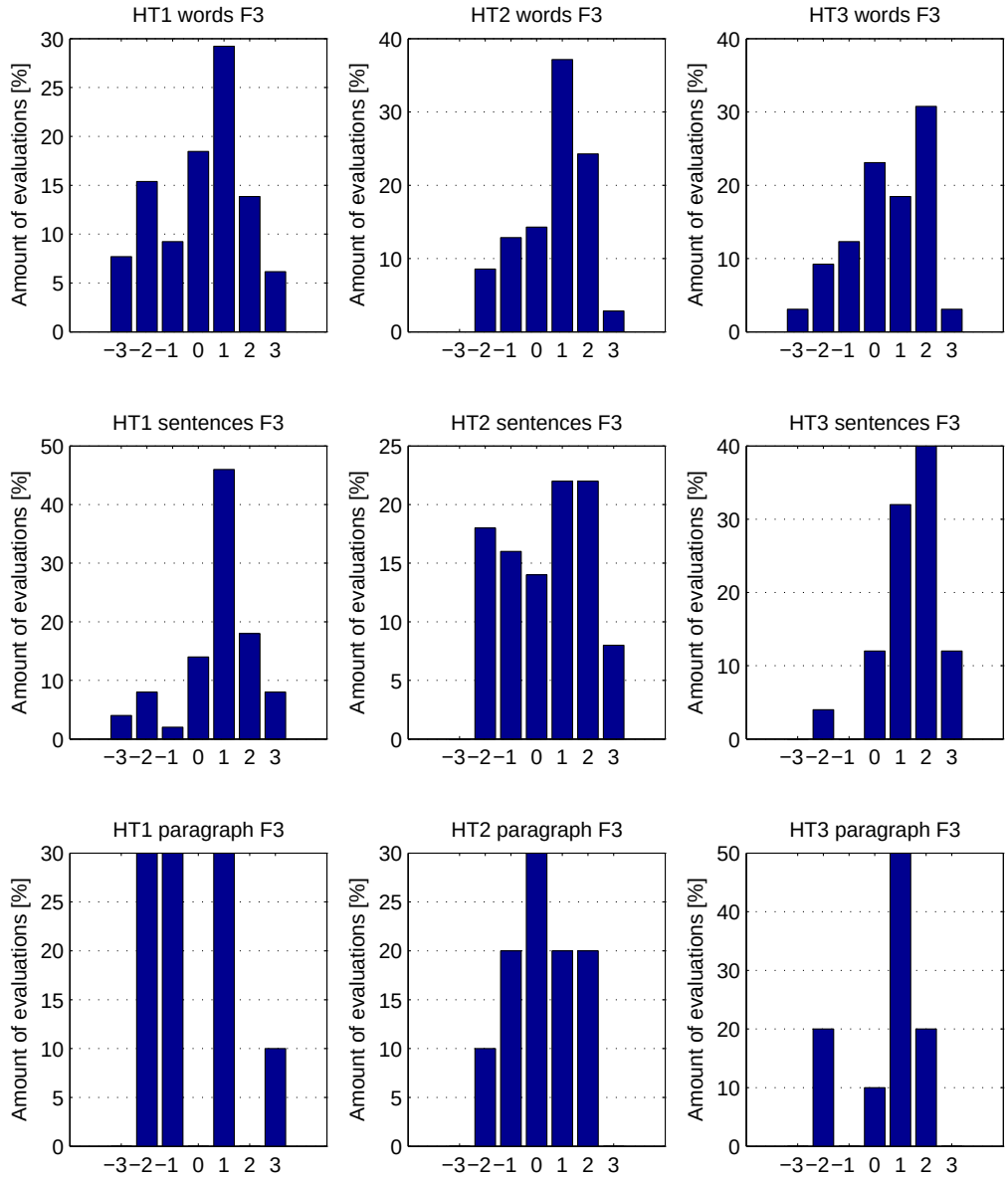


Figure B.3: Overview of the evaluation of female speaker F3. (200 evaluations for words, 150 evaluations for sentences and 30 evaluations for the paragraph separated into three parts of listening test)

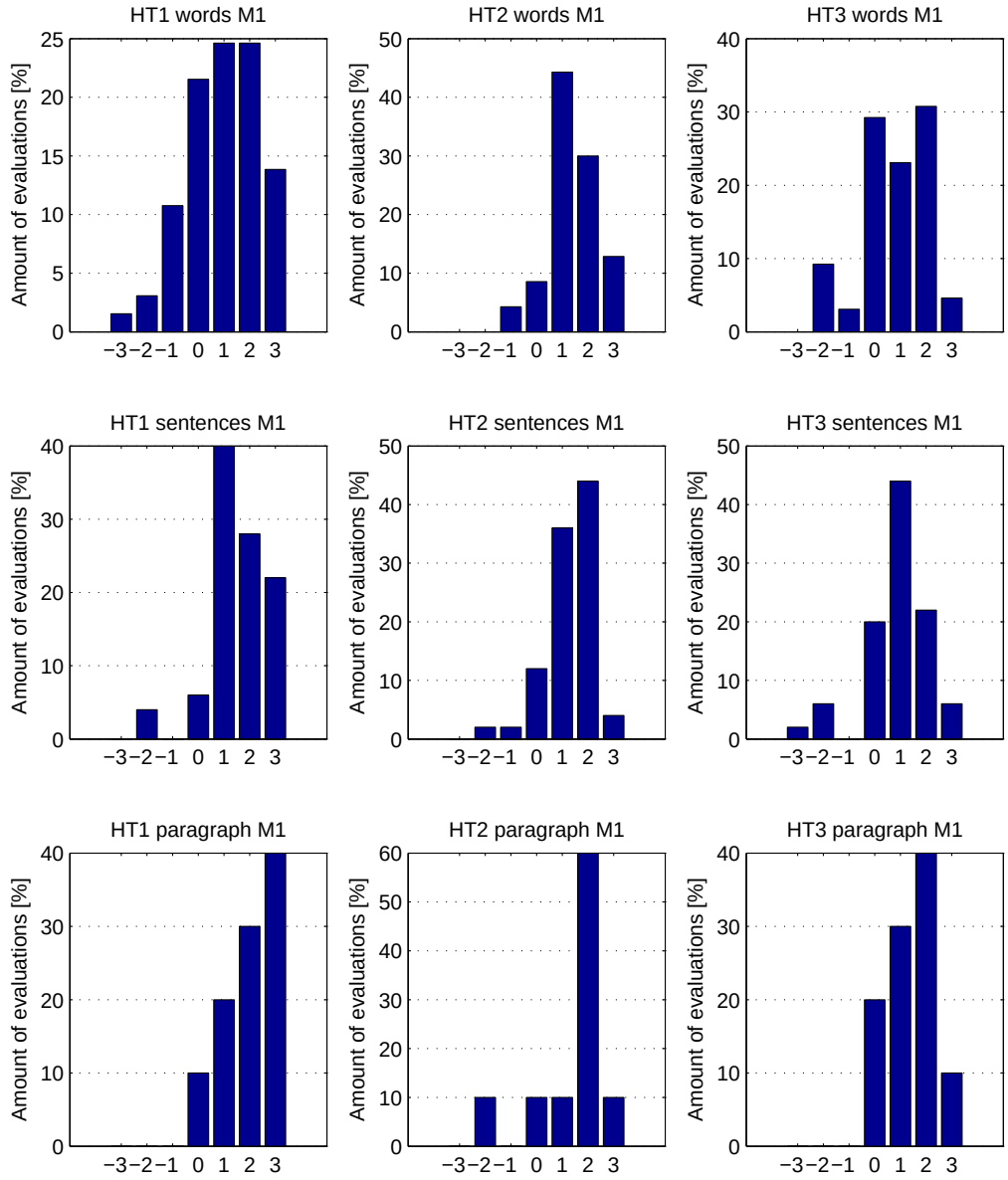


Figure B.4: Overview of the evaluation of male speaker M1. (200 evaluations for words, 150 evaluations for sentences and 30 evaluations for the paragraph separated into three parts of listening test)

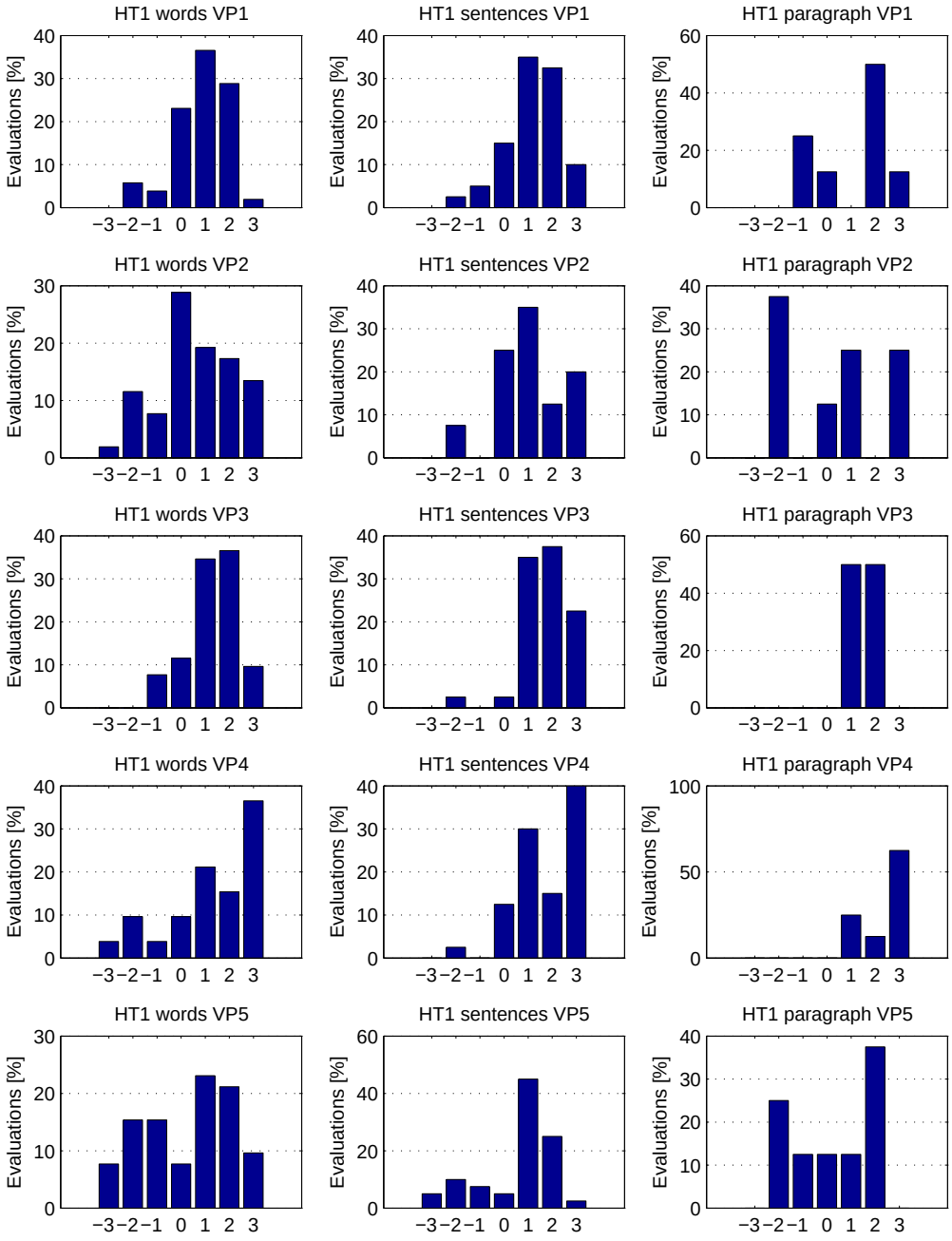


Figure B.5: Evaluation of listeners VP1 to VP5 of part 1 of the listening test. (each listener made 52 evaluations for words, 40 evaluations for sentences and 8 evaluations for the paragraph)

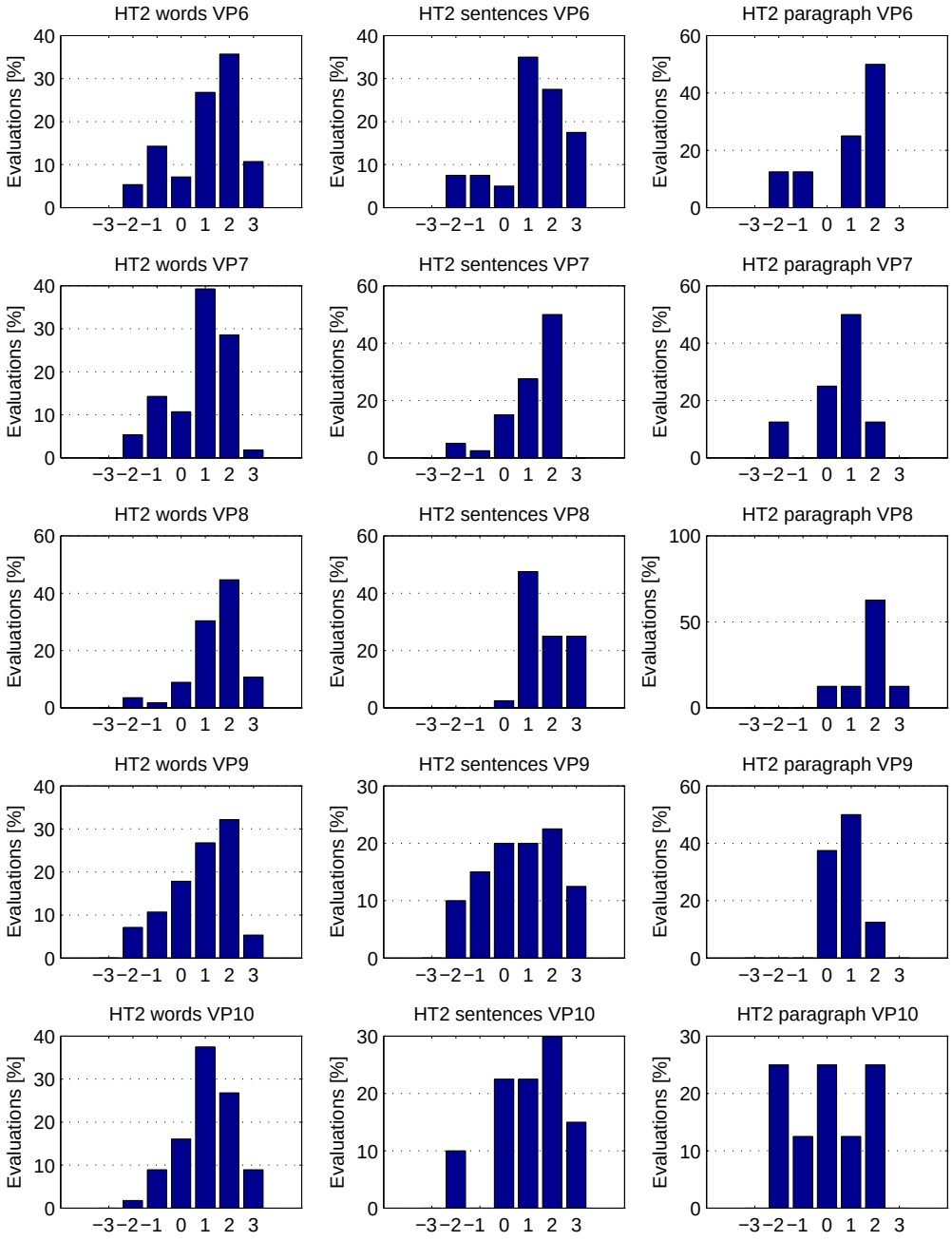


Figure B.6: Evaluation of listeners VP6 to VP10 of part 2 of the listening test. (each listener made 56 evaluations for words, 40 evaluations for sentences and 8 evaluations for the paragraph)

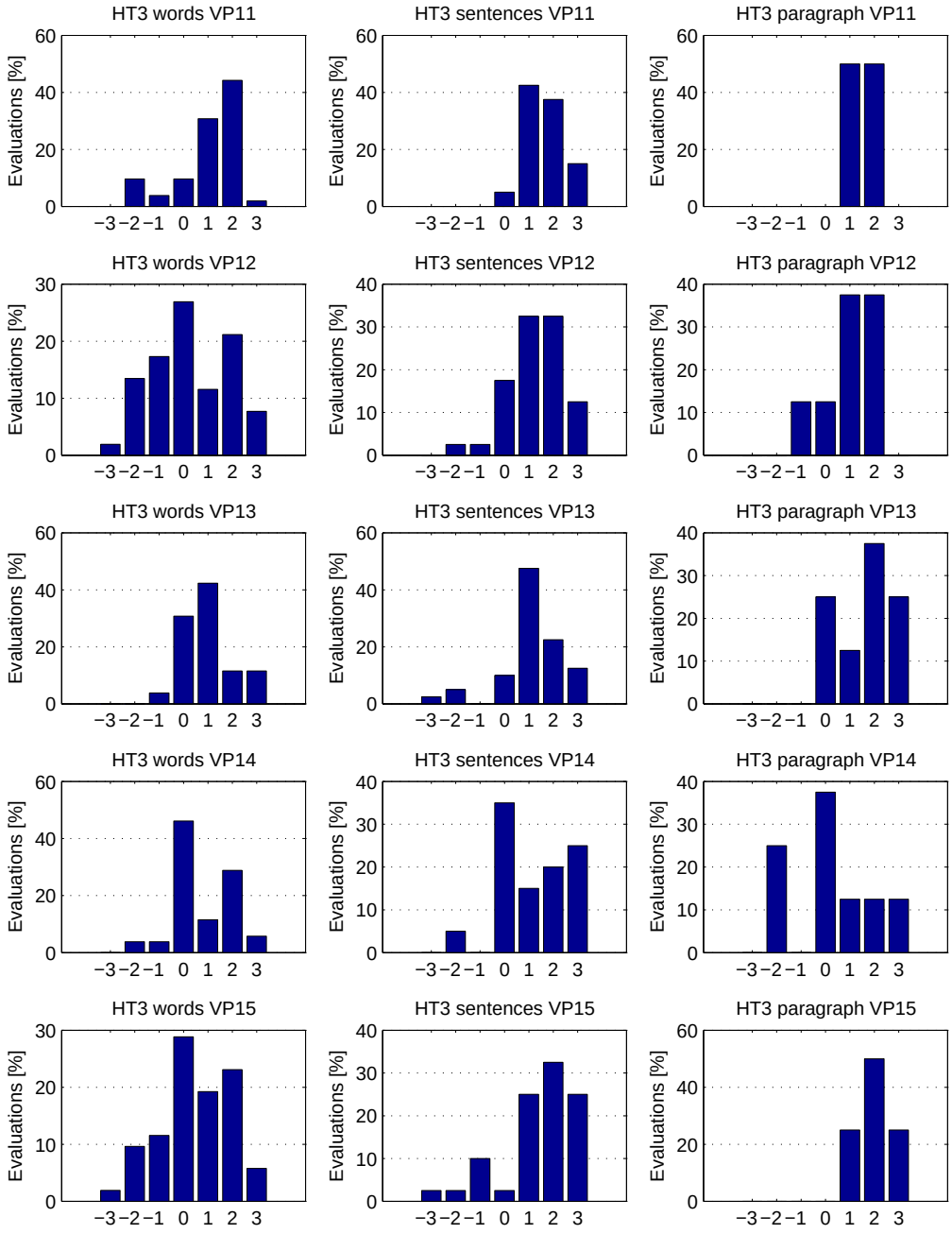


Figure B.7: Evaluation of listeners VP11 to VP15 of part 3 of the listening test. (each listener made 52 evaluations for words, 40 evaluations for sentences and 8 evaluations for the paragraph)

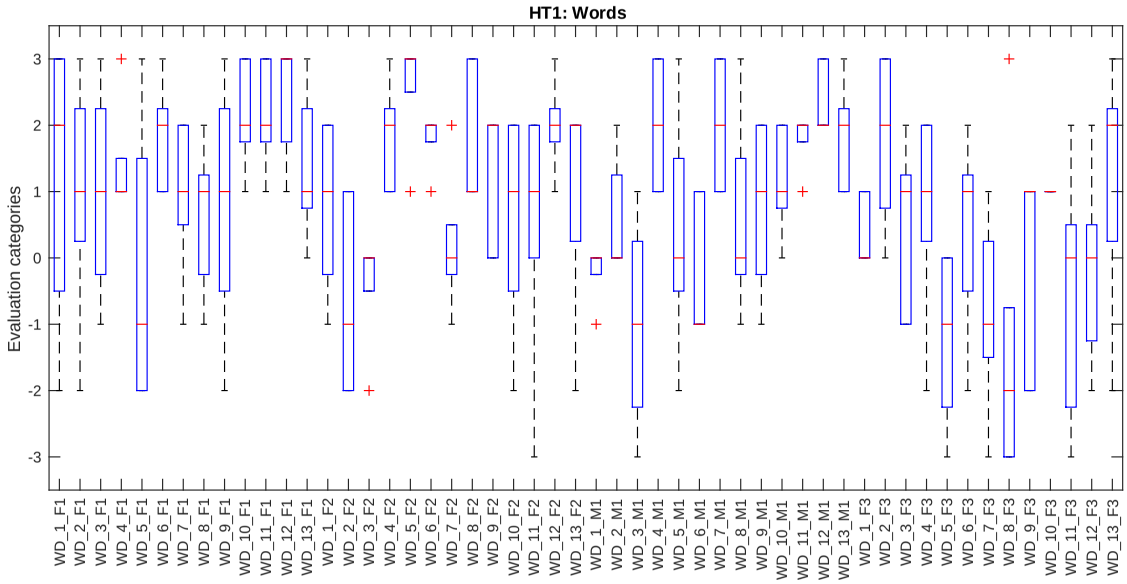


Figure B.8: Range of individual evaluations for each speech sample for words of the five listeners (VP1 - VP5) for listening test part 1 (HT1).

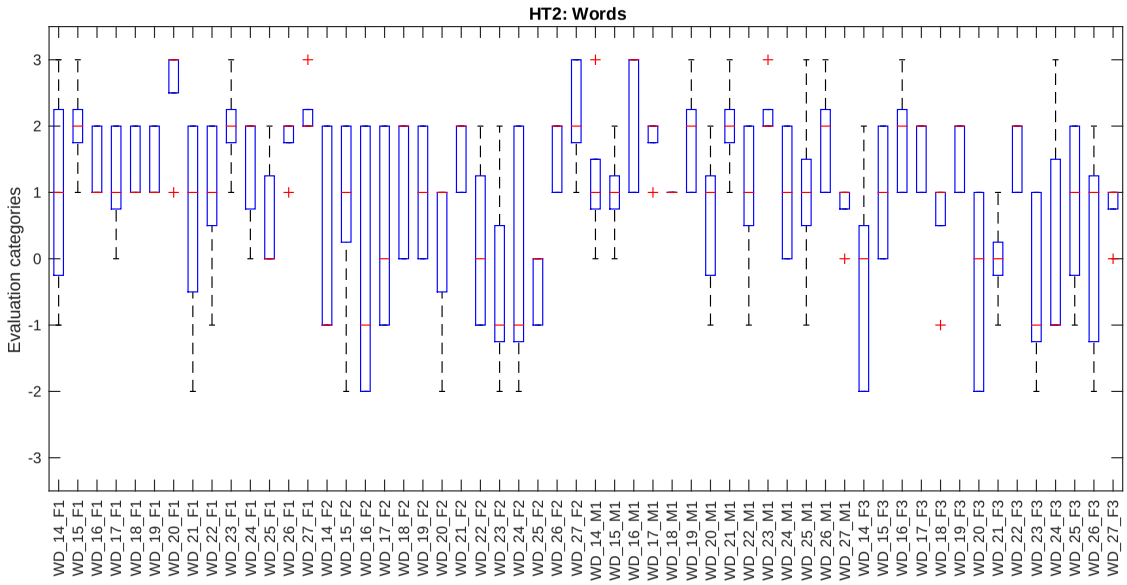


Figure B.9: Range of individual evaluations for each speech sample for words of the five listeners (VP6 - VP10) for listening test part 2 (HT2).

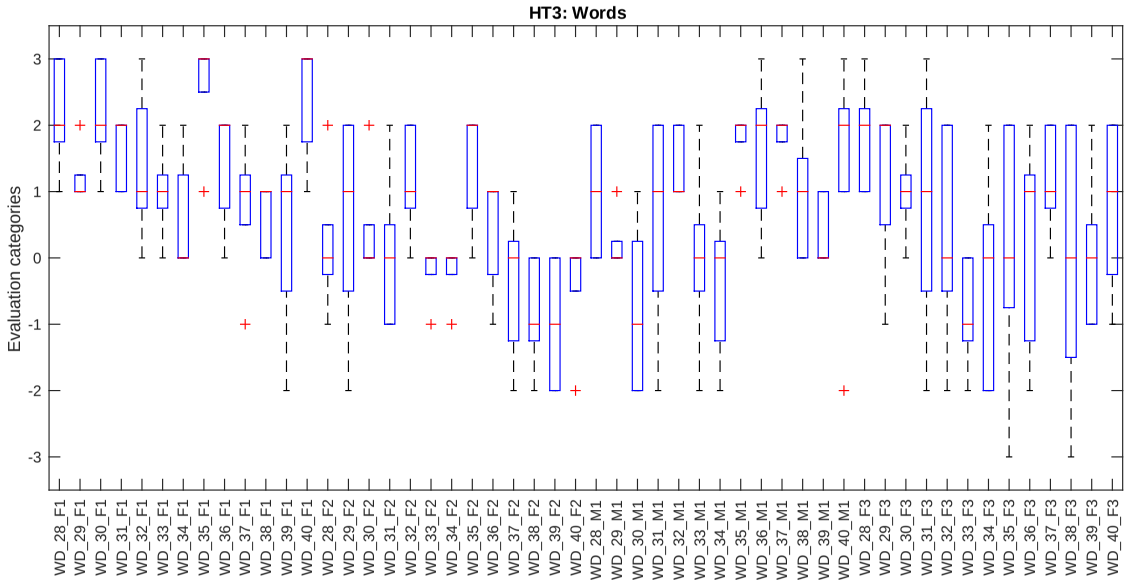


Figure B.10: Range of individual evaluations for each speech sample for words of the five listeners (VP11 - VP15) for listening test part 3 (HT3).

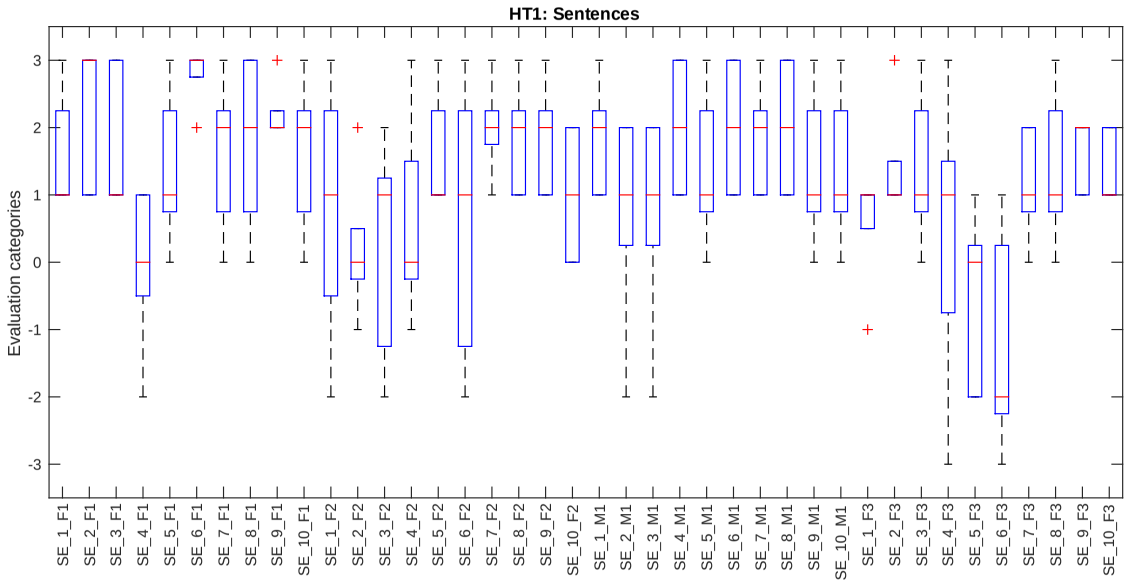


Figure B.11: Range of individual evaluations for each speech sample for sentences of the five listeners (VP1 - VP5) for listening test part 1 (HT1).

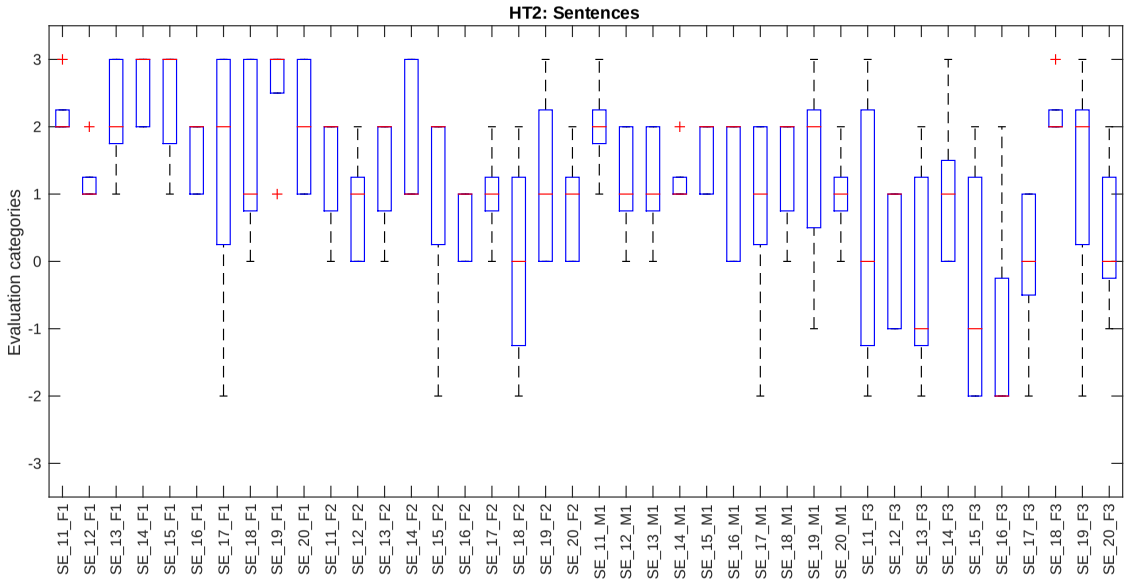


Figure B.12: Range of individual evaluations for each speech sample for sentences of the five listeners (VP6 - VP10) for listening test part 2 (HT2).

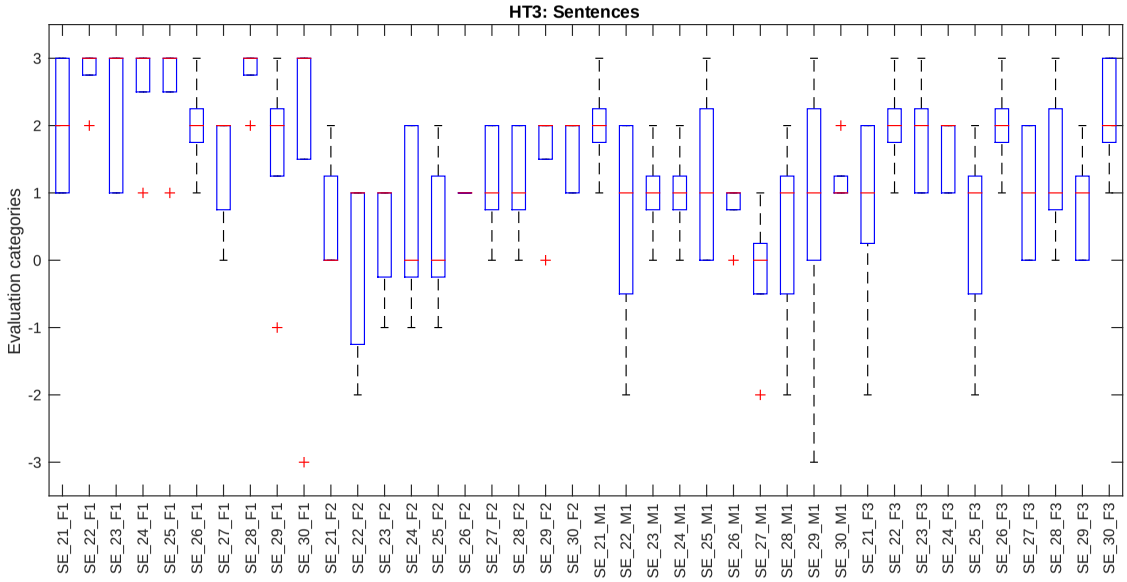


Figure B.13: Range of individual evaluations for each speech sample for sentences of the five listeners (VP10 - VP15) for listening test part 3 (HT3).

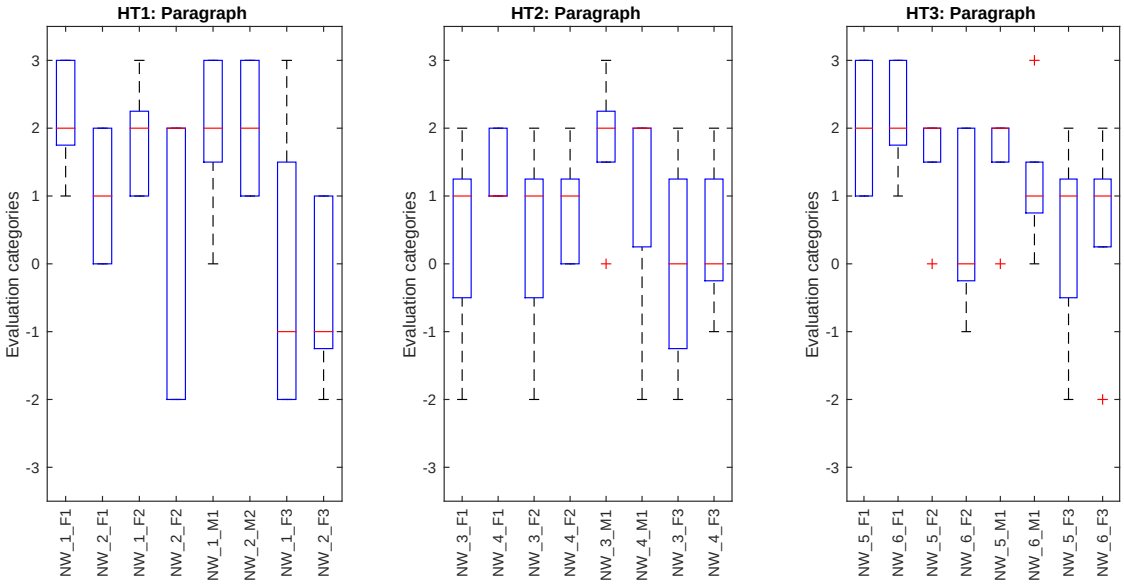


Figure B.14: Range of individual evaluations for each speech sample for the paragraph of the five listeners for all parts of the listening test (HT1: VP1 - VP5, HT2: VP6 - VP10 and HT3: VP11 - VP15).

Bibliography

[1] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[2] (2014) Comparison category rating (CCR) – ITU-T recommendation P.800. [Online]. Available: http://www.ntt.co.jp/qos/qoe/eng/technology/sound/03_3.html

[3] H. Liu and M. L. Ng, "Electrolarynx in voice rehabilitation," *Auris Nasus Larynx International Journal of ORL and HNS*, vol. 34, pp. 327 – 332, 2007.

[4] C. Amon, "Electrolarynx control using electromyographic signals," Master's thesis, Graz University of Technology, Austria, 2014.

[5] E. A. Goldstein, J. T. Heaton, C. E. Stepp, and R. E. Hillman, "Training effects on speech production using a hands-free electromyographically controlled electrolarynx," *Journal of Speech, Language, and Hearing Research*, vol. 50, pp. 335–351, 2007.

[6] E. A. Goldstein, J. T. Heaton, J. B. Kobler, G. B. Stanley, and R. E. Hillman, "Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 325–332, 2004.

[7] C. J. Van As-Brooks, F. J. Koopmans-van Beinum, L. C. W. Pols, and F. J. M. Hilgers, "Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech," *Journal of Voice*, vol. 20, pp. 355 – 368, 2006.

[8] C. E. Stepp, "Electromyographic control of prosthetic voice after total laryngectomy," Master's thesis, Massachusetts Institute of Technology, 2008.

[9] M. A. van Rossum, G. de Krom, S. G. Nooteboom, and H. Quené, ""pitch" accent in alaryngeal speech," *Jounal of Speech, Language, and Hearing Research*, vol. 45, pp. 1106 – 1118, 2002.

[10] J. E. Vinton, "Phonetic emphasis in Tamil," *UTA Working Papers in Linguistics*, vol. 1, pp. 95 – 107, 1994.

[11] (2014, April) Die Stimme: Anatomie & Physiologie. [Online]. Available: http://www.vocalis-projekt.de/html/stimme/anatomie.htm

[12] C. Jochum and P. Reiner, "Comparison of excitation signals for an electronic larynx," Master's thesis, Graz University of Technology, Austria, 2008.

[13] (2014, April) Organs and structures of the respiratory system. [Online]. Available: http://cnx.org/content/m46548/latest/?collection=col11496/latest

[14] H. L. Barney, F. E. Haworth, and H. K. Dunn, "An experimental transistorized artificial larynx," *The Bell System Technical Jounal*, vol. 38, pp. 1337 – 1356, 1959.

[15] F. Arifin, T. A. Sardjono, and M. H. Purnomo, "The relationship between electromyography signal of neck muscle and human voice signal for controlling loudness of electrolarynx," *Biomedical Engineering: Applications, Basis and Communication*, vol. 26, pp. 1 450 054–1 – 1 450 054–7, 2014.

[16] (2014, April) Head & neck cancer guide. [Online]. Available: http://www.headandneckcancerguide.org/adults/cancer-diagnosis-treatments/surgery-and-rehabilitation/cancer-removal-surgeries/laryngectomy/

[17] I. Brook. (2014, April) My voice. [Online]. Available: http://dribrook.blogspot.co.at/p/basic-skills-for-new-laryngectomees.html

[18] C. Drewes, "Electromyography: Recording electrical signals from human muscle," *Association for Biology Laboratory Education*, vol. 21, pp. 248 – 270, 2000.

[19] D. F. Stegeman and H. J. Hermens, "Standards for surface electromyography: the european project "surface EMG for non-invasive assessment of muscles (SENIAM)"," Institute of Neurology, University Medical Centre Nijmegen and Institute for Pathophysiology, Friedrich-Schiller-University Jena, Germany, Tech. Rep., 1996 – 1999.

[20] J. Blascovich, W. B. Mendes, E. Vanman, and S. Dickerson, *Social Psychophysiology for Social and Personality Psychology*. SAGE Publications Ltd, 2011.

[21] C. Draxler, *SpeechRecorder Quick Start and User Manual*, Institut für Phonetik und Sprachverarbeitung, Universität München.

[22] (2014, August) MathWorks MATLAB. [Online]. Available: http://www.mathworks.de/products/matlab/

[23] (2014, August) SpeechRecorder. [Online]. Available: http://www.bas.uni-muenchen.de/Bas/software/speechrecorder/

[24] (2014, August) REAPER Digital Audio Workstation. [Online]. Available: http://www.reaper.fm/

[25] F. Beritelli, S. Casale, and G. Ruggeri, "Performance evaluation and comparison of itu-t/etsi voice activity detectors," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICAASP)*, vol. 3, 2001, pp. 1425–1428.

[26] A. K. Fuchs, C. Amon, and M. Hagmüller, "Speech/non-speech detection for electro-larynx speech using EMG," Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria, Tech. Rep., 2014.

[27] C. Brell, J. Brell, and S. Kirsch, *Statistik von Null auf Hundert*. Springer-Verlag, 2014.