Graz University of Technology

Using sEMG for Disordered Speech Enhancement

Author: Klaus Huber, 0831542
Julia Ziegerhofer, 0873106

Date: Graz, December 15, 2014

Supervisors: DI Dr.techn. Martin Hagmüller
DI Anna Katharina Fuchs

Abstract

The larynx, which contains the vocal folds, is essential for the production of speech. If the larynx is surgically removed, typically due to cancer, the patient loses the ability to speak. One possibility to be able to communicate again is to use a hand-held, battery-driven device – the electro-larynx (EL). This device is held against the neck and its vibrations are transmitted through the neck. The patient can create speech by modulating the energy in the vocal tract.
A major drawback of the device is that the constant excitation signal sounds like robotic speech. To make the speech sound more natural a varying fundamental frequency needs to be introduced to the EL. One possibility is to use surface electromyography (sEMG) to control the pitch of an EL.
The main focus of attention in this work lays in the extraction and evaluation of reasonable features out of the sEMG signals. The used data consisted of various two-channel-recordings where the first channel was the sEMG signal and the second channel was either the healthy speech (HE) or the EL speech. The recordings were phonetically balanced German sentences. For the processing of the sEMG signals *Matlab* was used and features (e.g. the envelope of the sEMG signal) were calculated as described in this work. Afterwards those features which were best in appearence were chosen to be part of the listening test (LT). Furthermore the results of the LT were evaluated and discussed.
Summerized it can be said that EL signals modified by various features sound better than unmodified EL speech, even if the feature doesn't follow the determined fundamental frequency. This work only considers the offline approach which means that computer analysis and synthesis was performed with data from a data pool.
The advantage of an sEMG based EL is that ON/OFF switch as well as fundamental frequency variation can be controlled in an automatic way.
In a previous project the hardware to capture sEMG signals was developed.

Kurzfassung

Der Kehlkopf (Larynx), in welchem die Stimmlippen liegen, ist äußerst wichtig für die menschliche Sprachproduktion. Muss der Kehlkopf chirurgisch entfernt werden - häufig aufgrund einer Krebserkrankung - verliert der Patient die Fähigkeit zu sprechen. Eine Möglichkeit, die Sprachfähigkeit wiederherzustellen, ist die Verwendung eines batteriebetriebenen Gerätes (Elektro-Larynx, E-Larynx, EL) welches ein Signal generiert und somit die Hauptaufgabe des Kehlkopfes übernehmen kann.
Dieses Gerät wird an den Hals gehalten, wodurch die Vibrationen in den Hals übertragen werden. Der Patient kann nun durch Modulationen der Energie im Vokaltrakt wieder Sprache erzeugen. Ein großer Nachteil dieser Methode ist das konstante Anregungssignal des Gerätes, wodurch die Sprache künstlich und roboterähnlich klingt. Eine natürlichere Sprache könnte mit einer variierenden Grundfrequenz erzielt werden. Eine Möglichkeit, die Grundfrequenz eines EL-Gerätes zu steuern, stellen an der Hautoberfläche detektierte Muskelsignale (elektromyographische Signale, sEMG Signale) der Muskeln des Vokaltraktes dar.
Das Hauptaugenmerk dieser Arbeit liegt in der Extrahierung und Evaluierung verschiedener Merkamle (*Features*) aus den sEMG Signalen. Die verwendeten Daten bestanden aus Zweikanalaufnahmen wobei der erste Kanal das sEMG Signal und der zweite Kanal das gesunde (HE-) bzw. EL-Sprachsignal beinhaltete. Für das Sprachmaterial wurden phonetisch ausbalancierte deutsche Sätze verwendet. Die sEMG Signale wurden in *Matlab* verarbeitet und Features, wie z.B. die Einhüllende des sEMG Signals, wurden berechnet. Anschließend wurde ein Hörtest durchgeführt und dessen Ergebnisse ausgewertet und diskutiert.
Zusammenfassend kann gesagt werden, dass die durch verschiedene Features modifizierten EL Signale besser klingen als das unveränderte EL Signal, sogar wenn das Feature nicht dem ermittelten Grundfrequenzverlauf folgt. In dieser Arbeit wurde offline mit Datenbanken und Computeranalyse/-synthese gearbeitet.
Der Vorteil eines auf sEMG Signalen basierenden EL-Gerätes besteht darin, dass Ein- und Ausschalten genauso wie die Kontrolle der Grundfrequenz automatisch passieren können, der Patient muss das Gerät nicht per Hand bedienen.
In einem früheren Projekt wurde die Hardware für die Erfassung von sEMG-Signalen entwickelt.

Acknowledgement

We would like to thank our supervisors DI Dr.techn. Martin Hagmüller and DI Anna Katharina Fuchs for their guidance and help through the research and writing of this work. Furthermore, we want to thank all participants who took part in the listening test.

Contents

List of Figures

List of Tables

1 Introduction

In this chapter basic information about the theory of human speech and sEMG signals will be given. Furthermore, details about preparation and analysis of the existing database are given.

1.1 Motivation

Spoken language is probably the most important way of human communication. Not only objective content, but also information about the speaker - like information about one's mood - are transferred during the speaking process. Being excluded from that way of communcation implies difficulties in everyday life.
If the larynx has to be removed surgically - for instance during the treatment of cancer - as a consequence the person will have to breath through a *stoma* (a hole) in the neck and will lose the ability to produce normal speech [2]. However, if the vocal tract is intact, a prosthetic device - like the EL - can enable alaryngeal speech. The EL is a handheld battery-operated device usally with the size of a small electric razor. It is placed under the mandible and produces vibrations which replace the task of the vocal folds. Two things which EL users often describe as unconvenient and disturbing are firstly the necessary use of one hand in order to hold the device against the neck and secondly the monotonic and therefore unnatural sound of the produced speech [2].

1.2 Related Work

There are many papers and theses which deal with the topics of sEMG signals and sEMG controlled devices.
2008, C. Stepp [3] and 2004, Goldstein et al. [2] clarify the theory of sEMG signals and describe the implementation of a sEMG controlled EL device. 2009, A. Phinyomark et al. [4] presents several feature extraction methods.
This project is based on the diploma thesis *Using sEMG for Disordered Speech Enhancement* [5] written by Clemens Amon. In his master's thesis he developed a signal acquisation hardware for catching sEMG signals. Additionally he implemented and evaluated signal processing and activity detection methods for cotrolling ON/OFF-activity of the EL.
In this project different methods to control the fundamental frequency of speech, i.e. the EL's fundamental frequency through sEMG signals were implemented and tested. The whole signal processing was executed offline by using already existing data bases, computer analysis and computer synthesis.

1.3 Theory of the Human Speech Production

As described in [1] and depicted in figure 1.1 there are four main components in the human speaking process:

- lungs,
- trachea,

- larynx with vocal cords,
- vocal tract.

The lungs generate the necessary energy which is transported by the trachea. The larynx with the vocal cords is responsible for the excitation signal generation and the vocal tract, consisting of oral and nasal cavities and the pharynx, performs acoustic filtering (cf. source-filter model in signal processing theory).

> "The production of speech sounds involves the manipulation of an airstream. The acoustic representation of speech is a sound pressure wave originating from the physiological speech production system. [...] By contraction, the lungs produce an airflow which is modulated by the larynx , processed by the vocal tract , and radiated via the lips and the nostrils. The larynx provides several biological and sound production functions. In the context of speech production, its purpose is to control the stream of air that enters the vocal tract via the vocal cords." [1]

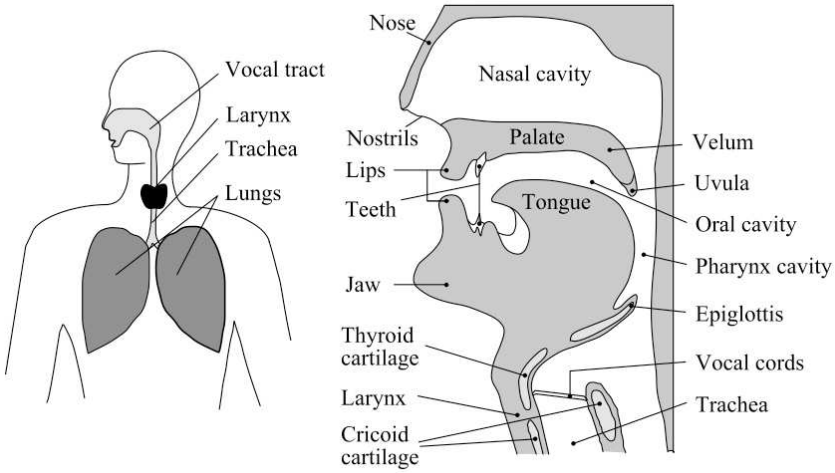

Figure 1.1: Organs of speech production [1]

1.4 Theory of sEMG Signals

In this project recorded surface electromyographic (sEMG) signals were used.

> "EMG is a technique, which investigates the muscular contractions [...]." [6]

> "As a nerve impulse from an alpha motor neuron reaches the motor end plates of muscle fibers comprising a motor unit, the fibers innervated by that neuron discharge nearly synchronously. The electric potential field generated by the depolarization of the outer muscle- fiber membranes essentially reflects the alpha motor neuron activity; the electromyogram (EMG) is a representation of this "myoelectricity" as summed over a number of motor units and measured at some distance. Tissues separating the EMG signal sources (depolarized zones of the muscle fibers) act like spatial low-pass filters on the potential distribution, and constitute a volume conductor. Therefore, the EMG may be measured intramuscularly or at the surface of

the skin, yielding different information based on the distance of the observation site from the muscle fibers. For surface detection particularly, the effect of the separating tissues becomes significant."[3]

As described in [4] noisy conditions, like electrode noise, electrode and cable motion artefacts, alternating current power line interference and broadband noise from electronic instruments, will lead to poor sEMG signal recognition results. Some types of noise can be removed by using band-pass filters, band-stop filters or the use of well electrode and instrument, but the removal of interferences of random noise that fall in sEMG dominant frequency regions is difficult.

1.5 Analysis and Preperation of the Available Speech Data

Speech data used in this project consisted of recorded phonetically balanced German sentences which were articulated by one male and one female native speaker. The sentences were two-channel-recordings, sEMG signals were stored on channel 1, the EL signals and the healthy (HE) speech signals, respectively, were stored on channel 2 (see figure 1.2 and figure 1.3). As there weren't persons without larynx available the sEMG signals were recorded from people with a healthy larynx. So there can't be made a general conclusion regarding the sEMG signals of persons without larynx.

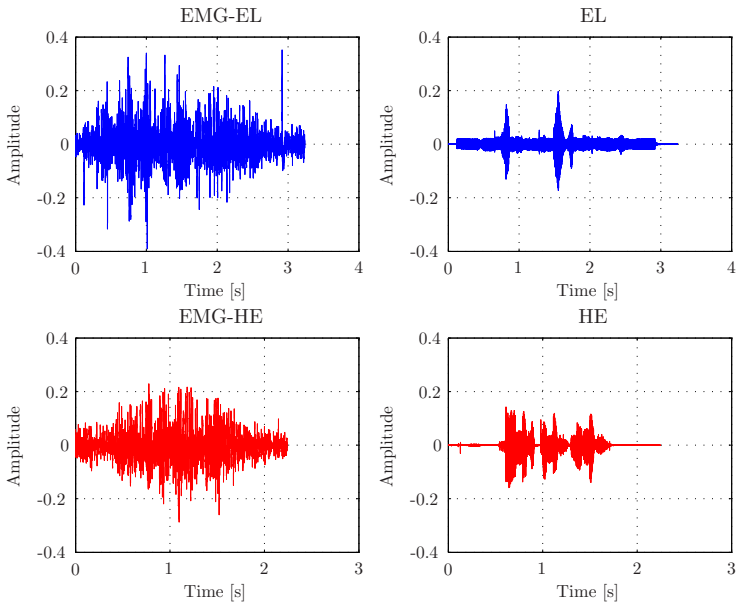

Figure 1.2: An example for the used speech data (female speaker)
blue: EL signal; red: HE signal
left side: sEMG signal; right side: Speech signal

In order to have nearly the same conditions for all sentences the whole speech data was edited with *sox* [7], a sound processing program, which was used for cutting off silence at the beginning and at the end of every sentence.

The provided data base consisted of sentences spoken by two speakers with a healthy larynx. Every sentence was recorded twice, once with EL speech and once with HE speech. So it was evaluated whether there are differences between sEMG signals produced by healthy speech

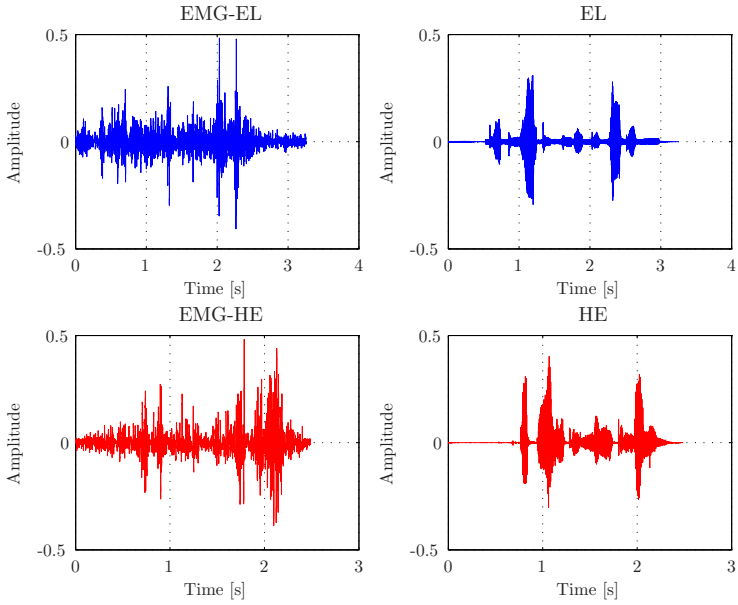

Figure 1.3: An example for the used speech data (male speaker)
blue: EL signal; red: HE signal
left side: sEMG signal; right side: Speech signal

(EMG-HE) and sEMG signals produced by talking with the EL (EMG-EL). Articulation variations produced time discrepancies in the signals for the same sentences. So both sEMG signals couldn't be compared directly. By the use of the *Dynamic Time Warping (DTW)* algorithm [8] the EMG-HE signal was fitted to the equivalent EMG-EL signal and comparison between the signals got possible.
DTW is a well known algorithm for time alignment between two temporal sequences which vary in time or speed. After performing DTW the sEMG signals were analysed. There weren't any noticeable differences in the sEMG signals related to the amplitudes (see figures 1.4 and 1.5). So the assumption was made that the vibrations of the vocal cords don't influence the sEMG signals and that motion of muscles is similar, whether EL speech or HE speech is used. However, generally it can't be said that sEMG signals from people without larynx are similar to sEMG signals from people with larynx who talk with an EL. For such an assumption it would be necessary to analyse sEMG signals taken from patients without a larynx. Analysis and plotting was performed in *Matlab* [9].

Another way for evaluating and comparing data was achieved by computing the SNR for male and female sEMG files, see figure 1.6. The SNR was obtained by equation 1.1 where the noise signal was extracted from the beginning and the end of the sEMG signal where no relevant information except the noise floor was existent.
It can be observed that female data has a lower SNR (about 10 dB) than male data (see figure 1.6).

$$SNR = 10 \cdot \lg\left(\frac{SIGNAL^2}{NOISE^2}\right) \tag{1.1}$$

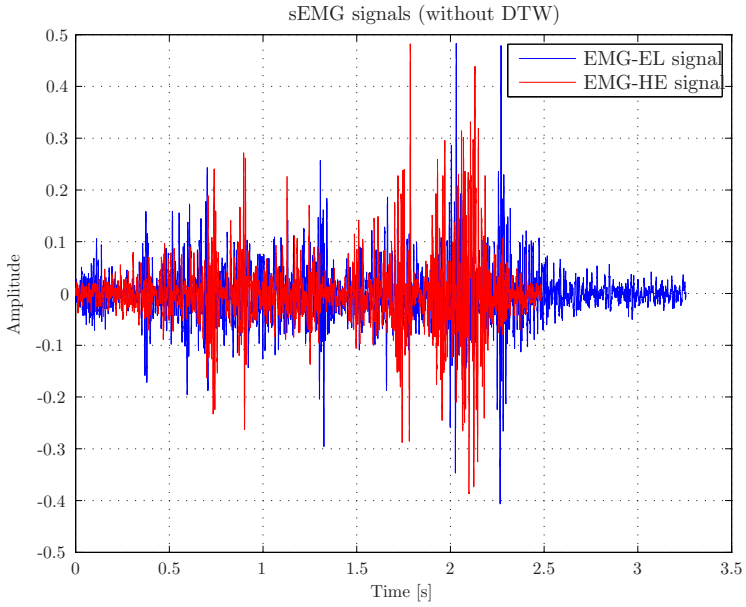

Figure 1.4: Comparison of the two sEMG signals without DTW (male speaker)

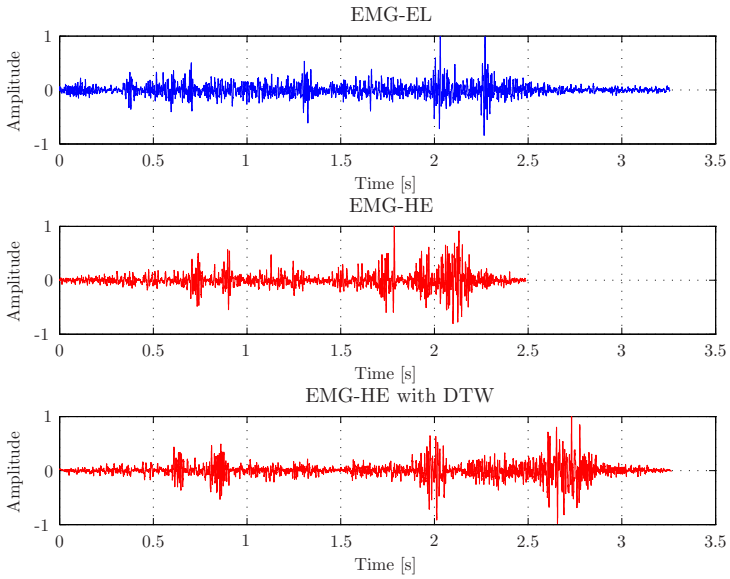

Figure 1.5: Comparison of the two sEMG signals with DTW (male speaker)

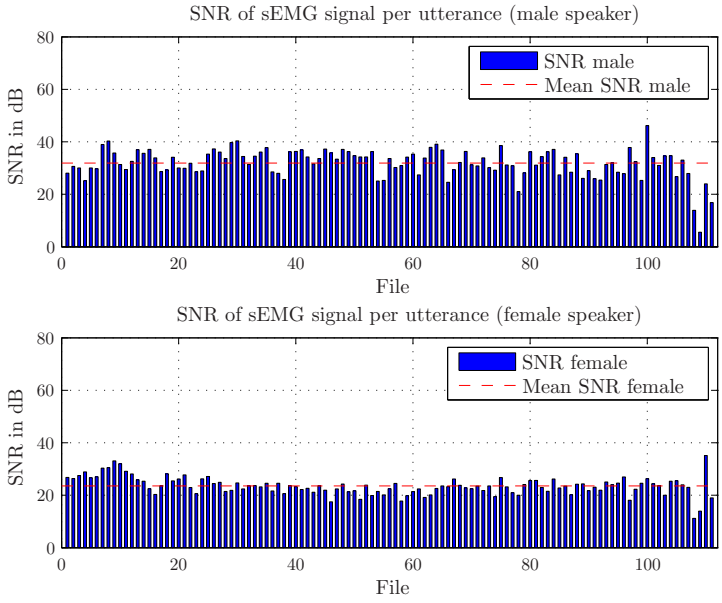

Figure 1.6: SNR of sEMG signals per utterance

2 Implementation

In order to analyse the given speech data and to calculate the features *Matlab* [9] was used. *Praat* [10] was used for the manipulation of the speech data. *Matlab* is a numerical computing environment, *Praat* is a free speech analysing software package.

2.1 The Proposed Features' Calculation

The features' calculation was performed partly according to the theory described in [2] and [4]. Primarily nine different features dependent on the sEMG signals were calculated. Additionally a random feature was created and the reference pitch was extracted from the appropriate HE signal.
So finally there were eleven features which are:

Table 2.1: Calculated features

Feature	Name
1	**Scaled Envelope**
2	Inverted Scaled Envelope
3	Zero Crossing (ZC)
4	Inverted Slope Sign Change (Inverted SSC)
5	**Modified Median Frequency (MMDF)**
6	Modified Mean Frequency (MMNF)
7	Combination of MMDF and MMNF
8	Maximum Energy
9	**Random Pitch**
10	Slow Envelope
Ref	**Reference Pitch**

The features used in the listening test (LT) were the Scaled Envelope Feature (feature 1, referred as Method B) and the Modified Median Frequency Feature (feature 5, referred as Method C). Equations 2.8, 2.10 and 2.12 were taken from [4]. Most of the calculated features were spread to get a larger frequency range. This was achieved by a simple multiplication with an aligned spreading factor α. In order to shift the calculated raw features to the estimated speaker's average fundamental frequency all of them were adjusted in the last step by β which is either the male or the female scaling value (see equation 2.1). The scaling values were computed by averaging the reference pitches. Note that in all equations bold symbols denote a vector.
Figures 2.2 to 2.11 show the calculated features before post processing (red) compared to the extracted fundamental frequency (grey) from the HE signal. In those segments, where the extracted fundamental frequency was zero, *Praat* didn't detect voiced speech and therefore no fundamental frequency. The calculated features were set to zero for these segments (see section 2.1.1).

$$\boldsymbol{Feature} = \left[\boldsymbol{X} - \overline{X}\right] \cdot \alpha + \beta \tag{2.1}$$

with

$$\overline{X} = \frac{1}{N}\sum_{n=1}^{N} X_n \tag{2.2}$$

- $\boldsymbol{Feature}$ ... Feature vector
- $\boldsymbol{X}$ ... raw feature vector $\boldsymbol{X} = [X_1, X_2, \dots X_N]$ (not spread and shifted)
- $\overline{X}$ ... mean value of $\boldsymbol{X}$ (see equation 2.2)
- α ... aligned spreading factor to get a larger frequency range
- β ... offset value to shift the feature to the appropriate average fundamental frequency
  193 Hz for female voice
  112 Hz for male voice
- N ... length of feature vector

Feature 1: Scaled Envelope

Feature 1 is the calculated envelope of the signal (see figure 2.1). First the signal was transformed with *Matlab's* hilbert transformation (*hilbert()*). Afterwards the absolute value of the result was smoothed. In addition the envelope was spread with a spreading factor α of 100 and finally shifted (see figure 2.2 and equation 2.3 to 2.4).

$$\boldsymbol{X1} = |\mathcal{H}(\boldsymbol{sig})| \tag{2.3}$$

$$\boldsymbol{env} = \left[\boldsymbol{X1} - \overline{X1}\right] \cdot \alpha + \beta \tag{2.4}$$

with

- $\boldsymbol{env}$ ... scaled envelope feature
- $\mathcal{H}$ ... Hilbert transformation
- $\boldsymbol{X1}$ ... raw envelope vector
- $\boldsymbol{sig}$ ... EMG signal
- $\overline{X1}$ ... mean value of $\boldsymbol{X1}$ (see equation 2.2)

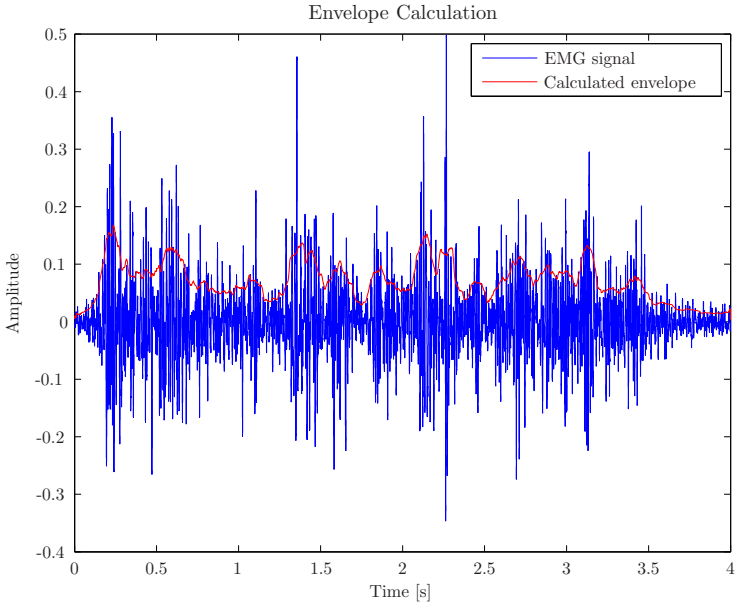

Figure 2.1: Calculated envelope of the EMG signal

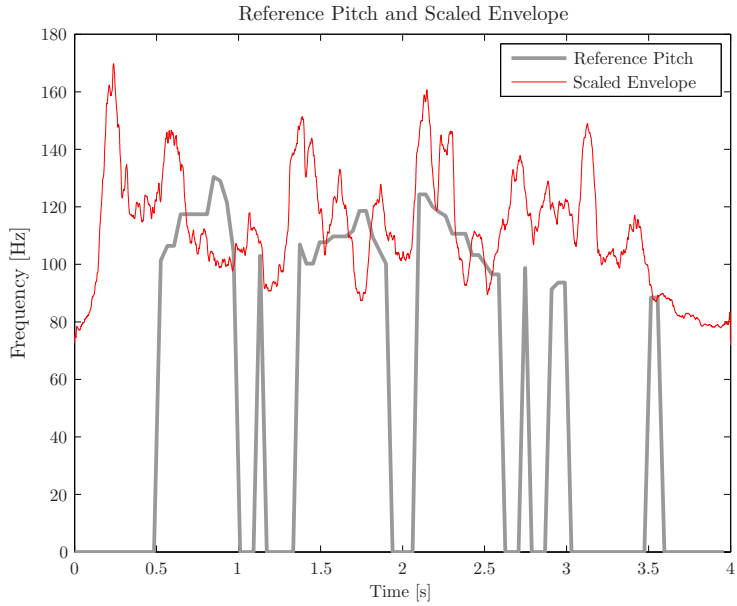

Figure 2.2: Reference Pitch and Scaled Envelope

Feature 2: Inverted Scaled Envelope

Feature 2 was obtained by inverting the scaled envelope (feature 1). It's obvious that the feature's graph in figure 2.3 is the inverted graph of feature 1 (figure 2.2).

$$\boldsymbol{inv} = [\boldsymbol{env} - \beta] \cdot (-1) + \beta \tag{2.5}$$

with

- $\boldsymbol{inv}$ ... inverted scaled envelope feature

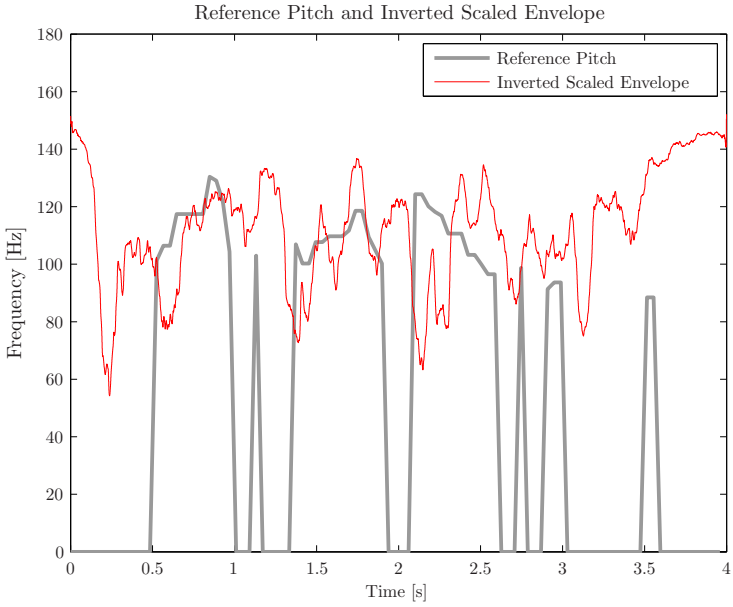

Figure 2.3: Reference Pitch and Inverted Scaled Envelope

Feature 3: Zero Crossing

The Zero Crossing Feature (ZC) uses the number of times that the amplitude value of the sEMG signal crosses the zero y-axis. The calculation of this feature is based on equation 2.6 and 2.7. In the *Matlab* implementation the indices where the signal crosses the zero y-axis were calculated first. Afterwards the differences of these indices were calculated, so there were high values for a low rate of zero crossings and low values for a high rate of zero crossings. In the next step this feature vector was smoothed and median filtered and inverted additionally. Finally the raw feature vector was shifted to the speaker's average fundamental frequency.

$$\boldsymbol{X3} = \Delta(\boldsymbol{ind}) \cdot (-1) \tag{2.6}$$

$$\boldsymbol{ZC} = \left[\boldsymbol{X3} - \overline{X3}\right] + \beta \tag{2.7}$$

with

- $\boldsymbol{ZC}$ ... zero crossing feature
- $\boldsymbol{X3}$ ... raw ZC feature vector
- Δ ... calculates the difference between adjacent indices
- $\boldsymbol{ind}$ ... indices of zero crossings
- $\overline{X3}$ ... mean value of $\boldsymbol{X3}$ (see equation 2.2)

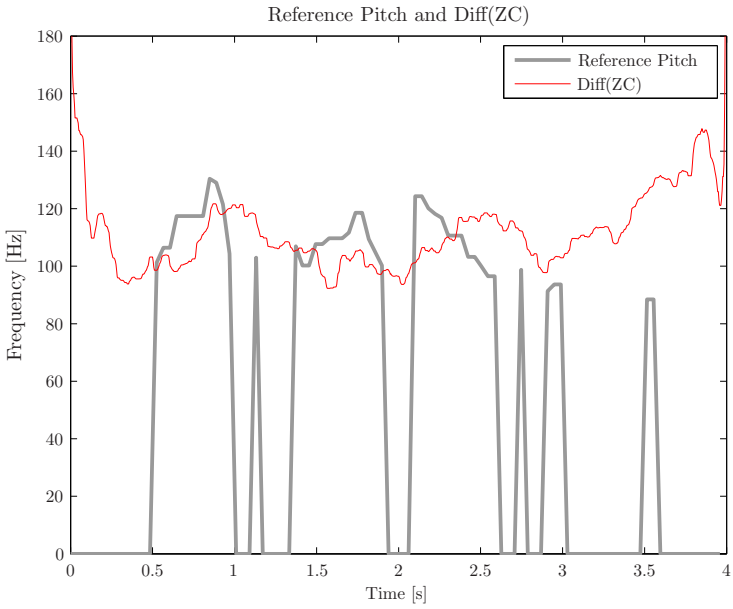

Figure 2.4: Reference Pitch and Difference of Zero Crossings

Feature 4: Inverted Slope Sign Change

The calculation of feature 4 is based on equation 2.8 and 2.9. In this feature the number of changes between positive and negative slope is calculated. In the implementation the SSC was obtained by dividing the sEMG signal in small windows and applying the particular function on each window. Afterwards this feature was smoothed, inverted and shifted.

$$X4 = \sum_{n=2}^{N-1} \Big[f\big[(y_n - y_{n-1}) \cdot (y_n - y_{n+1})\big] \Big] \tag{2.8}$$

$$f(y) = \begin{cases} 1, & \text{if y} \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\boldsymbol{X4} = [X4_1, X4_2, \ldots X4_M]$$

$$\boldsymbol{ISSC} = \big[\boldsymbol{X4} - \overline{X4}\big] \cdot (-1) + \beta \tag{2.9}$$

with

- $\boldsymbol{ISSC}$ ... inverted slope sign change feature
- $X4$ ... number of slope sign changes for a single window
- $\boldsymbol{X4}$ ... vector of slope sign changes for all windows
- y_n ... samples of sEMG signal
- $\overline{X4}$ ... mean value of $\boldsymbol{X4}$ (see equation 2.2)

- M ... number of windows
- N ... length of windows

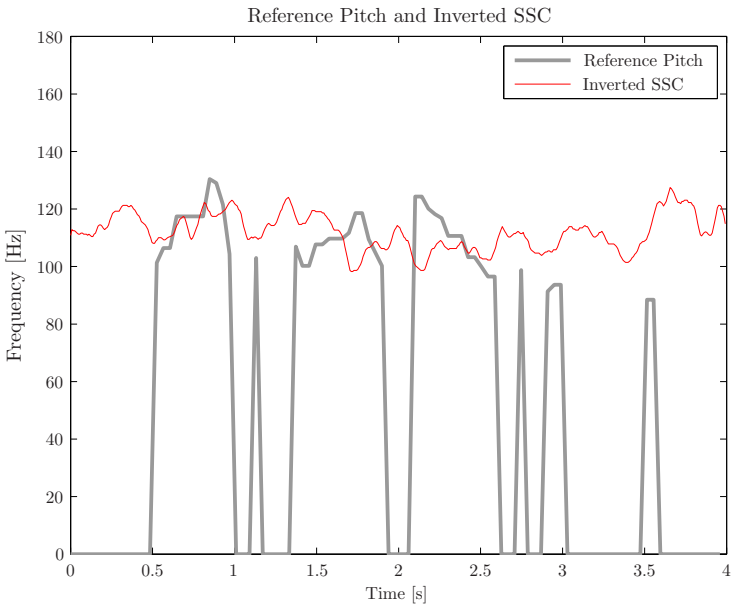

Figure 2.5: Reference Pitch and Inverted SSC

Feature 5: Modified Median Frequency (MMDF)[1]

The calculation of feature 5 is based on equation 2.10. Therefore the signal was transformed to the frequency domain and after the calculation this feature got smoothed. Finally the raw feature was spread with a spreading factor α of 5 and shifted to the average fundamental frequency (see 2.11).

$$\sum_{j=1}^{X5} A_j = \sum_{j=X5}^{N} A_j = \frac{1}{2}\sum_{j=1}^{N} A_j \tag{2.10}$$

$$\boldsymbol{X5} = [X5_1, X5_2, \ldots X5_n \ldots X5_M]$$

$$\boldsymbol{MMDF} = \left[\boldsymbol{X5} - \overline{X5}\right] \cdot \alpha + \beta \tag{2.11}$$

with

- $\boldsymbol{MMDF}$ ... modified median frequency feature
- $X5$ ... modified median frequency of a single window
- $\boldsymbol{X5}$ ... vector of modified median frequencies for all windows
- A_j ... sEMG amplitude spectrum at frequency bin j

[1] see section 3.5

- M ... number of windows
- N ... length of windows
- $\overline{X5}$ ... mean value of $\boldsymbol{X5}$ (see equation 2.2)

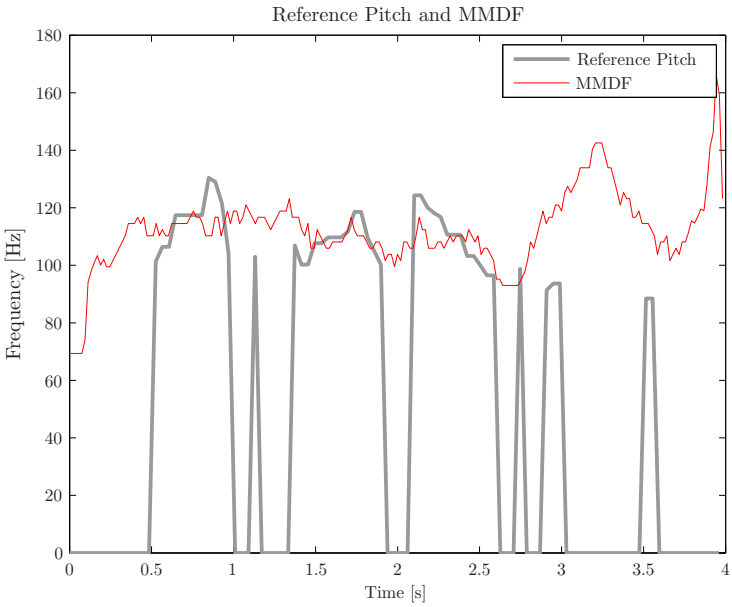

Figure 2.6: Reference Pitch and MMDF

Feature 6: Modified Mean Frequency (MMNF)

The calculation of feature 6 is based on equation 2.12 and 2.13. In addition to the calculation this feature was smoothed.

$$X6 = \sum_{j=1}^{N} f_j \cdot A_j \Big/ \sum_{j=1}^{N} A_j \tag{2.12}$$

$$\boldsymbol{X6} = [X6_1, X6_2, \ldots X6_M]$$

$$\boldsymbol{MMNF} = \left[\boldsymbol{X6} - \overline{X6}\right] + \beta \tag{2.13}$$

with

- $\boldsymbol{MMNF}$ ... modified mean frequency feature
- $X6$ ... modified mean frequency of a single window
- $\boldsymbol{X6}$ ... vector of modified mean frequencies for all windows
- A_j ... sEMG amplitude spectrum at frequency bin j
- f_j ... frequency of spectrum at frequency bin j

- M ... number of windows
- N ... length of windows
- $\overline{X6}$ ... mean value value of $\boldsymbol{X6}$ (see equation 2.2)

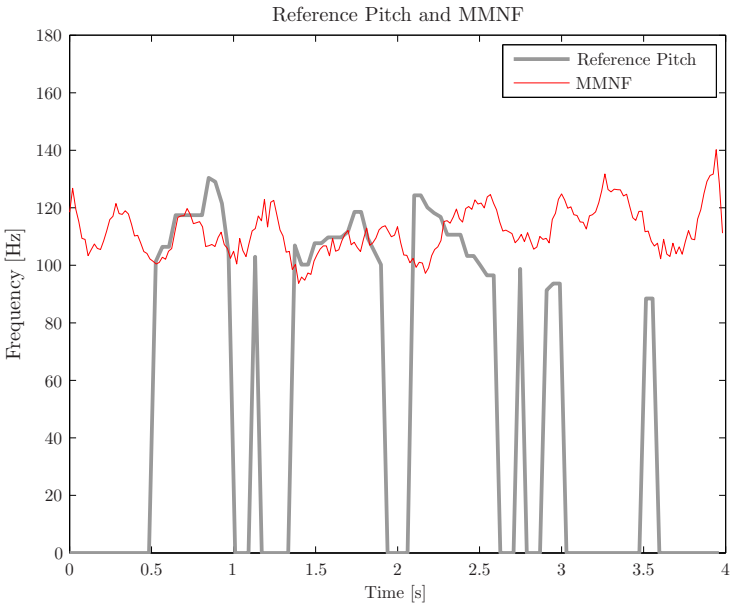

Figure 2.7: Reference Pitch and MMNF

Feature 7: Combination of MMNF and MMDF

Feature 7 is a combination of feature 5 and feature 6 and was calculated according to equation 2.14.

$$\boldsymbol{Comb} = \sqrt{\boldsymbol{MMNF} \cdot \boldsymbol{MMDF}} \tag{2.14}$$

with

- $\boldsymbol{Comb}$ ... Combination of MMNF and MMDF
- $\boldsymbol{MMNF}$ ... modified mean frequency feature
- $\boldsymbol{MMDF}$ ... modified median frequency feature

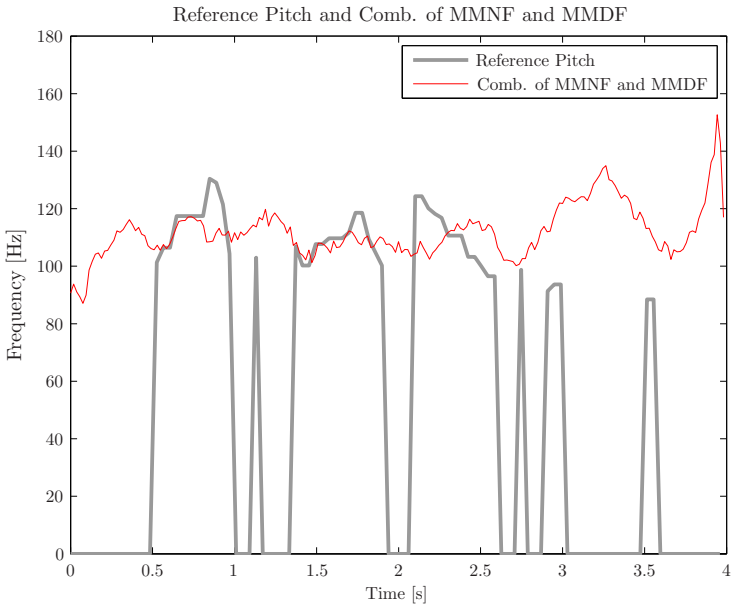

Figure 2.8: Reference Pitch and combination of MMNF and MMDF

Feature 8: Maximum Energy

In order to calculate feature 8 *Matlab's* spectrogram function (*spectrogram()*) was used. The calculation is based on equation 2.15 and 2.16. In every segment the frequency with the highest energy was detected. In the last step the raw feature was shifted to the estimated average fundamental frequency.

$$X8 = max(\boldsymbol{A}) \tag{2.15}$$

$$\boldsymbol{X8} = [X8_1, X8_2, \ldots X8_M]$$

$$\boldsymbol{Emax} = \left[\boldsymbol{X8} - \overline{X8}\right] + \beta \tag{2.16}$$

with

- $\boldsymbol{Emax}$ ... maximum energy feature vector
- $X8$ ... frequency with the highest energy in a single window
- $\boldsymbol{X8}$ ... vector of frequencies with highest energy of all windows
- $\boldsymbol{A}$ ... sEMG amplitude spectrum at each window
- $\overline{X8}$ ... mean value of $\boldsymbol{X8}$ (see equation 2.2)
- M ... number of windows

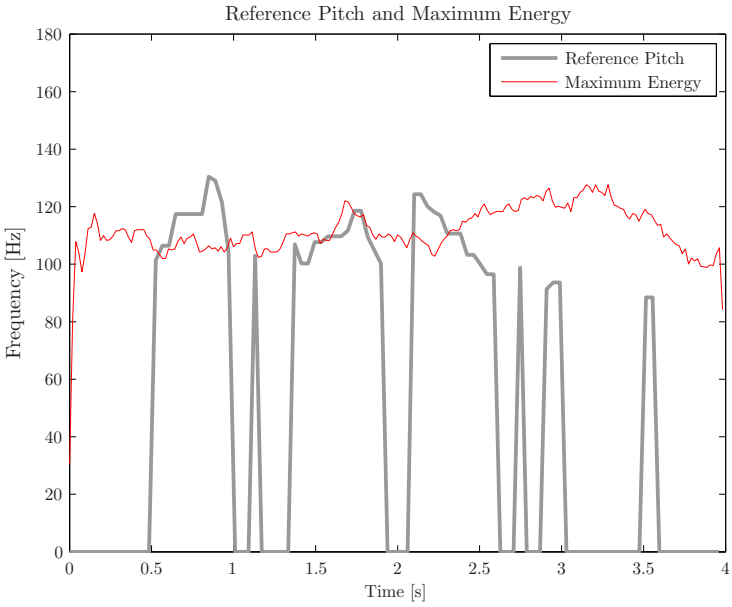

Figure 2.9: Reference Pitch and Maximum Energy

Feature 9: Random Pitch

To get the Random Pitch Feature a random sequence of numbers between 0 and 1 was generated with *Matlab's randn()*. Due to a multiplication the range was increased in order to get an appropriate frequency range for the field of human speech. After this the raw feature vector was shifted to the corresponding fundamental frequency (see 2.17).

$$\boldsymbol{Random} = \left[\boldsymbol{X9} - \overline{X9}\right] + \beta \tag{2.17}$$

- $\boldsymbol{Random}$ ... random feature vector
- $\boldsymbol{X9}$ ... raw random feature vector
- $\overline{X9}$ ... mean value of $\boldsymbol{X9}$ (see equation 2.2)

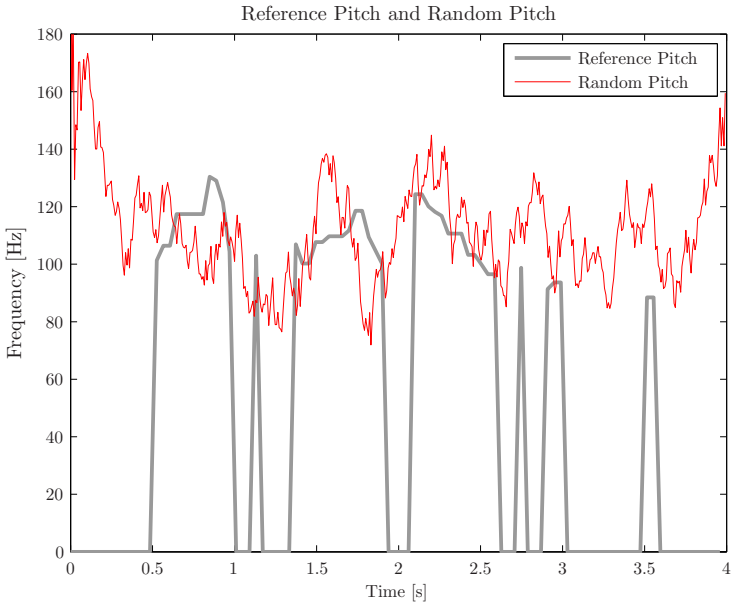

Figure 2.10: Reference Pitch and Random Pitch

Feature 10: Slow Envelope (Goldstein)

Feature 10 was obtained by a simple lowpass filter. Therefore a third order lowpass filter with cutoff frequency at $1Hz$ was used. This Slow Envelope Feature was extremely spread with a factor α of 50000 (see 2.18).

$$\boldsymbol{SE} = \left[\boldsymbol{X10} - \overline{X10}\right] \cdot \alpha + \beta \tag{2.18}$$

- $\boldsymbol{SE}$ ... slow envelope feature vector
- $\boldsymbol{X10}$ ... raw slow envelope feature vector
- $\overline{X10}$ ... mean value of $\boldsymbol{X10}$ (see equation 2.2)

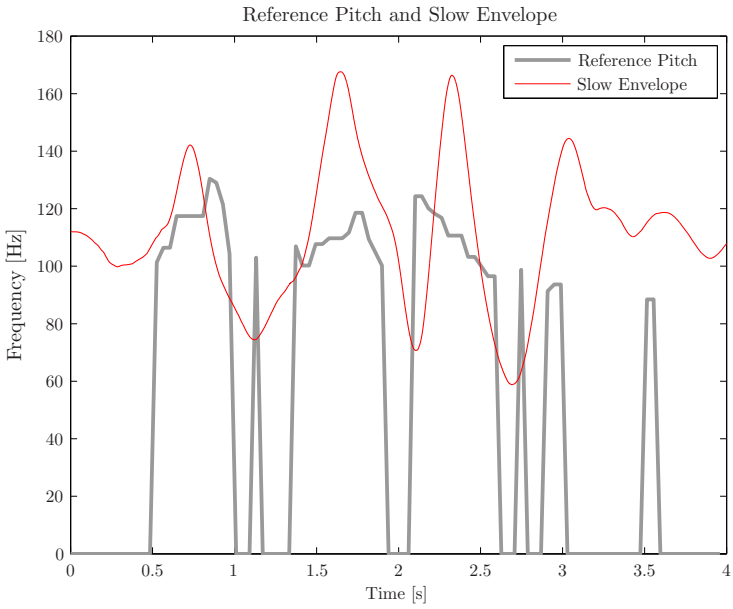

Figure 2.11: Reference Pitch and Slow Envelope

Reference Pitch

To obtain the reference pitch the HE speech signal was first modified through DTW. Then the fundamental frequency was received from *Praat* using *To Pitch(ac)...* . This algorithm performs an acoustic periodicity detection on the basis of an accurate autocorrelation method in order to create a pitch object.

2.1.1 Post Processing

Speech consists of voiced and unvoiced segments. In contrast to voiced segments, unvoiced parts of speech don't have a fundamental frequency because they don't have a detectable periodicity. In this project the assumption was made that there exist perfect voiced/unvoiced parts of speech. Due to this the elements in the feature vector which corresponded to the unvoiced parts of the EL signals were set to zero. Moreover these zero-parts were set to the offset value β which is the estimated average fundamental frequency for either female voice or male voice.

The EL speech was manipulated with the calculated features in order to obtain speech which sounds more natural than the original EL speech due to a varying fundamental frequency. The strategy of EL speech manipulation will be explained in more detail in section 2.2.

2.1.2 Spectral Subtraction

When the EL operates it produces an excitation noise and so the recorded EL speech was overlaid by this excitation noise. In order to get rid off this noise the Spectral Subtraction Method was applied. The recorded EL speech signal was windowed and Fourier transformed. In the frequency domain the noise was subtracted from the noise-corrupted signal and so removed from the recordings.

2.2 Manipulation of the EL Speech Data

The manipulation of the EL speech was performed in *Praat*. The calculated features were written to a textfile called *PitchTier*, an object *Praat* can handle.

> "A PitchTier object represents a time-stamped pitch contour, i.e. it contains a number of (time, pitch) points, without voiced or unvoiced information." [11]

The EL signal and the *PitchTier*-file were loaded to *Praat*. The EL signal was sent to *Praat's: To Manipulation...* and a *Manipulation object* was created. In the next step the frequency components in the *Manipulation object* were replaced by the *PitchTier*-frequencies with *Praat's Replace pitch tier*. Finally a resynthesis with overlap and add was done (with *Praat's Get resynthesis (overlap-add)*).

2.3 Features Used in the Listening Test

A LT shouldn't last too long because it is exhausting for the participants and as a result the attention level of the participants decreases with time. Due to this only some of the calculated features were chosen to be part of the LT.
Compared to the Reference Pitch the Scaled Envelope Feature and the Modified Median Frequency Feature showed a good performance regarding the fundamental frequency, which was the reason they were used in the LT.
In order to verify if the possible improvements due to sEMG based features are a consequence of the variation of the sEMG signals, or if just a - maybe arbitrary - variation in the pitch contour itself is the reason that makes speech sounding more natural, a randomly generated pitch contour was created.
The theoretically best case is given by the Reference Pitch Feature, where the EL signal is manipulated with the extracted fundamental frequency from HE speech, while the theoretically worst case is the original - but spectral subtracted - EL signal.
Finally as mentioned before and in section 2.1, the used features for the LT were the original, but spectral subtracted, EL signal, referred as Method A, the Scaled Envelope Feature, referred as Method B, the Modified Median Frequency Feature, referred as Method C, the Random Pitch, referred as method D, and the Reference Pitch which was called Method E in the LT.

2.4 Frequency Analysis

In order to obtain some knowledge about the calculated features and the manipulated LT files, frequency analysis was performed.

2.4.1 Frequency Ranges

With *Matlab's boxplot* function the features for all LT sentences were analysed graphically regarding the occuring frequencies, see figure 2.12 and figure 2.13. The central red mark of each box is the median, the blue edges of the boxes are the 25th and 75th percentiles and the black whiskers extend to the most extreme datapoints the algorithm considers to be not outliers. The outliers are plotted individually with red crosses. Note that the red labeled feature wasn't calculated correct (see section 3.5). Furthermore the red and blue labeled features were used in the LT.
Considering these two plots, one can see that the median of nearly every feature deviates only sligthly from the offset values, where the female offset value was 193 Hz and the male offset value

was 112 Hz. These little variations once derive from the shifting itself because the offset values have been calculated by using *Matlab's mean* function but *boxplot* shows the median value. On the other side the boxplot shouldn't be falsified by the values in the feature vectors corresponding to the unvoiced parts of speech, so these values were ignored for the *boxplot* calculation.

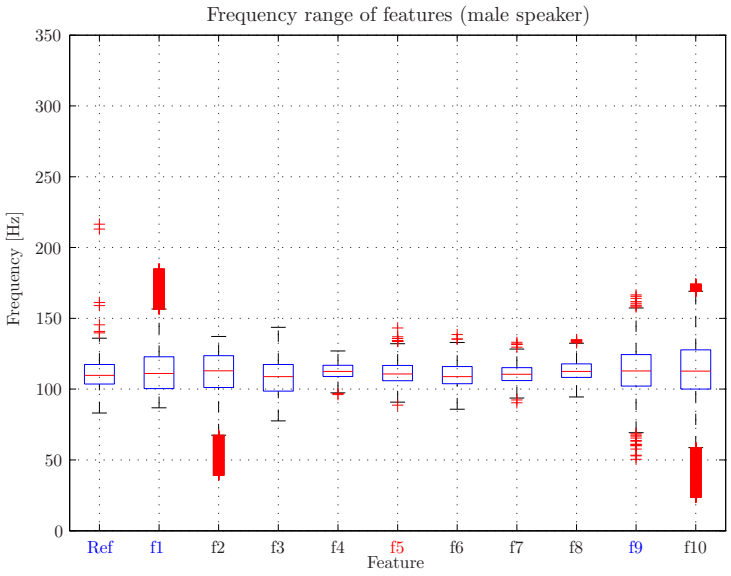

Figure 2.12: Frequency ranges of all features and over all LT sentences (male speaker); Ref... Reference Pitch (Method E); f1... Scaled Envelope (Method B); f2... Inverted Scaled Envelope; f3... ZC; f4... Inverted SSC; f5... MMDF (Method C); f6... MMNF; f7... Comb. of MMDF and MMNF; f8... Maximum Energy; f9... Random Pitch (Method D); f10... Slow Envelope

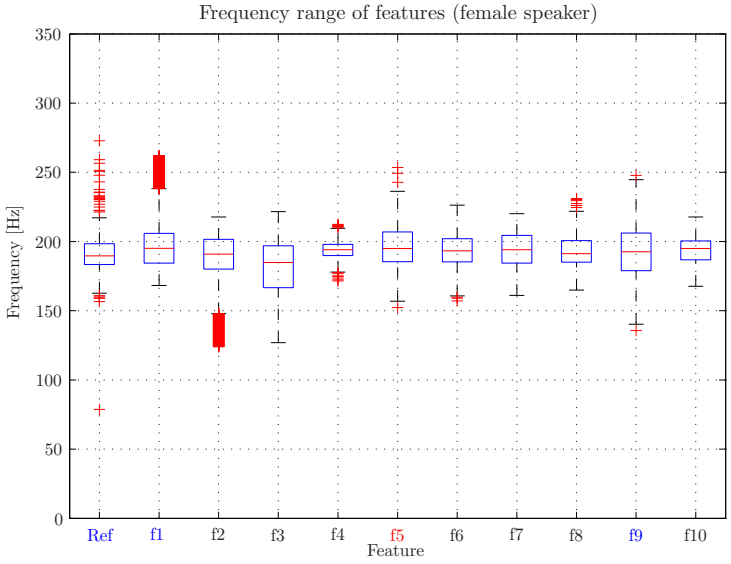

Figure 2.13: Frequency ranges of all features and over all LT sentences (female speaker); Ref... Reference Pitch (Method E); f1... Scaled Envelope (Method B); f2... Inverted Scaled Envelope; f3... ZC; f4... Inverted SSC; f5... MMDF (Method C); f6... MMNF; f7... Comb. of MMDF and MMNF; f8... Maximum Energy; f9... Random Pitch (Method D); f10... Slow Envelope

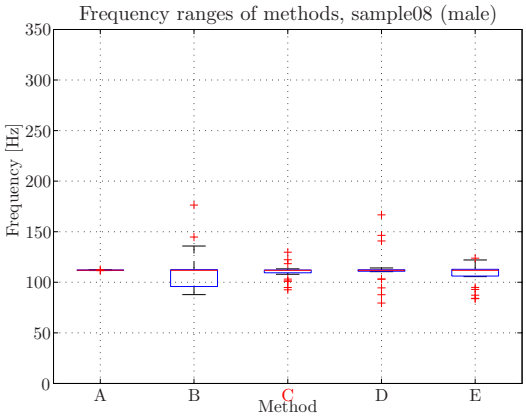

Figure 2.14: Frequency ranges of LT sample08 (utterance: 'Der Bär hat den Fisch gefangen.' (Male speaker)); A... Spectral Subtracted EL sentence; B... Scaled Envelope; C... MMDF; D... Random Pitch; E... Reference Pitch

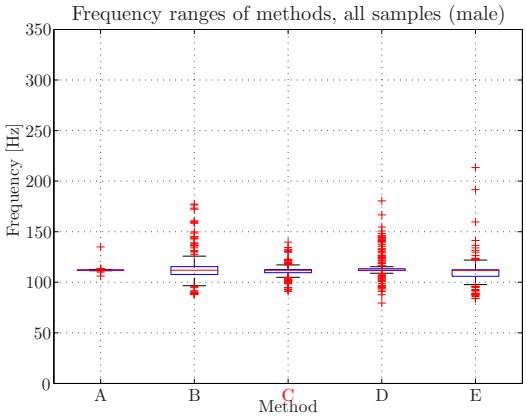

Figure 2.15: Frequency ranges of LT samples (male speaker); A... Spectral Subtracted EL sentence; B... Scaled Envelope; C... MMDF; D... Random Pitch; E... Reference Pitch

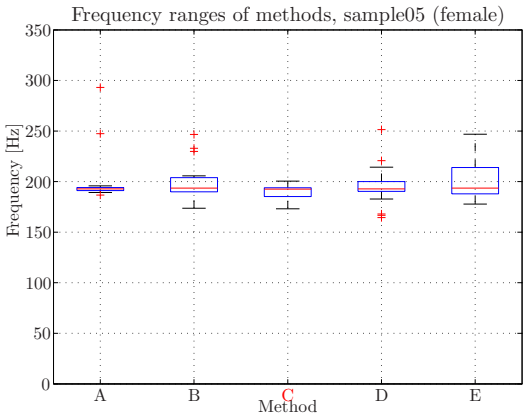

Figure 2.16: Frequency ranges of LT sample05 (utterance: 'Die Oma trinkt einen Kaffee.' (Female speaker)); A... Spectral Subtracted EL sentence; B... Scaled Envelope; C... MMDF; D... Random Pitch; E... Reference Pitch

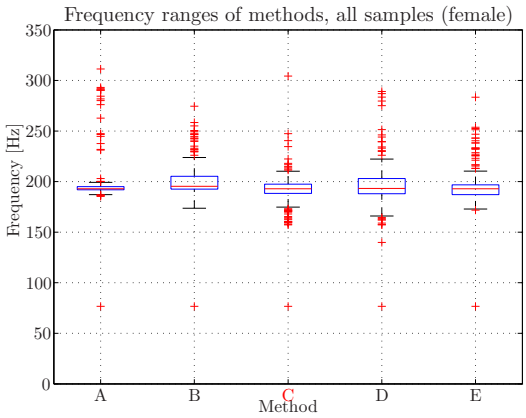

Figure 2.17: Frequency ranges of LT samples (female speaker); A... Spectral Subtracted EL sentence; B... Scaled Envelope; C... MMDF; D... Random Pitch; E... Reference Pitch

Figure 2.14 and figure 2.16 show the frequency ranges of a particular sentence spoken by the male and the female speaker. Additionally figure 2.15 and 2.17 show the frequency ranges over all used LT sentences. In contrast to figure 2.12 and figure 2.13 the sentences here are already modulated with the features and analysed with *Praat* again.
It clearly can be seen that Method A, which is the spectral subtracted EL signal, doesn't have any frequency fluctuations because of its monotonic and constant characteristics. The random pitch (Method D) has many outliers due to the fact that the generation of the random pitch frequency vector wasn't dependent on the given speech signal in any way.
In the *Matlab* code for the calculation of feature 5 (Method C, red labeled) an error appeared, see section 3.5, therefore comparisons with Method C are only possible to a certain degree.

2.4.2 Frequency of Modulated Files

Figure 2.18 and 2.19 are comparisons of the calculated Scaled Envelope Feature and the reference pitch frequencies respectively, with the frequencies of the corresponding manipulated EL speech file. The pitches of the manipulated EL files were obtained by analysing them with *Praat*. For a better arrangement the calculated features were plotted before setting the zeros (the non-voiced segments) to the offset values. The overshooting in figure 2.18 round 0 Hz is a result of the resampling operation.

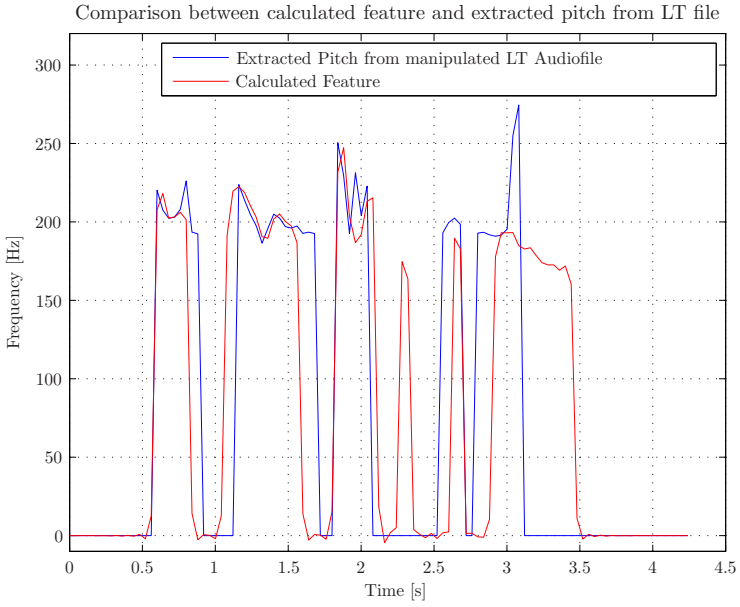

Figure 2.18: Comparison Sample 04 Method B (utterance: 'Radfahrer sausen vorbei.')

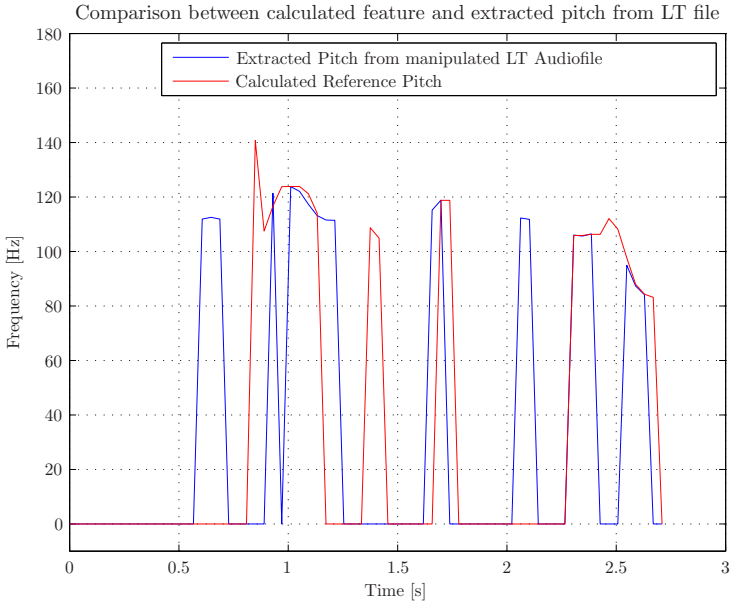

Figure 2.19: Comparison Sample 08 Method E (utterance: 'Der Bär hat den Fisch gefangen.')

The manipulation of the EL speech files was only possible in that speech sections where *Praat* detected voiced segments. Considering figure 2.18 and figure 2.19 the detected voiced segements appear where the blue graph is unequal to zero. If there is also a non-zero red graph in the same segment, the blue graph is the result of the manipulation of the EL signal with the calculated feature. If the red graph is equal to zero in that segment, the EL signal was set to the male or female offset value, which can easily be seen as constant sections of the blue graph in 2.19. If the red graph is unequal to zero while the blue graph is zero, *Praat* didn't detect a voiced segment in the EL signal and so the frequency manipulation of this segment wasn't possible.

3 Listening Test and Results

This chapter should clarify the developement of the listening test. In addition the evaluation of the results is presented.

3.1 Design of the Listening Test

Participants of the LT had to rank ten sentences which were manipulated as described in 2. Every sentence was provided five times in different manipulation realisations.
The different methods were linked to different sliders in every turn. Due to this random preparation of the LT sentences falsifications because of familiarisation of the participants with the sentences and the methods were avoided.

- Method A = original: Spectral Subtracted EL sentence
- Method B = feature 1: Scaled Envelope
- Method C = feature 5: Modified Median Frequency (MMDF)
- Method D = feature 9: Random Pitch
- Method E = feature 11: Reference Pitch

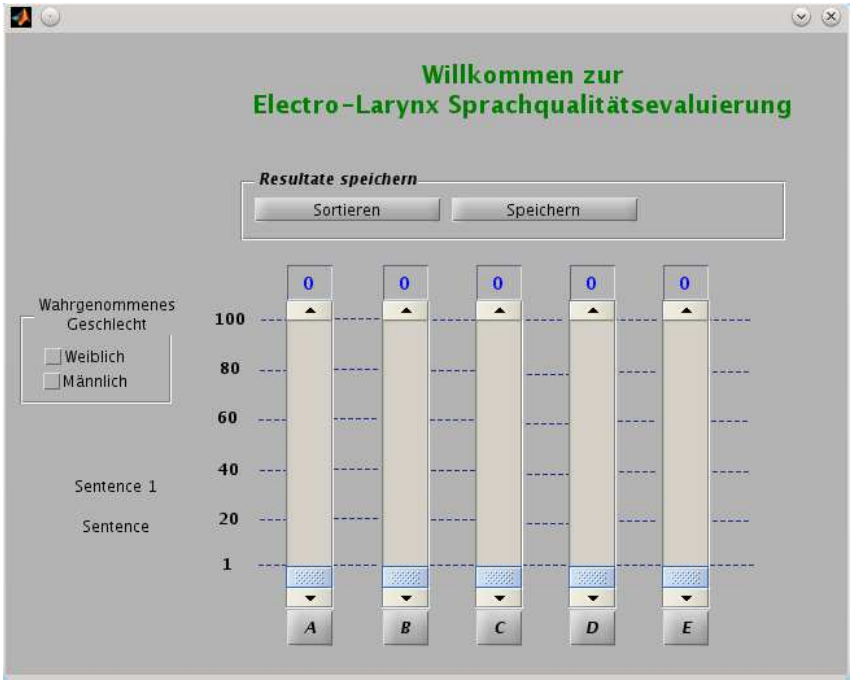

Figure 3.1: Listening Test GUI

The participants were instructed to evaluate the heard sentences regarding to their naturalness. Therefore they had to rate every sentence with a number between *1* and *100*. Additionally the

speaker's sex had to be specified. In figure 3.1 the LT's graphical user interface (GUI) is shown. Nine persons took part in the LT. Three of them had previous experiences with EL speech and therefore were regarded as expert listeners. The remaining six listeners were naive in the topic of EL speech.

Below there is a list of the sentences used in the LT.

- **Test sentences**
  - sample 01: Wer möchte noch Milch? (Female voice)
  - sample 02: Bald ist der Hunger gestillt. (Female voice)
  - sample 03: Achte auf die Autos! (Female voice)
  - sample 04: Radfahrer sausen vorbei. (Female voice)
  - sample 05: Die Oma trinkt einen Kaffee. (Female voice)
  - sample 06: Ich werde mit der Fähre nach Irland übersetzen. (Male voice)
  - sample 07: Ich werde den Text ins Englische übersetzen. (Male voice)
  - sample 08: Der Bär hat den Fisch gefangen. (Male voice)
  - sample 09: Der Kaffee dampft in den Tassen. (Male voice)
  - sample 10: Achte auf die Autos! (Male voice)
- **Sentences to make the testperson familiar with the LT sentences**
  - sample 11 Die Sonne lacht. (Female voice)
  - sample 12 Hans isst so gerne Wurst. (Female voice)
  - sample 13 Am blauen Himmel ziehen die Wolken. (Male voice)
  - sample 14 Messer und Gabel liegen neben dem Teller. (Male voice)

3.2 Realisation of the Listening Test

The LT was conducted on February, the 6th and 7th 2014, and took place in the Cocktail Party Room at the SPSC Institute of the Technical University of Graz. In order to expulse participants with a loss of hearing, the hearing levels of the participants were determined beforehand through an audiometric testing which took place in the Cocktail Party Room, too (see figure 3.2).

Following instruments were used for the audiometric test:

- Software and equipment: Audio-Ton

The audiometry test took place under same conditions for all participants and was undertaken by an instructor who sat opposite to the participant. Participants were instructed before the test. During the audiometric testing, the paticipants first got to hear a particular frequency with a level of 40dB. Then the same frequency was played again starting at -10dB and the level was increased until the participant gave a sign for noticing the testtone again. In the next step the same testtone was played starting from that noticed point and the volume level was decreased until the participant gave a sign, now for being unable to hear the testtone. While repeating these two steps a few times there occured an intersection which was drawn as perception treshold point into the audiogramm. This procedure was repeated for the speech relevant frequencies from 250 Hz to 8 kHz and for both ears. A participant with normal hearing level

should not reach values higher than 15 dB (see figure 3.2). On average the audiometry took about 15 minutes. One problem was that the location was barely perfect regarding background noise.

In case of a positive audiometric test (no values higher than 15 dB) the participant was able to take part in the LT. To create the same conditions for each participant, everyone had to read an instruction sheet first, which described the procedure and listed the test sentences used in the LT. Finally participants had to fill in a short feedback questionnaire. The LT took about 25 minutes, thus, considering also the time for the audiometry test, the concentration of the participants decreased in the end, which was mentioned in the feedback questionnaire. It was also mentioned that it wasn't that easy for the participants to compare and rank five different audio files among each another.

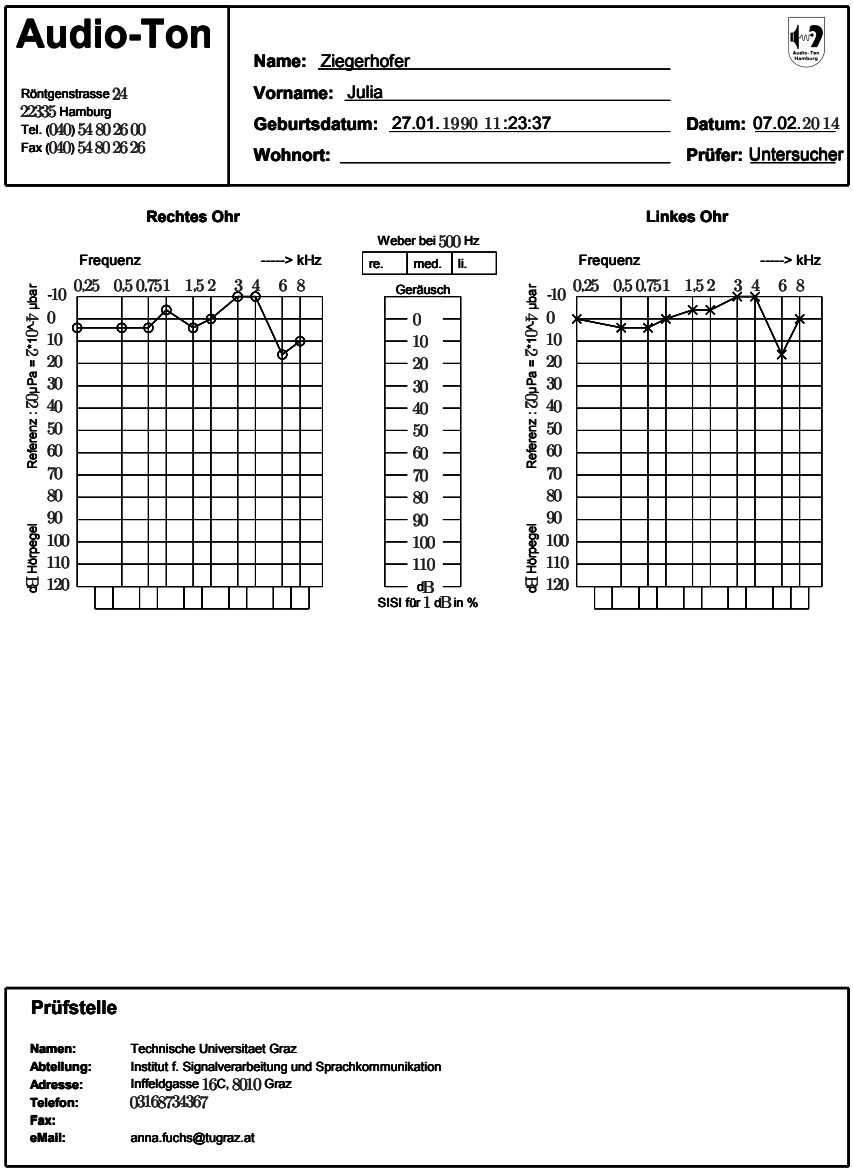

Figure 3.2: One result of the audiometric testing

3.3 Results

3.3.1 Results for Different Groups of LT Participants

In order to compare different groups of participants in the LT the test results were normalized and represented with *Matlab's boxplot*. Figure 3.3 shows the results for all participants. In figure 3.4 the results for the three expert listeners are plotted and figure 3.5 shows the results for the naive listeners. Considering figure 3.4 it is obvious that Method A was ranked worst most times because of the fact that expert listeners were already used to EL speech. In the same figure Method E (manipulation with healthy frequency values) was ranked best most of the time which was an expected result too. In figure 3.3 and 3.5, respectively, despite not as distinct results as in figure 3.4 it can be obtained that Method A was ranked worst most of the time.

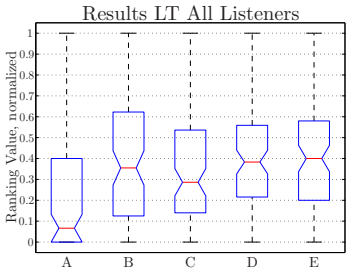

Figure 3.3: Result of the LT All Listeners

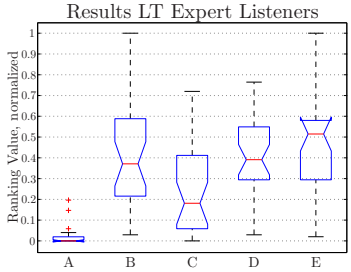

Figure 3.4: Result of the LT Expert Listeners

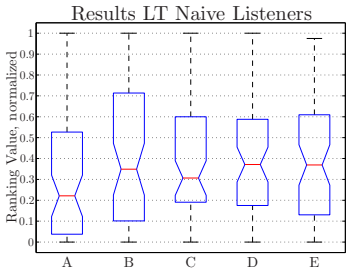

Figure 3.5: Result of the LT Naive Listeners

3.4 Evaluation

3.4.1 Evaluation of the Ranking Levels

Ranking of Preference for the Various Methods

In the following tables the LT answers are analysed. Tables 3.1 to 3.4 list the times Method XY was rated best. Additionally this value is expressed as a percentage. Table 3.5 to 3.8 shows how often Method XY was rated worst. In table 3.4 and table 3.8 three of the naive listeners whose LT answers were irreproducible and implausible were omitted, the remaining group was referred as certain listeners. The results of the participants in the LT were analysed together (see table 3.1 and table 3.5) and in different groups. Considering table 3.5 to table 3.8 it's obvious that EL speech was rated worst most frequently by all participants in the LT. However, considering 3.1 to 3.4 a similar distinct result can't be seen. As mentioned before in contrast to the other features the Reference Pitch Feature (Method E) wasn't calculated but obtained from the HE signal. So the assumption was made that manipulation with reference pitch would lead to the best - since sounding most natural - reachable result. However, against this assumption, Method E was only rated as the best method by the expert listeners (table 3.2), while naive listeners rated Method B and Method C, respectively, best.

3.4.2 Evaluation with Student's T-Distribution

It was assumed that the LT testpersons' ratings were distributed normally, so a one-sided t-test was used to determine the statistical significance between those answers.

Table 3.1: Times Method XY was rated best All Listeners (9 Persons)

Method	Quantity	Percentage
Method B	24	26.67%
Method E	19	21.11%
Method C	18	20%
Method D	17	18.89%
Method A	12	13.33%

Table 3.2: Times Method XY was rated best Expert Listeners (3 Persons)

Method	Quantity	Percentage
Method E	12	40%
Method B	10	33.33%
Method D	5	16.67%
Method C	3	10%
Method A	0	0%

Table 3.3: Times Method XY was rated best Naive Listeners (6 Persons)

Method	Quantity	Percentage
Method C	15	25%
Method B	14	23.33%
Method D	12	20%
Method A	12	20%
Method E	7	11.67%

Table 3.4: Times Method XY was rated best Certain Listeners (6 Persons)

Method	Quantity	Percentage
Method B	19	31.67%
Method E	16	26.67%
Method C	12	20%
Method D	12	20%
Method A	1	1.67%

Table 3.5: Times Method XY was rated worst All Listeners (9 Persons)

Method	Quantity	Percentage
Method A	48	53.33%
Method C	14	15.56%
Method D	11	12.22%
Method E	9	10%
Method B	8	8.89%

Table 3.6: Times Method XY was rated worst Expert Listeners (3 Persons)

Method	Quantity	Percentage
Method A	27	90%
Method C	3	10%
Method B	0	0%
Method D	0	0%
Method E	0	0%

Table 3.7: Times Method XY was rated worst Naive Listeners (6 Persons)

Method	Quantity	Percentage
Method A	21	35%
Method C	11	18.33%
Method D	11	18.33%
Method E	9	15%
Method B	8	13.33%

Table 3.8: Times Method XY was rated worst Certain Listeners (6 Persons)

Method	Quantity	Percentage
Method A	40	66.67%
Method C	9	15%
Method E	5	8.33%
Method B	3	5%
Method D	3	5%

The Student t-distribution was chosen due to the fact that the sample size was small (only nine testpersons) and the standard deviation was unknown. For a large number of samples the t-distribution resembles the normal distribution.

> "A one-sided test is a statistical hypothesis test in which the values for which we can reject the null hypothesis H_0, are located entirely in one tail of the probability distribution. In other words, the critical region for a one-sided test is the set of values less than the critical value of the test, or the set of values greater than the critical value of the test." [...]
>
> "The significance level of a statistical hypothesis test is a fixed probability of wrongly

rejecting the null hypothesis H_0, if it is in fact true." [12]

For the one sided t-test the significance level was set to 0.05% which yielded to a confidence level of 0.95%.
Then it was tested whether the testpersons' ratings for one certain method had a significant higher mean value than the ratings for the other methods.
Figures 3.6 to 3.9 show the results for the Student t-test analysis. It's obvious that Methods B,C,D and E are significantly better than Method A.

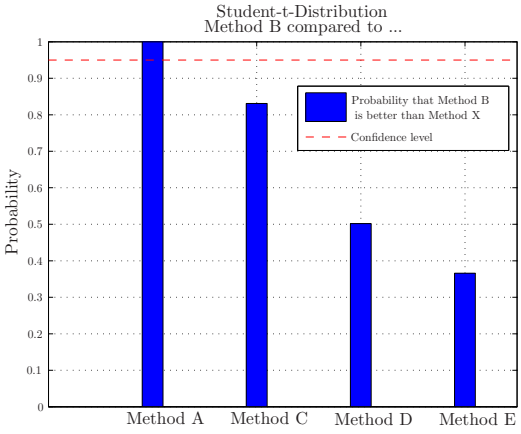

Figure 3.6: Comparison Method B with Method X

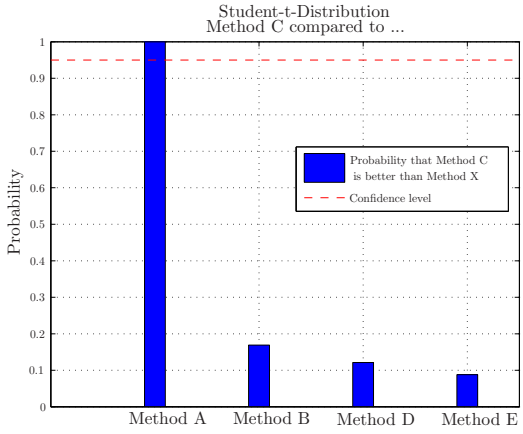

Figure 3.7: Comparison Method C with Method X

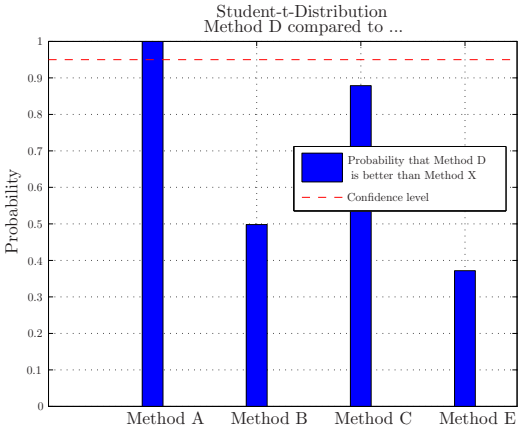

Figure 3.8: Comparison Method D with Method X

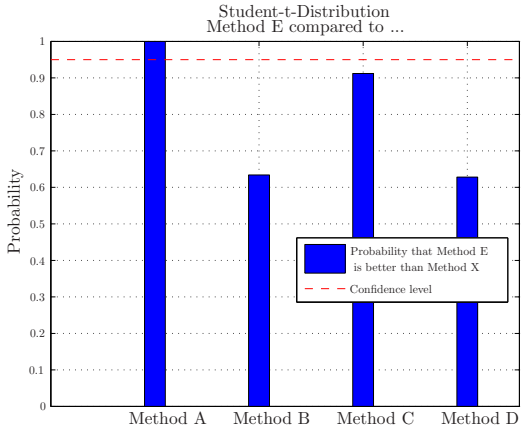

Figure 3.9: Comparison Method E with Method X

3.5 Bug

In the function which calculated feature 5, the Modified Median Frequency Feature, there was a fatal bug so that this function didn't calculate the feature to the actual input sEMG signal, but always took one sEMG signal for all features. So every EL signal was manipulated with the same, fixed feature vector. Unfortunately this mistake was not found out until the LT was completely finished. As a result the LT files for that particular feature are not valid.
However, there is some knowledge to withdraw from this bug. The LT results show that, no

matter what happens with the EL signal, everything is better than the original EL signal. For instance even the random feature deliverd better results than the original EL signal.

4 Conclusion

The focus of this project was to improve the perception of the EL speech referred to its naturalness. After reading several papers in order to get familiar with the topics of EL speech and EMG signals, data was analysed and features were calculated. Then it was decided which features were used and the EL signal was manipulated with them. In the next step a LT was designed and undertaken. In the LT manipulated sentences had to be ranked. The results were evaluated and illustrated.
The evaluation of the LT results showed that the frequency manipulated EL speech was rated better than the spectral subtracted EL speech. However, there weren't that expected relevant differences in the rating results for features and reference pitch. But, one have to say that when there is discourse on the "original" EL signal, which was used in the LT, then it is barely true because it has already been spectral subtracted. That's the reason why this "original" EL signal would probably lead to a better performance compared to EL speech as it sounds in real life.
Retrospectivly there are a few improvements which can be made. First the available speech data should be checked with respect to the speech quality and clarity. To verify if results for sEMG signals of speakers with a healthy larynx also hold for speakers without a larynx, sEMG data of speakers without a larynx should be available. Considering the design of the LT there are some modifications which probably would lead to better results. The test sentences should be chosen carefully with regard to the preprocessing. For some sentences the depression of the constant EL frequency worked better than for others. In some cases there was a disturbing musical noise, in other cases the filtered signal sounded very artificial, especially for test sentences spoken by the male speaker. There was a cracking noise in the lower frequency regions.
Participants in the LT noted that it was very exhausting to compare five methods to each other. So the layout of the GUI used for the LT could be changed for the better by inquiring only pairs of methods (paired comparison). Regarding the rating values it maybe would be better to use a nominal scale instead of a numerical scale.

Bibliography

[1] P. Vary and R. Martin, *Digital Speech Transmission Enhancement, Coding and Error Concealment*. The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England: John Wiley & Sons Ltd, 2006.

[2] 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, no. 2, 2004.

[3] C. E. Stepp, "Electromyographic Control of Prosthetic Voice after Total Laryngectomy," Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 2008.

[4] A. Phinyomark, C. Limsakul, and P. Phukpattaranont, "A Novel Feature Extraction for Robust EMG Pattern Recognition," *Journal of Computing*, vol. 1, 2009.

[5] C. Amon, "Using sEMG for Disordered Speech Enhancement," Master's thesis, Signal Processing and Speech Communications Laboratory Graz University of Technology, Austria, 2014.

[6] S. Kumar, D. K. Kumar, M. Alemu, and M. Burry, "EMG Based Voice Recognition," *Intelligent Sensors, Sensor Networks and Information Processing Conference*, pp. 593–597, 2004, DOS: 10.1109/ISSNIP.2004.1417528.

[7] S. S. eXchange, "Sound processing program," http://sox.sourceforge.net/.

[8] D. Ellis, "Dynamic time warping (matlab code)," dpwe@ee.columbia.edu, 2003.

[9] Matlab, "Numerical computing environment," http://www.mathworks.de/products/matlab/.

[10] Praat, "Software package for the analysis of speech in phonetics," http://www.fon.hum.uva.nl/praat/.

[11] P. Boersma and D. Weenink, "Manual praat 5.3.16," 1992-2012.

[12] V. J. Easton and J. H. McColl, "Statistics Glossary v1.1 from the STEPS Project," http://www.stats.gla.ac.uk/steps/glossary/, 1997.