TU Graz

Graz University of Technology

Dipl.-Ing. Anna Katharina Fuchs

The Bionic Electro-Larynx Speech System

Challenges, Investigations, and Solutions

Dissertation

for the degree of:

Doctor of Technical Sciences

submitted at

Graz University of Technology

Supervisor and First Examiner:

Gernot Kubin, Professor

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

Second Examiner:

Tomoki Toda, Associate Professor

Augmented Human Communication Laboratory
Nara Institute of Science and Technology, Japan

Graz, August 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 used sources.

date	(signature)

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Graz, am	(Unterschrift)

Abstract

Humans without larynx need to use a substitution voice to re-obtain speech. The electro-larynx (EL) is a widely used device but is known for its unnatural and monotonic speech quality. Previous research tackled these problems, but until now no significant improvements could be reported. The EL speech system is a complex system including hardware (artificial excitation source or sound transducer) and software (control and generation of the artificial excitation signal). It is not enough to consider one separated problem, but all aspects of the EL speech system need to be taken into account. In this thesis we would like to push forward the boundaries of the conventional EL device towards a new bionic electro-larynx speech system.

We formulate two overall scenarios: a closed-loop scenario, where EL speech is excited and simultaneously recorded using an EL speech system, and the artificial excitation signal is controlled based on the preceding recordings and sent back to the EL speech system to excite the vocal tract; and an open-loop scenario, where signal processing algorithms are used to enhance and improve recorded EL speech, and a loudspeaker is used for playback. Although we emphasize the first scenario, because it is closer to natural speech production, the latter is capable of significant improvements in terms of naturalness and can be used in telecommunication applications.

We record a German parallel electro-larynx speech – healthy speech database in order to carry out our experiments. Moreover, we provide algorithms for signal-to-noise ratio calculations and analyses of the data. We propose an algorithm to estimate a changing fundamental frequency from the speech spectral envelope. Listening tests show that a changing fundamental frequency improves the perceived naturalness of EL speech. Moreover, our proposed estimation algorithm increases the naturalness significantly compared to constant and random fundamental frequency contours. Furthermore, we study electromyographic (EMG) signals to analyze their suitability for on/off control of the EL speech system and investigate learning effects of naive users. Listening tests show that, after training, EMG controlled EL speech is significantly more pleasant to listen to than before training. We propose a new transducer for the EL speech system based on electro-magnetic mechanisms. The technical properties of the new transducer show significant advantages over the conventional electro-dynamic transducer. We design a housing for the transducer, and suggestions for an optimal coupler disk and the waveform of the excitation source are given. Listening tests serve as a proof of concept for the resulting EL speech system which means that the proposed system turns out to be promising.

For the open-loop scenario we perform statistical voice conversion (SVC) which leads to improvements in terms of naturalness, but intelligibility suffers. SVC is very promising to improve EL speech, but more investigations will need to be carried out.

In conclusion we propose a new, bionic electro-larynx speech system which uses a new, electro-magnetic transducer with a proposed waveform as artificial excitation signal, including a mechanism to estimate a changing fundamental frequency and providing a hands-free control of the on/off signal for the speech system.

Kurzfassung

Menschen ohne Kehlkopf müssen eine Ersatzstimme verwenden, um wieder sprechen zu können. Die elektronische Sprechhilfe (Elektro-Larynx – EL) ist ein weit verbreitetes Gerät. Es ist jedoch bekannt für die Unnatürlichkeit und Monotonie der resultierenden Sprache. Obwohl sich bisherige Forschung mit der Problematik beschäftigt hat, konnten bis heute keine signifikanten Verbesserungen berichtet werden. Die EL ist ein komplexes System, das aus Hardware (künstliche Anregungsquelle bzw. Schallwandler) und Software (Steuerung und Erzeugung der künstlichen Anregungssignale) besteht. Es ist nicht genug ein spezifisches Problem zu behandeln, es müssen alle Aspekte gleichrangig untersucht werden. Darum wollen wir in dieser Arbeit die Grenzen der gängigen EL vorantreiben und eine neue bionische Sprechhilfe vorstellen.

Wir gehen von 2 Szenarien aus: die geschlossene Regelschleife bei der produzierte Sprache simultan aufgenommen wird. Anschließend wird mit den, unmittelbar zuvor aufgenommenen Daten, das künstliche Anregungssignal gesteuert und an die EL zurückgeschickt, um dort den Vokaltrakt anzuregen; und die offene Regelschleife, bei der Algorithmen der Signalverarbeitung verwendet werden, um die aufgenommene EL Sprache zu verbessern und über Lautsprecher wiederzugeben. Unser Fokus liegt auf dem ersten Szenario, weil es mehr der natürlichen Sprachproduktion entspricht. Das zweite Szenario ist jedoch fähig, weitgehende Verbesserungen der Natürlichkeit zu erreichen und in der Telekommunikation angewendet zu werden.

Wir nehmen eine deutschsprachige, parallele Datenbank mit EL-Sprache und normaler Sprache gesunder SprecherInnen auf, um unsere Experimente durchführen zu können. Fernerhin stellen wir Algorithmen zur Berechnung des Signal-Rausch-Abstandes und zur Analyse zur Verfügung. Anschließend formulieren wir eine Methode zur automatischen Schätzung einer veränderlichen Grundfrequenz, basierend auf der spektralen Einhüllenden von Sprachsignalen. Hörversuche zeigen, dass eine sich verändernde Grundfrequenz die wahrgenommene Natürlichkeit von EL Sprache verbessert. Außerdem erhöht der vorgestellte Schätz-Algorithmus, im Vergleich zu konstanter und sich zufällig verändernder Grundfrequenz, die Natürlichkeit signifikant. Wir untersuchen elektromyographische (EMG) Signale auf ihre Eignung für die Steuerung des Ein-/Aus-Signals der EL und untersuchen auch den Lerneffekt bei naiven EL-BenutzerInnen. Hörversuche zeigen, dass die EMG gesteuerte EL Sprache nach dem Training signifikant angenehmer klingt als vor dem Training. Wir stellen eine neue Hardware vor, die ein elektro-magnetisches Wandlerprinzip verwendet. Die technischen Eigenschaften des neues Wandlers zeigen signifikante Vorteile zum herkömmlichen, elektro-dynamischen Wandler. Für den neuen Wandler wird ein Gehäuse entworfen und wir geben Empfehlungen für die optimale Koppelungsplatte und die Wellenform des künstlichen Anregungssignals. Hörversuche bestätigen die Machbarkeit des EL Sprachsystems und untermauern seine vielversprechenden Eigenschaften.

Für die offene Regelschleife verwenden wir Algorithmen der statistischen Sprachkonvertierung, die Verbesserungen hinsichtlich der Natürlichkeit jedoch schlechtere Verständlichkeit liefern. Statistische Sprachkonvertierung ist ein vielversprechender Ansatz zur Verbesserung von EL Sprache, es müssen jedoch noch mehr Untersuchungen durchgeführt werden.

Zusammenfassend stellen wir eine neue bionische Sprechhilfe vor, die einen neuen elektromagnetischen Wandler und ein veränderliches künstliches Anregungssignal verwendet, wobei ein automatischer Algorithmus die künstliche Grundfrequenz schätzt. Außerdem verwendet die neue bionische Sprechhilfe Muskelspannungen um das Gerät ein- und auszuschalten ohne die Hände zu benötigen.

Acknowledgments

The following people are owed a great deal of thanks for helping me directly and indirectly to complete this thesis:

I would like to express my appreciation to Gernot Kubin. Many thanks for being my supervisor and supporting me throughout the creation of my thesis. His support, guidance and expertise helped me not to get lost in details. His keen insight always reminded me of the big picture. Many thanks to Martin Hagmüller for an open door and for many hours of discussion and constant guidance through my work.

This work was funded by HEIMOMED Heinze & Co. KG. They made it possible for me to work on this interesting field of research. My gratitude to Marcus Raguse and Robert Recknagel for the support and the regular meetings.

During my work I had the possibility to go to Nara Institute of Science and Technology in Japan. Thanks to Tomoki Toda for not only being my supervisor during my research stay abroad, but also for agreeing to be my examiner. Thanks for allowing me to use the SVC framework for conducting parts of my experiments. His perceptive and valuable comments and hints were precious to further improve my thesis.

Juan Cordovilla helped me with the ASR experiments and was a faithful companion during lunch breaks. I had the possibility to work together with very motivated and interested students who provided invaluable input for my research: Daniel Pellicer Garcia, Clemens Amon, Theresa Loss, Julia Ziegerhofer, Klaus Huber, Gabriel Hülser, Felix Rothmund, Linda Lüchtrath and Hartwig Klammer. A special thanks to Philipp Aichinger for many telephone conversations and discussions and to all the volunteers who partook in the numerous experiments throughout this thesis.

I enjoyed going to work every day. Thanks to my colleagues: Andreas Läßer for providing solutions for many practical problems; Markus Köberl for maintaining my working environment; my office fellows Wolfgang Truppe and Bernd Lässer for many coffee breaks; all other colleagues for discussions in the kitchen, many birthday cakes and for the motivating and inspiring environment.

Thanks to my precious friends: Siggi, for being a talented and inspiring companion; Christoph, for being a great steersman; Julia, for many years of friendship and proof-read pages. My biggest thank however goes to my siblings and family: Thomas, Much, Cle, Dani and Ursula. Thank you for unconditional love and support in every possible way, in good times and in difficult times. Last but not least, thank you Michi, for supporting me and for coming to Graz so many weekends when I preferred to go to work. Thank you, for willingly listen to my moans when things did not go as planned and sharing my joy in times of success. Thank you, for not letting me forget what the real important things are in life!

Contents

Acronyms

AC Alternating Current or Aperiodic Components.

ACD Aperiodic Component Distortion.

ASR Automatic Speech Recognition.

BDR Block Detection Ratio.

BEE Back End Error.

CCR Comparison Category Rating.

CELP Code-Excited Linear Prediction.

DC Direct Current.

DREL Directly Radiated Electro-Larynx Noise.

DREL-L Directly Radiated Noise of the Electro-Larynx Device Level.

DT Double Threshold.

DTW Dynamic Time Warping.

DUT Device Under Test.

EL Electro-Larynx.

ELHE EL + HE.

EMG Electromyography.

EMG-EL EL Device Controlled by Electromyography.

FE Feature Extraction.

FEE Front End Error.

FFT Fast Fourier Transform.

FN False Negative.

FP False Positive.

GMM Gaussian Mixture Model.

GV Global Variance.

HE Healthy Speech.

HGS Hanquinet-Grenez-Schoentgen.

HIL Hilbert Transform.

HMM Hidden Markov Model.

HPF High Pass Filter.

IIR Infinite Impulse Response.

Lar Laryngograph.

LF Liljencrants-Fant.

LP Low-Pass.

MAP Maximum a Posteriori.

MCD Mel-Cepstrum Distortion.

MF Modulation Filtering.

MFCC Mel Frequency Cepstral Coefficients.

MLLR Maximum Likelihood Linear Regression.

MOS Mean Opinion Score.

MSE Mean Square Error or Mid Speech Error.

NAM Non-Audible Murmur.

NDS Noise Detected as Speech.

NL Noise/Silence Level.

PCA Principal Component Analysis.

PE Pharyngoesophageal.

PESQ Perceptual Evaluation of Speech Quality.

PLP Perceptual Linear Prediction.

PSD Power Spectral Density.

PSOLA Pitch Synchronous Overlap and Add.

QOL Quality of Live.

RASTA Relative Spectral Transform.

RMS Root Mean Square.

RMSE Root Mean Square Error.

ROC Receiver Operating Characteristics.

SAMPA Speech Assessment Methods Phonetic Alphabet.

SE Sensitivity.

SL Speech Level.

SNR Signal-to-Noise Ratio.

SP Specificity.

SS Spectral Subtraction.

ST Single Threshold.

STFT Short-Time Fourier Transform.

SVC Statistical Voice Conversion.

TN True Negative.

TP True Positive.

UVV Unvoiced-to-Voiced Error.

VD Vowel Duration.

VIT Vowel Initiation Time.

VTT Vowel Termination Time.

VUV Voiced-to-Unvoiced Error.

Nomenclature

A Factor to influence spectral richness for *HGS* model.

B Induction of magnetic field.

$CI_{95\%}$ 95 %-confidence interval.

D Number of deletions.

E_e Absolute value of glottal flow derivative at t_e.

F Force.

F_L Lorentz force.

H Magnetic field strength.

I Number of insertions or Current.

K Number of components in GMM.

N Number of words.

$N_{\hat{f}_{0,incorr}}$ Number of incorrectly estimated f_0 values according to a threshold δ.

N_{f_0} Number of f_0 values.

S Number of substitutions or Cross-section of the air gap.

T_0 Fundamental period.

U Voltage.

W_{Acc} Word accuracy in [%].

Δ Dynamics of a feature in terms of its first or higher-order difference.

Φ Magnetic flux.

Σ (Full) covariance matrix of GMM.

α Weight of GV likelihood.

δ Threshold for gross and fine error calculation.

$\dot{U}_g(t)$ Glottal flow derivative.

ϵ Constant (slightly larger than 1) for SNR calculation.

γ Smoothing constant for SNR calculation.

$\hat{f}_{0,corr}$ Correctly estimated f_0 values according to a threshold δ.

$\hat{f}_{0,t}$ t-th value of estimated fundamental frequency contour.

λ Contains parameter of GMM $\{(b_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k); k = 1, 2, \ldots, K\}$.

$\mathbf{W}$ Matrix to extend static into joint static and dynamic vectors.

$\mathbf{X}$ Source feature vector sequence.

$\mathbf{X}_t$ Static and dynamic source feature vector at frame t.

$\mathbf{Y}$ Target feature vector sequence.

$\mathbf{Y}_t$ Static and dynamic target feature vector at frame t.

$\mathbf{Z}_t$ Joint feature vector of static and dynamic source and target features at frame t.

$\hat{\mathbf{y}}$ Converted feature sequence.

$\mathbf{c}_t$ Concatenated feature vector at frame t.

$\mathbf{v}(y)$ GV of the target static feature vector.

$\mathbf{x}_t$ Spectral feature vector of source speaker at frame t.

$\mathbf{y}_t$ Feature vector of target speaker t.

$\mathbf{z}_t$ Joint feature vector at frame t.

len Absolute length of sentences.

$sSNR$ Averaged segmental SNR.

μ Mean value or Permeability.

$\overline{len}$ Averaged length of sentences.

$\overline{n^2[k]}$ Short-term power of noise signal.

$\overline{s^2[k]}$ Short-term power of speech signal.

$\overline{y^2[k]}$ Short-term power of noisy speech signal.

ρ Correlation coefficient.

σ Standard deviation.

σ^2 Variance.

$\tilde{M}$ Harmonics of Fourier series.

$\mathit{\Theta}$ Phase.

a_m Fourier coefficients of Fourier series of m-th harmonic.

ac_{conv} d-th cepstral coefficient of converted speech of frame t.

ac_{tar} d-th cepstral coefficient of target speech of frame t.

b_k Weights of each component of GMM.

b_m Fourier coefficients of Fourier series of m-th harmonic.

f Frequency in [Hz].

f_s Sampling frequency.

f_0 Fundamental frequency.

l Length of current-carrying conductor.

mc_{conv} d-th cepstral coefficient of converted speech of frame t.

mc_{tar} d-th cepstral coefficient of target speech of frame t.

$n[k]$ k-th value of noise signal.

$s[k]$ k-th value of clean speech signal.

t_a Time of exponential closure.

t_e Instant of flow derivative negative peak.

t_p Instant of maximum glottal flow.

thr Iterative changing threshold for SNR calculation.

$thr_{d,off}$ Double threshold (termination).

$thr_{d,on}$ Double threshold (activation).

$thr_{s,on-off}$ Single threshold (activation and termination).

v Velocity.

x_0 Distance between diaphragm and magnet.

$x_\sim$ Displacement of diaphragm.

$y[k]$ k-th value of noisy speech signal.

$\mathbf{X}_t'$ Segmental feature at frame t.

$\Phi_=$ Magnetic flux of AC field.

$\Phi_\sim$ Magnetic flux DC field.

$\hat{f}_{0,incorr}$ Incorrectly estimated f_0 values according to a threshold δ.

1 Introduction

1.1 Technical Background and Problem Definition

Human speech production is the result of an interaction of the lungs, the trachea, the glottis and the vocal tract (nasal and oral cavity). The main parts involved in speech production are illustrated in Figure 1.1.

An air stream (pulmonary pressure) coming from the lungs travels through the trachea until it reaches the larynx including the glottis. The glottis is the gap between the vocal folds. It is influenced by the muscles controlling the vocal folds as well as the airstream. Depending on the position of the muscles the vocal folds vibrate periodically or build a canal through which the air travels. This means that the air stream is either interrupted periodically or swirled before traveling to the vocal tract. There it is modulated according to the position of the articulators. Speech can be classified into voiced and unvoiced sounds. These properties describe the level of vibration of the vocal folds. Vowels like [a], [e], [i], [o], [u] are voiced, plosives [p], [t],... or fricatives [ʃ],... are unvoiced sounds. The acoustic wave is filtered by the vocal tract and

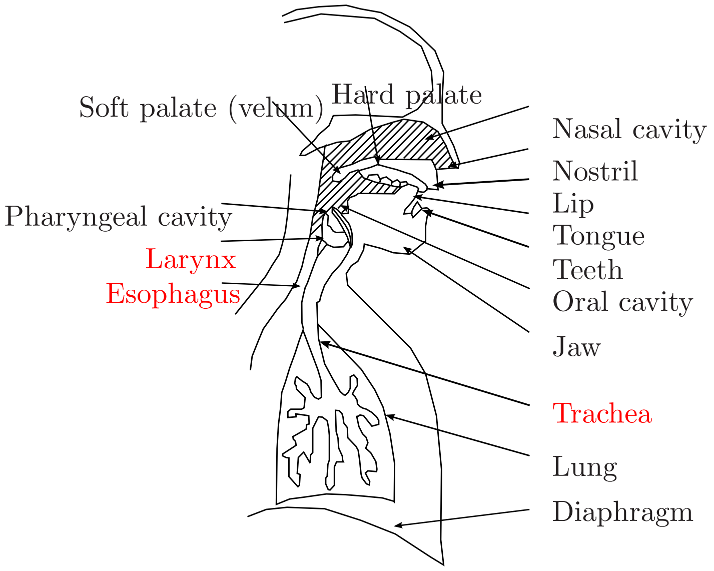


Figure 1.1: Anatomy of a human being: speech production (inspired by [Shanxi Datong University, 2015]).

radiated from the lips resulting in audible and comprehensible speech. Mathematically the speaking process can be described using the source-filter model (Figure 1.2). The main source signal is the glottal excitation signal and the filter is the vocal tract with its resonance frequencies depending on the position of the articulators.

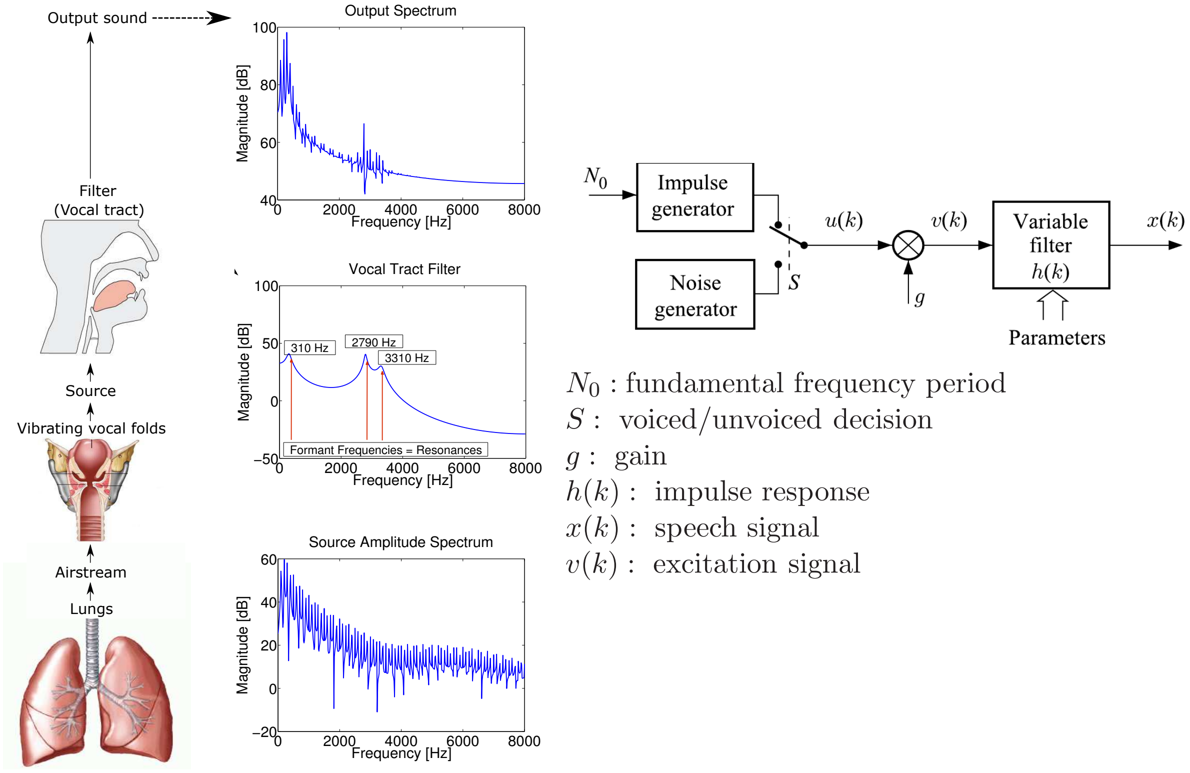


Figure 1.2: Human speech production (right part from [Vary and Martin, 2006]).

The abuse of alcohol and tobacco products is the main reason for cancer of the larynx. For people who suffer from laryngeal cancer or similar diseases, a total laryngectomy, which means the complete removal of the larynx (see Figure 1.3 – (a) and (b)) is inevitable in many cases. If the larynx is removed surgically, the anatomy is changed dramatically. The trachea ends at the so called tracheostoma at the neck. As a result, the vocal tract is shortened. Therefore, it has not the same structure than a healthy vocal tract anymore. According to the source-filter model, the larynx and the vocal folds can be seen as the source of sound energy and the vocal tract as the filter. If people lose their larynx they will consequently also lose their ability to speak.

Currently there are around 20 000 – 25 000 people who underwent a laryngectomy in Germany. Annually around 2 000 laryngectomies [Schiefer and Hagen, 2000] take place in Germany. These numbers also reflect the situation in Austria (in a scaled manner). There are three alternatives for people to re-obtain their speech. The first method is the *esophageal voice*. Within this method air is firstly gulped and then released in a controlled manner. Instead of the vocal folds, the tissue of the pharyngoesophageal (PE) segment in the pharynx vibrates. The second method is the *tracheo-esophageal voice*, where a shunt valve is placed between trachea and esophagus. Due to the shunt valve, speech can be generated with the air coming from the lungs (see Figure 1.3 (c)). The third method is the transcutaneous *Electro-Larynx device* (EL) (see Figure 1.4). This is a small, hand-held and battery driven device. The cross-section of a conventional device can be seen in Figure 1.4(a). The vibrating coupler disk of the device is held against the neck. The signal of the coupler disk is carried into the vocal tract and filters the signal in a similar way as in healthy speech production. Another non-electric artificial speaking device is the pneumatic device, e.g., the Tokyo Artificial Larynx [Welch, 2015]. It is a tube which

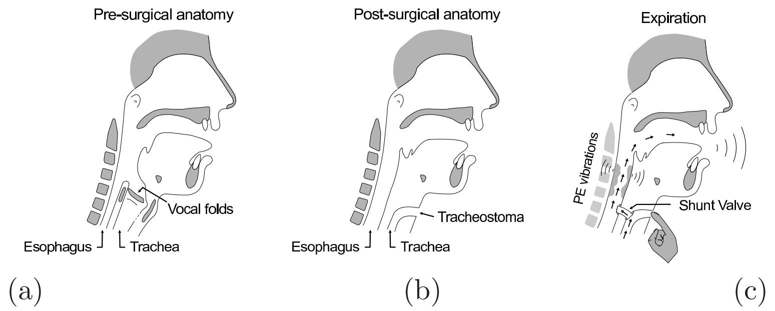
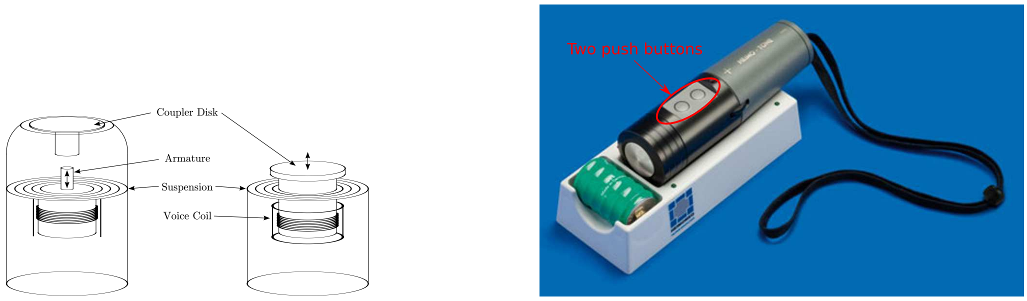


Figure 1.3: Schematics of pre-surgical anatomy (a) , post-surgical anatomy (b) and tracheo-esophageal voice production (c) (from [Lohscheller, 2003]).

connects the tracheostoma with the mouth of a speaker. In the middle of the tube there is a rubber diaphragm which will start to vibrate when air is transmitted from the lungs through the tracheostoma and thus, through the tube.



(a) Cross-section of conventional EL device.

(b) Conventional EL device.

Figure 1.4: Cross-section of a conventional Electro-Larynx (EL) device and commercially available device from company Heimomed [Heimomed, 2015].

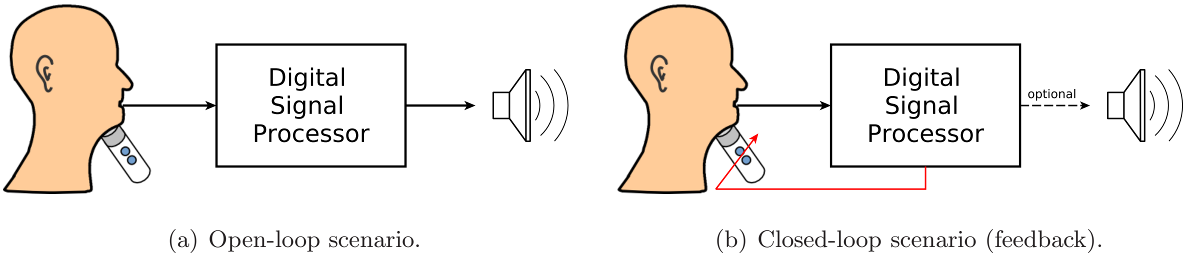
The disadvantage of all these kinds of substitution voices is the poor quality of the resulting speech. The preferred substitution voice depends on the user, his/her ability to learn how to use the substitution voice and other factors such as degree of operation (what and how much of the affected muscles and tissue needed to be removed) or the length and type of radiation therapy [Rosso et al., 2012]. In western Europe the speech valve is the primary method of speech rehabilitation. In the US more than 50 % of laryngectomees are using the EL [Cox and Doyle, 2014].

The esophageal (ructus) voice is particularly difficult to learn. Furthermore, the number of words which can be produced in a row is limited depending on the amount of gulped air. As a result, the temporal structure of talking is unnatural. Most of all, speaking sounds like belching, which can be embarrassing for the speaker. The tracheo-esophageal voice results in a good quality of substitution voice, but the necessary operation is highly invasive and can cause additional health problems: in the course of time bio-film grows on the shunt valve. Therefore, the shunt valve needs to be replaced every other month. The replacement is equivalent to a small operation.

According to [Meltzner and Hillman, 2005] the major drawbacks of EL speech are 1. the directly radiated electro-larynx noise (DREL) of the device itself, 2. the unnatural, monotonous quality of speech and 3. the need of one hand to operate the device. DREL reduces intelligibility and disturbs the speech quality. Regarding device operation, it is inconvenient to use one hand to operate the device, but otherwise this aspect does not have any effect on neither speech quality nor intelligibility. Although the use of an EL is limited due to the condition of the user's neck muscle and tissue after the radiotherapy/chemotherapy and the user's ability and willingness to adopt available technologies, the EL device is, to a certain amount, easy to use and is often the first post operative method to re-obtain speech. Therefore, this work will focus on this kind of substitution voice.

All over the world there are different devices available. The Servox Digital from Servona [Servona, 2015] and Heimo-TONE [Heimomed, 2015] are market leaders in Europe, the TruTone from Griffin laboratories [Griffin, 2015] is very popular in the US and in Asia the MyVoice from SECOM [SECOM, 2015] and the YourTone from Densei [YourTone, 2015] are the leading products. All products are very similar in terms of design and technology. They use a push-button which serves as the user interface. Then, the artificial excitation signal is generated using an oscillator, which generates a square wave with fixed frequency and amplitude. The artificial excitation signal of the Servox Digital is a train of impulses alternating in sign. Afterwards, the excitation signal drives a moving coil which hammers against the coupler disk. The coupler disk, in turn, is held against the neck. For all products it is possible to program the device's fundamental frequency. The TruTone has the feature of a pressure sensitive device's button with which the fundamental frequency can be controlled manually. The Servox Digital offers two push-buttons for two different preset frequencies (see Figure 1.4(b)). Both of them can be controlled individually so as to apply a coarse prosody.

An important aspect is to name the target application of the EL speech improvement. In general two kinds of setups can be distinguished (see Figure 1.5): 1. An **open-loop** where EL speech is recorded, processed and presented using loudspeakers or telecommunication applications; 2. A **closed-loop** (**feedback method**) where EL speech is recorded, processed and then an artificial excitation signal is generated to operate the EL device. These two approaches have different possibilities, can use different techniques and will, for sure, obtain different results. As we would like to get as close as possible to natural speaking, we focus on the closed-loop method. It must be mentioned that most of our experiments are based on recorded speech data. Nevertheless, the approaches and algorithms are designed in a way to work in a closed feedback system. Furthermore, it must be noted that our aim is EL speech improvement, which includes speech enhancement [Vary and Martin, 2006] (e.g., removing DREL of EL speech) as well as intelligibility enhancement [Sauert and Vary, 2006] [Kleijn et al., 2015] (e.g., improving intelligibility of speech presented in background noise). To improve EL speech we can influence two things: the artificial excitation signal, e.g., the waveform and frequency of the signal; and/or the device itself, e.g., the electro-mechanical transducer. In this thesis we would like to overcome known problems and present a new generation of EL devices, the *bionic EL speech system*, which is capable of decreasing the gap between mechanical devices and the maintained biological anatomy.



(a) Open-loop scenario.

(b) Closed-loop scenario (feedback).

Figure 1.5: Two different kinds of setup in this thesis.

Why is the proposed EL bionic?

There are two well known areas: biomimicry (or biomimetic) [Lakhtakia and Martín-Palma, 2013] and bionics [Nachtigall, 2002]. The first one describes and adopts phenomena from nature to technologies, whereas the latter deals with merging biological and technical properties and performances. The best known examples for biomimicry are airplanes which are inspired by flight characteristics of birds or Velcros which are inspired by burdocks. In terms of bionics, which is compounded by the words **bio**logy and tech**nic**, cyborgs (people who are half human, half machine) are a popular example. Existing technologies include artificial organs (cochlear implants, bionic eyes, portable pancreas) and artificial limbs (prosthetic arms and feet). It is even possible to replace parts of the brain with a chip that learned the mapping between input and output of the hippocampus which was one of the breakthrough technologies in 2013 [MIT, 2015]. The definition of bionic is according to [Merriam-Webster, 2015]: *"having normal biological capability or performance enhanced by or as if by electronic or electromechanical devices"*.

We would like to add the adjective bionic to our newly developed EL because we improve biologic structure using an electronic device. Compared to related work in EL speech improvement, we tackle the problem as a whole and not only certain aspects. Therefore, the whole EL system, including all single improvement steps, will be called bionic EL speech system. In future we also would like to work on the field of biomimicry in order to transfer the knowledge of natural movement of the vocal folds to the excitation signal of the bionic EL. We would like to learn from healthy vocalization and approximate the excitation signal of the bionic EL in order to sound more intelligible and natural.

1.2 Overview on EL Speech Improvement

The aim of this thesis is to process EL speech in order to obtain more natural speech, similar to healthy (HE) speech. In order to work on the differences and problems, EL speech and HE speech need to be compared and analyzed. For visual inspection, HE speech and EL speech with and without DREL suppression in the time and frequency domains are shown in Figure 1.6 and 1.7. For DREL suppression we used the modulation filtering approach proposed by [Hagmüller, 2009]. For more details concerning this approach see Section 2.2.2. Only if the drawbacks of EL speech are known, we can focus on removing these problems.

Figure 1.6 shows the fact that speaking with the EL device takes much longer than speaking with HE voice. The EL user needs to articulate the words and syllables very carefully in order to produce intelligible speech. This can be seen in the example sentence of Figure 1.6, where the duration of the HE sentence is 3 s, whereas it takes 4.5 s to utter the same sentence using an EL device. In this example EL speech results from a healthy person who simulates EL speech. We also analyzed sentences spoken by a laryngectomee and found out that the average sentence length is closer to the average HE speech sentences. Figure 1.7 shows the corresponding spectrograms of the spoken sentences. The fundamental frequency f_0 of HE speech is changing. In EL speech f_0 is constant according to the constant excitation signal of the device. This is clearly visible in Figure 1.7(b). We formulate the hypothesis that a changing f_0 contour of EL speech leads to an improvement of the quality of speech. With a changing contour the EL spectrogram will get closer to the HE speech spectrogram. Consequently, the resulting speech will sound more natural. Operating the EL device often means that there is no differentiation between voiced and unvoiced phonemes because the EL device is not constantly turned on and off while speaking. On the one hand, switching all the time would annoy listeners, but on the other hand the constant voiced sound leads to confusions especially between word initial voiced and unvoiced stop consonants [Weiss et al., 1979].

The sound producing mechanism and the quality of the resulting speech has hardly improved since the EL was first introduced in the 1960s. Prior studies investigated the problems of EL speech which include: improper source spectrum such as the reduction of low frequency energy, lack of fine control over f_0, amplitude and voice on/offset (i.e., constant f_0; no variation in the harmonic structure), interference of DREL, reflections in the vocal tract due to the changed

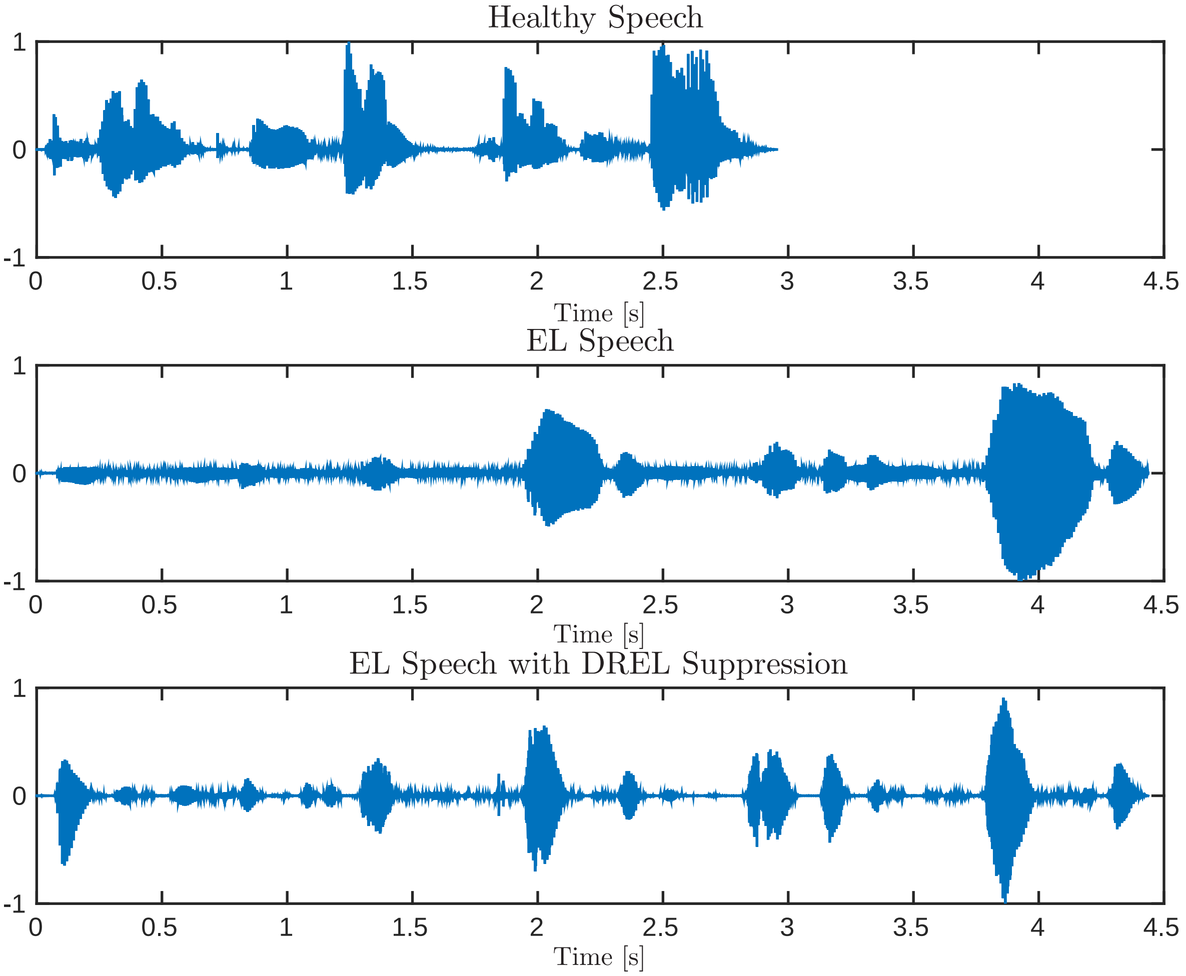


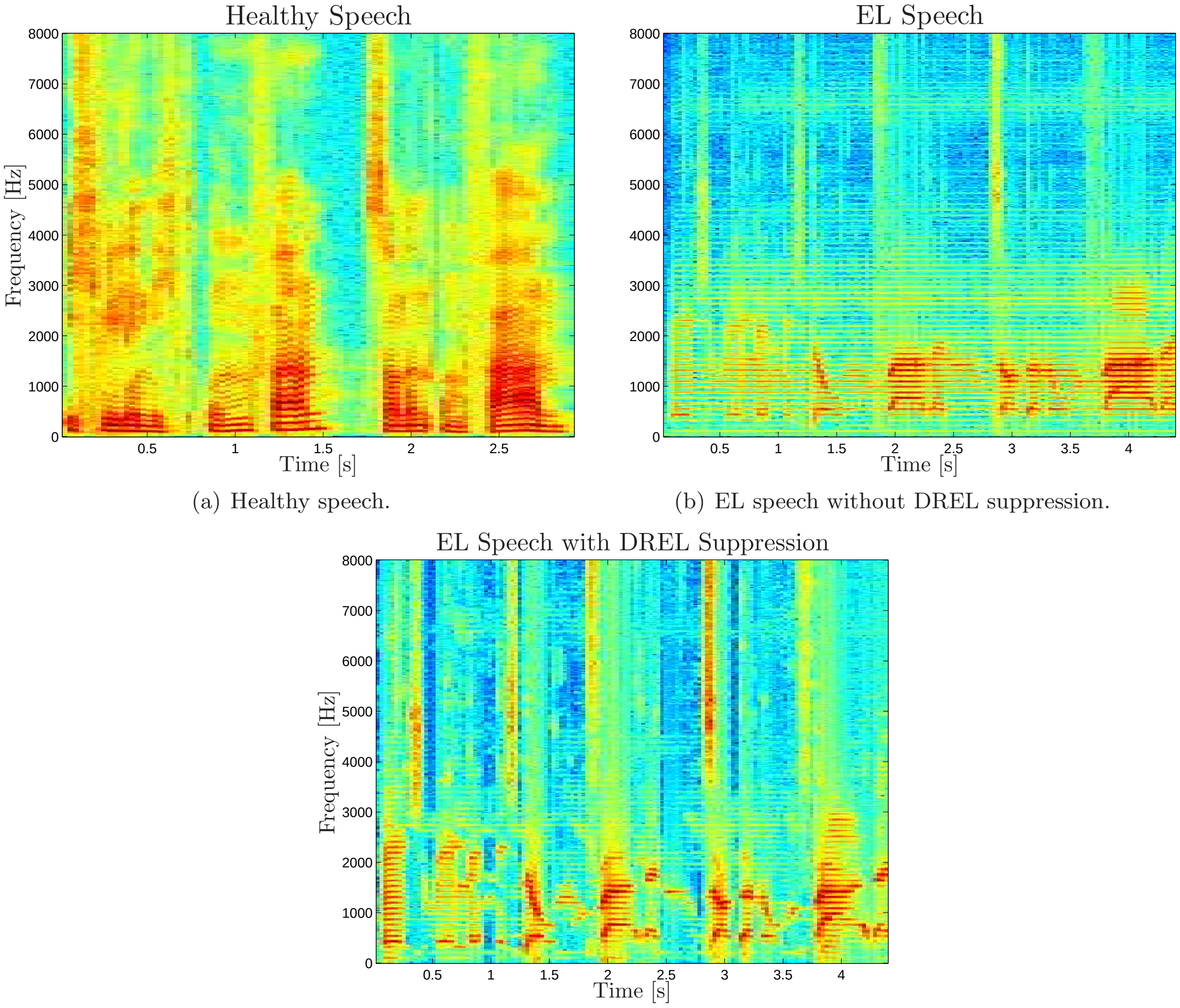
Figure 1.6: Speech signal in time domain; Male speaker: "Ich will ihn nicht umfahren, sondern umfahren." translated: " I do not want to knock him over but to go around him."; The semantic is entirely changed when changing the main accent from first to second syllable of the word "umfahren".

anatomy, the resulting reduced intelligibility concerning confusion between voiced and unvoiced consonants as well as vowel intelligibility.

The investigation of [Ng, 1996] emphasized the differences between HE and EL speech for male speakers: 10 laryngeal, 10 esophageal and 10 electro-laryngeal were compared in terms of average f_0, identification of the six Cantonese tones, f_0 contour of the six tones, speech intensity level of the six tones and average vowel duration values of the six tones. The results showed that, during average reading, f_0 values of esophageal speakers were higher than the values of laryngeal and electro-laryngeal speakers of Cantonese and esophageal and laryngeal speakers of English. Identification of the six tones was best for laryngeal, followed by esophageal and electro-laryngeal speech. Speech intensity was highest for laryngeal and average vowel duration was longer for laryngeal speakers than the other two.

[Meltzner, 2003] reported that monotonous speech is, to a certain amount, more intelligible than EL speech with varying f_0 but reinforces the unnatural characteristic of the speech. Additionally, changes of the vocal tract anatomy produce narrower formant bandwidths and spectral zeros (dependent on EL position) altering the spectral properties of EL speech.

[White, 1994] investigated acoustical and perceptual characteristics of electronic artificial larynges in more detail. In his thesis he extracted the volume velocity waveform for healthy subjects and for five different EL devices. A modified version of the Klatt formant synthesizer was used to synthesize generated HE and EL source waveforms and construct a computerized model of an EL. With this model EL speech can be produced synthetically using different excitation signals. In a listening experiment synthetic EL syllable stimuli were presented, their



(a) Healthy speech.

(b) EL speech without DREL suppression.

(c) EL speech with DREL suppression.

Figure 1.7: Speech signal in frequency domain; Male speaker: "Ich will ihn nicht umfahren, sondern umfahren." translated: " I do not want to knock him over but to go around him."; The semantic is entirely changed when changing the main accent from first to second syllable of the word "umfahren"

pulse parameters were manipulated and listener preference was measured. The final goal was to find out how these parameters can be modified and incorporated into EL hardware in order to improve naturalness. It was found out that the open quotient of the glottal volume velocity is different between HE and EL excitation pulses. The EL parameters impacting the perceived voice quality were identified to be f_0 variation and short duration open quotients. However, it is important to note that these conclusions were drawn from synthetic EL speech in a noise free condition without reverberation. In normal listening environments, noise as well as reverberation interferes with speech perception and can shift listener's perceptual judgment due to the masking effect which limits acoustical information.

Research shows that there is a relation between f_0 and intelligibility as well as f_0 and perceived gender. [Nagle et al., 2012] investigated the differences using listening tests with which speech samples from 34 healthy adults using EL devices set at 75 Hz, 130 Hz, and 175 Hz were evaluated. The results show that the intelligibility is highest for an f_0 of 75 Hz. Furthermore, the wrong gender was recognized in a mismatched case (male – 175 Hz; female – 75 Hz). This means that best intelligibility can be reached for low f_0 values which is a contradiction when appropriate gender solutions are to be designed. It is important to find a solution, where EL speech is as intelligible as possible and, at the same time, fulfills the claims of the speakers gender.

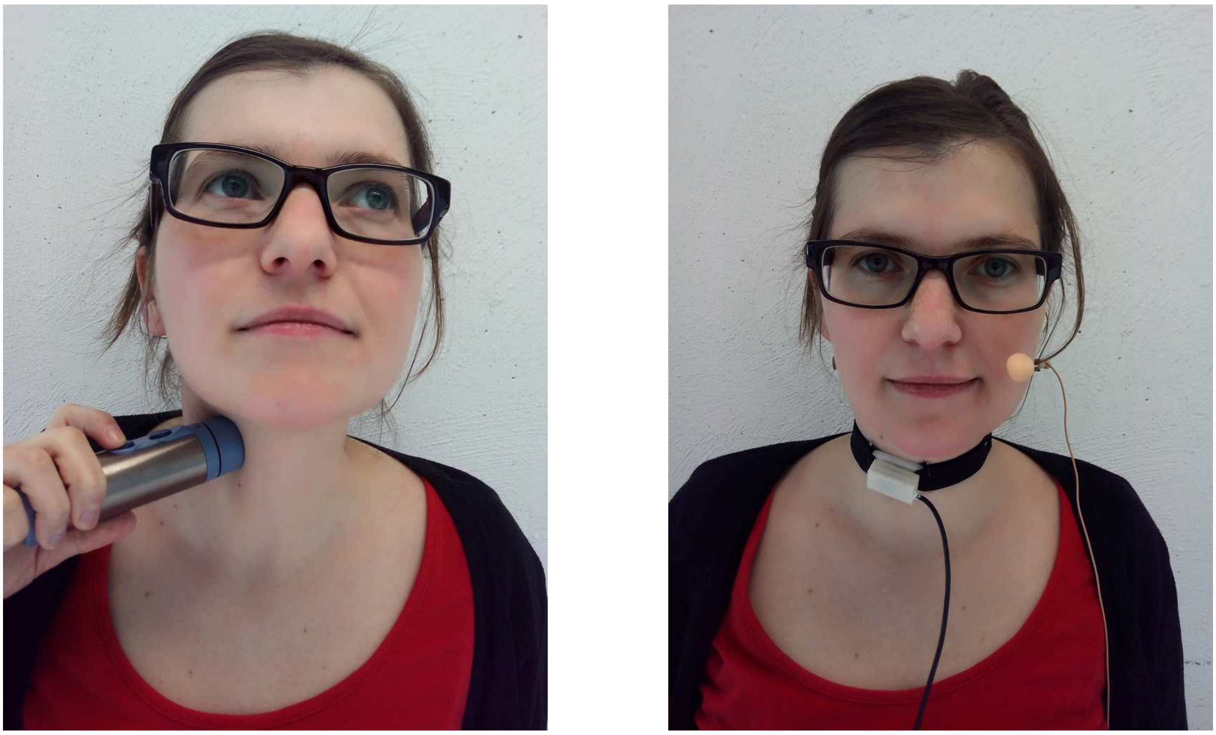
Furthermore, the relation between intelligibility and acceptance needs to be considered. Speaker identity is an important aspect to consider, as speech conveys a variety of information. This information can be linguistic and para-linguistic. It is known that para-linguistic information is the most important factor for speaker identity. Linguistic cues involve dynamic manipulation of articulations (variation due to language, dialect, etc.), whereas para-linguistic cues can be separated into source characteristics (laryngeal anatomy and physiology, f_0 and its dynamics, etc.) and filter characteristics (anatomy and physiology of the pharyngeal, oral, and nasal cavities constrain the range of resonances, ...). The anatomy of the vocal tract affects the spectral content of the resulting speech. Sociological factors may influence the speaking style (e.g., the prosodic factors) such as f_0, breathiness of the speech, duration of syllables, words, and pauses, etc. The most important acoustic features characterizing speaker individuality include the third and the fourth formant, f_0 and the closing phase of the glottal wave. [Brown and Feinstein, 1977] reported that listeners can still distinguish male and female users based on formant spacing. The work of [Perrachione et al., 2014] dealt with the process of identification of a speaker using the EL device in order to separate source and filter characteristics. They conclude that individuals are uniquely identifiable from EL speech, but training talker identification on one source mechanism (HE speech) does not generalize to the speech from the other mechanism (EL speech). There is no reason to assume that female speakers are not perceived as female as long as f_0 is high enough. [Coleman, 1976] examined the speaker identity within EL speech. In his work two experiments were carried out and he could confirm that 1. The correlation between f_0 and the degree of maleness and femaleness in the voice is high, whereas the correlation between vocal tract resonances and degree of maleness and femaleness of the voice is less strong; and 2. In experiments where vocal tract resonances and f_0 are investigated using an EL device, female f_0 was a weak indicator for female voice when combined with male vocal tract resonance and male f_0 is a strong indicator for male voice. One of the suggestions in this work is that it is easier to produce a more natural sounding male voice with an EL device than a female voice. This means that speech from a female vocal tract with low f_0 is likely to be misjudged as male and that the f_0 cue dominates in such a mixed case. It proves that a female vocal tract and high f_0 is judged as female and a male vocal tract and low f_0 is judged as male. [Nagle et al., 2012] showed that female speech with low f_0 was more penalized and that listeners were unable to identify female EL speakers as female when they used an EL device with low f_0. There are still open questions concerning the differences in speaking between males and females and how this is related to intelligibility. We will further cope with these aspects in the outlook and in a follow-up project of this work.

Furthermore, speech intelligibility and acceptability depend on the age of the listener group [Law et al., 2009]. While for younger listeners alaryngeal speech is more intelligible than for older listeners, the level of acceptance is lower compared to older listeners. In Japan the pneumatic device is more popular than in Western countries. This kind of alaryngeal substitution voice receives best intelligibility ratings, but laryngectomees did not perceive themselves as having the best communication-related quality-of-live (QOL). Tracheo-esophageal speech has been found out to perform best in communication-related QOL, whereas EL speech was perceived the poorest. Concerning lexical tone intelligibility: the performance of the pneumatic device and tracheo-esophageal turned out to be the best choice. Recent studies conducted by [Cox and Doyle, 2014] claim that prior studies may not accurately reflect the potential value of the EL device because there was a substantial variability in self-perceived QOL based on physical and social-emotional factors.

This work wants to investigate all facets of the complex EL speech production system meaning that problems related to the device itself (hardware) as well as the structure of the artificial excitation signal need to be addressed. Our goal is to build a real-time system which is able to produce more natural speech compared to the conventional system. Our motivation to reach this aim is to prevent people from social isolation. People around a laryngectomee are probably used to the sound of the device and the resulting speech, but talking to strangers is often an insurmountable obstacle for the laryngectomees. Two properties are fundamental in terms of quality: naturalness and intelligibility. Speech can be natural but not intelligible and vice versa, robotic voice is intelligible, but all people would agree that it is the opposite of natural.

The task of EL speech improvement is overlapping with related tasks in speech processing. For example, the techniques for reconstructing natural speech for laryngectomee can also be used to restore natural pitched speech in telephone communication when one party talks in a whispering mode for privacy or security reasons. This is addressed in the thesis of Sharifzadeh [Sharifzadeh, 2011]. Another approach to tackle this problem is to use electromyography based speech recognition [Schultz and Wand, 2010] [Meltzner et al., 2011]. The main field of application of the electro-larynx is, of course, speech rehabilitation after laryngectomy. However, the EL device was also found useful for orally intubated patients [Girbes and Elbers, 2014]. The author reports that a patient underwent a video-assisted bilobectomy of the right lung and due to complications, mechanical ventilation needed to be continued. The patient *"consented to the plan to use the electrolarynx, and to his surprise — and ours — the device immediately returned the gift of speech to him, without the passage of air through the vocal cords"*. This is less surprising for us, but nevertheless it is another field of application motivating us even more to improve the quality of the resulting speech.

Previous attempts to revolutionize the artificial speech aid did not succeed in creating a new electro-larynx device. This is due to the fact that it is important not to focus on one specific problem but to look at the system as a whole and investigate how different parts are influencing each other. Creating a new overall system involves several quite distinct work packages, like electro-acoustic aspects, prototype design, or signal processing. What we learned from previous work on this topic, and also during our project is that it is not enough to include a changing f_0 or to find a way to optimally suppress DREL of the device. Many small single steps do not change the existing system, but many small connected steps might yield a solution. Of course, there is not one best way to find solutions to the existing problems because one solution might be good for one problem but inadequate for another one. The first big decision to make is whether to use a closed-loop or an open-loop approach. Both approaches can use different methods to solve the same problems. In this work we vindicate the point of view that the speech production of a laryngectomee is as natural as possible which means that we would like to stick to a closed-loop approach. Regarding loudspeaker playback: this can be used in closed-loop as well as in open-loop approaches and resulting speech might benefit from the increased volume. For telecommunication on the other hand, an open-loop solution might be a better choice. In the figures below (Figure 1.8) we show the conventional and the proposed system and in the thesis we explain the proposed changes and improvements and highlight our considerations and suggestions for the design. In the course of this thesis we developed and investigated concepts which are very different from the existing device.



(a) Conventional EL device: Servox Digital (Servona).

(b) Proposed EL speech system.

Figure 1.8: Conventional EL devices and proposed EL speech system.

1.3 Milestones in the Literature

According to [Shute, 2003] EL users are not satisfied with the quality of EL speech at all. Most of all a major improvement for telephony is needed. Furthermore, the volume is often a problem. The research group in Boston investigated major drawbacks and differences of EL speech. According to them the poor quality of EL speech is a result of the limited performance of conventional EL devices, the loss of the fine control of pitch, amplitude, and voice onset and offset timing. They confirmed a deficit in voice-related segmental (e.g., voiced-unvoiced distinctions for consonants) and supra-segmental (e.g., intonation, syllabic stress) speech parameters. In the thesis of [Meltzner, 2003] listening tests, acoustic analysis and acoustic modeling was carried out to investigate the properties of EL speech (perceptual as well as acoustical). 10 listeners judged the addition of pitch information to be the most important benefit. Removing DREL and correction for a lack of low frequency energy would also improve the speech.

When talking about EL speech improvement there are different parts for fine-tuning. Thinking back to the source-filter model of speech production the source is changed completely, whereas the filter is influenced by the speaker him- or herself and remains approximately unimpaired. Therefore, EL speech improvement methods based on signal processing approaches focus on the artificial excitation signal. We can categorize the different techniques for EL speech improvement as follows:

A) methods for generating and manipulating the artificial excitation signals

B) removing the directly radiated noise of EL device itself

C) changing the artificial excitation signal producing element itself

For A) only the artificial excitation signal will be synthesized (open-loop), whereas in B) we can use existing speech synthesis and voice conversion methods (closed-loop). C) is beneficial for both, open-loop and closed-loop approaches. The aims of improvement strategies are to make EL speech

- more intelligible
- more natural
- more individual
- more acceptable
- easy to use
- wearable, not only portable.

A summary of the literature is described in the following.

A) Approaches for a Hands-Free Device, Fundamental Frequency f_0 Control and Artificial Excitation Signal

The attempts to propose a hands-free device and to control f_0 are often overlapping. A new generation of EL devices should be as simple as possible. Therefore, the sensors used should be able to perform several tasks at once. Several surveys confirm that, besides the unnatural quality of the resulting EL speech, the handling of the device would benefit from a hands-free design.

[Goldstein et al., 2004] developed a neural interface for an automatic on/off control of the EL device using electromyographic (EMG) signals from neck strap muscles. Reaction time experiments with an EMG controlled EL device (EMG-EL), conventional EL device and tracheo-esophageal were conducted and compared to healthy voice. It turned out that: voice initiation was faster for EMG-EL than for healthy voice; voice initiation was faster for EMG-EL than for conventional EL for a healthy subject; voice initiation was faster for EMG-EL than for tracheo-esophageal for a laryngectomee; voice initiation was slower for EMG-EL than for conventional EL for a laryngectomee; voice termination was slowest using the EMG-EL. The authors could show that EMG is an appropriate signal to switch on/off an EL device.

[Kikuchi and Kasuya, 2004] designed a four-directional switch which is able to control on/off and f_0 at the same time. Up and down movement of the arm turns the device on and off, whereas left and right movement regulates pitch control. The approach was implemented and tested on five female healthy students speaking 2 sentences. In the first experiment pitch control was investigated: Four of the speakers concluded that pitch control in terms of absolute magnitude and pitch variation range is possible. In the second experiment pitch control as well as on/off control was performed simultaneously: The control went well although pitch adjustment was more difficult.

The idea of an intra-oral transducer was introduced by [Takahashi et al., 2005]. The tiny electro-magnetic transducer generates an excitation signal inside of the oral cavity. A fingertip switch was responsible for fundamental frequency control. In a follow-up, [Takahashi et al., 2008] extended the intra-oral transducer to an intra-oral pressure sensor in order to distinguish between voiced and unvoiced consonants. Suspending the transducer during unvoiced consonants resulted in an improved intelligibility. There work's conclusion is that the intra-oral pressure sensor can serve as an automatic intention detector for vibration.

[Hashiba et al., 2007] used a thermo-plastic brace on which a thin transducer was mounted to develop a hands-free version. The wireless on/off switch consists of a small fingertip push-button switch and a wireless transmitter set designed as a wristband. Good results were reported when performing usability and intelligibility tests. However, the brace is sensitive to movements of the head, which is very cumbersome. In a next step the f_0 control mechanism that allows laryngectomees to control voice intonation using their exhalation from [Uemi et al., 1994] will be implemented.

Also [Pineda-Rico et al., 2008] picked up the EMG based on/off control. A switching capacitor CMOS based device was implemented. For activation and termination the same method as in [Heaton et al., 2011] was used: amplified, rectified and low-pass filtered ($c_f = 3$ Hz) envelope and single threshold implemented as voltage comparator. The focus was on the implementation and on the advantages of switching capacitor circuits which are: excellent time constants, relative precision, simple design elements, minimum power waste and reduced size on chip.

The idea of using a brace was the base for the doctoral thesis of Madden [Madden, 2013]. In this thesis a complete new system was built instead of focusing on the development of one single facet of the existing EL. A new transducer was mounted on a neck brace. An accelerometer placed on the chest and abdomen captured breathing and seemed to be a promising solution for a hands-free control. Although it is known that speech and breathing is decoupled for EL speakers the connection can be retrained again. Additionally the pitch can be controlled using hand gestures (i.e., by an accelerometer attached to a person's hand). This work is inspiring and confirms our approach to work on a whole solution for EL speech improvement and not only on specific topics.

A very different approach in terms of excitation signal was proposed by [McLoughlin, 2014] who used an ultrasonic chirp to detect voice activity. This idea was recently published and investigated voice activity detection in noise. Within this approach a wide-band ultrasonic excitation signal is emitted from near the lips which is reflected from the mouth region. The resonance patterns of the reflected excitation signal give information about lip opening/closing. Evaluation showed that, the system is relatively insensitive to sensor placement and highly insensitive to background noise.

Off-line methods like voice conversion or whisper-to-speech conversion do not need to deal with the problem of an occupied hand because improved EL speech is output via loudspeakers.

[van Rossum et al., 2002] investigated how tracheo-esophageal and esophageal speakers are able to convey accent. 10 laryngeal, 10 tracheo-esophageal and 9 esophageal speakers participated in the study. English and Dutch are stress-accent languages, which means that important information is placed on the syllables in the word that is stressed. Used cues for stress-accent are speaker-dependent. Peak intensity was used consistently. For esophageal speakers the rhythm of the speech was unnatural. Therefore, listeners had difficulties using temporal information as a cue to accent and esophageal speakers performed worse. Alaryngeal speakers had difficulties with f_0, but nevertheless they were able to articulate an accent. There was unexpected concentrated acoustic energy in the formant frequencies. A surprising finding of this work is that people who are not able to use f_0 do not compensate using non-melodic cues (peak intensity, spectral tilt, pause, word duration). It remains uncertain, if it is only easy to control other cues when it is possible to control f_0. A pitch perception experiment confirmed that tracheo-esophageal and esophageal speakers who had no control over f_0 were, nevertheless, able to produce pitch movements. [Giet, 1956] described that it is possible to hear movement of f_0 within whispered utterances. [Meyer-Eppler, 1957] investigated whispering in tone languages. Chinese language, although being a tonal language, can still be understood if whispered. The author concluded that two substitutes exist for periodic pitch movement, as in some vowels, gaps in the higher frequencies were filled with noisy components and in other vowels, formants shifted upwards. Thus, in laryngeal speech, pitch movements can be produced independently of f_0. The influence of the electronic device is not able to be correlated with these findings. The peak intensity cue cannot be controlled. Pauses and word durations remain as dominant cues because spectral tilt is suppressed by the electronic component.

The approaches for generating an artificial f_0 contour which have been proposed so far either require manual interaction to change f_0 or provide some predefined f_0 contours only. There are some commercially available approaches such as using a switch from the standard f_0 level to a different level to mark accentuation. A more flexible approach is to use a pressure sensitive button, which allows a continuous f_0 contour to be produced [Griffin, 1998].

[Uemi et al., 1994] used an air-pressure sensor that is put on the tracheostoma with the result that f_0 can be regulated using lung pressure. They showed that laryngectomees are able to control the fundamental frequency using expiration after a short period of training. They developed a transform function, which transforms expiration pressure in Pascal into Hertz and as a result were able to report improved naturalness. Another possibility is to create an artificial f_0 contour from the speech energy envelope [Loscos and Bonada, 2006]. The energy envelope is one possibility to convey accentuation if no f_0 is available. In this paper, the energy contour had been scaled and offset with the result that the transformed contour matched the average f_0 and the dynamic range of the speaker. This did not produce a linguistically correct intonation, but it did give useful f_0.

[Saikachi, 2009] developed and perceptually evaluated f_0 control in EL speech. The proposed algorithm estimated a more natural f_0 using the root mean square amplitude (RMS): $f_{0,est} = k1 + k2 \cdot \mathrm{RMS}$. The prosodic control, in f_0 modified EL speech, was examined concerning contrastive stress (OBJect vs. objECT; BLACKboard vs. black BOARD) and sentence mode (question and statement). In the conducted listening test stimuli were synthesized using the Klatt formant synthesizer. In amplitude-based f_0 estimation, perception of contrastive stress is improved and perception of sentence mode is degraded. Naturalness was improved, but prosodic structures were difficult to control due to the high number of free parameters. However, in this work only limited speech material (only a few sentences) was used. There was more acoustic energy in [a] than in [i] which leads to problems for amplitude-based f_0 estimation.

One chapter in the dissertation of [Hagmüller, 2009] is dedicated to prosody for alaryngeal speech, e.g., the influence of f_0. f_0 is estimated using the formants because formants are influenced by prosody. The author concluded that artificial changing of f_0 could not improve the perception of contrastive stress or sentence mode, but the overall subjective quality was improved significantly.

More recent work was carried out by Nakamura *et al.* [Nakamura, 2010] who also used the air-pressure sensor to improve f_0 estimation before enhancing EL speech using a statistical voice conversion technique. This work was extended by [Tanaka et al., 2014b] who implemented a hybrid approach to improve naturalness of voice conversion results. f_0 was estimated using a Gaussian mixture model (GMM) based voice conversion approach and correlation coefficients of around 0.58 were reported. In [Tanaka et al., 2014a] the approach was extended to a direct control of f_0 of the EL device. The processing delay of f_0 prediction was investigated, but no significant differences were found. In [Tanaka et al., 2014c] the authors extended previous work to multiple speakers (2 laryngectomees, 1 healthy subject) and the direct control of f_0 was compared to baseline algorithms. Furthermore, the problem that in training constant f_0 is available and in the conversion process, estimated changing f_0 patterns are analyzed, was tackled. Correlation coefficients showed convergency, subjective intelligibility, listenability and naturalness was preserved.

[De Armas et al., 2014] used support vector machines to estimate the f_0 of eight healthy subjects using EMG of the neck strap muscles. They used several features (root mean square value, mean amplitude value, auto-regressive coefficient and waveform length) and trained a Support Vector Machine classifier and regression function to estimate discrete tones and the fundamental frequency contour of sentences. The performance was very good with the squared correlation coefficient of $\rho^2 = 0.93 \pm 0.03$ for tones and $\rho^2 = 0.78 \pm 0.04$ for sentence intonation. Also [Ahmadi et al., 2014] investigated electromyography for estimating the intended f_0. In their work, healthy subjects were investigated and a GMM based voice conversion approach was used to convert features of the EMG signal into f_0 values. The same features than in [De Armas et al., 2014] were used. Furthermore, meaningful features were added (cepstrum coefficients, histogram, zero crossing, slope sign change, modified mean frequency and wavelet energy). Results in terms of squared correlation coefficient was able to be improved to $\rho^2 = 0.95 \pm 0.03$ for discrete tones and $\rho^2 = 0.87 \pm 0.03$ for sentence intonation. It must be noted that these results were based on voiced as well as unvoiced f_0 values. As described later on, this will increase the correlation significantly and we expect much lower coefficients if only voiced frames are used. Furthermore, the results were based on healthy speech signals only.

Another recent study of [Matsui et al., 2014] built a multi-agent technology for prosody control of the EL. They used a micro electro-mechanical systems (MEMS) accelerometer integrated in a smart phone. Hand gestures were used to control pitch contour (two types – linear mapping and f_0 template based method). Subjective evaluations revealed that prosody of EL output using the fundamental frequency control technique is more natural and, at the same time, as intelligible and stable as the conventional system. In addition, the implementation of a mobile device offers the advantage of being able to be expanded to a more complex system with the disadvantage that a connection to the server (either by local network or mobile internet) is needed.

One main problem of the EL device is the different point of excitation. Instead of the vibrating vocal folds, a coupler disk is vibrating and mechanical energy needs to be transmitted through the neck tissue into the vocal tract at a different location (supra-glottal). Due to the acoustic energy losses through the tissue the question arises if it is possible to approximate the artificial excitation signal of the EL device through a prototype oscillation of the healthy vocal fold vibration. [Wu et al., 2013] tackle this problem and propose a new supra-glottal excitation signal. The authors undertook simulations as well as measurements based on an implemented systems. They were able to improve the low frequency energy and correct the shifted formants as well as eliminate the visible spectral zeros. The compensation worked in two steps: 1. compensate the vocal tract (abnormal formants) and 2. compensate back cavity (abnormal spectral zeros). The authors motivate that the vocal tract is changed in terms of its length (it is reduced due to the removal of the larynx) and the point where the excitation signal is applied. The supra-glottal excitation signal was evaluated synthetically and with an EL system and recording from a laryngectomee and a healthy subject. The first three formant frequencies and their amplitudes, the low frequency energy and the frequencies of visible spectral zeros were taken into account. The results showed that the new supra-glottal excitation signal is able to eliminate the abnormal acoustic properties of EL speech for healthy and laryngectomy conditions. The same authors also investigated on/off control of an EL device using a video-based system which recorded lip deformation [Wan et al., 2012].

[McLoughlin, 2014] who introduced ultrasonic sounds to perform voice activity detection also extended the work to use it as an artificial excitation signal for non-audible voice restoration for future use.

B) Approaches for a New Transducer and Device

The conventional EL design is very limited in terms of practicability and usability. Due to the coil-coupler disk setup the artificial excitation signal can only be changed to a certain extent. A new transducer should be efficient in terms of power consumption and energy production at the same time. [Ooe et al., 2000] proposed the use of an intra-tracheal device that uses a piezoelectric transducer ceramic as its sound source. Due to the fact that the sound source excites the pharynx region locally from within the trachea it does not need to produce such an intense mechanical vibration as a regular EL. Nevertheless, the transducer is driven using a saw-tooth waveform. The f_0 of the output is approximately 120 Hz. [Sugio et al., 2007] introduced a piezoelectric transducer which is able to generate enough volume in the low frequency range. This kind of transducer should be implanted inside of the body.

[Houston et al., 1999] together with [Meltzner, 2003] propose a linear, electro-dynamic EL transducer. The new EL sounds more natural than commercially available EL devices, but due to problems concerning energy supply no further work was published.

[Merlo et al., 2008] proposed an intra-oral device which is powered remotely and controlled wirelessly. The device is composed of a dental unit, an external unit and controlling electronics.

An audio amplifier drives an electromagnet which is located at the cheek near the position of the intra-oral transducer (located in a dental appliance). A sine wave produces an alternating magnetic-field generated by the electromagnet. The permanent magnet in the dental transducer oscillates and strikes an acoustic coupler disk to produce audible sound. A proof of concept was given and different distances between the electromagnet and the intra-oral transducer were tested. Another intra-oral device was developed by [Takahashi et al., 2005] which was introduced earlier in part A. In his doctoral thesis, [Madden, 2013] developed a novel EL transducer, i.e., a light-weight miniature pager motor, typically found in common mobile phones. In [Madden et al., 2010] the novel transducer was compared to a conventional EL device. Three healthy speakers (two male, one female) recorded 48 monosyllabic words and four listeners evaluated the intelligibility. Intelligibility tests showed significantly better results than the old system.

[Yan et al., 2014] proposed a new miniature transducer system to replace the conventional EL design. The advantage of this device is that it can incorporate a varying driving signal. Within perceptual experiments they examined listeners' intelligibility and acceptability of the newly proposed device. The results included suggestions for the improvement of the artificial excitation signal (impulse, impulse + Gaussian noise, pulse train + square wave, etc.) and the best frequencies (108, 105, 111, 120 and 140 Hz). Very recently [Madhushankara et al., 2015] proposed a low-power frequency oscillator for driving the vibration head of an EL device. Due to the new design the power supply voltage was reduced. This paper analyzed the new oscillator in terms of design considerations but does not give any information about its performance with EL speech.

C) Approaches for DREL Removal and Other Individual Open-Loop Approaches

The EL device needs to be designed in a careful way as a balance between damping the artificial excitation signal so as to avoid leakage into the surrounding and efficiency in terms of bringing the maximum mechanical energy into the vocal tract. DREL is concentrated in the same range than speech occurs which causes confusion in vowel separation due to the auditory masking of the vowel formants [Liu et al., 2006], [Pandey et al., 2002], [Weiss et al., 1979]. It can be either removed using effective shielding of the vibration within the device or using noise subtraction [Azarnoush et al., 2007] or (adaptive) filtering algorithms.

[Norton and Bernstein, 1993] investigated the effect of placing acoustic shielding around the EL in order to reduce the DREL. As it turned out, the shielding is only able to reduce the noise to a certain amount. Also [Espy-Wilson et al., 1998] reported that acoustic shielding leads only to a minor improvement and that such an approach might be impractical. However, [Madden, 2013] focuses on developing a hands-free device and removing DREL through the hardware. Compared to the Servox Digital EL device the amplitude of the DREL noise was lower. The author concluded that the reduced DREL noise is a factor for the favorable scores in the conducted intelligibility study.

Concerning noise subtraction methods it is an important task to estimate the noise, this can be done using minimum statistics or quantile based noise estimation. [Cole et al., 1997] used a hybrid noise reduction method using both, spectral subtraction and root cepstral subtraction procedures. Test material included utterances of three male EL speakers, taken from conversational speech. The hybrid algorithm was evaluated by 30 listeners using a Mean Opinion Score (MOS) test ranging from 1 (unsatisfactory) to 5 (excellent). The authors claimed to reduce the direct buzzing sound with no apparent reduction in intelligibility, but the MOS results were not conclusive (2.3 for original EL speech and 2.8 for improved EL speech). However, no statement was made on how well this technique compares to other methods. [Espy-Wilson et al., 1998] succeeded in applying adaptive filters to remove the leaking DREL noise, resulting in improved naturalness while preserving intelligibility. Adaptive filtering assumes additive, uncorrelated noise and a reference signal that is uncorrelated with the desired signal which is not true for EL speech. Therefore, it is surprising that adaptive filtering can be used to remove DREL noise. Besides the signal processing technique, e.g., adaptive filtering, a microphone was used to measure the DREL to improve the output speech.

[Pandey et al., 2002] applied spectral subtraction, the noise was estimated using recordings of the operating device on the neck with closed lips. The average magnitude spectrum of noise was subtracted from the magnitude spectrum of the noisy speech using the original phase spectrum. The parameters subtraction factor α and spectral floor factor β as well as the exponent factor γ were optimized. α is set to 1: complete subtraction of magnitude speech of noise, β to zero (no noise floor), and γ to 2, meaning subtraction of power spectrum. In [Pratapwar et al., 2003] this work was extended using quantile based noise estimation. In [Pandey and Basha, 2010] different phase spectra were analyzed. Additionally, introduction of jitter and shimmer into the excitation signal was investigated. The different phase spectra were: noisy phase, zero phase, randomly selected phase, phase set for continuity across the frames and phase spectrum estimated from spectral subtracted magnitude spectrum using the assumption of a minimum-phase signal. Informal listening tests showed that no noise estimation strategy is better than taking the noisy phase. Introduction of shimmer did not help either, peak-to-peak jitter of 6 % and spectral compensation increased the quality. A compensation filter was designed as a linear-phase finite impulse response filter to approximate the long-duration averaged spectrum because the excitation signal (impulse train) emphasis high frequencies. Low frequencies are attenuated due to the transmission of the vibrations through the neck tissue. For investigating the effect of jitter and shimmer a Linear Predictive Coding (glslinkLPCLPC) based analysis-synthesis was used before creating a real-time system in [Basha, 2011] and [Basha and Pandey, 2012]. No listening tests or Signal-to-Noise Ratio (SNR) calculations were reported. [Liu et al., 2006] aimed on reducing radiated noise from the EL during phonation. The authors used a perceptual weighting technique which was developed to adapt spectral subtraction parameters to design a computationally efficient algorithm. [Azarnoush et al., 2007] and later [Kabir et al., 2008] used spectral subtraction and minimum statistics to improve EL speech and reduce the DREL signal. A Discrete Cosine Transform was introduced which has the advantage that no voice activity detector is needed. Noise estimation based on minimum statistics was also applied by [Hagmüller, 2009] who formulated an alternative approach of noise reduction in the modulation frequency domain. The so-called multi-path signal separation algorithm is an important part in our work to pre-process EL speech. It describes a new speech model for EL speech and modulation filtering is compared to spectral subtraction. Objective measures are used for evaluation as well as listening tests, but the results are not significant.

In our experiments we realized that the spectral subtraction algorithm fails when working on reverberant signals instead of clean signals. This leads to the question which of the above mentioned methods were also evaluated in a real-world scenario and to what amount they failed in reality.

In the PhD thesis of [Sharifzadeh, 2011] a solution for the conversion of whispers to fully-phonated speech with a modified framework of a Code-Excited Linear Prediction (CELP) codec is proposed (reconstruct natural sounding speech from whispers). The proposed framework is applied to the reconstruction of natural speech for laryngectomees. The system is based on mapping whispered speech phonemes to healthy speech phonemes. Problems were reported if there was no counterpart in whispered speech, and if a phoneme was mapping to two different phonemes. Voiced nasals add zeros to the frequency spectrum. The proposed method is mainly based on peak finding and peak smoothing. In the case of nasals, not the poles of the system should be smoothed but the valleys. Moreover, a vowel formant space for whispered speech and a comparison with corresponding phonated samples was presented. Vocal tract parameters for whispered speech have a much higher variance than those of healthy speech. These parameters need to be improved in order to prepare whispered speech for pitch insertion. Two methods for spectral improvement were developed, one based on a line spectral pair narrowing technique and the other based on a novel approach of applying probability mass functions to find the formant trajectories (formant smoothing) where the latter showed more accurate and reliable results. It is known that spectral locus is of greater perceptual importance than formant bandwidth. Two methods of pitch generation/insertion for the voice regeneration were pointed out, one based on the basic long term prediction filter and the other for pitch variation based on formant locations and amplitudes (the final system was tested using the basic long term prediction filter). For evaluation, subjective (MOS) and objective (log-likelihood ratio) tests were conducted. To summarize: This thesis used a CELP codec to analyze, modify and reconstruct speech. The system included a whisper activity detector, whispered phoneme classification, spectral improvement and pitch insertion based upon a CELP codec. The reconstruction is performed using a Vocoder which leads to an open-loop (telecommunication) approach. The average MOS of the regenerated speech was found to be 2.1 and 1.6 using the EL.

A similar but more efficient approach is proposed in the thesis of [Nakamura, 2010]. Two kinds of speaking-aid systems are described: 1) EL-speech to healthy speech conversion (estimation of target spectra and f_0 contour) and 2) EL-speech to whispered speech conversion (avoid estimating artificial f_0 contour). Another sound source unit was employed that generates a small poser source signal. Then, a non-audible murmur (NAM) microphone [Nakajima et al., 2006] recorded the speech and converted the small-powered EL speech to (a) healthy speech or (b) whispered speech. The naturalness has been improved dramatically, whereas the intelligibility slightly degraded. This work presents an open-loop setup which results in a promising way to reach a higher level of naturalness of EL speech. Follow-up work by [Doi et al., 2014] extended speech conversion using one-to-many eigenvoice conversion in order to reproduce speaker identity. One-to-many eigenvoice conversion is used to convert speech from a specific source speaker into speech of an arbitrary target speaker. Thus, the speaker individuality can be controlled by manipulation of a small number of parameters, or can be adapted using a small number of given utterances of a target speaker. Thus, the identity of an arbitrary speaker can be represented as a unique weight factor in the eigenvoice speaker space. Within their work they could show that eigenvoice conversion is capable of effectively adjusting the voice quality using only one arbitrary utterance of the target voice. Follow-up work was done by [Tanaka et al., 2014b] who performed speech conversion using statistical voice conversion (SVC) for EL speech enhancement. They showed with using their hybrid approach it was possible to improve naturalness without decreasing intelligibility. Spectral features were improved using spectral subtraction, whereas excitation parameters were estimated using GMM based SVC algorithms. As conventional EL speech is normally always voiced, there is no degradation of the converted speech when all speech frames are regarded as voiced. The authors conclude that voiced-to-unvoiced errors have minor effects on naturalness and intelligibility. Therefore, continuous f_0 contours are estimated without voiced/unvoiced decision. Furthermore, the continuous f_0 using a low-pass filter for smoothing was able to increase naturalness.

A quite different approach is to use a system based on magnetic field variations. In [Hofe et al., 2013] a number of magnets was implanted into the laryngectomee's mouth. These magnets produced a variation in the magnetic fields surrounding the mouth during speech. Using these signals isolated word recognition was performed and a word accuracy of over 98 % was reported. Based on the proof of concept the principles are currently extended to develop a new system and move on to patient trials. A new prototype is presented in [Cheah et al., 2015]. This system shows comparable recognition performance to the previous system but provides much desired hardware improvements such as portability, hardware miniaturization, desirable appearance and lower cost.

To the best of our knowledge there are currently the following active academic groups working on EL speech improvement:

1. Massachusetts General Hospital, Boston: This group works with electromyographic signals to control on/off signals and changing f_0 [Kubert et al., 2009]. The group also focused

on developing a new electronic device using an electro-dynamic transducer but did not continue in that direction due to problems with energy supply [Houston et al., 1999].

2. Nara Institute of Science and Technology: This group uses statistical voice conversion to convert EL speech to HE speech [Nakamura et al., 2011]. More recently they also focused on converting f_0 only to generate an artificial excitation signal for a closed-loop method [Tanaka et al., 2014b].

3. Nanyang Technological University, Singapore: Aim of this group is to regain natural sounding speech of laryngectomy and other voice-loss patients using non-surgical, non-invasive and non-intrusive electronics [McLoughlin, 2015]. They are working on whisper-to-speech conversion using a CELP-based reconstruction system where the quality of phonemes can be improved [Sharifzadeh et al., 2010]. Extensions to sentences performed poorly. An interesting idea is to use ultrasonic signals to resonate inside the vocal tract which results in a non-audible artificial excitation signal.

4. University of Hull in cooperation with University of Sheffield: The aim of this project is to develop a wearable sensing system which is capable of detecting movement in the vocal apparatus based on the principal of magnetic sensing and allows the reconstruction of speech without acoustic signals [Hofe et al., 2013]. The proposed technique is called Permanent-Magnetic Articulography. For this approach an array of magnetic sensors which are attached onto a wearable headset capture the movement of articulators where permanent magnets are implanted. They reach good results concerning speech recognition using these kinds of data. This method might be far too invasive and may be difficult to establish in medical practice.

5. Graz University of Technology, Austria: Aim of this group is to make a major revision of existing EL devices. They propose a novel approach to reduce DREL noise which is based on the removal of constant frequency parts [Hagmüller, 2009]. Furthermore, prosody for EL speech is investigated and f_0 estimation algorithms are proposed. They also deal with the mechanical transducer and investigate EMG methods to make the device hands-free. The main contributions are summarized in the thesis at hand.

1.4 Scientific Contributions and Publications

This thesis covers the following contributions:

- A German parallel electro-larynx speech – healthy speech database is recorded and prepared for simulations. Within these recordings, criteria for a good database are established. Furthermore, methods for post processing and analysis are developed. For instance, dynamic time warping for time aligning the speech data and signal-to-noise estimation for electro-larynx speech have been formulated. Furthermore, the database is evaluated in an Automatic Speech Recognition scenario where we are able to show that conventional Automatic Speech Recognition systems are capable to deal with the abnormal spectral structure of EL speech.

- A procedure for converting electro-larynx spectral features to a changing f_0 contour using a Gaussian mixture model based voice conversion approach is presented. Listening tests confirmed that a changing f_0 leads to a more natural sounding speech quality.

- A new EL design based on an electro-magnetic transducer is investigated and a prototype is developed. It includes a housing, a coupler disk and an elastic band with which the transducer can be fixed around the neck. Listening tests show that this new device is not optimized in terms of leakage of noise, but the small, light-weight and power efficient transducer leads to a promising outlook.

- Electromyography is used to control the electro-larynx device in a hands-free manner. A bio-signal capturing device is designed, built and evaluated in listening tests. Furthermore, we evaluated the training effect using this device and were able to show that people are capable of learning the use of their muscles in order to operate the EL device in an optimal way.
- Statistical voice conversion is investigated in terms of applicability. It turned out that voice conversion is a good method to generate natural speech, but the question of intelligibility is still unsolved.
- Based on the above mentioned contributions we developed an overall system which combines all small steps towards an improved bionic EL speech system.

The contributions in this thesis are divided between the author and co-workers as follows: The ideas concerning speech improvement of electro-larynx speech in terms of database recordings and estimation of fundamental frequencies originate from discussions with Gernot Kubin and Martin Hagmüller. Speech recognition results are obtained with the help of Juan A. Morales Cordovilla, Gabriel Hülser and Felix Rothmund [Hülser and Rothmund, 2015]. Input for a new transducer and its analyses and evaluation came from Wolfgang Truppe and Linda Lüchtrath [Lüchtrath, 2015]. The work on electromyographic control of the device was carried out together with Clemens Amon [Amon, 2014] and Hartwig Klammer [Klammer, 2015]. The work on statistical voice conversion was carried out during my research stay at the Augmented Human Communication Laboratory at Nara Institute of Science and Technology in Japan under the supervision of Tomoki Toda. The following articles have been published during the course of this thesis:

- Anna K. Fuchs and Martin Hagmüller, "Learning an Artificial F0-Contour for ALT speech," *13th Annual Conference of the International Speech Communication Association (Interspeech)*, Portland, USA, pp. 70-73, September 2012.
- Anna K. Fuchs and Martin Hagmüller, "A German Parallel Electro-Larynx Speech – Healthy speech corpus," *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (Maveba)*, Florence, Italy, pp. 55-58, December 2013.
- Anna K. Fuchs, Juan A. Morales Cordovilla and Martin Hagmüller, "ASR for Electro-Laryngeal Speech," *IEEE Workshop on Automatic Speech Recognition and Understanding Workshop (ASRU)*, Olomouc, Czech Republic, pp. 234-238, December 2013.
- Anna K. Fuchs, Martin Hagmüller and Gernot Kubin. "Artificial Fundamental Frequency Contour for Electro-Larynx Speech," *Proceedings der 40. Jahrestagung für Akustik (DAGA)*, Oldenburg, Germany, pp. 507-508, March 2014.
- Anna K. Fuchs, Clemens Amon and Martin Hagmüller, "Speech/Non-Speech Detection for Electro-Larynx Speech using EMG," *Proceedings of Biosignals - 8th International Conference on Bio-Inspired and Signal Processing (BIOSIGNALS)*, Lisbon, Portugal, pp. 138-144, January 2015.

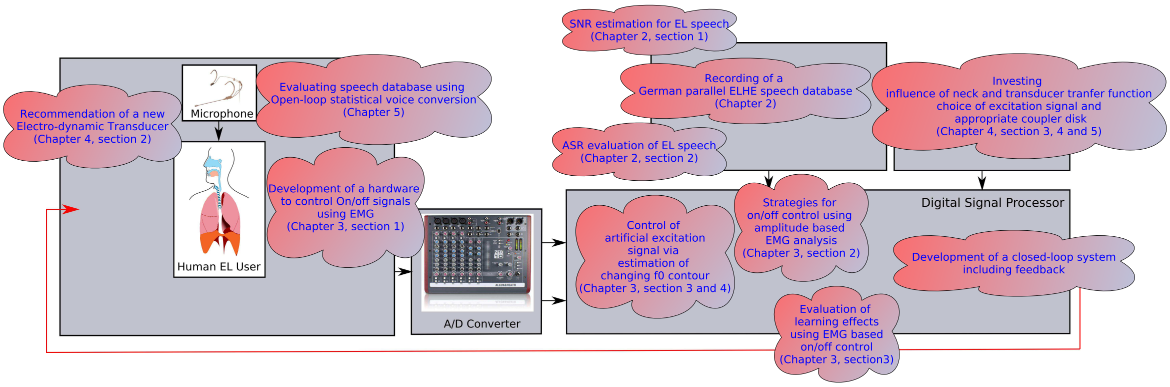


Figure 1.9: Overall structure of the thesis with main contributions.

1.5 Outline of the Thesis

In this chapter we summarized an overview of relevant information about electro-larynx speech and explained advantages and challenges related to this kind of substitution voice (Section 1.1 and 1.2). Furthermore, we summarized the state-of-the-art research (Section 1.3). Our positioning and the steps towards a new bionic electro-larynx speech system is explained in Section 1.4. We would like to improve EL speech in a closed-loop setup where we can solely influence the excitation source itself (waveform, f_0, etc.) and/or the artificial excitation signal producing hardware device, i.e., the artificial excitation source. The following parts of the thesis are structured as follows and summarized in Figure 1.9:

In Chapter 2 we discuss and evaluate the recorded speech material. Speech recordings which are needed to analyze and develop appropriate algorithms are introduced with the German parallel electro-larynx (EL) speech – healthy (HE) speech (ELHE) database (Section 2.2). We evaluate this database using a proposed SNR estimation algorithm which takes the properties of EL speech into account (Section 2.2.1). Furthermore, we applied it to an Automatic Speech Recognition system in Section 2.2.2.

Chapter 3 focuses on the possibilities for controlling the artificial excitation source. Section 3.2 deals with the shape and the parameters of the artificial excitation signal. We propose an estimation strategy based on Gaussian mixture models (GMMs) which is inspired by voice conversion models (Section 3.2.3). We show a correlation between estimated f_0 and natural f_0 and compared different f_0 estimation approaches using formal listening tests (Section 3.2.4). Section 3.3 deals with an approach to control hands-free on/off signals for the EL device. We build a hardware based on electromyographic (EMG) signals and evaluated the errors using a small database (Section 3.3.2). Furthermore, we investigate the learning effects, i.e., the ability of people to focus on EMG signal to willingly control the EL device and the increased pleasantness for listeners (Section 3.3.3).

Chapter 4 describes a proposed transducer for the EL device. The measured transducer transfer functions in section 4.3 reveals properties of the proposed transducer. Due to the advantages in terms of efficiency and size we optimize this new transducer in terms of an appropriate coupler disk (Section 4.4) and waveform of the artificial excitation signal (Section 4.5). In the last experiment of this chapter we compare the conventional device with the proposed one in a listening test (Section 4.6).

In Chapter 5 we investigate an open-loop approach for EL speech improvement in terms of voice conversion. We carry out several experiments to confirm the speaker-dependent results and the properties of our recorded German parallel ELHE speech database. Furthermore, we want to reveal the influence of the training material on the conversion results. Number of utterances in training (Section 5.3), number of GMMs for the conversion functions (Section 5.4) and criteria to choose the best sentences for training (Section 5.5) are compared in terms of objective error measures. To verify that efficient time alignment is crucial for the performance we also conducted an experiment with synchronized data (Section 5.6).

Chapter 6 concludes this thesis and provides a perspective on future work.

2 Database Development and Evaluation

2.1 Introduction

This chapter focuses on the recording of speech data for evaluation of EL speech and HE speech. We organize the chapter as follows: First, the German parallel electro-larynx (EL) speech – healthy (HE) speech (ELHE) database, its properties and challenges are presented. The nature and thus, the properties of EL speech is different from HE speech. Therefore, an SNR estimation algorithm which takes the nature of EL speech into account is proposed. We compare SNR results of EL speech and HE speech to emphasis the different properties. The different kinds of utterances allow analyzing EL speech according to prosodic properties. Thus, the mapping of prosodic properties from healthy speech to EL speech can be implemented using machine learning strategies. With this database, we can evaluate the influence of different improvement strategies as well as directly compare speech recognition results between HE and EL speech.

Second, we evaluate the database in an Automatic Speech Recognition (ASR) scenario. We want to show that EL speech can be applied to a standard ASR engine. We provide evaluation in terms of adaptation. Furthermore, we compare two DREL reduction algorithms. Finally, we investigate different features (MFCCs, PLP and RASTA-PLP) for ASR.

2.2 Database

In this section we introduce a new German parallel ELHE speech database that consists of utterances spoken by healthy speakers, one time with an EL device and one time with their normal voice (HE). Additionally to the latter, ground truth signals for the vocal fold oscillations from a laryngograph are provided. In the following, we describe the recording setup and material of this database and provide statistical information on f_0. SNR estimation is carried out in order to emphasize the spectral differences between healthy speech and EL speech.

Recording Details

To record the German parallel ELHE speech database, parallel utterances were spoken and recorded one time with healthy voice and one time with the EL device. The speech material of the German parallel ELHE database consists of almost 500 different phonetically rich utterances per speaker. The utterances were organized in 10 sessions with approximately 50 utterances per session. Prosodic differences can be investigated, because utterances with main focus on intonational and contrastive stress are included (statement vs. question, emphasis on different parts of the word, etc.). In total, this database consists of around 500 utterances spoken by 7 speakers which were recorded two times (HE and EL).

The Austrian German native speakers consisted of healthy subjects with an average age of 26 years (female) and 36 years (male). Although the anatomy of laryngectomees is rather different, we recorded healthy subjects who also produced EL speech in order to build up this parallel ELHE database. According to [Hagmüller, 2009], who carried out listening tests, there are no significant perceptual differences between EL speech produced by a laryngectomee or by a healthy subject. The subjects used a Servox Digital, a widely used device in Europe and the US. Three female (F01, F03 and F07) and four male speakers (M02, M04, M05 and M06) were recorded. The f_0 of the device was adjusted to a comfortable level for each speaker individually. All recordings were carried out on-site at the recording studio of the Signal Processing and Speech Communication Laboratory at Graz University of Technology. During the recordings each test subject was overseen by a supervisor. The supervisor monitored the speech recording software in order to control and modify the recording process immediately. This is necessary when, e.g., something is misread. The software used for recording was SpeechRecorder [Draxler and Jänsch, 2004] which had been designed to record speech corpora. The test subjects were recorded sitting in a sound proof recording room. The supervisor observed the test subject through a glass window. The test subject had to read utterances displayed on a screen. The speech material was recorded with a headset microphone AKG HC 577 L with omni-directional pick-up pattern. The head-mounted high-quality condenser microphone was chosen to ensure a consistent recording quality, since it guarantees a constant distance of about 2 cm from the corner of the mouth.

Additionally, a laryngograph was used to provide a ground truth signal for f_0 of healthy speech signal. The laryngograph measures time varying trans-glottal impedance corresponding to the healthy phonation. To be more precise, it captures the vibration of the vocal folds. The speaker has to carry a neck band with the laryngograph electrodes. The electrodes are loaded with a high frequency current. The measured impedance is proportional to the contact area of the closing and opening vocal folds (if the vocal folds are fully open, no relevant information can be extracted). Due to the vertical movement of the larynx a low frequency component disturbs the high-frequency oscillations. Therefore, a band-pass filter with linear phase response and with a lower cut-off frequency of 50 Hz and an upper cut-off frequency of 2000 Hz is recommended. Both, the microphone and the laryngograph signals, were sampled at 48 kHz with 16 bit resolution.

2.2.1 Signal-to-Noise Ratio (SNR) Estimation

SNR is often used as a simple objective measure for speech quality [Benesty et al., 2008]. EL speech is corrupted with DREL noise, thus, SNR compared to HE speech is decreased. Furthermore, we want to classify the quality of EL speech based on SNR values. Many SNR estimation techniques are based on the knowledge of the clean speech signal and the noisy signal. We need to estimate SNR when only the noisy signal is available.

In healthy speech there are (ideally) two levels: Speech Level (SL) and Noise/Silence Level (NL). Thus, the SNR for HE speech is defined as the ratio of these two levels → $\mathrm{SNR} = \frac{\mathrm{SL}}{\mathrm{NL}}$. Within EL speech we have to deal with three levels: speech level (SL), noise level of directly radiated EL noise (DREL-L) and noise/silence level (NL). This is illustrated in Figure 2.1. DREL-L is masking the resulting EL speech. Therefore, the SNR is defined as the ratio between SL and DREL-L because this represents the human perception. To find DREL-L, a threshold *thr* needs to be implemented. We used an iteratively changing one which is able to find DREL-L automatically.

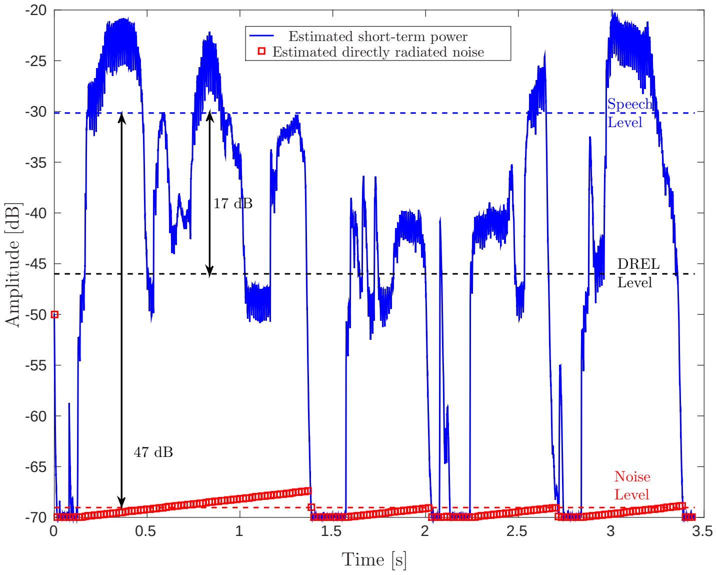


Figure 2.1: Short-term power $\overline{y^2[k]}$ and directly radiated noise $\overline{n^2[k]}$ of one utterance spoken by a male speaker with the EL device; three levels: Speech Level (SL), DREL Level (DREL-L) and Noise Level (NL); SL - NL: 47 dB; SL - DREL-L: 17 dB.

One standard SNR definition for speech signals is the averaged segmental SNR ($sSNR$). We calculate the arithmetic average of logarithmic SNR:

$$sSNR = \frac{1}{L}\sum_{k=0}^{L-1}\left(10 \cdot \log_{10}\frac{\overline{s^2[k]}}{\overline{n^2[k]}}\right). \tag{2.1}$$

Within this notation $\overline{s^2[k]}$ is the k-th value of the averaged speech signal power, and $\overline{n^2[k]}$ is the averaged k-th noise power value. We use a segment length of 1, which means that we carry out a sample-wise processing. Therefore, L is equal to the length of the signal $s[k]$. In order to obtain values for significance testing we increase the segment length (e.g., in Section 4.5). Furthermore, $sSNR$ is evaluated only on parts of the signal where speech is active, which is determined via another fixed threshold. Values below this thresholds, which are associated to non-speech parts of the utterance, are neglected (see Figure 2.2, VAD block).

We can only observe the noisy speech signal $y[k]$, which is the sum of the speech signal $s[k]$ and the noise signal $n[k] \rightarrow y[k] = s[k] + n[k]$. To obtain $\overline{s^2[k]}$ in (2.1) we have to subtract the estimate of $\overline{n^2[k]}$ from the estimate of $\overline{y^2[k]}$. To estimate the short-term power of the signal $\overline{y^2[k]}$ and of the noise $\overline{n^2[k]}$, we use first-order infinite impulse response (IIR) smoothing [Hänsler and Schmidt, 2004]. Whereas this algorithm works well for healthy speech, EL speech needs to be handled with more care.

$$\overline{y^2[k]} = \begin{cases}(1-\gamma[k])y^2[k] + \gamma[k]\overline{y^2[k-1]}, & \text{if } \overline{y^2[k]} \geq thr;\\ \overline{y^2[k-1]}, & \text{otherwise;}\end{cases} \tag{2.2}$$

with γ being a smoothing constant, which differs for rising and falling signal edges in order to detect rising signal powers very rapidly. To estimate the directly radiated noise level, the short-term power of the signal computed in (2.2) is used:

$$\overline{n^2[k]} = \min\{\overline{y^2[k]}, \overline{n^2[k-1]}\}(1+\epsilon) \tag{2.3}$$

ϵ is a constant slightly larger than one to avoid that the result of the minimum operator is freezing at a global minimum.

For HE speech only the first part of (2.2) is used (see also Figure 2.2). The threshold *thr* is only used for EL speech. The results of (2.2) and (2.3) depend on *thr*. Therefore, it is possible to find the optimal value for *thr* with an iterative loop. The start value of *thr* and other parameters are chosen empirically. This was done by plotting the estimates of short-term power and directly radiated noise and optimize the levels visually.

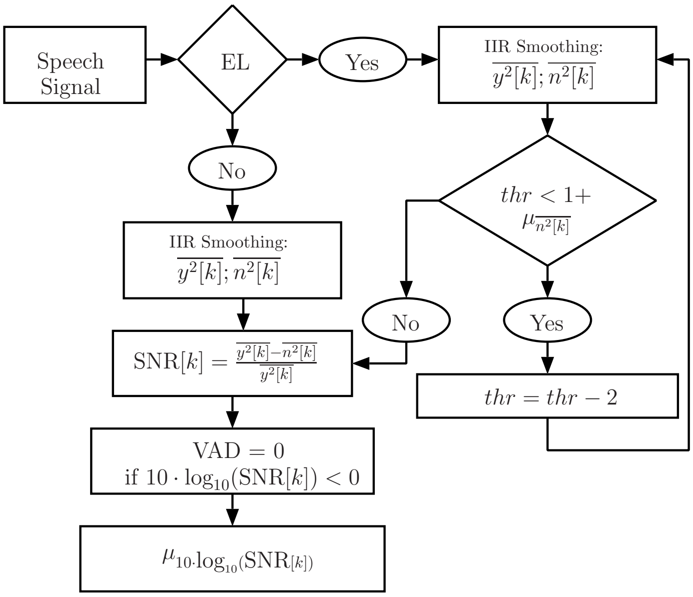

Speech Signal
EL
Yes
IIR Smoothing: $\overline{y^2[k]}; \overline{n^2[k]}$
No
$thr < 1 + \mu_{\overline{n^2[k]}}$
IIR Smoothing: $\overline{y^2[k]}; \overline{n^2[k]}$
No
Yes
$\mathrm{SNR}[k] = \frac{\overline{y^2[k]} - \overline{n^2[k]}}{\overline{y^2[k]}}$
$thr = thr - 2$
VAD = 0
if $10 \cdot \log_{10}(\mathrm{SNR}[k]) < 0$
$\mu_{10 \cdot \log_{10}(\mathrm{SNR}[k])}$

Figure 2.2: IIR smoothing for EL and HE speech with iteratively changing threshold thr (initialized with a fixed starting value); $\mu_{\overline{n^2[k]}}$ – estimated mean value of (2.3).

Descriptive Statistics of Database

The whole database has been analyzed with respect to important speech related properties. In Table 2.1, one can see that speaking with the EL takes much longer than with normal voice because careful articulation improves intelligibility. Although the same utterances were spoken, we recorded 4 h30 min of EL speech but only 2 h42 min of HE speech.

The mean fundamental frequency $\overline{f_0}$, as well as the standard deviation σ_{f_0}, were estimated for each speaker and type of speech. Praat [Boersma and Weenink, 2007] was used to extract f_0. $\overline{f_0}$ of healthy speech depends on the speaker. Due to the anatomy of the vocal folds female speakers have a higher $\overline{f_0}$ than male speakers and also σ_{f_0} is larger. For EL speech, $\overline{f_0}$ depends on the adjustable EL device. Additionally to our German parallel ELHE database speech material of a laryngectomee (M08) was taken into account. Looking at Table 2.2 we can see that only 35 utterances were available for this speaker who also used the Servox Digital (his own device). f_0 of the device was set at very low 87 Hz which might be beneficial for intelligibility and handling the device. The SNR is 17.79 dB which is a typical SNR for male speakers compared to our database. We expect that a low SNR will result in a high number of confusions and, therefore, in a reduced intelligibility.

An important and necessary task during speech improvement is to evaluate the improved speech. Subjective methods, e.g., listening tests, are the preferred method to evaluate results, but due to their relatively high effort, objective measures of speech quality are easier to apply. However, if we conduct listening tests we verified that the listeners are normal-hearing subjects. See Appendix B.1 for the standards to perform audiometry. These measures are commonly based on the SNR or on some distance between the original speech and the ”improved“ speech.

Preliminary SNR measurements based on IIR smoothing show that due to the DREL level, SNR values of EL speech are around 17 dB (with the exception of only two female speakers), whereas HE speech produces an SNR close to 50 dB.

ID	Age		# Utterances	Length	$\overline{f_0}$	σ_{f_0}	SNR
F01	28	EL	503	45 min 28 s	192	7	17.95
		HE	503	29 min 57 s	198	27	46.57
		Lar			196	27	
F03	31	EL	250	19 min 51 s	199	6	9.02
		HE	250	13 min 48 s	175	28	49.08
		Lar			174	28	
F07	18	EL	503	48 min 39 s	199	2	8.28
		HE	503	26 min 53 s	209	33	48.18
		Lar			209	33	
M02	38	EL	503	36 min 30 s	99	4	16.97
		HE	503	24 min 55 s	113	17	46.52
		Lar			112	17	
M04	50	EL	503	52 min 10 s	93	1	18.83
		HE	503	30 min 5 s	140	30	52.62
		Lar			136	30	
M05	28	EL	503	45 min 56 s	93	0	20.63
		HE	503	26 min 2 s	138	28	52.31
		Lar			136	27	
M06	29	EL	250	19 min 32 s	94	1	16.61
		HE	250	12 min 58 s	119	20	53.32
		Lar			117	20	
Sum			6030	7 h 12 min 44 s			

Table 2.1: Statistical analyses of ELHE database together with recordings from laryngograph (Lar): Number and length of utterances; mean value and standard deviation of $f_0 - \overline{f_0}$, σ_{f_0}; signal-to-noise ratio (SNR).

ID	Age		# Utterances	Length	$\overline{f_0}$	σ_{f_0}	SNR
M08	68	EL	35	2 min 16 s	87	3	17.79

Table 2.2: Statistical analyses of additional speech material: Number of utterances; mean value and standard deviation of $f_0 - \overline{f_0}$, σ_{f_0}; signal-to-noise ratio (SNR).

Power Spectral Density (PSD) estimation was carried out on both speaking modes (EL and HE) and averaged per-utterance and per-gender. For PSD calculation we used Welch's method where the data is split into segments of length 70 ms, without overlap (Hamming window is applied), periodograms were computed and averaged. In Figure 2.3 the PSD averaged over all utterances is illustrated for each gender. The spectral structure for EL and HE is completely different. The low frequency deficit up to around 500 Hz, which was also reported by [Qi and Weinberg, 1991], can be seen for both genders. Moreover, the harmonics of the constant frequency excitation signal, which are responsible for the monotonic EL speech, can be seen.

Additionally, the database has been analyzed according to its word statistic: The utterances contain 3961 word tokens, without counting multiple occurrences there are 1444 word types. 1210 words only occur once. In order to implement a speech recognizer based on the recorded data, a dictionary, that lists the phonetic transcriptions needs to be set up manually. In Figure 2.4 the symbol distribution according to the extended speech assessment methods phonetic alphabet (X-SAMPA) [Wells, 1995] is illustrated for the given utterances. The distribution is typical for the German language where the most common phonemes are: the alveolar voiceless plosive [t], nasal [n] and fricative [s], as well as the open centered vowel [a], the close-mid front vowel [e], and the schwa [@], among others [Lasarcyk, 2005]. Note that German phoneme distribution in this example clusters phonems differently, e.g., open-mid front vowel [E] and close-mid front vowel [e]; schwa [@] and open-mid schwa [6].

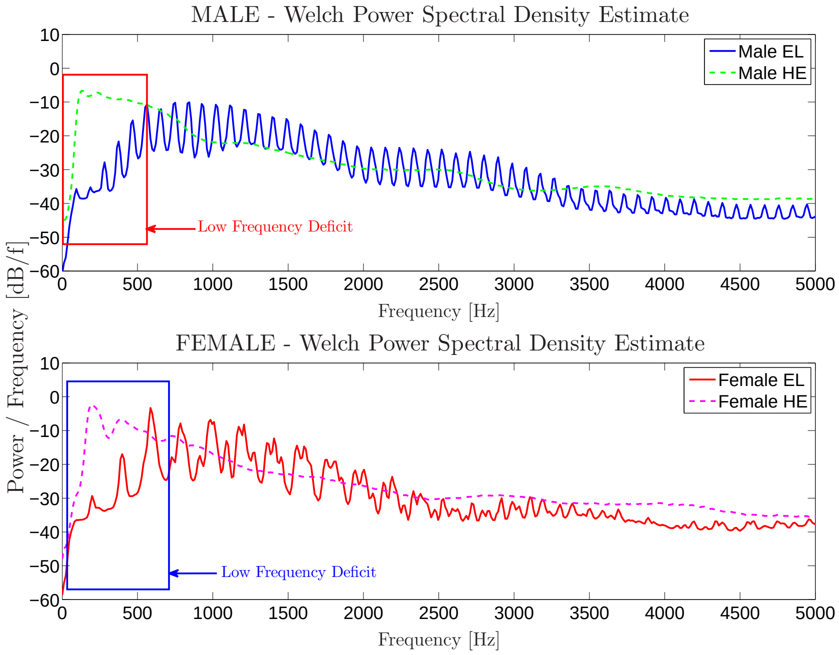


Figure 2.3: Power Spectral Density of EL speech and HE speech averaged over all male (upper plot) and all female speakers (lower plot).

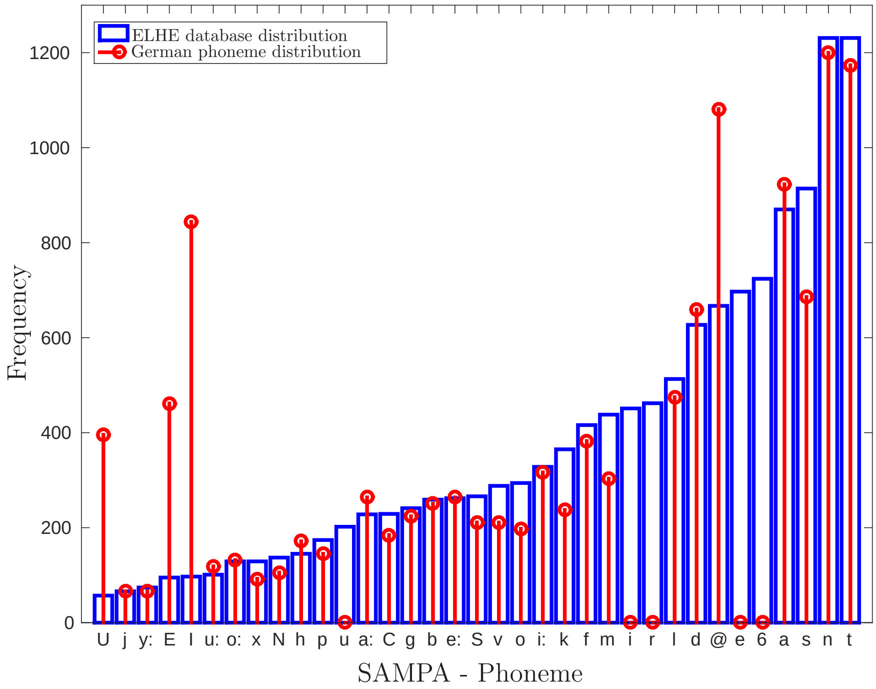


Figure 2.4: Phoneme distribution of the ELHE database (blue) compared to typical German phoneme distribution (red).

2.2.2 Experiment I: Automatic Speech Recognition (ASR)

The motivation to apply ASR on disordered speech is twofold. On the one hand, ASR systems could be used to control assistive technologies, whereas on the other hand, ASR systems can also be used for evaluation purposes. Based on the word accuracy rate, speech intelligibility can be quantified. In existing ASR systems under controlled conditions, the word recognition accuracy is very high (above 90 %) [Benesty et al., 2008]. These ASR frameworks often use a large amount of (continuous) speech for training. For disordered speech like dysarthric voice, where the ability to articulate is drastically reduced, building ASR systems is quite a problem, especially, because the amount of available speech material is much smaller than for healthy speech. For patients with speech problems, speech recordings are significantly more exhausting and difficult than for healthy speakers. State-of-the-art systems train triphone models and for this reason large amounts of speech material are needed [Benesty et al., 2008]. So far, ASR for disordered speech has been addressed by few authors, but it is an increasingly active research area. Some work is done on speech recognition of dysarthric speech, which can have a very profound influence on speech intelligibility and thus, on the recognition results.

In [Christensen et al., 2012], a database of dysarthric speech was used. This database is still much smaller than typical speech databases. The study investigated the influence of fundamental training and adaptation techniques on the ASR system. Different data sets (dysarthric and healthy speech) and adaptation (Maximum a Posteriori (MAP) approach) on different targets were investigated. The speech material consisted of 15 speakers with 250 unique words per speaker and approximately 50 minutes of speech per speaker. 12 perceptual linear prediction (PLP) features were used to train the acoustic model. The best average word accuracy rate of around 54 % was obtained by MAP adaptation of the dysarthric speech model, where the test speaker was also present in the training. Although the difference between healthy speech and dysarthric speech is large, MAP can deal with it to some extent. Results strongly depend on the speaker and on the severity of the dysarthric speech.

ASR has also been applied in a completely different approach in [Meltzner et al., 2011]. Instead of acoustic speech, the muscle activity of the facial and neck musculature was recorded and this “silent speech” was recognized. Electromyography (EMG) can capture sufficient speech information to feed an ASR system. Moreover, EMG can be used as an improvement strategy for disordered speech. Although it is not yet clear whether mel-frequency cepstral coefficients (MFCCs) are the most effective parametrization for EMG signals, they were used in this study. Using MFCC features with a combination of muscle co-activation levels, an average recognition rate of 86.7 % on a 65 isolated word vocabulary for 9 speakers could be achieved.

In [Maier et al., 2010], the authors focused on speech material from patients suffering from head and neck cancer. A standard text read by 41 German laryngectomees (using tracheo-esophageal substitution voice) and 49 German patients who had suffered from oral cancer was evaluated. The results were compared to a control group of 40 speakers without speech pathology. The word recognition rate was then compared to perceptual ratings by a panel of experts. As an outcome it is shown that ASR is a good measure with low effort to objectify and quantify intelligibility of disordered speech. Several language models were investigated. The ASR system was non-adapted. The word accuracy results for the control group (76 % ± 7) is significantly higher compared to the laryngectomees group (48 % ± 19). The agreement, calculated using Spearman’s correlation coefficient ρ, between word recognition rate and the mean scores of the perceptual ratings is very high in both patient groups with -0.83 and -0.9, respectively.

In his doctoral thesis, Nakamura investigated a speech aid system for electro-laryngeal speech using statistical voice conversion [Nakamura, 2010]. Within his thesis he also carried out a case study of speech recognition for electro-laryngeal speech. He employed phonetically tied-mixture acoustic models. Maximum likelihood linear regression (MLLR) was the employed adaptation technique to transform the speaker independent model into a speaker dependent one. Two sets of speech data were used: 1) EL speech of a laryngectomee (native Japanese, 50 utterances for adaptation, 30 for test) and 2) speech of other types of speaking-impaired people (10 speakers (cerebral palsy, hearing-impaired, etc.)). The used speech material comprises words, digits and short utterances. MFCC features were employed. For the second group around 20 to 40 utterances were taken for adaptation and around 20 for test. For 1) the accuracy for EL speech is almost 80 % with adaption. Using the voice conversion technique word, accuracy is around 75 % without adaptation and does not surpass the values of EL speech (without conversion). The word accuracy for 2) is around 20 % depending on the kind of disorder and increased to around 60 % after the MLLR adaptation.

Experimental Setup

We used the German parallel ELHE database described in Section 2.2 for evaluation using an ASR system. The speech utterances were sampled at 48 kHz and 16 bit amplitude resolution and re-sampled to a sampling frequency of 16 kHz for the speech recognition task. In the following experiments, we used two versions of the German parallel ELHE database. For **Experiment I, Part A** and **Experiment I, Part B**, speakers F01, F03, M02, M04, M05 and M06 according to the description in Table 2.1 were evaluated. 445 (192) utterances per speaker composed a phonetically rich set for training and 58 utterances per speaker are kept for testing (different for each speaker). The number of different words in the training set was 579 and for the test 2404. Additionally, around 2500 clean utterances of the Bavarian Archive for Speech Signals PHONDAT-1 database [Schiel and Baumann, 2006] sampled at 16 kHz were used. These utterances correspond to 25 different speakers from both genders resulting in around 100 utterances per speaker. These subjects were native German speakers and produced the same speech material as the speakers of the German parallel ELHE speech database. In **Experiment I, Part B**, the main focus is on the evaluation of two noise reduction algorithms as explained below.

For **Experiment I, Part C**, additionally F07 and M08 (see Tables 2.1 and 2.2) were taken into account, but no PHONDAT-1 data was used. Furthermore, the test utterances were changed to be the same 60 utterances for every speaker. Speaker M08 is a laryngectomee with 35 utterances which were only used for testing.

EL Speech Enhancement Strategy

Listening tests carried out by [Meltzner and Hillman, 2005] have shown that EL speech can be most improved by removing DREL sound. We used two simple enhancement strategies to reduce the DREL sound: 1) spectral subtraction (SS) and 2) modulation filtering (MF).

The first method was SS (see also literature review in Section 1.3). The DREL noise of an EL is only slowly changing. SS is based on estimating the noise power spectrum and then subtracting this spectrum from the signal power spectrum. Although SS has the problem that the direct noise is not additive, this method was able to reduce DREL to a large extent.

The second method, MF, filters out the DREL sound in the modulation frequency domain. This approach introduced by [Hagmüller, 2009] takes advantage of the different properties of the EL speech sound and the DREL sound. The directly radiated component of the EL energy is not modulated by the articulatory organs but transmitted over the air to the human ear on a direct path. Therefore, this signal is only modulated at very low frequencies, if at all, and can effectively be assumed to be time-invariant. If we consider that the speech sound is a time and frequency dependent modulation of the excitation signal, in our case the EL sound, then we only have to suppress the signal path which is constant. In order to do so, a notch filter is placed at a modulation frequency of 0 Hz. The filter removes spectral constant parts from speech. Although this can produce audible distortions the spectral domain is enhanced. Therefore, the method is highly beneficial for pre-processing.

The spectrograms of HE speech, EL speech and enhanced EL speech using MF are illustrated in Figure 2.5. It can be seen that there is a mismatch between the HE and EL domain in terms of high- and low frequency deficit as well as differences in the position, the bandwidth and the acoustic energy of the formants [Saikachi, 2009].

Used Features: MFCCs, PLP and RASTA-PLP

MFCC features:
Mel-frequency cepstrum coefficients (MFCCs) are a compact representation of the frequency spectrogram and are standard features for ASR systems. The most important parameters of the front end were: 32 ms frame length and 100 Hz frame rate; 26 triangular filters for the mel-

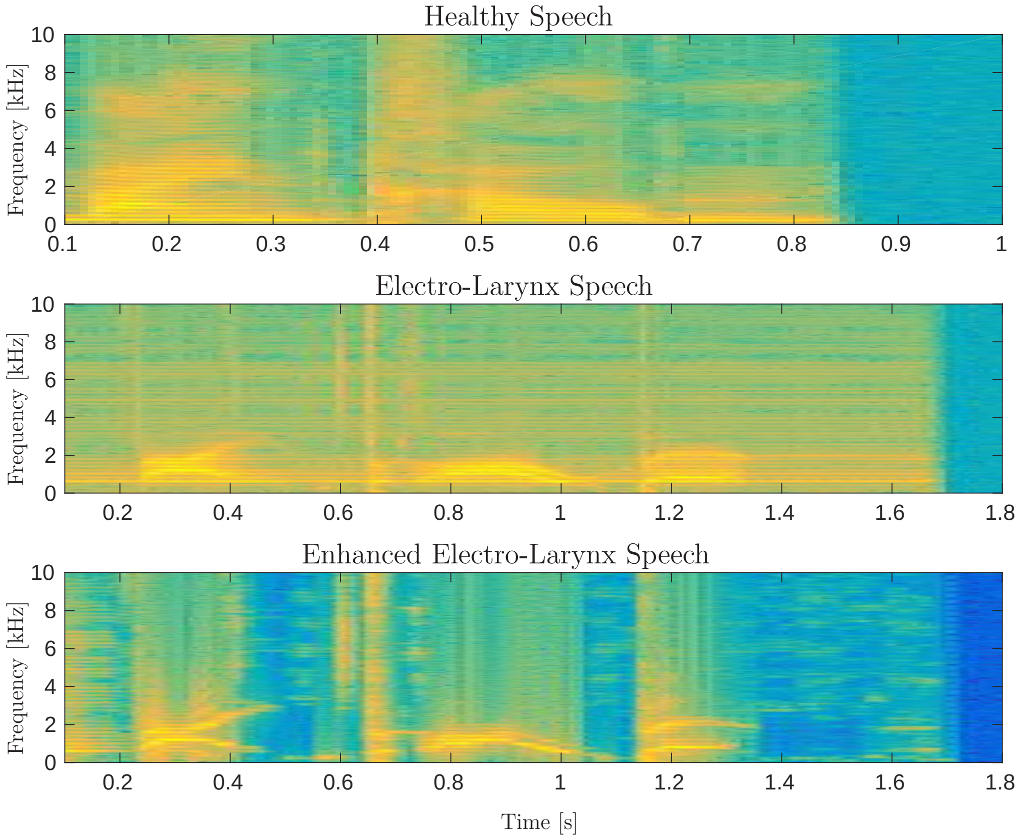


Figure 2.5: Spectrograms of the word "Weintrauben" (translation: "grapes"); Healthy (HE) speech (upper plot), Electro-Larynx (EL) speech (middle plot) and enhanced EL speech using modulation filtering (MF) (lower plot).

spectrum; 13 MFCCs and cepstral mean normalization. Δ and $\Delta - \Delta$ features with a window length of 5 were also appended, obtaining a final feature vector with 39 components.
PLP and RASTA-PLP features:
Perceptual linear prediction (PLP) coefficients are similar to MFCCs but differ in some details. The mel-frequency warping from MFCC analysis is replaced with grouping into critical bands according to the bark scale. In the next step, the signal is weighted using a simulated equal loudness curve and compressed to approximate the perceived intensity loudness (cubic root). Finally the linear prediction coefficients are calculated using a parametric model.

Relative Spectral Transform (RASTA)-filtering [Hermansky and Morgan, 1994] is used to support PLP features. The human hearing is relatively insensitive to slow changes in speech signals. This is considered within RASTA processing because it enables removing slowly varying frequency components. To do so, the critical-band power spectrum is transformed into the logarithmic spectrum. Then, a bandpass filter is applied to each frequency band.

Linear distortions as caused by the recording equipment are represented by an additive constant due to the logarithmic spectral domain. In case of the EL speech, the direct noise of the excitation signal generate such additive constants, but they can be easily filtered. In the next step, the filtered speech is transformed back into the spectral domain (inverse discrete Fourier transform) and the feature vector is then derived as in conventional PLP [Hermansky and Morgan, 1994]. Figure 2.6 shows the basic steps for the calculation of PLP and RASTA-PLP features.

Figure 2.7 gives an overview on the different extracted features. The first spectrogram shows the unprocessed speech signal sampled at 16 kHz from 0 to 8000 Hz. The second and third

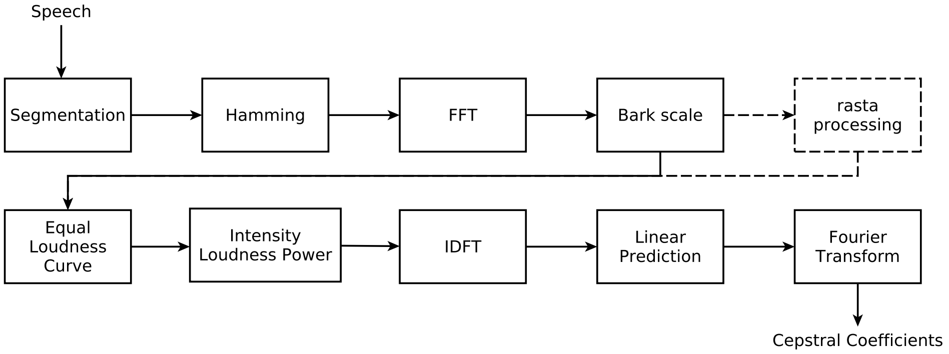


Figure 2.6: PLP feature extraction with optional RASTA processing; FFT and IDFT are the Fast Fourier Transform and Inverse Discrete Fourier Transform.

figures show PLP and RASTA-PLP, respectively. The constant spectral components caused by the EL device can be most effectively suppressed using RASTA-PLP. For the PLP features we used a frame length of 32 ms, a frame shift of 10 ms, 26 triangular filters for bark spectrum and a model order of 12 for the linear prediction. Also Δ and Δ − Δ features were appended.

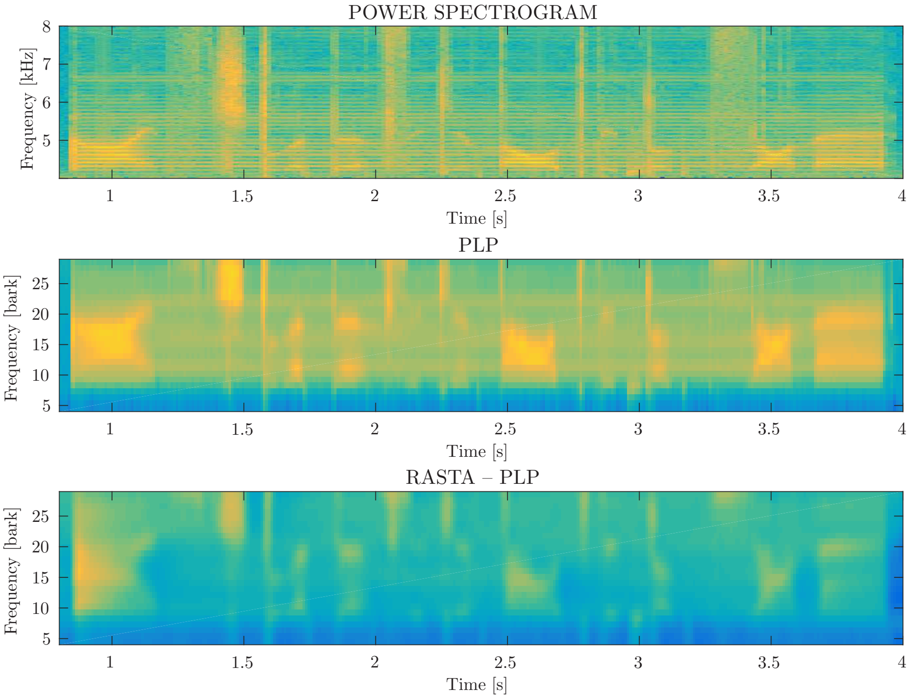


Figure 2.7: ”Einst stritten sich Nordwind und Sonne” (translation: “The North Wind and the Sun were disputing“) – female EL speaker; different feature extraction methods.

Automatic Speech Recognition System

In standard speech recognition systems the phonemes are represented using hidden Markov models (HMMs). Although it is necessary to estimate a large number of parameters compared to monophone HMMs, we built an ASR system based on HMM triphones. The advantage is that the size of the lexicon can easily be increased in the future. Therefore, it is most useful in real applications. Also, the variability of phonetic realization is reasonably well expressed with triphones. Both the front end and the back end have been derived from the standard base-line recognizer employed for the Aurora-4 database [Hirsch, 2002]. The language model is a bi-gram model derived from the test corpus with low perplexity.

We conducted three experiments: Experiment I, Part A) Standard ASR with adaptation strategies using standard MFCC features; Experiment I, Part B) Standard ASR with EL improvement strategies using standard MFCC features and Experiment I, Part C) Standard ASR using PLP and RASTA-PLP features.

To train the triphones, the back end employed a transcription of the training database based on 34 SAMPA-monophones. This transcription was derived from a more detailed monophone transcription (based on 44 SAMPA-monophones) by means of a careful clustering of the less common monophones. Each triphone was modeled by an HMM of 6 states and 8 Gaussian-mixtures/state. By means of a monophone classification created with the help of a linguist, a tree-based clustering of the states was also applied to reduce the complexity and the lack of training data. Tree-based clustering also allowed the creation of triphone models which have not been observed in the training stage. We used a bi-gram language model. The perplexity of the language model is around 3.5 for Experiment I, Part A and Part, B. For Experiment I, Part C we changed the sentences to be the same for all speakers. Therefore, we expect a further decrease of perplexity. The higher the value of the perplexity is, the worse the prediction in the test set is and the worse the recognition results are. In [Young et al., 2006] the authors mention a perplexity of around 131 for a bi-gram language model using the 'Sherlock Holmes' books.

One of the most powerful and popular adaptation techniques is maximum likelihood linear regression (MLLR) [Leggetter and Woodland, 1995]. We applied two kinds of adaptation: **1) Speaker dependent MLLR adaptation** based on a class tree regression and **2) Domain MLLR adaptation** based on retraining the model using new data from the target domain.

Results are presented using the word accuracy rate W_{acc}. According to [Young et al., 2006] the percentage accuracy is defined as

$$W_{acc} = \frac{N - S - D - I}{N} \cdot 100\% \tag{2.4}$$

where N is the number of words, S is the number of substitutions, D is the number of deletions and I is the number of insertions.

Experiment I, Part A: Standard ASR using MFCC Features and Adaptation Techniques

Three types of sets were used to train the triphones: 1) speech data from healthy individuals (H_H or H_E), 2) electro-laryngeal speech data (E_{E1} or E_{E2}) and 3) a combination of healthy and electro-laryngeal speech data (EH_{E1} or EH_{E2}).

Using this notation, H stands for healthy, and E for electro-laryngeal. The capital letter indicates the training, and the subscript indicates which data is tested. The difference in the subscript between E_{E1} and E_{E2} (HE_{E1} and HE_{E2}) indicates whether some speech material of the tested speaker occurs in the training or not (1 if it does not, 2 if it does). It must be noted that the amount of training material differs for the different types and speakers. For the training and test of speakers F01, M02, M04 and M05, around 500 utterances are available, whereas for F03 and M06 there are only around 250.

In Table 2.3 and 2.4, W_{acc} rates are shown for the different setups. The W_{Acc}, when training material only consists of healthy speech (PHONDAT-1 as well as healthy speech material from the 6 speakers) and we test on healthy speech, is 98.96 % (Baseline - H_H). When the test is carried out on electro-laryngeal speech, the performance is very low (5.53 %; Baseline - H_E) due to the mismatched domain (differences in the position, the bandwidth and the acoustic energy of the formants between healthy and electro-laryngeal speech – see also Figure 2.5).

Speaker ID	Baseline [%]					
	H_H	H_E	E_{E1}	E_{E2}	EH_{E1}	EH_{E2}
F01	97.39	5.22	15.36	78.55	27.25	67.25
M02	98.89	28.61	64.71	91.45	83.33	88.70
F03	99.56	0.22	31.43	62.58	39.00	55.31
M04	99.27	-6.57	47.93	84.67	37.24	71.83
M05	98.95	3.66	42.93	75.39	57.74	70.03
M06	99.33	5.37	55.96	83.20	45.03	75.43
Average	98.96	5.53	42.62	79.31	48.02	70.84

Table 2.3: Results of ASR for Baseline setup; H_H and H_E – Training: healthy, Test: healthy, electro-laryngeal; E_{E1} and E_{E2} – Training: electro-laryngeal, Test: electro-laryngeal (1 - speaker is not included in training, 2 - speaker is included in training); EH_{E1} and EH_{E2} – Training: mixed healthy and electro-laryngeal, Test: electro-laryngeal (1 and 2 as before); values in grey are explained in the text.

Speaker ID	Speaker MLLR Adaptation[%]			Domain MLLR Adaptation[%]
	H_E	E_{E2}	EH_{E2}	dH_{E2}
F01	51.50	81.74	83.77	81.16
M02	61.40	91.32	89.72	88.44
F03	1.64	80.71	71.68	54.68
M04	63.04	84.43	80.66	85.40
M05	37.40	81.15	76.70	80.63
M06	15.82	89.19	87.70	83.67
Average	36.82	84.76	81.70	79.00

Table 2.4: Results of ASR for different setups; 1) Speaker adaptation using MLLR and 2) Domain adaptation using MLLR; H_E – Training: healthy, Test: electro-laryngeal; E_{E2} – Training: electro-laryngeal, Test: electro-laryngeal (2 - speaker is included in training); EH_{E2} – Training: mixed healthy and electro-laryngeal, Test: electro-laryngeal (2 as before); dH_{E2} – Training: healthy, Adaptation: to electro-laryngeal domain, Test: electro-laryngeal; values in grey are explained in the text.

The performance of speaker M02 is consistently good for each setup. Even in the mismatched domain (training: healthy; test: electro-laryngeal) this speaker performs rather well (28.61 %; Baseline - H_E). This speaker is most accustomed to using the EL device. Speaker F03 performs worse than anybody else. Informal listening tests verified that this speaker is less intelligible than the others when speaking with the EL device. This is one reason of her low performance. Another reason is that this speaker is female, and female speakers are less represented than male speakers in the parallel ELHE database. The same applies to speaker F01 in the baseline experiments for E_{E1} (15.36 %).

When speech material of electro-laryngeal speech is added to the healthy training, (EH_{E1} and EH_{E2}) the word accuracy rate improves to 70.84 % regarding H_E. The results improve even further using only electro-laryngeal speech for training (79.31 %; Baseline - E_{E2}). Considering that only 2164 utterances were used in the training this result is so good due to the low perplexity of the language model.

Using speaker MLLR adaptation, results for the healthy-disordered mixed training (EH_{E2}) reach a value of 81.70 %, and increase to 84.76 % for the training with electro-laryngeal speech (E_{E2}) only.

Furthermore, we investigate the case when only little data is available. Connected to that we also show that it is possible to obtain a robust electro-laryngeal model starting from a healthy speech model. For this reasons we apply domain MLLR adaptation of healthy speech to electro-

Speaker ID	**Domain MLLR Adaptation[%]**					
	dH_{E1} – Adapted material					
	F01	M02	F03	M04	M05	M06
F01	78.59	-36.86	0.89	-17.07	-0.91	0.88
M02	10.03	82.72	0.28	8.17	3.10	0.28
F03	4.41	-1.81	16.56	2.61	10.68	0.00
M04	-18.97	-33.33	1.01	82.43	-3.89	-6.72
M05	3.14	-29.92	5.65	-3.48	73.82	0.52
M06	-2.75	0.93	0.00	8.47	2.26	27.29

Table 2.5: Results of ASR for the setup: Domain adaptation using MLLR; dH_{E1} – Training: healthy, Adaptation: to speaker dependent electro-laryngeal domain, Test: electro-laryngeal; values in grey are explained in the text.

laryngeal speech. Within this adaptation EL speech from all speakers is used to adapt the speaker-independent HMM trained with healthy speech. The classic EM algorithm is used to re-estimate the model parameters using the MLLR criterion. With this approach we reach a word accuracy of 79.00 %, which is of the same order as the electro-laryngeal speech model (E_{E2}). Additionally, we applied domain MLLR adaptation to a specific speaker. Looking at the results in Table 2.5 we can see that only the speakers included in the training also perform well in the test. The test utterances are different from the training and adaptation utterances. Also the baseline results of E_{E1} (42.62 %) and E_{E2} (79.31 %) of Table 2.3 confirm this circumstance. These results show that the EL models for ASR are strongly speaker dependent probably due to the different ways to articulate EL speech.

Experiment I, Part B: Standard ASR using MFCC Features and Enhanced EL Speech

For this experiment, we applied two basic enhancement strategies, explained previously. These strategies were tested on the E_{E2} model. Results can be seen in Table 2.6. Electro-laryngeal speech was first enhanced using the two strategies, SS and MF, then the models were trained and tested using these signals (Baseline - E_{E2}, $E_{E2_{SS}}$, $E_{E2_{MF}}$). For the domain MLLR adaptation, we took the enhanced speech utterances to adapt to the healthy speech model (dH_{E2}, $dH_{E2_{SS}}$, $dH_{E2_{MF}}$). We can observe that both enhancement algorithms improve the results regarding the baseline E_{E2}. In general MF outperforms SS because the multi-path approach to reduce the DREL noise reflects the true nature of DREL noise better. For the domain MLLR adaptation results, the changes in the average word accuracy rate are -0.97 % and 2.41 % regarding $dH_{E_{E2}}$. This suggests that the domain adaptation can deal with EL speech as well as enhanced EL speech. Nevertheless, MF enhancement is still beneficial.

Experiment I, Part C: Standard ASR using Different Feature Extraction Techniques

The following experiments originate to a great amount from the joint work with and bachelor's thesis of [Hülser and Rothmund, 2015]. Within the experiments, there were small changes in the database. As explained before two speakers were added to the database. We used 60 utterances per speaker for testing. A main difference, which is also visible in the results, is that the test utterances are the same textual utterances for each speaker. This makes the recognition task easier due to the simpler language model. Note that due to the limited amount of available utterances of speaker M08 all of them are only used in test and not in training.

To evaluate the performance of PLP features in comparison to MFCC features we trained the recognizer with 1) healthy speech only, 2) EL speech only and the 3) combined database as above but without PHONDAT-1 data. We evaluated each training scenario with both healthy

Speaker ID	**Baseline** [%]			**Domain MLLR adaptation** [%]		
	E_{E2}	$E_{E2_{SS}}$	$E_{E2_{MF}}$	dH_{E2}	$dH_{E2_{SS}}$	$dH_{E2_{MF}}$
F01	78.55	89.86	87.25	81.16	88.99	83.48
M02	91.45	93.61	94.72	88.44	88.89	92.50
F03	62.58	62.75	60.78	54.68	53.16	51.85
M04	84.67	85.16	89.54	85.40	83.94	88.70
M05	75.39	83.25	84.03	80.63	78.01	82.46
M06	83.20	82.77	92.17	83.67	75.17	89.49
Average	79.31	82.90	84.75	79.00	78.03	81.41

Table 2.6: Results of ASR for different setups; 1) Baseline and 2) Domain adaptation using MLLR ; E_{E2} – Training: electro-laryngeal, Test: electro-laryngeal (1 - speaker is not included in training, 2 - speaker is included in training); $E_{E2_{SS}}$ – enhanced electro-laryngeal speech using spectral subtraction (SS); $E_{E2_{MF}}$ – enhanced electro-laryngeal speech using modulation filtering (MF); dH_{E2} – Training: healthy, Adaptation: to electro-laryngeal domain, Test: electro-laryngeal; $dH_{E2_{SS}}$ – enhanced electro-laryngeal speech using SS; $dH_{E2_{MF}}$ – enhanced electro-laryngeal speech using MF; values in grey are explained in the text.

and EL speech. All training and testing was performed using all of the three discussed feature extraction methods, namely MFCC, PLP and RASTA-PLP. All speakers except for M08 were included in the training process. In Tables 2.7 and 2.8 W_{Acc} is listed for the different speakers and training scenarios.

Good results can be reported for the matched case (H_H). All three feature extraction methods obtain high average W_{Acc} values: 99.61 % for MFCC features, 99.55 % for PLP features and 98.83 % for RASTA-PLP. For the mismatched case (H_E) the performance decreases significantly, especially for PLP features (0 % – 2.28 %). In this scenario speaker M08 stands out with relatively good 20.66 %. Mediocre recognition rates can be presented for the MFCC features (35.18 %). RASTA filtering improves the performance of the PLP features, but the MFCC features still outperform the RASTA-PLP features. As already reported in Experiment I, Part A, speaker F03 stands out with poor performance. As mentioned before, this speaker has little experience using the EL device and the intelligibility of the resulting speech is lower compared to the other speaker. Again, M02 and laryngectomee M08 are experienced users and the average W_{Acc} are high.

Speaker ID	**MFCC**				**PLP**				**RASTA-PLP**			
	H_H	H_E	E_H	E_{E2}	H_H	H_E	E_H	E_{E2}	H_H	H_E	E_H	E_{E2}
F01	99.54	51.90	11.26	95.89	99.39	1.83	1.98	86.76	99.70	28.92	4.57	84.17
M02	99.85	59.82	7.15	99.54	99.70	1.83	0.91	95.13	99.85	55.40	3.81	91.48
F03	100.00	3.96	11.87	74.58	100.00	0.76	2.44	40.79	98.93	3.96	4.57	51.75
M04	99.85	23.29	38.51	95.89	99.70	0.91	2.28	88.74	99.09	14.00	9.13	82.95
M05	98.17	14.16	4.11	94.82	99.09	0.00	0.91	87.06	98.63	11.72	5.02	80.82
M06	100.00	19.48	3.81	96.19	99.85	2.28	1.98	82.50	99.54	19.33	3.81	72.60
F07	99.85	15.83	13.85	95.74	99.09	2.28	1.52	83.26	96.04	1.07	6.24	77.78
M08	-	57.85	-	78.93	-	20.66	-	40.50	-	31.82	-	19.01
Average	99.61	35.18	12.94	91.45	99.55	3.81	1.72	75.59	98.83	20.78	5.31	70.07

Table 2.7: Results of ASR for different feature extraction methods; H_H and H_E – Training: healthy, Test: healthy, electro-laryngeal; E_{E2} – Training: electro-laryngeal, Test: electro-laryngeal; (2 - speaker is included in training); values in grey are explained in the text.

The recognition rates for the matched case (E_{E2}) are satisfying for the MFCC features with a W_{Acc} of over 90 %. The performance decreases for PLP features (75.59 %) and even more when applying RASTA filtering (70.07 %). In this experiment we added the mismatched case (E_H) which shows insufficient results for all three feature extraction methods, especially for the PLP features (1.72 %). In Table 2.8 results are presented for the scenario when HE as well as EL speech is used for training. Good results are obtained when testing on healthy speech. These values are comparable to those of Table 2.7 where only healthy data was used for training. Testing on electro-larynx speech decreases the rates only slightly: 87.44 % using MFCC and 83.26 % using PLP. Again, RASTA filtering does not improve the performance of PLP feature extraction.

Speaker ID	**MFCC**		**PLP**		**RASTA-PLP**	
	EH_H	EH_{E2}	EH_H	EH_{E2}	EH_H	EH_{E2}
F01	99.09	95.28	98.48	91.93	97.56	84.93
M02	98.78	97.56	99.24	98.17	98.02	94.22
F03	99.54	63.77	100.00	54.64	93.46	45.21
M04	100.00	95.28	100.00	93.00	97.56	84.78
M05	98.93	94.52	99.09	90.11	93.91	81.13
M06	99.54	90.56	99.54	90.87	98.93	79.91
F07	98.02	96.04	99.54	79.15	91.32	77.17
M08	-	66.53	-	68.18	-	35.95
average	99.13	87.44	99.41	83.26	95.82	72.91

Table 2.8: Results of ASR for different feature extraction methods; EH_H and EH_{E2} – Training: mixed healthy and electro-laryngeal, Test: healthy, electro-laryngeal; (2 - speaker is included in training); values in grey are explained in the text.

For all training scenarios the speaker-independent case (M08) is particularly interesting. Although speaker M08 was not included in training, the performance is remarkable. For the matched case (Table 2.7 – E_{E2}) W_{acc} is 78.93 %, which is around 20 % lower compared to the other speaker (except F03). However, for the mismatched case (Table 2.7 – H_E), M08 outperforms the other speakers by around 40 % when using the PLP features. This might be due to the fact that this speaker is a laryngectomee and uses the EL device as the main means of communication and. Although he used the same kind of device (Servox Digital), he used his own device. All other speakers in the database used exactly the same device.

2.3 Summary and Discussion

Within this chapter we introduced the German parallel ELHE speech database in Section 2.2. We recorded 7 healthy subjects who were able to produce the phonetically rich utterances with their healthy voices as well as with the conventional EL device. To analyze the database we proposed a strategy to calculate the SNR. This strategy takes the nature of EL speech into account, which means that there are three levels to differentiate: level of speech (SL), level of directly radiated background noise (DREL-L) and level of noise/silence. We define the SNR of EL speech as the ratio between SL and DREL-L. This gave us values between 8 dB and 20 dB, depending on the quality of the produced EL speech. The SNR of a laryngectomee was 17.79 dB which confirms the relation between quality of the produced EL speech and SNR values. Furthermore, we assume a relation between the absolute f_0 of the device and SNR. If f_0 is low, resulting speech is louder and thus, SNR reaches higher values. The HE speech utterances reached SNR values of around 50 dB in clean conditions.

Furthermore, we evaluated our database using ASR. For the first experiment we obtained very good results with EL speech in training and with adaptation of HE speech models to fit EL data (word accuracy rates around 80 %). Our database is dominated by male speakers. This might be a reason for the performance of the female speakers, which was lower.

In a second experiment we evaluated two different enhancement algorithms for EL speech, in terms of DREL noise suppression: spectral subtraction (SS) and modulation filtering (MF). For both algorithms the word accuracy rates increased with clear advantage for MF. This suggests that ASR can be used to evaluate the intelligibility of enhanced electro-laryngeal speech.

Moreover, we compared different feature extraction methods without DREL suppression in a third experiment. The used features are: MFCC features and PLP features with and without RASTA processing. Generally MFCC features led to higher recognition rates than PLP. The idea was that their might be other features which are optimal for EL speech and RASTA-PLP eliminates the effect of the DREL to some extent, but it seems that not enough speech related information is left back. While comparable results can be achieved for matched training and testing, the performance of PLP features significantly decreases for mismatched cases. Also a RASTA filtering did not consistently improve the recognition performance, although slightly better results were obtained for the mismatched cases. Even though there is a great mismatch between the two domains, MFCC and PLP feature can achieve relatively high word accuracy rates when both, healthy speech as well as EL speech, are used in training.

For all ASR experiments it must be mentioned that the language model has huge influence on the recognition results. Whereas in the first two experiments the test utterances were different for each speaker, the test utterances had the same text in the third experiment. Therefore, the perplexity of the language model decreased and recognition results increased.

One important conclusion is that the recognition results for electro-laryngeal speech are high. The reasons for this are that people tend to articulate very clearly in order to be understandable and that the stationary noise of the electro-larynx device can be modeled in ASR training. Another important conclusion is that the EL model for ASR strongly depends on the speaker and with a speaker dependent MLLR adaptation strategy, electro-laryngeal ASR results are nearly as high as for healthy speech. Although there is a large mismatch between the two domains, as soon as we include electro-laryngeal speech in the training, the ASR system performs well. If some pre-processing is done, EL users can make use of ASR technologies. We stated that ASR can be used to evaluate the quality of EL speech and enhanced EL speech in terms of intelligibility but did not confirm this using listening tests. Although we applied EL speech in an ASR system the connection between word accuracy rate and perceived intelligibility still remains an open question. Although the results are very promising from the current point of knowledge it is still not clear if ASR can be used as an evaluation tool for the quality of (enhanced) EL speech.

3 Control of Artificial Excitation Signal

3.1 Introduction

This chapter focuses on how to control the artificial excitation source, e.g., the device itself. In particular, we will focus on two main problems: the constant fundamental frequency f_0 and the unhandiness of occupying one hand to operate the device. The shape of the excitation signal has huge influence on the quality of the resulting EL speech. We suggest different kinds of waveform models and analyze their frequency distribution. Afterwards, an f_0 estimation algorithm based on statistical models is proposed. In the context of the estimation algorithm it is important to time aligning the features of EL speech and HE speech. The time alignment is crucial for the performance of the estimation strategy. We introduce objective measures to evaluate the estimated f_0 contours and show results based on our speech data. Listening tests are carried out in order to evaluate our proposed estimation algorithms. Furthermore, we show the importance of a changing f_0 contour in terms of naturalness which confirms the work of other researches. The last part of this chapter deals with an electromyography (EMG) based approach to control the on/off signal of the device. We investigate different approaches to obtain an on/off message for the EL device and confirm the feasibility within a real-time system. The real-time system is also used to gain insight into learning effects.

3.2 Waveform Shape and Fundamental Frequency

In this section, we estimate an artificial f_0 contour for EL speech using a machine learning approach. The goal of this approach is to improve EL speech in terms of naturalness and intelligibility by introducing variations in the f_0 contour, which means changing the fundamental frequency of the EL excitation signal. The approach for such a variation needs to be of low complexity in order to implement it on a real-time platform. A strong requirement is that the learning procedure is automatic and that users do not need to actively change speaking parameters. Therefore, our intention is to perform an estimation of the f_0 contour using Gaussian mixture models (GMM) as well as Hidden Markov models (HMMs). Furthermore, the waveform shape of the excitation signal needs to be selected, too, as will be discussed in the following subsection.

3.2.1 Shape of Excitation Signal

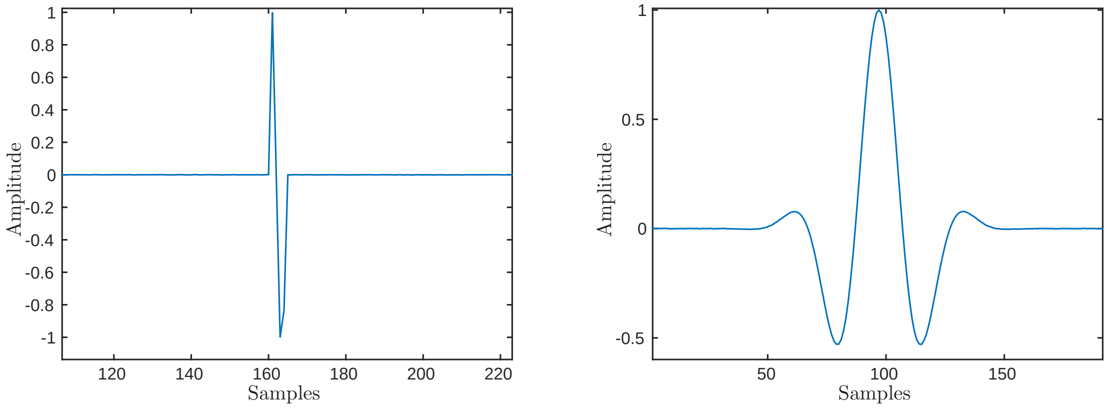
As mentioned earlier the conventional EL device is not designed to dynamically change the waveform of the excitation signal (or even f_0). In order to test and evaluate f_0 estimation strategies we need to set up an experimental prototype which is able to output the estimation results. The excitation signal of the conventional Servox Digital device strongly depends on the mechanics of the device. We also have to consider which type of excitation signal corresponds best to a healthy excitation signal and whether the excitation signal can be optimized in terms of a newly proposed system. We implemented different types of excitation signal which are described in the following. These prototypes were additionally superimposed with white noise to create random fluctuations of the amplitude (shimmer). Further investigation on the artificial excitation signal should include random variation of the frequency as well. The white noise was added to the prototype. Therefore, the white noise was repeated periodically which should be investigated and if necessary improved.

The Pulse Model

The *Pulse* excitation signal is obtained using the prototype vector [1, -1] and is filled up with zeros until the next frequency peak (see Figures 3.1(a) and 3.3).

The Gauss Pulse Model

This model consists of a Gaussian modulated sinusoidal pulse. It is approximately narrow-band, because frequencies above 1000 Hz are heavily damped (see Figure 3.1(b)).



(a) Time signal of a prototype *Pulse*.

(b) Time signal of a prototype *Gauss Pulse*.

Figure 3.1: Example prototypes of different artificial excitation sources.

Liljencrants-Fant (LF) Model

The *LF* model was proposed by G. Fant and J. Liljencrants in 1985 [Fant et al., 1985] and is a parametric representation of the glottal flow derivative $\dot{U}_g(t)$ in the time domain. The two parts of the equation (3.1) characterize the glottal opening from zero to the maximum negative amplitude and the glottal closing phase.

$$\dot{U}_g(t) = \begin{cases} E_0 e^{\alpha t} \sin(\omega_g t), & \text{for } 0 \leq t \leq t_e \\ -\dfrac{E_e}{\varepsilon t_a} \left[\mathrm{e}^{-\varepsilon(t-t_e)} - \mathrm{e}^{-\varepsilon(T_0-t_e)} \right], & \text{for } t_e < t < T_0 \end{cases} \tag{3.1}$$

where $w_g = \pi/t_p$. One cycle with duration equal to the fundamental period T_0 of the *LF* model is defined by four parameters (see Figure 3.2). One parameter corresponds to the overall amplitude (E_e). The three timing parameters concern the spectral content of the pulse (t_p, t_e, t_a). The parameters α, E_0 and ε are derived from 1) the zero energy balance and 2) amplitude continuity constraint. This easy and representative model of the vocal fold movement is very popular and widely used.

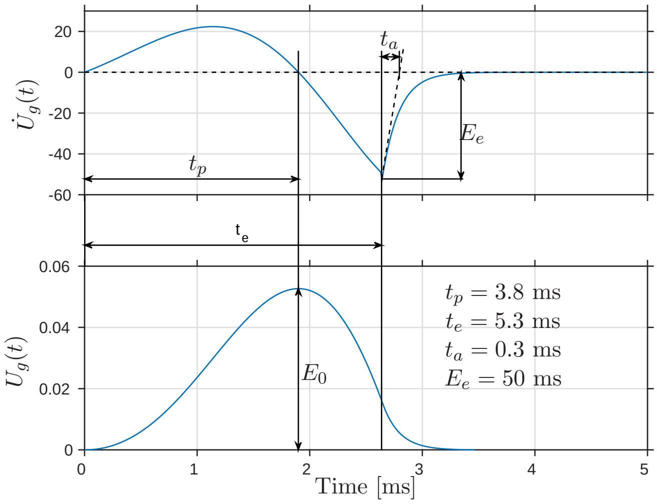


Figure 3.2: Waveforms generated by the Liljencrants-Fant model, with the four wave-shape parameters: t_p instant of maximum glottal flow, t_e instant of flow derivative negative peak, t_a time of exponential closure, and E_e absolute value of glottal flow derivative at t_e (based on [Veldhuis, 1998]).

The Hanquinet-Grenez-Schoentgen (HGS) Model

The *HGS* model was introduced by J. Hanquinet, F. Grenez and J. Schoentgen as a phonatory excitation model particularly suitable for the synthesis of disordered speech [Hanquinet et al., 2006]. [Jochum and Reiner, 2008] summarized the model as follows: It utilizes a shaping function to transform a trigonometric driving function into a desired waveform whereby the amplitude and the fundamental frequency of the driving function are used to control the instantaneous frequency and the spectral richness of the output signal independent from each other. While the driving function is represented by a cosine function, the shaping function is defined as an equivalent polynomial formulation of the Fourier series,

$$U_g(t) \approx \frac{1}{2}a_0 + \sum_{m=1}^{\tilde{M}} a_m A^m \cos(m\Theta_t) + b_m A^m \sin(m\Theta_t) \tag{3.2}$$

which is truncated after $\tilde{M}$ harmonics.

While A is used to modify the Fourier coefficients a_m and b_m to influence the spectral richness of the synthetic source signal and is limited by $0 \leq A \leq 1$, the phase Θ_t defines its instantaneous fundamental frequency. The possibility to vary these two parameters asynchronously and continuously make this particular model very flexible in terms of jitter, micro-tremor and shimmer implementations.

The FFT of the four different types of excitation signal can be seen in Figure 3.3. The properties of the spectra are very different. The artificial *Pulse* excitation is broad-band with most acoustic energy around 3000 Hz. However, the sound is very sharp and quite annoying. The artificial *Gauss Pulse* excitation is narrow-band and acoustic energy is concentrated in the low frequency domain. The sound is very soft, but higher frequencies are missing. This makes

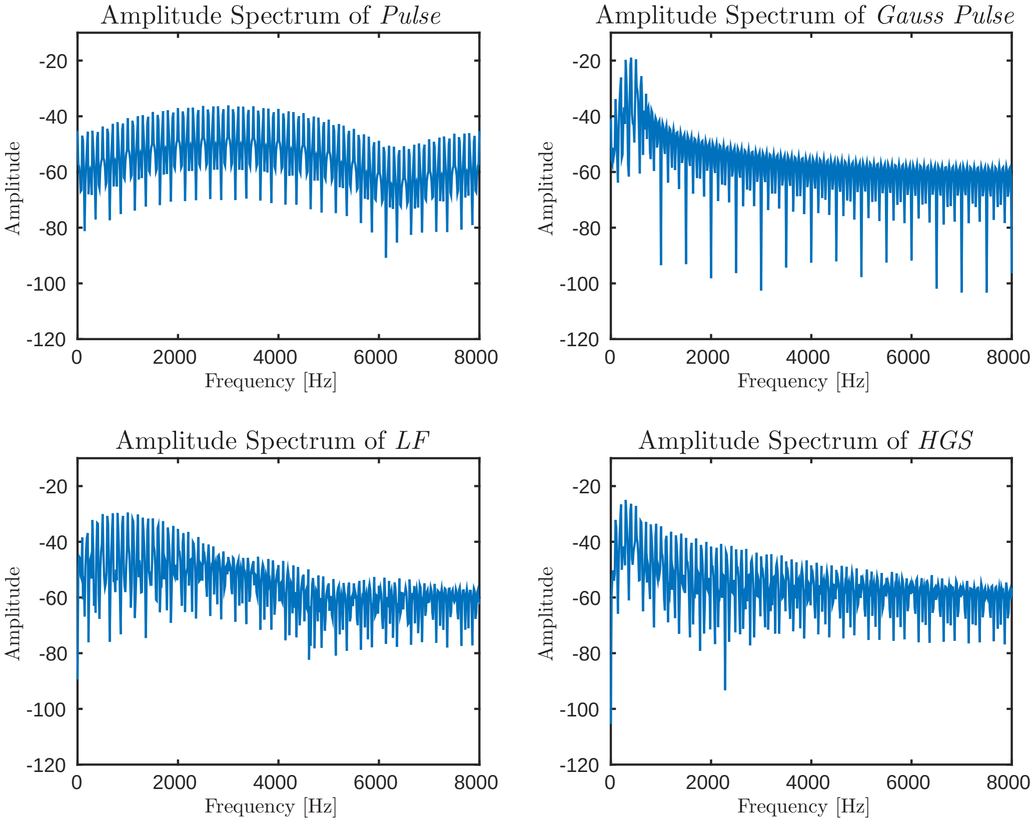


Figure 3.3: Fourier transform of different artificial excitation sources.

a negative impact on the perceived loudness of the excitation signal. Artificial *LF* and *HGS* are similar in their spectral structure. In future the waveform of the excitation signal needs to be designed with much care. In Chapter 4 the neck transfer function and the transducer transfer function of the EL device are analyzed and compensated. Furthermore, extensive listening tests must confirm the pleasantness and the ability to appropriately model the gender of the resulting speech. We will further discuss the optimal excitation signal for the proposed EL device in Section 4.5.

3.2.2 Error Measures of f_0 Estimation

Objective evaluation of the excitation signal is not straightforward. Standard measures for speech enhancement algorithms are segmental SNR [Hansen and Pellom, 1998], PESQ [ITU, 2001], LPC-based objective measures including the log-likelihood ratio, Itakura-Saito distance measure, frequency-weighted segmental SNR [Tribolet et al., 1978]. These measures are only hardly useful for evaluating EL speech in terms of improvements of the excitation signal. Especially f_0 contours are challenging because we are not only interested in how close the contours are to healthy contours but also how the estimated contours sound with speech. Minor differences in the f_0 contours in general do not affect the perception. Within these error measures values need to be compared to reference values and some kind of distortion is calculated. We use the following objective measures in order to be able to compare f_0 contours:

(Pearson's linear) Correlation Coefficient (ρ)

The correlation coefficient is based on the standard Pearson's linear correlation coefficient:

$$\rho = \frac{\sum_{n=1}^{N} \left(f_0(n) - \mu_{f_0}\right)\left(\hat{f}_0(n) - \mu_{\hat{f}_0}\right)}{\sqrt{\sum_{n=1}^{N}(f_0(n) - \mu_{f_0})^2}\sqrt{\sum_{n=1}^{N}(\hat{f}_0(n) - \mu_{\hat{f}_0})^2}}. \tag{3.3}$$

N is the number of f_0 values in the contour. For calculating ρ the reference f_0 contour is compared to the estimated f_0 ($= \hat{f}_0$) contour for voiced frames of the reference f_0 contour. Figure 3.4 (first plot) shows an example of f_0 in blue and $\hat{f}_0$ in red ($\rho = 0.01$). In the second plot all unvoiced frames of f_0 are neglected. Thus, the blue line is a concatenation of all values of f_0 which are not zero. $\hat{f}_0$ is plotted at the same frames. ρ has the same value than before because the same values are taken to calculate the correlation coefficient. As can be seen, there are voiced-to-unvoiced errors for $\hat{f}_0$ (all unvoiced-to-voiced errors are discarded). To obtain reasonable values for ρ, $\hat{f}_0$ is interpolated in order to provide a smooth trajectory. Thus, ρ reaches a value of 0.28. This measure captures the correlation between the reference f_0 and $\hat{f}_0$ for voiced frames only where μ_{f_0} and $\mu_{\hat{f}_0}$ are the mean value of the fundamental frequency contour of reference speech and estimated speech, respectively.

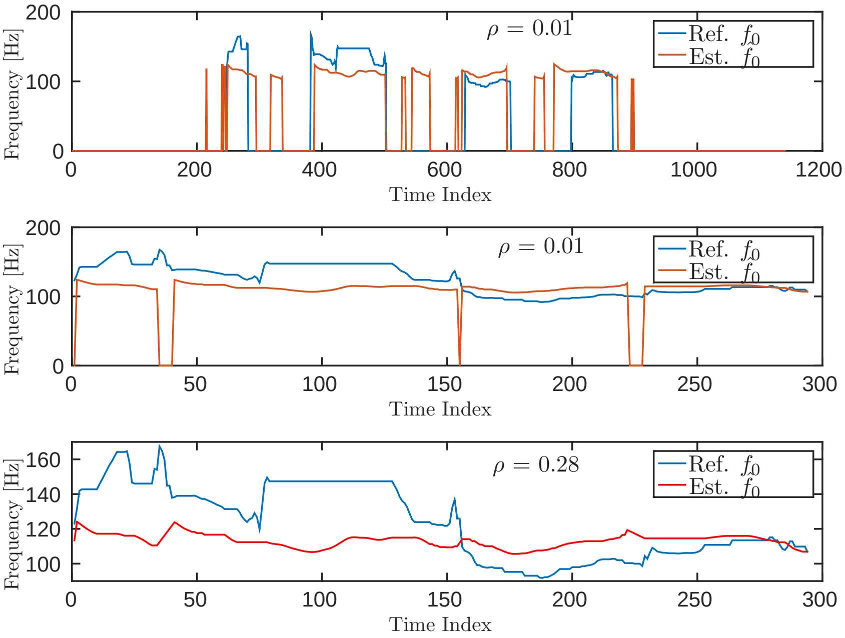


Figure 3.4: Calculation of correlation coefficient ρ; top: f_0 contours to compare, middle: values for voiced frames of reference (Ref.) f_0, down: interpolation of estimated (Est.) f_0.

In our experiments we reached values for ρ up to 0.3. To verify that this correlation is not random we carried out following experiment: We took the fundamental frequency from healthy speech (reference) and calculate ρ compared to a sequence of random numbers. Repeating this experiment n times and investigating min and max values we can see that ρ can take values between ±0.3, but most of the values are around 0, in fact only around 5 % are outside the interval ±0.1. For $\hat{f}_0$ most values are above ±0.1 (see Figure 3.5).

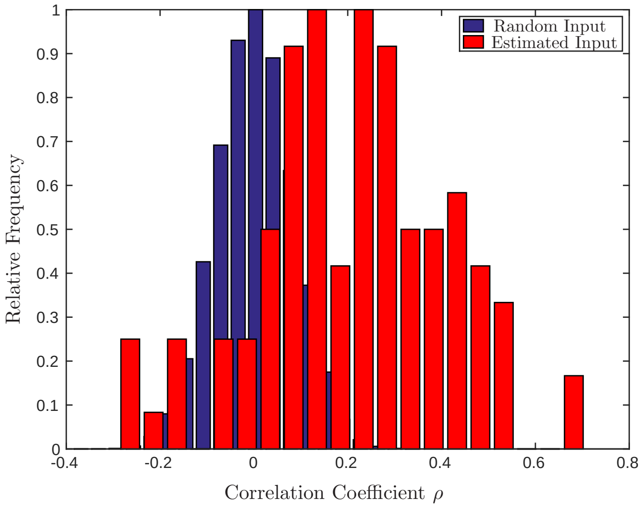


Figure 3.5: Histogram of the correlation coefficient ρ between random f_0 values and healthy f_0 (blue) and $\hat{f}_0$ values and healthy f_0 (red).

(Root) Mean Square Error – (R)MSE

The (root) mean square error is the simple Euclidean distance between f_0 and $\hat{f}_0$ at frame n.

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(f_0(n) - \hat{f}_0(n))^2}{N}} \tag{3.4}$$

Human beings can only discriminate two frequencies if the differences are bigger than three semitones [Hart, 1981]. Therefore, (R)MSE measures are often presented in semitones. For evaluation f values of the f_0 contours are transformed into semitone values according to:

$$f_{st} = 69 + 12 \cdot \log_2\left(\frac{f}{440}\right). \tag{3.5}$$

Gross and Fine f_0 Error

Text-to-Speech technologies focus on producing a natural prosody. From this field of research we know that besides listening tests, measures like correlation coefficient ρ and (R)MSE, also the so-called gross and fine f_0 errors are used to evaluate the voiced f_0 values. These measures are believed to provide to gain a deeper insight into the error [Rabiner et al., 1976]. To calculate them we differentiate between errors which are bigger or smaller than a certain threshold δ.

$$\hat{f}_{0,corr} = |\hat{f}_0 - f_0| < \delta \tag{3.6}$$

$$\hat{f}_{0,incorr} = |\hat{f}_0 - f_0| > \delta \tag{3.7}$$

The gross f_0 error is the ratio of incorrect estimated frames to all frames [Nakatani and Irino, 2004].

$$\text{Gross } f_0 \text{ error} = \frac{N_{\hat{f}_{0,incorr}}}{N_{f_0}} \tag{3.8}$$

where N_{f_0} is the number of frames with both, $\hat{f}_0$ and f_0 consider to be voiced, $N_{\hat{f}_{0,incorr}}$ is the number of frames for which (3.7) holds.

The fine f_0 error is the normalized RMSE between correct estimated f_0 values $\hat{f}_{0,corr}$ and reference f_0 values [Nakatani and Irino, 2004].

$$\text{Fine } f_0 \text{ error} = \sqrt{\frac{\sum_{n=1}^{N}\left(\frac{\hat{f}_{0,corr}(n)-f_0(n)}{\hat{f}_{0,corr}(n)}\right)^2}{N}} \tag{3.9}$$

For our experiments we set δ to 0.05. Furthermore, we used semitone f_0 values.

3.2.3 Experiment II: f_0 Estimation based on GMMs

In the literature different approaches have been proposed to increase naturalness of EL speech using a changing fundamental frequency. Bio-signals like EMG [Goldstein et al., 2004] or air pressure sensors [Uemi et al., 1994] can be used. A changing f_0 can also be produced using a rule based method based on the audio signal [Saikachi et al., 2009]. We investigate the power of voice conversion procedures. Our work is inspired by the idea of [Giet, 1956] who proposed that although whispered speech is unvoiced by its nature, humans can perceive an absolute tone of the whispered words. We assume that the vocal tract carries information about f_0 and learn this relation using mapping strategies known from voice conversion algorithms which are based on GMMs.

We establish a method which does not require manual control of the f_0 contour and an action by the user. Therefore, an automatic procedure for adjusting an artificial f_0 contour is a considerable advantage.

GMM based f_0 Estimation Approach

Basically statistical models (GMMs) are matched to static input. This means that features are calculated on a frame-by-frame basis. However, this does not correspond to the nature of speech, because phonemes are not independent but strongly depend on the phoneme before and afterwards (co-articulation). This fact can be considered using dynamic first or higher-order differences ($\boldsymbol{\Delta}$ features). Mixture models are probabilistic models. They are useful to model arbitrary probability distributions. A GMM consists of a weighted linear combination of K multivariate Gaussian probability density functions given by

$$P(\mathbf{z}_t) = \frac{1}{\sqrt{\det(\boldsymbol{\Sigma})}(2\pi)^{D/2}} \mathrm{e}^{\left(-\frac{1}{2}(\mathbf{z}_t-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z}_t-\boldsymbol{\mu})\right)}, \tag{3.10}$$

with D-dimensional mean value vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The mathematical representation is given by

$$P(\mathbf{z}_t|\lambda) = \sum_{k=1}^{K} b_k\mathcal{N}(\mathbf{z}_t;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = \sum_{k=1}^{K} b_k c_k. \tag{3.11}$$

K is the number of Gaussian components, $\mathbf{z}_t$ is an observation of a D-dimensional random vector at frame t, b_k are the weights for each component and c_k is a single, multivariate Gaussian distribution. For b_k the assumption $\sum_{k=1}^{K} b_k = 1$ holds. With λ= $\{(b_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k); k = 1,2,\ldots,K\}$, the whole mixture model is described. The parameters λ of a model are estimated using the expectation-maximization (EM) algorithm [Bishop, 2007]. The estimation of an artificial f_0 contour using GMMs is, amongst others, inspired by [Milner and Shao, 2007] who used mel-frequency cepstral coefficients (MFCCs) for prediction of f_0 and voicing in unconstrained speech. The input for training the GMM is the joint feature vector $\mathbf{z}_t = [\mathbf{x}_t, y_t]^T$. $\mathbf{x}_t$ contains the MFCCs for one frame for EL speech and y_t the corresponding f_0 for healthy speech. To create $\mathbf{x}_t$ and y_t, static and dynamic, e.g., Δ features, are taken into account. Furthermore, subsequent feature vectors are stacked to capture the behavior in time. For this approach full covariance matrices $\mathbf{\Sigma}_k$ need to be trained. The mean value vector $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ can be seen as a concatenation:

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_k^x \\ \boldsymbol{\mu}_k^y \end{bmatrix}, \quad \mathbf{\Sigma}_k = \begin{bmatrix} \mathbf{\Sigma}_k^{xx} & \mathbf{\Sigma}_k^{xy} \\ \mathbf{\Sigma}_k^{yx} & \mathbf{\Sigma}_k^{yy} \end{bmatrix}.$$

In order to estimate f_0, the nearest component to an unknown $\mathbf{x}_{in}$ is calculated: $m^* = \arg\max_K \; P(\mathbf{x}_{in}|c_m) \cdot b_m$, where $P(\mathbf{x}_{in}|c_m)$ is the marginal distribution of the MFCC vector for the m-th cluster c_m. The final result is obtained through

$$\hat{f}_i = \boldsymbol{\mu}_{m^*}^f + \mathbf{\Sigma}_{m^*}^{fx} \cdot (\mathbf{\Sigma}_{m^*}^{xx})^{-1}(\mathbf{x}_{in} - \boldsymbol{\mu}_{m^*}^x) \tag{3.12}$$

which is the MAP solution.

Dynamic Time Warping

In order to create the input vectors for training, spectral features need to be extracted from EL speech and f_0 from HE speech. These utterances are not of the same length. Therefore, dynamic time warping (DTW) needs to be introduced. We used a dynamic time warping method based on dynamic programming as suggested by [Ellis, 2003]. Although it has been largely superseded by hidden Markov models, early speech recognizers used dynamic programming to accommodate differences in timing between sample words and templates.

Within DTW a sample utterance should be aligned to a template utterance. In order to do so the local matches between STFT magnitudes are calculated using the inner product between two vectors. Dynamic programming is then used to find the lowest-cost path between the opposite corners of the cost matrix. Within this search the minimum-cost alignment gives information on how well the sample matches the template. In other words, we find a path through the cost matrix of EL and HE spectral features that maximizes the local match between the aligned frames. The basic principle is to allow a range of steps in the space (time frames in sample, time frames in template) and to find the path through that space that maximizes the local match between the aligned time frames. The total 'similarity cost' found using this algorithm is a good indication of how well the sample and template match, which can be used to choose the best-matching template. For this algorithm we use the spectrograms of EL and HE speech. As these spectrograms are very different the algorithm fails in many cases. Especially DREL of the device disturbs the cost matrix due to the high spectral energy. If the EL device starts to radiate noise before speech or ends after speech, this can lead to serious performance problems. Therefore, EL speech needs to be enhanced in terms of DREL, before performing DTW. We used the enhancement technique MF to remove DREL described in Section 2.2.2.

To verify the statement above the following experiment was carried out: We performed DTW and calculated the minimum cost, one time between HE and EL and a second time between HE and enhanced EL speech. Comparing the results per utterance we confirm that the minimum cost is smaller (except for M05) when we aligned HE and enhanced EL (see Table 3.1 and Figure 3.6).

Experiment II, Part A: Preliminary Results

Preliminary experiments were carried out on a subset of the German parallel ELHE database explained in Section 2.2 (100 utterances of speakers F01 and M02).

The pre-processing step, as shown in Figure 3.7, re-sampled the data (HE and EL speech) to a sampling frequency of f_s =16000 Hz. A high pass filter (HPF) removed DC and very low-frequency components and DREL noise was removed from the EL speech using a modulation

Speaker	DTW cost (HE-EL)	DTW cost (HE-EL$_{enh}$)
F01	0.83±0.08	0.76±0.08
M02	0.79±0.08	0.73±0.08
F03	0.87±0.11	0.85±0.12
M04	0.93±0.08	0.85±0.09
M05	1.01±0.09	1.01±0.10
M06	0.88±0.09	0.84±0.09
F07	0.87±0.11	0.89±0.11

Table 3.1: Minimum DTW cost for each speaker between healthy speech and electro-larynx speech (HE-EL) and healthy speech and enhanced electro-larynx speech (HE-EL$_{enh}$).

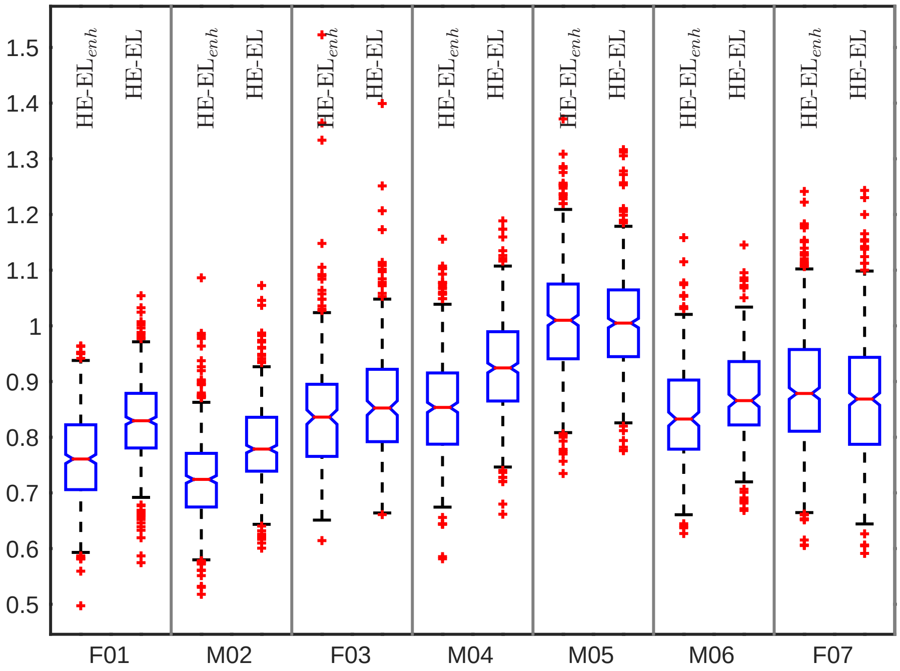


Figure 3.6: Boxplot of minimal DTW cost between healthy speech and electro-larynx speech (HE-EL) and healthy speech and enhanced electro-larynx speech (HE-EL$_{enh}$) for each speaker.

filtering technique [Hagmüller, 2009] (25 ms frame length and 300 Hz frame rate). In the training step 28 MFCCs (including 0-th order cepstral coefficient, the log energy and Δ coefficients) features were extracted from the EL speech using a Hann window with a frame length of 50 ms. 23 filters in the filter bank were used. The healthy f_0 of the HE speech was tracked using the autocorrelation algorithm provided by the Praat speech analysis software [Boersma and Weenink, 2007]. The features and the f_0 values were time aligned using DTW [Ellis, 2003] because people need more time to articulate with an EL than with healthy voice. Then, the frame-wise extracted MFCC feature vector concatenated with the f_0 value for the corresponding frame created the joint feature vectors with the dimension $D = 29$ as described previously. Afterwards, GMMs with full covariance matrix and $K = 16$ components were used for training. The statistical model is speaker-dependent. In the test scenario, the MFCC features from an unknown EL input utterance were calculated in the same way as in the training scenario. Afterwards, a voiced/unvoiced decision was carried out and the f_0 value was estimated only for the voiced

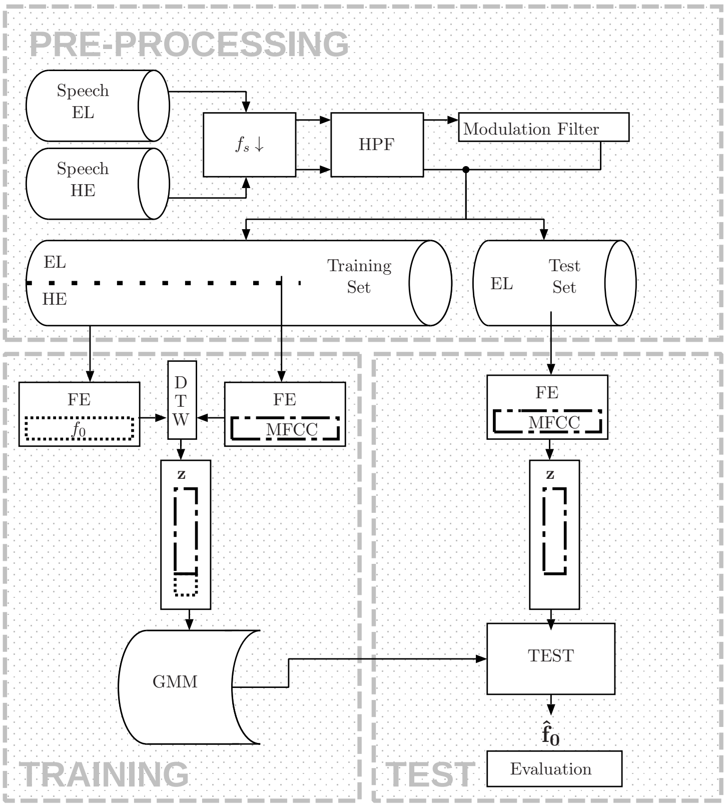


Figure 3.7: Schematic overview of experimental setup: PRE-PROCESSING with downsampling to f_s = 16000 Hz, high pass filter (HPF) and removal of directly radiated noise (modulation filter framework); TRAINING and TEST with feature extraction (FE), dynamic time warping (DTW), concatenation of $\mathbf{z} = [\mathbf{x}, y]^T$.

frames according to the method described in the previous section. For evaluation resulting f_0 contours were compared to HE f_0 contours.

Figure 3.8 shows an example of speaker M02. ρ compares the HE f_0 contour and the original

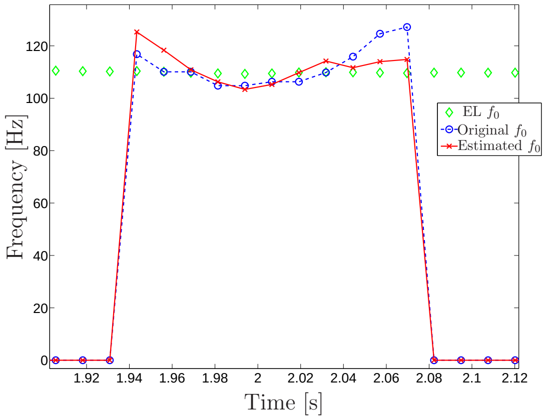


Figure 3.8: One voiced phrase of speaker M02: ◊ – original constant f_0 from EL (without processing); f_0 from healthy speech (o); $\hat{f}_0$ (×).

(= constant) f_0 contour for the EL speech (ρ_{orig}) on the one hand, and the artificial f_0 contour from the enhanced speech (ρ_{est}) on the other hand (enhanced in terms of f_0 estimation). For one arbitrary utterance from the male speech database ρ_{orig} is 0.41 and the enhanced speech reaches a very good value of $\rho_{est} = 0.91$.

For this experiment we took the absolute mean error $|\epsilon|$ between reference f_0 contour and $\hat{f}_0$ of length N. It is calculated as

$$|\epsilon| = \frac{\sum_{t=1}^{t=N} |f_0(t) - \hat{f}_0(t)|}{N}. \tag{3.13}$$

For evaluation frequency values of the f_0 contours are transformed into semitones according to (3.5).

For the above utterance, the absolute mean error between the healthy and the original EL f_0 contour is $|\epsilon|_{st,orig} = 12.2$ and decreases to $|\epsilon|_{st,est} = 0.7$ for $\hat{f}_0$.

Furthermore, we used 4-fold cross validation. The training data was randomly split into four sets. Three parts were then used in training and the fourth for validations. Then, three others were picked and so on. Following this, results could be obtained separately for speakers F01 and M02. ρ and $|\epsilon|$ were calculated for each utterance and the mean value over all utterances was determined. The results for ρ are shown in Table 3.2. Figure 3.9 shows the boxplot of $|\epsilon|$.

	M02	F01
ρ_{orig}	0.29	0.23
ρ_{est}	0.92	0.90

Table 3.2: Correlation coefficient obtained from 4-fold cross validation.

The values are listed in Table 3.3. It can be seen that the improvement in terms of the absolute mean error is significant.

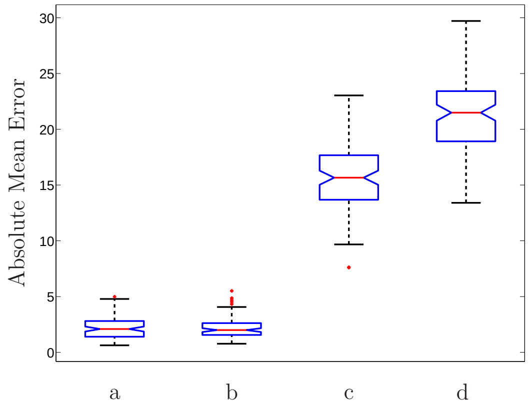


Figure 3.9: Absolute mean error of 4-folds cross validation for M02 (a and c) and F01 (b and d): a and b: $|\epsilon|_{st,est}$ – error between healthy and estimated f_0 contour; c and d: $|\epsilon|_{st,orig}$ – error between healthy and original f_0 contour.

Additionally, the EL speech was manipulated with the $\hat{f}_0$ contour using the Pitch synchronous overlap and add (PSOLA) method [Moulines and Charpentier, 1990], which results in an improved listening quality compared to original EL speech. Note that the above reported correlation coefficients ρ were calculated using all frames instead of voiced frames only. This increased ρ.

Experiment II, Part B: Extended Results

In this extended evaluation we took the recorded speech material from the German parallel ELHE database explained in Section 2.2. Speakers F03 and M06 recorded less data than the other speakers. 380 (130) utterances were taken for training and the same 100 utterances for each speaker were taken for test. The previously explained experimental setup stays the same

	$q1$	$q2$	$q3$
	M02		
$\lvert\epsilon\rvert_{st,est}$	1.40	2.09	2.81
$\lvert\epsilon\rvert_{st,orig}$	13.68	15.66	17.69
	F01		
$\lvert\epsilon\rvert_{st,est}$	1.56	1.99	2.62
$\lvert\epsilon\rvert_{st,orig}$	18.92	21.49	23.42

Table 3.3: 25th (q1), median (q2) and 75th (q3) percentiles of 4-folds cross validation for database; $|\epsilon|_{st,est}$ – error between healthy and estimated f_0 contour; $|\epsilon|_{st,orig}$ – error between healthy and original f_0 contour.

except for the frame length which depends on the gender in this experiment. 40 ms for male speakers and 25 ms for female speakers. We changed the frame length because the frame length in the f_0 extraction algorithm depends on the minimum frequency we would like to extract. This, however, depends on the gender because in our database f_0 of female speakers is higher than f_0 of male speakers. In this experiment the frame length and frame rate was also changed for MFCC calculation to match the parameters from f_0 extraction. EL speech is always voiced, i.e., when the EL device is turned on there is always voiced speech. We assume that EL speech will degrade if we distinguish between voiced and unvoiced phonemes in the same way as in HE speech. Therefore, we changed the training of the GMM from Figure 3.7 and took voiced as well as unvoiced frames for training. We interpolated f_0 contours and set f_0 not to zero but to a constant value. For reference, we carried out the experiment based on HE speech only, which means that we extracted spectral features from HE speech and not from EL speech. Also f_0 values are taken, as usual, from the HE utterance. As a result there was no need to apply DTW. Results are presented in Table 3.4. In the main experiment, we changed from HE spectral features to spectral features extracted from EL speech. This results in the values of Table 3.5.

	Speaker ID						
	F01	M02	F03	M04	M05	M06	F07
ρ	0.52±0.16	0.40±0.21	0.55±0.17	0.52±0.19	0.43±0.25	0.53±0.19	0.61±0.21
$\sigma^2_{\hat{f}_0}$	11.48	4.61	11.93	11.34	9.08	7.98	18.66
$\sigma^2_{f_0}$	23.27	11.57	23.05	22.36	21.64	16.26	27.08
MSE ($\hat{f}_0$-f_0)	3.73±1.82	3.58±1.86	4.00±2.11	6.09±2.33	6.74±4.78	4.08±2.38	3.38±2.27
Gross $\hat{f}_0$ error	0.14	0.22	0.17	0.31	0.32	0.23	0.10
Fine $\hat{f}_0$ error	1.85	1.22	1.72	1.27	1.06	1.14	1.68

Table 3.4: Results of correlation coefficient (mean value and variance), MSE and Gross/Fine error for f_0 estimated from HE files.

From the results we can conclude that:

- The correlation coefficient ρ is much higher (up to 0.61) for features extracted from healthy speech than for for features extracted from EL speech.
- The variance of the estimated fundamental frequency $\sigma^2_{\hat{f}_0}$ is smaller than for the healthy f_0. This can be seen in both experiments.
- Gross and Fine f_0 error is for both experiments approximately the same.
- The results of Table 3.5 correspond to the SNR results of Table 2.1 except for speaker M05.

	Speaker ID						
	F01	M02	F03	M04	M05	M06	F07
ρ	0.20±0.19	0.23±0.23	0.16±0.20	0.24±0.20	0.10±0.26	0.22±0.26	0.08±0.19
$\sigma^2_{\hat{f}_0}$	6.31	3.62	5.42	7.32	4.93	6.10	3.82
$\sigma^2_{f_0}$	27.40	13.35	22.21	24.03	21.51	16.21	27.21
MSE ($\hat{f}_0$-f_0)	6.18±2.82	5.49±4.32	5.49±3.25	9.48±3.39	9.19±4.25	6.34±3.46	6.30±3.60
Gross $\hat{f}_0$ error	0.27	0.29	0.25	0.44	0.45	0.34	0.26
Fine $\hat{f}_0$ error	2.11	1.33	1.96	1.40	1.24	1.29	2.10

Table 3.5: Results of correlation coefficient (mean value and variance), MSE and Gross/Fine error for f_0 estimated from EL files.

Experiment II, Part C: Subjective Results

Additionally, we conducted a listening test which compared the pleasantness of healthy f_0 to the pleasantness of $\hat{f}_0$ without enhancement strategy. In this experiment speakers directly produced EL speech using pre-determined f_0 patterns. The f_0 patterns are created as follows: On the one hand we took HE f_0 extracted from HE speech, on the other hand we used our proposed algorithm which learns the mapping from EL speech data to HE f_0. HE f_0 pulse trains were time-aligned in order to have same duration than $\hat{f}_0$ pulse trains. We convolved the HE f_0 pulse trains and $\hat{f}_0$ pulse trains with an *LF* pulse to obtain an *LF* excitation signal. It is not easy to control speaking speed according to the pre-determined excitation signal, but after some minutes of training the speakers were able to do so. Two speakers (F01 and M02) uttered the paragraph "Nordwind und Sonne" [Association, 1999] with both kinds of excitation signals. The used device was our newly proposed EL device explained in Chapter 4. The sentences were recorded in an office environment using an omni-directional headset condenser microphone AKG HC 577 L at a distance of about 2 – 3 cm from the mouth corner. The software SpeechRecorder [Draxler and Jänsch, 2004] was used for the recordings.

13 volunteers participated in the listening test. We used an AB test to compare EL utterances produced with HE f_0 contours with EL utterances produced with $\hat{f}_0$ contours. Sentence order as well as AB order are randomized. We told the listeners to rate whether A or B sounds more pleasant to them. More details to this kind of listening test can be found in Appendix B.

The overall evaluation of the listening test for all listeners can be seen in Figure 3.10. The excitation signals with a healthy f_0 contour outperforms the estimated ones with 75 %. In 19 % the samples with $\hat{f}_0$ contour are rated to be better. 6 % rate both devices to be the same.

The results for each listener can be seen in Figure 3.11. Generally, healthy f_0 is rated more pleasant than the estimated one. Listeners 2, 6, 9 and 13 rate clearly in favor of the healthy f_0, whereas listeners 5, 7, 10 and 12 are uncertain. Evaluation results depending on the speaker are shown in Figure 3.12. Listeners judge female and male speakers slightly, although not significantly, different. The results point more clearly towards healthy f_0 for the female speaker than for the male speaker.

Furthermore, the listeners were asked about the difficulty of the test. On average the listeners found it moderate difficult to choose a value of preference and the differences between samples were perceived easily. All listeners complained about the DREL noise of the device which leads to the conclusion that DREL noise influences the decision of the listeners. Although they were asked to rate the samples based on the pleasantness 9 of 13 listeners answered that their primary focus was on speech intelligibility. Other parameters were humanity, softness and nuisance. This means that intelligibility has a main impact on the perceived pleasantness of EL speech. Furthermore, no preference regarding gender was noticed. Only two listeners prefer the male over female voice.

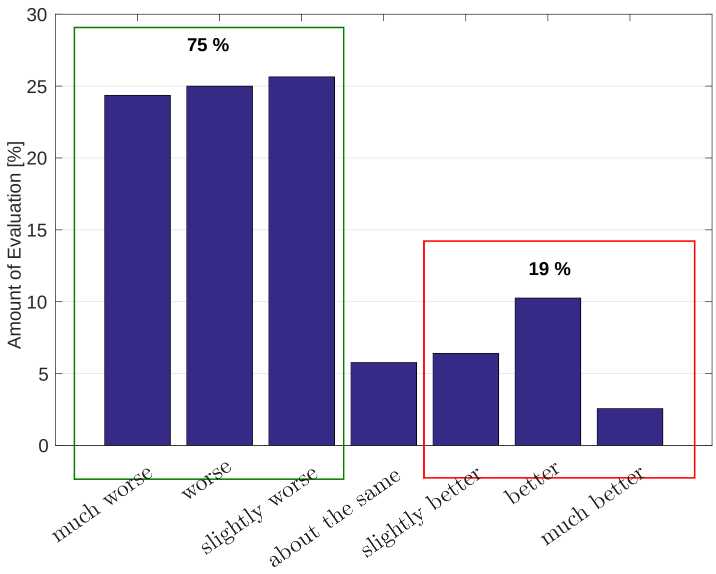


Figure 3.10: Histogram of the listening test results; "better" refers to the excitation signal with healthy f_0 as being perceptually better than $\hat{f}_0$; averaged results over all speakers and listeners.

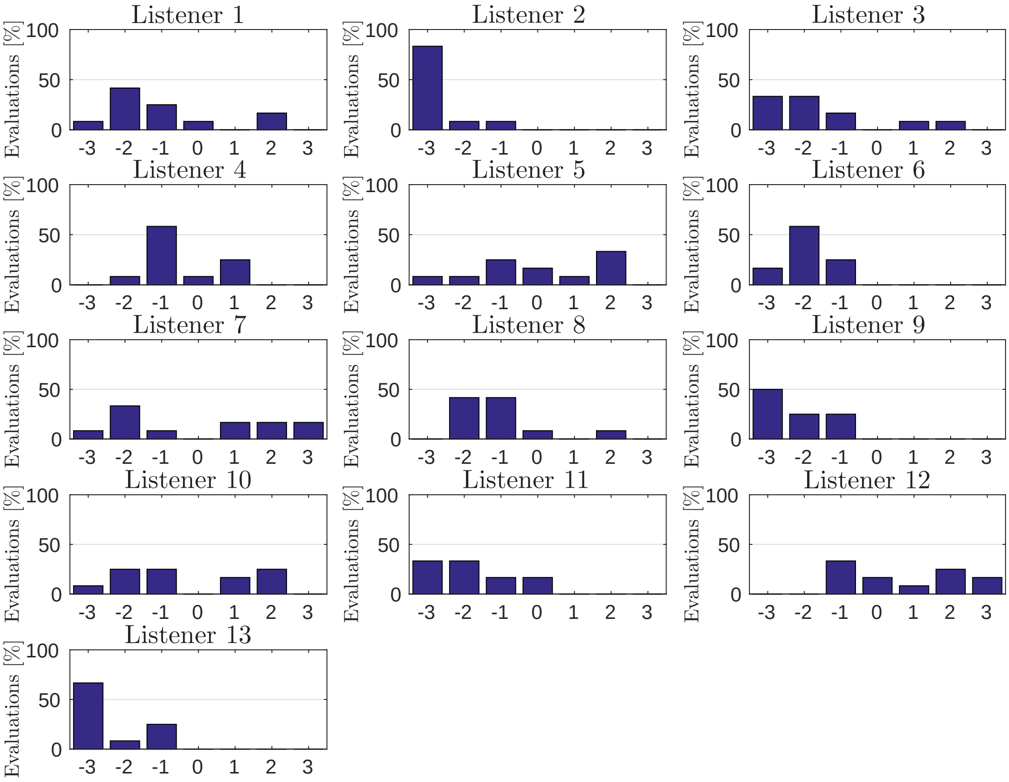
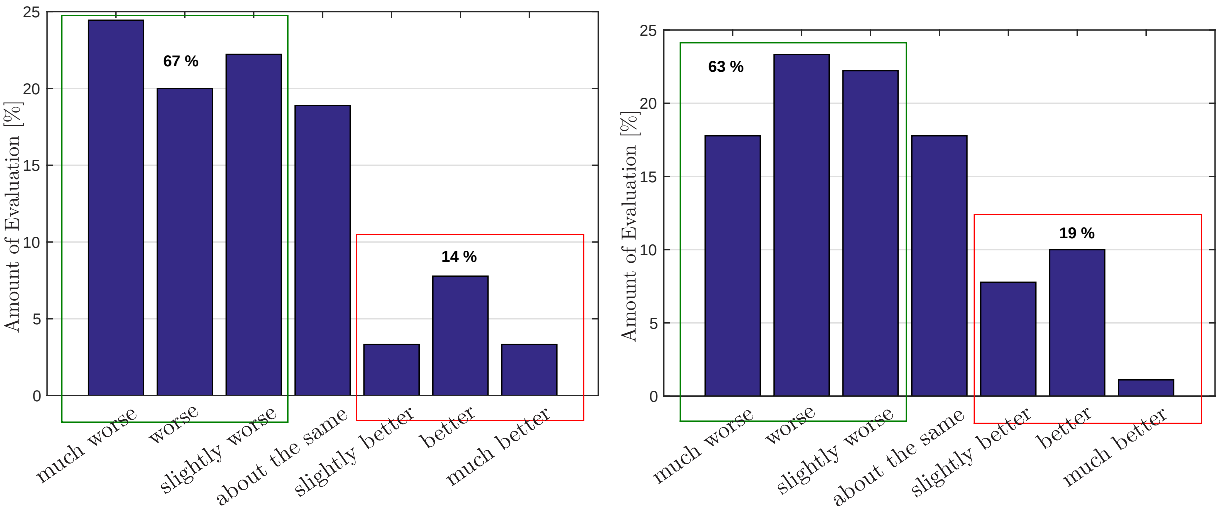


Figure 3.11: Histogram of the listening test results for each listener; scale from -3 ("much worse") to 3 ("much better"); 3 ("much better") refers to the excitation signal with healthy f_0 as being perceptually much better than $\hat{f}_0$; listener 5, 7, 10 and 12 were uncertain to choose a preference.



(a) Speaker F1

(b) Speaker M1

Figure 3.12: Histogram of the listening test results for each speaker; "better" refers to the excitation signal with healthy f_0 as being perceptually better than $\hat{f}_0$; slight preference towards the male speaker.

3.2.4 Experiment III: f_0 Estimation based on HMMs

In this experiment we investigated an HMM approach to estimate a changing f_0. We briefly introduce the model and afterwards compare different strategies for a changing fundamental frequency using a formal listening tests. In order to analyze different strategies, we implemented a framework as shown in Figure 3.13.

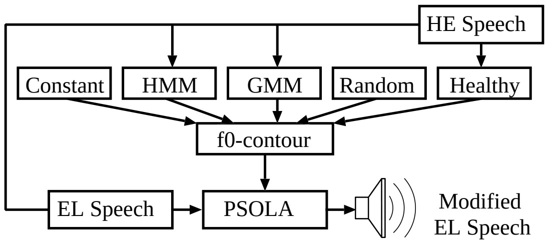


Figure 3.13: Synthesizing approach to evaluate algorithms for f_0 estimation based on modification using PSOLA.

For this purpose we used speech material from the German parallel ELHE speech database (see Section 2.2) to synthesize speech samples. PSOLA was used to modify f_0 of the speech files [Valbret et al., 1992]. It must be mentioned that this technique inserts additional artifacts. The following strategies for estimating a changing f_0 were implemented A) constant f_0, B) HMM approach [Wohlmayr and Pernkopf, 2010] (see Figure 3.14), C) GMM approach, (D) random f_0 and E) healthy f_0 from healthy speech. For A, D and E no training is needed and their implementation is straight forward. B and C are based on statistical models and, therefore, need prior training. Whereas C is explained in detail in Section 3.2.3, we would like to explain B: First, EL speech files were pre-processed, i.e., the DREL of the EL device itself was removed. For this experiment we used spectral subtraction in order to fulfill this task. Then, the fundamental frequency was extracted from HE speech (f_0) using the auto-correlation method implemented in Praat. The next step was to time align f_0 to the EL speech using a dynamic time warping algorithm. Mel-frequency cepstral coefficient (MFCC) features were extracted

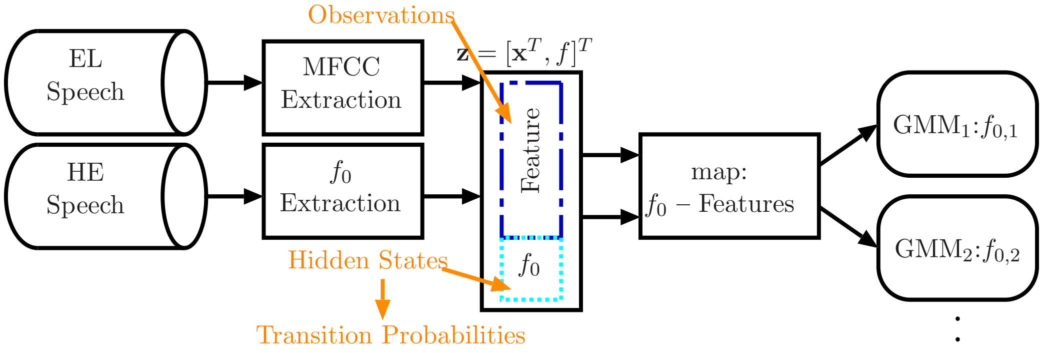


Figure 3.14: Feature extraction and training for B) HMM approach; estimation of f_0 based on best Viterbi path.

from EL speech. Finally, we grouped all features to their corresponding f_0 value and estimated a GMM for each of these values. Now one HMM is fully described using its observations (MFCC features), hidden states (f_0), prior probability of each state (histogram from f_0 – see Figure 3.16), transition probabilities (histogram from f_0) and emission probabilities (estimated GMMs per hidden state). In Figure 3.15 the transition matrix is depicted. To obtain the values of the matrix f_0 features from the training material is analyzed. We count how often one frequency value follows another one, e.g., how often $f_0 = 0$ is followed by $f_0 = 200$. Note, that frequency indices are listed, not absolute frequency values in Hz. From the same matrix we obtain the prior probability of each state (see Figure 3.16). To estimate f_0 from an unknown EL sentence, we can use the off-line estimated HMM and estimate the best Viterbi path.

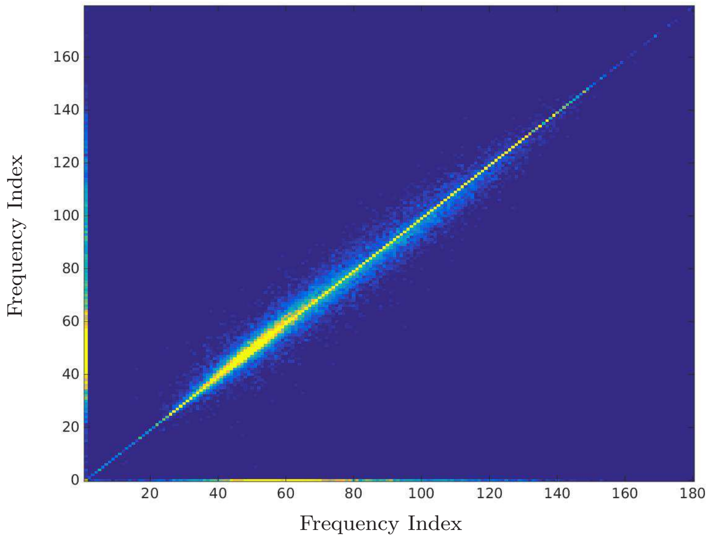


Figure 3.15: Transitions probabilities calculated from histogram of f_0 from training data of one speaker.

We used 2 sentences from 4 healthy speakers (2 female, 2 male). The estimation of fundamental frequency was done using the strategies explained before. Table 3.6 shows ρ between estimated values and f_0 values from HE speech. ρ confirms the good results for approach E (healthy f_0). ρ for approach E is not exactly 1, because f_0 is extracted from the manipulated EL sentence and compared to the original f_0 values. Approach D (random f_0) has a high ρ which is due to the calculation method of ρ. The result for approach B (HMM) is very low, but approach C

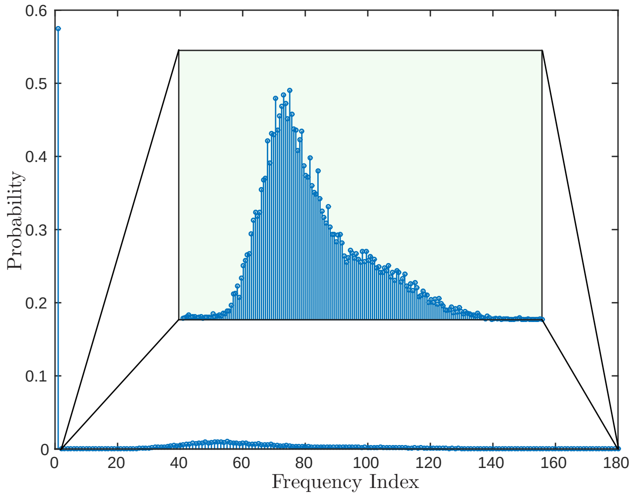


Figure 3.16: Prior probability of each state from histogram of f_0 from training data of one speaker; being in state $f_0 = 0$ (unvoiced) has highest probability (frequency index 0).

(GMM) reaches good results. We evaluate the sentences from the listening test. 6 sentences per method, however, might be not enough for the evaluation of ρ.

	Method				
	A	B	C	D	E
ρ	0.04±0.28	0.04±0.30	0.26±0.43	0.16±0.36	0.77±0.21

Table 3.6: Results of correlation coefficient ρ for $\hat{f}_0$ estimated from EL files.

A listening test was carried out with 8 normal hearing listeners who rated naturalness of the 5 approaches (A - E) using comparison category rating where the listeners are presented with a pair (A, B) of speech stimuli to evaluate their naturalness. The listener had to rate if stimulus A is “much better”, “better”, “slightly better”, “about the same”, “slightly worse”, “worse”, “much worse” than B as described in Appendix B. Furthermore, we asked for the perceived gender.

Results

The results are shown in Figure 3.17 and Table 3.7. We present the results in order of overall preference. As expected method A (constant f_0) is rated worst and method E (healthy f_0) is rated best. D is significantly better than A and significantly worse than B, C and E. We could not show any significant differences between B, C and E. We could show that the accuracy of

rank	model	μ	$CI_{95} - \mu$	σ^2
1	A (const)	-0.79	±0.16	1.27
2	D (rand)	-0.09	±0.16	1.33
3	C (GMM)	0.23	±0.16	1.34
4	B (HMM)	0.28	0.17	1.40
5	E (nat)	0.37	±0.18	1.42

Table 3.7: Overall preference μ with variance σ^2 and 95 % confidence interval $CI_{95\%}$.

the perceived gender is 98.24 %. This means that the listeners had no problems to distinguish between gender.

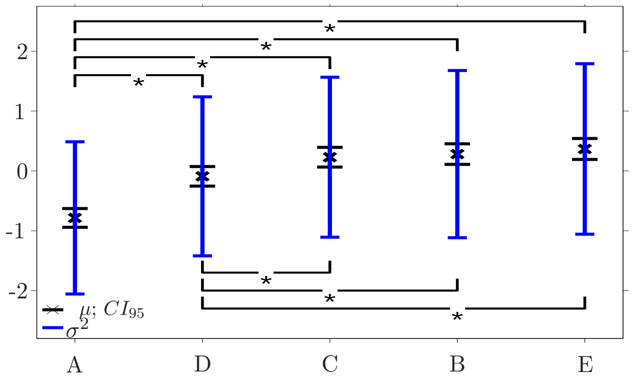


*Figure 3.17: Overall preference; * indicates significant difference according to Wilcoxon rank sum test. A) constant f_0, B) HMM approach [Wohlmayr and Pernkopf, 2010], C) GMM approach, (D) random f_0 and E) healthy f_0 from healthy speech.*

3.3 On/Off Control

The work on the following section was carried out together with [Amon, 2014] (3.3.2) and [Klammer, 2015] (3.3.3) and is documented in their master's theses.

According to [Shute, 2003] the most popular device is the Servox Digital, which has no possibility for hands-free control of on and off switching. This kind of device normally is tube like and fits into one hand. While speaking, people have to hold it against the neck and press a button to turn the device on and off. Therefore, many EL user would benefit from a new EL device which can be controlled in a hands-free manner.

We chose electromyography (EMG), i.e., the activity of the muscles, as a promising possibility to enable hands-free control of the EL device (EMG-EL). Using EMG signals, under the assumption that some of the muscles which controlled the larynx are partially unimpaired by the removal of the larynx, is a possibility to detect intended speech activity which can be used to control the EL device. In the following section we want to remove the drawback of utilizing a hand to control the EL. The usability of the device can be improved using EMG to detect speech activity and provide hands-free interaction. EMG is a common non-invasive method to measure motor unit action potentials. Motor units are the basic functional units for excitation and contraction in muscles. They can be activated and controlled voluntarily by the nervous system 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.

After developing a hardware to capture EMG signals we recorded a database consisting of EL speech and EMG signals. Furthermore, we systematically analyzed different strategies to the control on/off signal of the EL device. We employed different strategies to smooth the EMG envelope and developed a threshold based method (single and double) and a statistical method (GMMs) to detect voice activity and evaluate its performance. Afterwards, we implemented the strategy to perform evaluation in a real-world scenario to investigate possible learning effects. It turned out that there is a huge difference between measured EMG signals during ordinary speaking and EMG signals where the speaker is concentrating to control the device consciously. Therefore, a simple technique to change the signal is enough as long as the SNR of the EMG signal is sufficient high.

Our work builds on previous efforts to deal with on/off control for EL speech: In [Goldstein et al., 2004] the EMG signal was band-pass filtered ($10 - 500$ Hz), amplified, rectified, and low-pass filtered ($1 - 9$ Hz cut-off frequency) for the creation of envelopes that track EMG activity. This approach was implemented in an EMG-EL prototype and reaction time experiments were conducted and compared to conventional push-button EL and healthy voice. The developed analog version of the EMG-EL device had a controllable activation threshold. This single threshold was set to a value of 10 % of the amplitude range. The termination threshold was based on an internal (fixed) activation-threshold-dependent hysteresis band. The optimized threshold using recorded sentences was not significantly different than the threshold they chose in their informal tests. In follow-up work a new wireless version of the EMG-EL was presented [Heaton et al., 2011]. Within this work the EL could be either manually controlled or in a hands-free manner using the EMG signal (same processing as described above). Later, the same group used a computer based system with two controllable thresholds (double threshold) [Kubert et al., 2009]. In this work it is confirmed that EMG signals can serve as an intuitive and effective control source for EL voice activation and termination as well as f_0 modulation.

In [Ooe, 2012] the authors confirmed these findings. They used absolute values of the EMG signals. After smoothing with a moving average filter, a single threshold converts the envelope into the activation/termination signal. The time delay between EMG signal and speech start/stop was measured and it was confirmed that EMG signals can be used to control on/off signals for the EL device.

Also [Pineda-Rico et al., 2008] picks up the EMG based on/off control approach. They implemented a switched capacitor CMOS based device. For activation and termination the same method as in [Heaton et al., 2011] was taken: amplified, rectified and low-pass filtered ($c_f = 3$ Hz) envelope and single threshold implemented as voltage comparator. Their focus was on the implementation and on the advantages of switched capacitor circuits which are: excellent time constants, relative precision, simple design elements, minimum power waste and reduced size on chip. [Arifin et al., 2014] investigated the relation between EMG signals and human voice signals in terms of loudness and suggested that EMG signals are more appropriate to control loudness than fundamental frequency.

3.3.1 Data Acquisition Hardware

We developed a data acquisition hardware in order to reduce costs and size. The requirements for the bio-signal acquisition system were to be small in order to be portable, battery-operated and real-time capable. It consists of three main parts: the sensor straps, the bio-signal shield and an ARDUINO DUE micro-controller board. The board serves as a host for the connected strap and the shield (see Figure 3.18).

The strap is designed to be worn around the neck to ensure correct electrode position at the surface of the sternohyoid muscle. This muscle is a long, thin muscle which is located along the length of the front of the human neck. The functions of this muscle include, depression of the hyoid bone, head and neck movement, and speech. The strap holds three silver/silver-chloride electrodes. Two of them are used to detect the EMG signal, the third one serves as a reference electrode to improve the common-mode rejection ratio. The strap is connected to the instrumentation amplifier which is followed by an operational amplifier. The gain of this amplifier can be modified manually. After a low-pass filter where high frequency noise is suppressed, the positive and negative half-wave are split and fed to two discrete analog inputs of the micro-controller. Using this method, a higher bit resolution (i.e., 13 bit) of the digitized signal amplitude can be achieved. Then, the signal is converted from analog to digital. In the following experiments, the micro-controller board was connected to the computer via USB which served to power the shield via the micro-controller board (5 V). The sampling rate f_s of the ARDUINO DUE ADC was set to 8 kHz. This is enough as most of the frequency content of EMG signals is between 0 and 500 Hz.

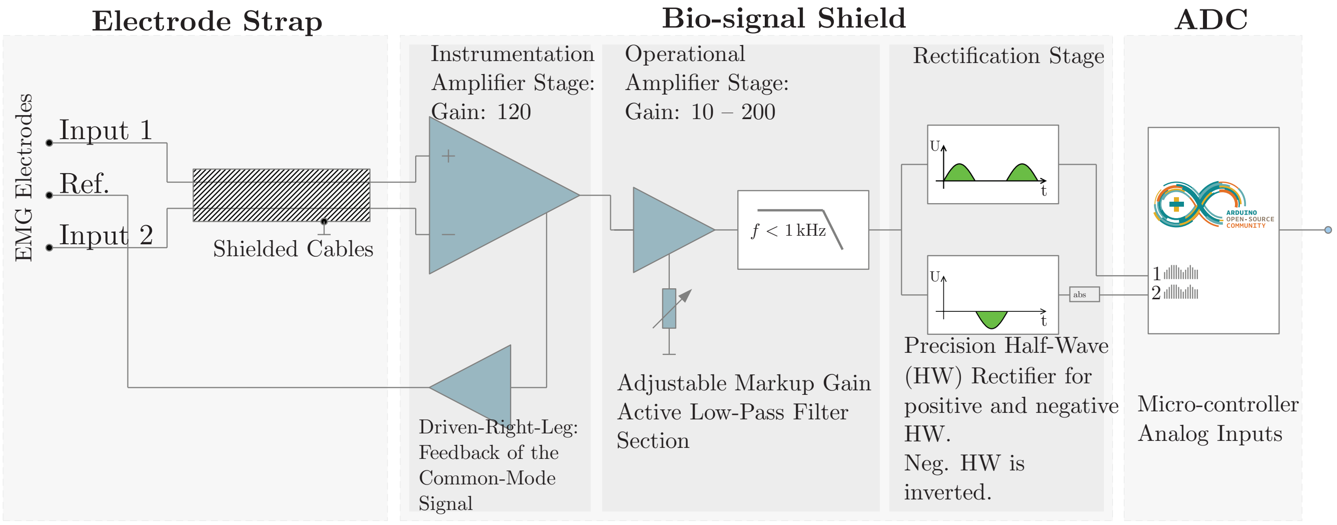

Electrode Strap
Bio-signal Shield
ADC
EMG Electrodes
Input 1
Ref.
Input 2
Shielded Cables
Instrumentation Amplifier Stage: Gain: 120
+
−
Operational Amplifier Stage: Gain: 10 – 200
$f < 1$ kHz
Rectification Stage
U
t
U
t
abs
1
2
Driven-Right-Leg: Feedback of the Common-Mode Signal
Adjustable Markup Gain Active Low-Pass Filter Section
Precision Half-Wave (HW) Rectifier for positive and negative HW. Neg. HW is inverted.
Micro-controller Analog Inputs

Figure 3.18: Schematic overview: Block diagram of the developed hardware system consisting of the electrode strap, the bio-signal shield which is compatible to the ARDUINO DUE micro-controller which serves as a host [Amon, 2014].

3.3.2 Experiment IV: Preliminary Investigation

Recorded Database

To evaluate different approaches for on/off control, a database was recorded and simulations were done off-line using the recordings.

We used around 100 phonetically rich utterances of a female and a male speaker. The skin surface EMG sensors were positioned on the neck and were attached to our processing hardware. EMG and speech signals were recorded using both, the bio-signal shield connected to an audio interface (RME Fireface 800), and a head-mounted microphone AKG HC 577L with omni-directional pickup pattern. The audio interface ensured a high quality digital signal. The sampling rate of the audio interface was set to 44100 Hz. Compared to the used sound card (24 bit), the micro-controller system is able to convert the input signal with a resolution of 13 bit. This is enough to perform all processing steps which are proposed in this work, without drawbacks in respect to signal detection. We analyzed the recordings manually and annotated speech and non-speech sections in order to obtain the ground truth.

All in all we recorded 18 min 45 s of data. The mean SNR for the male EMG signals is 16.7 dB and for the female 12.6 dB. For SNR calculations we used first order IIR smoothing (see Section 2.2.1). This difference in SNR will also influence the thresholds for on/off control. The main energy of the EMG signal is between 0 Hz and 500 Hz, in fact, over 90 % of the energy can be found in this range. The ratio of speech to non-speech in the database is 63 % to 37 %.

Pre-processing, Envelope Calculation, On/Off Control

In order to clean and prepare the recorded EMG signal some pre-processing needed to be applied. We used adaptive noise cancellation to remove crosstalk of the EL excitation signal and the EMG signal. Afterwards, a notch filter reduced interferences from electric hum caused by magnetic fields close to the sensors and amplification unit. The database was split into three parts in order to perform 3-fold cross validation. The sentences were chosen randomly for each validation set.

We chose three different envelope calculation methods: 1) root mean square, 2) Hilbert transform and 3) low-pass filtered rectified signal which was also suggested by [Goldstein et al., 2004].

1) Root mean square (RMS): In EMG analysis, the RMS of the signal is a common envelope calculation method to obtain the power of the signal. The RMS value for a windowed signal x with window length N is defined as $RMS\{x\} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x[n]^2}$.

2) Hilbert transform (HIL): Another method to calculate the power of the EMG signal is to use the Hilbert transformed signal and rectify it. If we compare the Hilbert envelope to a simply rectified EMG signal, it can be seen that this method works as an amplitude follower and provides, even un-smoothed, an envelope which is not touching the zero line on the x-axis. Both, RMS and Hilbert envelope use a moving average filter for smoothing. A latency related to the length of the window is expected in real-world applications.

3) Low-pass filter (LP): The signal is rectified and a low-pass filter with a cut-off frequency of 5 Hz is applied. An LP smoothes the envelope. Implemented as a 3-pole IIR filter it produces a potential delay of up to 150 ms.

Using these different envelope calculation methods, we find on/off control messages using a) single threshold (ST), b) double threshold (DT) and c) classification using Gaussian mixture models (GMM).

Classification using a) single threshold and b) double threshold is straightforward. The calculated envelope was compared to a threshold. As soon as the envelope exceeded this threshold speech was detected and vice versa, when the envelope fell below the threshold we determined the message for non-speech (see Figure 3.19 - upper plot). In case of the double threshold, speech was detected when the envelope surpasses the first threshold and non-speech was detected if the thresholds dropped below the second threshold (see Figure 3.19 - lower plot). A GMM is fully described with the parameters $\lambda = (b_k, \mu_k, \sigma_k^2); k = 1, 2, ..., K$. For the GMM on/off control two GMMs were trained, one for speech and one for non-speech. The used features were the calculated envelopes. The number of components in a GMM can be selected using the Bayesian information criterion. The criterion is composed by the log-likelihood and a complexity penalty term. We estimated the parameters of the GMM with changing number of components and decide for the number of components were the Bayesian information criterion was lowest. The number of components lay between 8 and 32. In the test we computed the probability density function for an unknown input for both GMMs and the maximum results in the on/off control

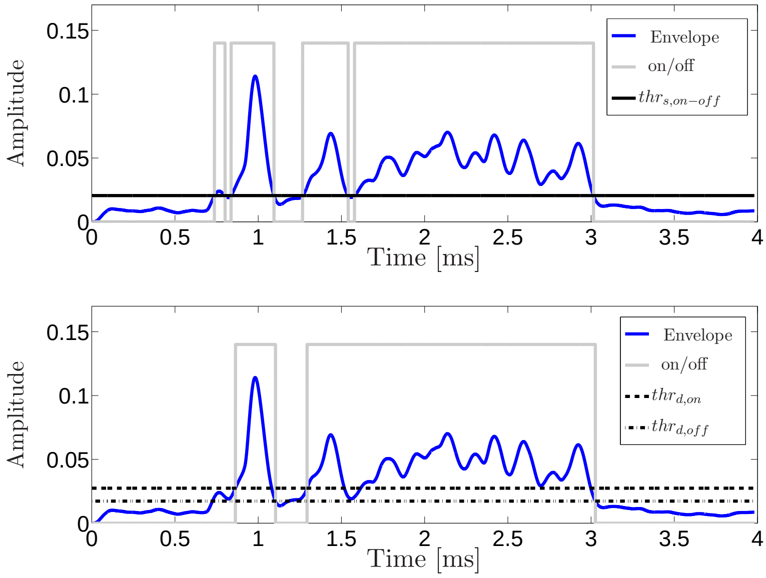


Figure 3.19: On/off control using single threshold (upper plot) and double threshold (lower plot).

classification. We trained the parameters (b, μ and σ) on two validation sets and tested on the third. All three combinations assured that we used all utterances in the test.

Threshold Determination

The single threshold as well as the double threshold for on/off control were determined using receiver operating characteristics (ROC). To prepare ROC curves we have to set up a cross table (see Table 3.8) where the columns denote target condition and the rows the test outcome. In our case the condition is whether speech is present or not. True positive (TP) and true negative (TN) are desirable outcomes. In TP target speech is present and the test detects speech. For TN it is vice versa: target speech is absent and the test detects no speech. False positive (FP) and false negative (FN) are errors. Based on this table we define sensitivity (SE = true positive rate) as

		Target Condition	
		True	False
Test outcome	Positive	TP	FP
	Negative	FN	TN

Table 3.8: Example of cross table with true positive (TP), true negative (TN), false positive (FP) and false negative (FN).

$$SE = \frac{TP}{TP + FN}, \tag{3.14}$$

and specificity (SP = false positive rate) as

$$SP = \frac{TN}{TN + FP}. \tag{3.15}$$

SE is the probability for a positive test outcome, given the target condition is present and the specificity is the probability for a negative test result, given the target condition is absent.

ROC curves are used to illustrate test performance of a threshold classifier with respect to its threshold. They are applied to determine the optimal threshold. Under the assumptions of signal detection theory, we interpolated the ROC points of the different values of the threshold. There are different methods for optimizing the threshold. We chose to maximize the sum of SE together with SP. Figure 3.20 (left) shows an example of SE in blue, SP in red and the sum of the two curves in yellow. The maximum value is given with a red circle and its index represents the optimal threshold. In the right part of the figure SE is plotted against 1-SP. Each point represents a certain threshold. The threshold with the index number of the left plot is chosen to be the optimal value with corresponding SE and SP. SE and SP were calculated analyzing the database sentences with thresholds going from 1 % to 100 % in steps of 1 %. For single threshold the activation threshold $thr_{s,on-off}$ (= termination threshold) was based on the maximal envelope amplitude, for double threshold the termination threshold $thr_{d,off}$ was relative to the activation threshold $thr_{d,on}$. The thresholds were chosen speaker dependent.

Error Calculation

The on/off control results were compared to the speech ground truth. Errors were calculated regarding the correctly detected activation time and taking into account the interruptions of the detection. The absolute error was classified into the following parameters [Freeman et al., 1989]: front end error (*FEE*), back end error (*BEE*), middle speech error (*MSE*) and noise detected

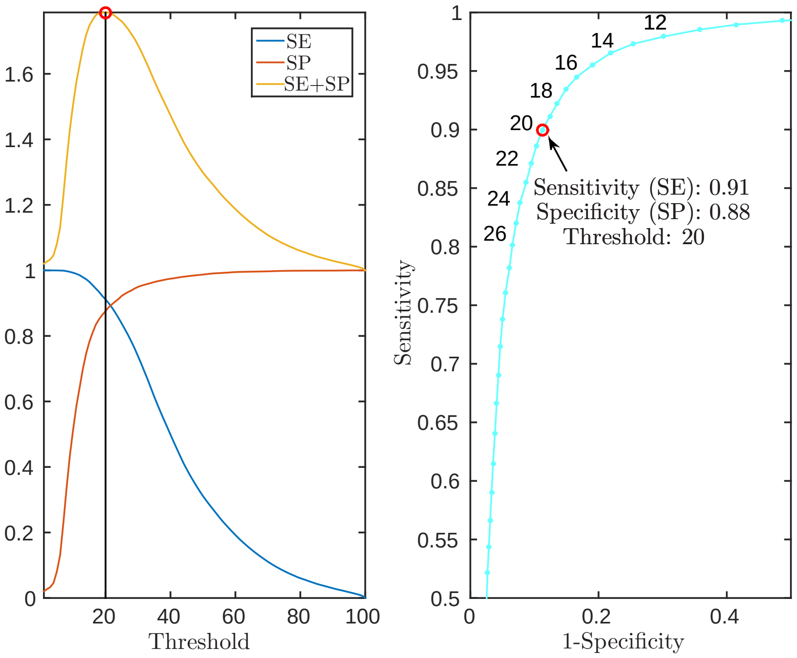


Figure 3.20: Receiver operating characteristics (ROC); left plot: sum of sensitivity (SE) and specificity (SP) is maximized, corresponding threshold in ROC curve (right plot).

as speech (*NDS*) (see Figure 3.21). The 4 error types were normalized to the total length of the analyzed sentence.

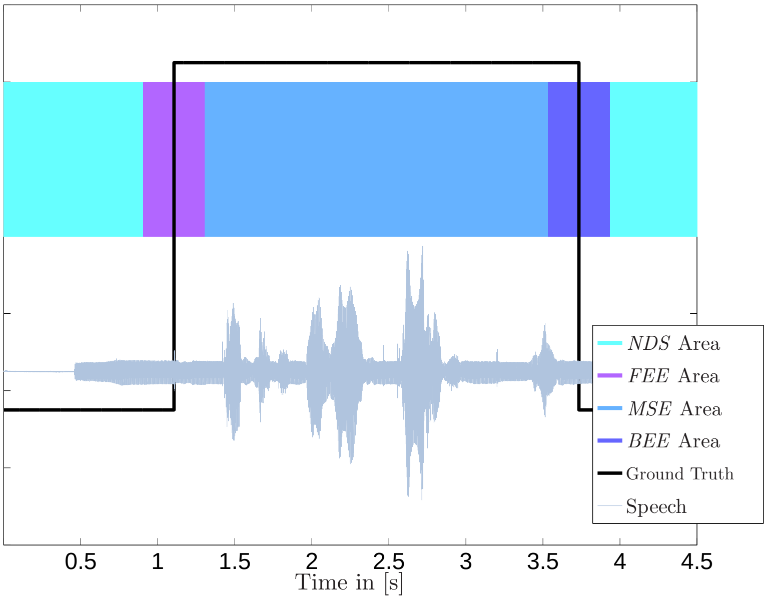


Figure 3.21: Error regions for noise detected as speech (NDS), front end error (FEE), mid speech error (MSE) and back end error (BEE) of an EL sentence [Amon, 2014].

If there is no triggering when the person wants to say something, information is lost. This error influences the speech quality much more than unwanted triggering when the person does not want to say anything. This is annoying and must also be avoided. Therefore, we assume that *MSE* and *FEE* are the most important errors.

The averaged absolute error for one method was calculated as the averaged sum of each separate error for a number of N sentences i:

$$\overline{|er|} = \frac{1}{N} \sum_{i=1}^{N} FEE_i + BEE_i + MSE_i + NDS_i. \tag{3.16}$$

Moreover, the relation between the correct number of interruptions inside a sentence and the unwanted interruptions due to wrong detection was presented as an indicator for interruptions of the on/off control. The block detection ratio (*BDR*) 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 perfect detection in terms of this interruption indicator would result in a block detection ratio of 1.

Time Constant Detection Smoothing

For post-processing of short-time on/off control, it is common to apply detection smoothing or long time detection algorithms to the detection output. This is done to avoid short interruptions of the detection. In live scenarios this is problematic as for offset events the length of the following interruption cannot be predicted. The proposed time constant detection smoothing algorithm treats two issues: 1. detection results might suffer from short interruptions and 2. EMG during speech shows a pre-activation of about 40 ms compared to the speech signal [Atkinson, 1978]. The algorithm is using a time constant to smooth the detection output in a real-time application and, therefore, avoids small interruptions of detection and, at the same step, the algorithm is compensating EMG pre-activation. The time constant detection smoothing is delaying every on- and offset by the time constant 40 ms.

Objective Evaluation

Threshold Determination

Single thresholds $thr_{s,on-off}$ for the female are lower than for the male speaker. The thresholds for LP and Hilbert envelope are similar but for RMS it is lower: F: LP - 25 %, HIL - 24 %, RMS - 20 %; M: LP - 18 %, HIL - 17 %, RMS - 12 %. For double threshold the activation threshold $thr_{d,on}$ is higher than $thr_{s,on-off}$, the termination threshold $thr_{d,off}$ is approximately the same for all envelope calculation strategies and both genders (see Table 3.9). Sensitivity is between 88 % and 96 % and the specificity between 77 % and 89 %. This means that in each method around 90 % is detected correctly and the false positive rate (1-specificity: speech detected, no speech in ground truth) is up to 20 % which leads to algorithms with very good performance.

	F		M	
	$thr_{d,on}$	$thr_{d,off}$	$thr_{d,on}$	$thr_{d,off}$
RMS	26 %	66 %	18 %	53 %
HIL	32 %	67 %	21 %	60 %
LP	33 %	63 %	25 %	60 %

Table 3.9: Speech/non-speech thresholds for female (F) and male (M) and for different envelope calculation methods: root mean square (RMS), Hilbert transform (HIL) and low-pass filtered (LP); $thr_{d,on}$ in % based on maximal envelope, $thr_{d,off}$ in % based on the activation threshold $thr_{d,on}$ (see Section 3.3.2).

Error Analysis

The individual error results and the block detection ratio values are shown in Figure 3.22 for the female speaker and in Figure 3.23 for the male speaker. The averaged absolute error $\overline{|er|}$ together with the standard deviation (σ) and the 95 % confidence interval ($CI_{95\%}$) for the mean are presented in Table 3.10. The presented results are processed with the proposed time constant detection smoothing. It is possible to improve $\overline{|er|}$ around 3 % regarding no time constant detection smoothing where especially *FEE* and *MSE* is reduced. The time constant detection smoothing also improves the block detection ratio.

- $\overline{|er|}$ for the male speaker (6.4 % – 12.0 %) are better than for the female speaker (9.5 % – 12.2 %).
- Double threshold outperforms single threshold for the male and the female speaker.
- LP envelope outperforms the RMS and Hilbert envelope in all tested scenarios (Single threshold, double threshold and GMM based approach) except for female HIL-GMM which slightly outperforms (0.1 %) LP-GMM.
- In GMM classification the *MSE* and *NDS* reaches low values, but the *FEE* errors are increased.
- GMM is comparable with LP-DT for the female speaker. For the male speaker LP-GMM is the best method, but there are huge differences between the GMM methods.
- Block detection ratio *BDR* a measure for interruptions, results are improved with the time constant detection smoothing and present good results for both speakers (ca. 3 for female; ca. 2 for male).

With our tuned parameters, LP envelopes with double threshold perform best for female (second: HIL-GMM) and LP envelopes with GMM for male speaker (second: LP-DT).

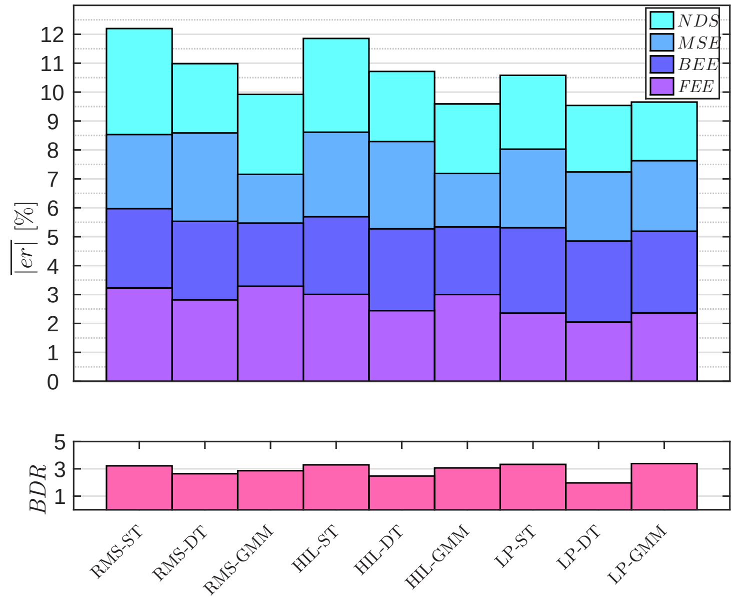


Figure 3.22: FEMALE: Averaged absolute error $\overline{|er|}$ for all envelope calculation strategies: (root mean square (RMS), Hilbert envelope (HIL) and low-pass filter (LP) together with classification methods: single threshold (ST), double threshold (DT) and GMM (upper plot); block detection ratio (BDR) (lower plot).

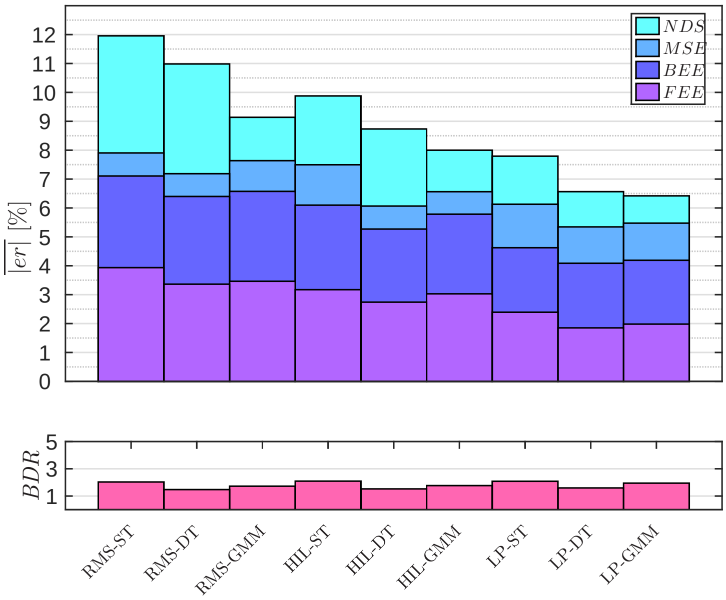


Figure 3.23: MALE: Averaged absolute error $\overline{|er|}$ for all envelope calculation strategies: (root mean square (RMS), Hilbert envelope (HIL) and low-pass filter (LP) together with classification methods: single threshold (ST), double threshold (DT) and GMM (upper plot); block detection ratio (BDR) (lower plot).

Method	$\overline{\lvert er\rvert}$ (σ) [%]		CI_{95} [%]	
	F	M	F	M
RMS-ST	12.2 (6.7)	12.0 (6.3)	10.9 – 13.5 *	10.8 – 13.2 *
RMS-DT	11.0 (6.8)	11.5 (6.8)	9.7 – 12.3	9.7 – 12.3 *
HIL-ST	11.9 (6.8)	9.9 (5.7)	10.6 – 13.1 *	8.8 – 10.9 *
HIL-DT	10.7 (6.6)	8.7 (6.7)	9.5 – 12.0	7.5 – 10.0 *
LP-ST	10.6 (5.3)	7.8 (4.9)	9.6 – 11.6	6.9 – 8.7 *
LP-DT	9.5 (5.6)	6.6 (4.7)	8.5 – 10.6	5.7 – 7.4
RMS-GMM	9.9 (5.7)	9.1 (5.3)	8.9 – 11.0	8.1 – 10.1 *
HIL-GMM	9.6 (5.5)	8.0 (5.0)	8.6 – 10.6	7.1 – 8.9 *
LP-GMM	9.7 (5.6)	6.4 (5.8)	8.6 – 10.7	5.3 – 7.5

*Table 3.10: Averaged absolute error $\overline{|er|}$ for female (F) and male (M); Best and second best methods in light grey (female) and dark grey (male); * indicates significant difference to the best method (F: LP-DT; M: LP-GMM).*

To summarize: We developed a data acquisition hardware for EMG signals which serves as a prototype in a real-time implementation. We found out that Hilbert envelope and LP envelope together with double threshold and GMM yield the best results in terms of error calculation. Due to the strong dependency on the SNR the thresholds need to be adjustable during speaking. The time constant detection smoothing turned out to be a good method for smoothing on/off control results and can also be implemented in real-time.

3.3.3 Experiment V: Learning Effects

Following experiment was carried out within the master's thesis of [Klammer, 2015]. We implemented the RMS envelope calculation method together with the single and double threshold detection in a real-time system. Thus, we did not use recorded data and carried out closed-loop experiments in the following. RMS turned out to deliver more robust signals than HIL and

LP in the closed-loop scenario. We would like to reveal the influence of training effects on an EMG controlled EL device (EMG-EL). Furthermore, we investigate the effects of training on intelligibility and naturalness of EL speech. In order to improve pleasantness of EMG-EL speech we would like to decrease interruptions of the EL, improve speaking rate, improve the conscious stops of speech at phrase boundaries between two clauses and the way of articulation.

The main reference to these investigations is [Goldstein et al., 2007] who examined the training effects using an EMG controlled EL device. Seven tasks were trained: vowel initiation, vowel duration, vowel termination, words, sentences, paragraphs and intonation contrasts. The participants were three total laryngectomees (male) and four healthy subjects (two male, two female). The setup of their experiment consisted of a condenser microphone, a video screen to present the test material and a photo cell on the video screen to measure the time delays. They evaluated recordings of the first three sessions and recordings of the last three sessions in terms of reaction time of vowel initiation, duration and termination. The same authors evaluated vowel initiation time and vowel termination time of EMG-EL speech compared to tracheo-esophageal speech, manual EL and healthy speech [Goldstein et al., 2004]. Voice initiation time does not show any differences between speech types, voice termination time is longer for EMG-EL compared to the other types. In the thesis of [Stepp, 2008] electrode positions were investigated. Furthermore, the relation between radiation therapy and reduced muscle integrity was discussed, but no conclusive relation could be determined.

Experimental Conditions

A training protocol was created and participants were recruited to perform the training in nine consecutive training sessions divided over two weeks in order to improve the handling of an EMG-EL system (see Table 3.11). We used the same tasks suggested by [Goldstein et al., 2007] without intonation contrast task. In the first session the participants could familiarize themselves with the device in terms of how to produce intelligible speech. The second and last session was used to record pre- and post-training recordings of the speech material. After the second session training started where participants had to repeat exercises for approximately 50 to 60 minutes. A supervisor oversaw the training sessions and provided feedback about the performance of the participants. Moreover, the supervisor gave instructions for improvements. Our study involved four participants, three women (F1, F2 and F3) and one man (M1). All participants are healthy subjects without voice disorders at the average age of 29 (female) and 40 (male) years.

Session	Exercise
1	Familiarize with the device
2	Pre-training recording
3	Vowel initiation
4	Vowel duration
5	Vowel termination
6	Words
7	Sentences
8	Paragraph
9	Post-training recording

Table 3.11: Training Protocol with six training sessions, two recording sessions and one initialization session.

The following training criteria are based on [Goldstein et al., 2007] but were adapted for this experiment:

Voice Initiation. This session was used to practice vowel initiation in order to concentrate on muscle activities to produce the vowel [a]. The participants learned how to produce neck muscle activity adequately to turn on the device quickly and consistently. For the evaluation 40 vowels [a] were recorded once at the pre- and once at the post-training.

Voice Duration. This session was used to produce continuous vowels for different time intervals (2 s, 2.5 s and 3 s). After the session the participant should be able to sustain vowels without any interruptions during the whole time interval. Again, for the evaluation 40 vowels [a] were recorded pre- and post-training.

Voice Termination. This session was used to learn to relax the neck muscles to stop the device between words and to produce and control pauses during sustained vowels. While not speaking the participant learned how to relax the neck muscles to prevent to turn on the EL unintentionally. For evaluation 40 vowels [a] were recorded pre- and post-training.

Words. In this session the participant read words from a printed list. The words used in the recordings are phonetically rich with a focus on the nasals [m] and [n] at the beginning of the word, because for these words it is more difficult to produce an appropriate EMG signal. For training 200 different words were used. These words are different from the 40 words used in pre- and post-training recordings to have same conditions for all speakers during recording.

Sentences. In this session the complexity was extended from words to sentences. For training a list with approximately 350 sentences was used. For the recordings 30 sentences, different from training, were used.

Paragraph. For training the paragraphs “Unser Garten” [Bergauer and Janknecht, 2011] and “Die Buttergeschichte” [Vieregge et al., 1996] were used. Pauses between words and phrases are allowed. The participants were motivated to produce speech as naturally as possible. For the evaluation of the paragraph the participant read “Nordwind und Sonne” from a printed list.

For training and recordings the software SpeechRecorder [Draxler and Jänsch, 2004] was used. Within this software the utterance (vowel, word) was displayed on a screen and a green traffic light gave the signal when to start and stop to pronounce (see Figure 3.24). In order to evaluate activation and de-activation errors at the beginning and the end of the recording phase, actual recordings already started at the beginning of the first yellow phase and ended after the second yellow phase. In order to avoid the effect of habituation, we defined different pre-recording, recording and post-recording times for the optical signal of the traffic light (see Table 3.12). Additionally, the traffic light was needed as a baseline for the ground truth.

Measurement	Pre-Recording [s]	Recording [s]	Post-Recording [s]
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

Table 3.12: Recording phase times for VIT, VD and VTT.

The speech material recorded before and after the training was evaluated in terms of objective and subjective measures. The **objective measures** were obtained in terms of Vowel Initiation Time (VIT), Vowel Duration (VD) and Vowel Termination Time (VTT). Furthermore, the errors (FEE, BEE, MSE and NDS) are presented as explained in 3.3.2. In order to obtain error measure, the recordings were compared to the ground truth. The ground truth is defined by the recording phase times. It is the intended starting point and endpoint of the articulation of the speaker and in our case was given by the traffic light. VIT and VTT were measured using the recorded time signal and the pre- and post recording times of Table 3.12. These errors include

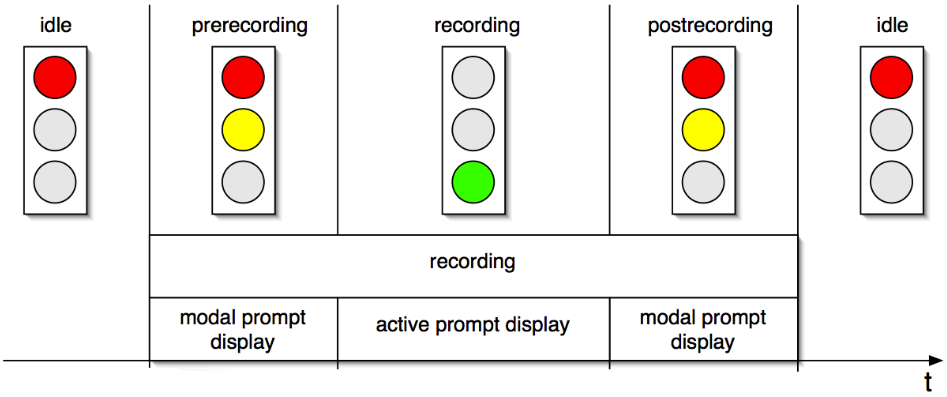


Figure 3.24: Recording phases of the traffic light of the recording software SpeechRecorder*. First yellow phase: pre-recording, green phase: recording, second yellow phase: post-recording [Draxler and Jänsch, 2004].*

the reaction time of the speaker from the visual cognition to the begin of articulation and the delay of the system. Within this experiments we can only measure reaction time including delay of the system together. Figure 3.25 shows a recorded vowel with 1 s pre-recording time, 2.5 s recording time and 1 s post-recording time. VIT and VTT were measured time intervals in relation to the ground truth signal. To turn on the device on purpose within a short time the VIT has to be as short as possible. VTT measures the ability to turn off the EMG-EL on purpose within a short time. VD is a binary measure: If the EL signal is not continuous during the whole recording but has at least one interruption, VD is counted as an error.

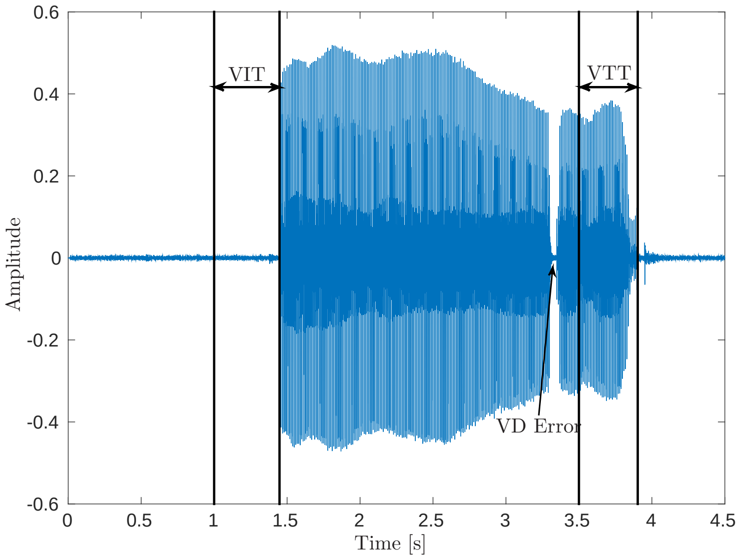


Figure 3.25: Recorded vowel [a] with the definition of VIT, VTT and VD.

Objective Evaluation

Results concerning VIT can be seen in Figure 3.26 for all four speakers for pre-training (left plot) and post-training (right plot). The mean values of speakers F1 and F3 decreased, whereas the mean values of speakers F2 and M1 increased. Nevertheless, all mean values are within a range of 400 ms to 480 ms, depending on the alertness and condition of the speakers at the recording sessions. The main difference between pre- and post-training is the variance. Before training the values vary between 100 ms and 750 ms which means that participants did not have an adequate control of the device. After training the values are within an interval of around 200 ms. This decrease of variance leads to the conclusion that the handling of the EMG-EL can be improved in terms of VIT.

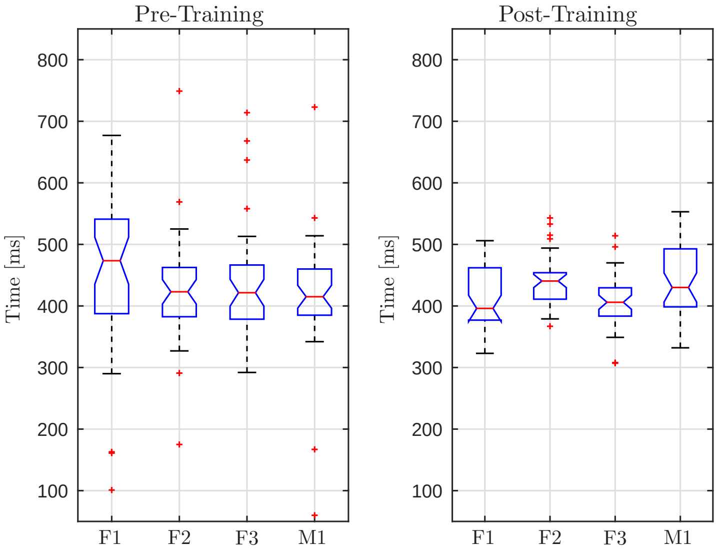


Figure 3.26: Voice initiation time (VIT) for pre- and post-training for three female speakers (F1, F2, F3) and one male speaker (M1).

Figure 3.27 shows the passed/failed results for the 40 pre- and post-training recorded samples of the vowel [a] for all four speakers. It is evident that training has an influence in maintaining a proper EMG signal to produce continuous vowels. Before training between 38 % (F2) and 95 % (M1) of all samples have one or more interruptions. After the training sessions all participants can obviously reduce unwanted interruptions.

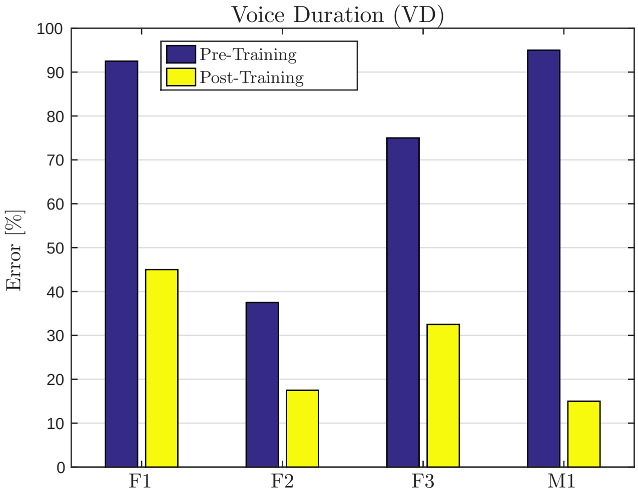


Figure 3.27: Errors of VD for pre- and post-training for three female speakers (F1, F2, F3) and one male speaker (M1).

Figure 3.28 shows a comparison of VTT for all four speakers. The mean values are reduced by around 100 ms and also the variances decrease. For all speakers it was quite difficult to produce a proper EMG signal to obtain a speech signal without interruptions and to relax their neck muscles immediately after uttering a word or sentence to turn off the device. These measured results indicate an improvement of EL handling.

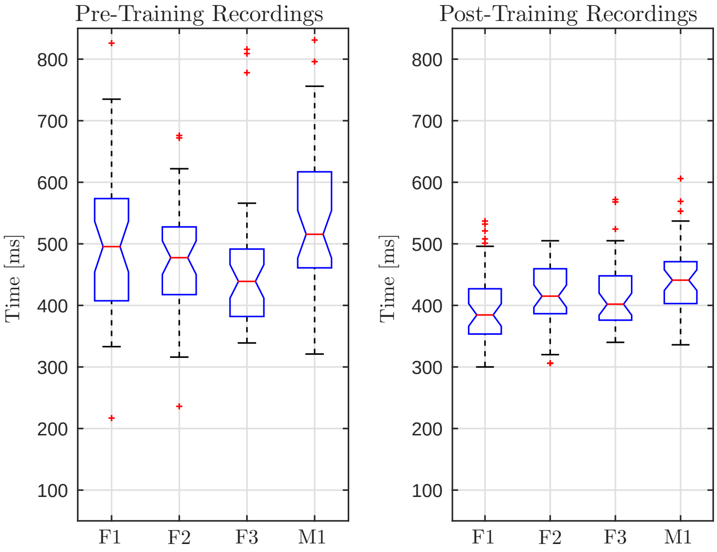
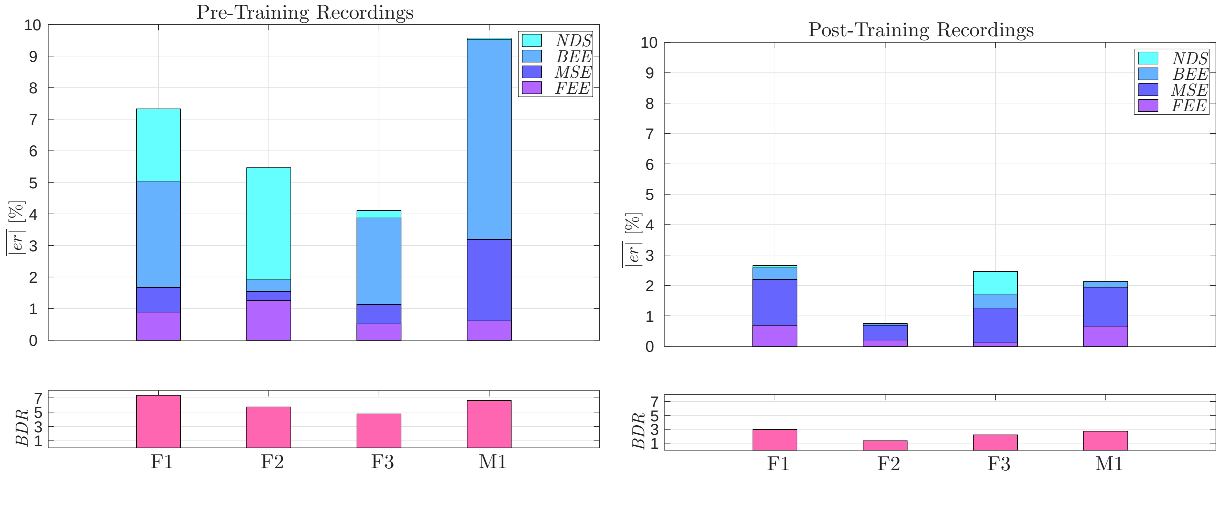


Figure 3.28: Voice termination time (VTT) for pre- and post-training for three female speakers (F1, F2, F3) and one male speaker (M1).

A comparison of the pre-training errors in Figure 3.29(a) with the post-training errors in Figure 3.29(b) shows a decrease of the total error for all four participants. Contrary to the error results before in Figure 3.27 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 NDS error of speakers F1 and F2 is caused by an increased threshold. We had to adapt the threshold over time. Before training the EMG signal was weak. Thus, ground noise and unwanted muscle activities (e.g., strong breathing, swallowing) already activated the EL. After training the threshold could be put on a higher level. The influence of the disturbing noise decreased and the participants were still able to produce more continuous speech. This can be seen in the decreased values of MSE. The detection of interruptions during ground truth activity causes a BDR higher than the optimum of 1. The decreased MSE leads to a decrease of the BDR.

A comparison with the results in the previous experiment 3.3.2 reveals some interesting improvements. In the preliminary experiments we reported a total error of 12.2 % for a female speaker and 12 % for a male speaker using the RMS envelope calculation method and ST. In this experiment female speakers have 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. The main difference is the real-time system. In the preliminary experiments the EMG signal was not consciously changed in order to provide an on/off signal for the EL device. Moreover, in the previous experiment we used whole sentences for the evaluation of errors which, in general, are more difficult to articulate than the vowels used in this evaluation.



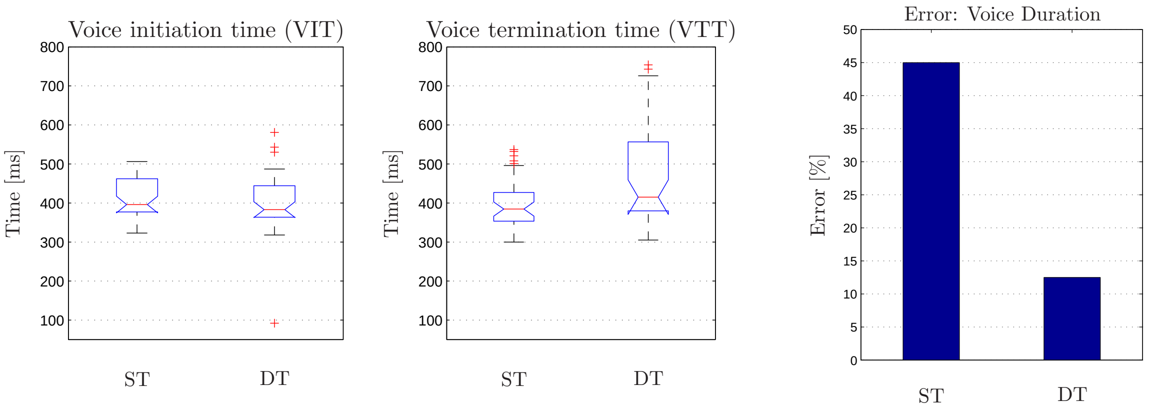
(a) Errors pre-training.

(b) Errors post-training.

Figure 3.29: Averaged absolute error $\overline{|er|}$ and separate errors: FEE, MSE, BEE, NDS and BDR before and after training for three female speakers (F1, F2, F3) and one male speaker (M1).

Comparison between single threshold (ST) detection and double threshold (DT) detection

Based on limited material we evaluated the difference between ST and DT in the real-time environment. The speech material consisted of 40 vowels [a] recorded by the female speaker F2 with each of the two detection algorithms. For the participant it was much easier to produce a continuous EL speech with DT but with the drawback that turning off the device was more difficult. Since DT supports the production of continuous speech it is more difficult to turn off the EL as can be seen in the VTT plot on the right side of Figure 3.30(a). The mean value and the variance of the VTT increases. For VD the error for ST is 45 %, whereas this error is reduced with DT to a value of only 12 %.



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

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

Figure 3.30: Differences between single (DT) and double threshold (DT) detection for VIT, VTT and VD.

Subjective Evaluation

For the **subjective measure** an AB listening test using comparative category rating was performed. The listening test consisted of three different parts: 1. evaluation of words, 2. evaluation of sentences and 3. evaluation of paragraphs. The used speech material originated from all four speakers. According to [van Rossum et al., 2002] the attributes naturalness and intelligibility can be determined by a pairwise comparison of examples. Here, the pre- and post-training speech samples of the same word, sentence and paragraph are compared to detect the perceptual influence of training. Since there is not much experience available in evaluation of EL speech this standard was established to be used in our case.

All in all thirteen listeners attended the test, three women and ten men (average age: 25 female, 26.7 male). Two of them have been experienced EL users, the rest were naive listeners. The speech material consists of 160 words, 120 sentences and 24 longer sentences of the paragraph. To ensure that all samples have the same volume we performed RMS adjustment. The listening test was carried out in an ordinary office where the stimuli were presented via a laptop and AKG K 271 headphones. Stimuli order as well as AB order were randomized.

The criterion for evaluation was 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 EL produces a continuous excitation signal. As a result continuous speech is possible. If the amplitude of the measured EMG signal is not high enough there will be unpredictable interruptions in the speech which cause difficulties in intelligibility and increase annoyance. We expect that speech interruptions play a major role in the evaluation. In the case of uncertainty the example with better intelligibility will reach better results. In order to make the listening test less exhaustive we took care that the test does not take longer than 30 minutes. All in all there are 800 evaluation results for words, 600 evaluation results for sentences and 120 evaluation results for the paragraph, altogether 1520 evaluations.

The overall results of the listening test is presented in Figure 3.31 which illustrates an improvement of EMG-EL speech due to training. 71 % of the listeners rate post-trainings as “slightly better”, “better” or “much better” than the recordings before training and only 13 % of the post-trainings are evaluated with “slightly worse”, “worse” or “much worse”. For the remaining 16 % no differences are recognizable.

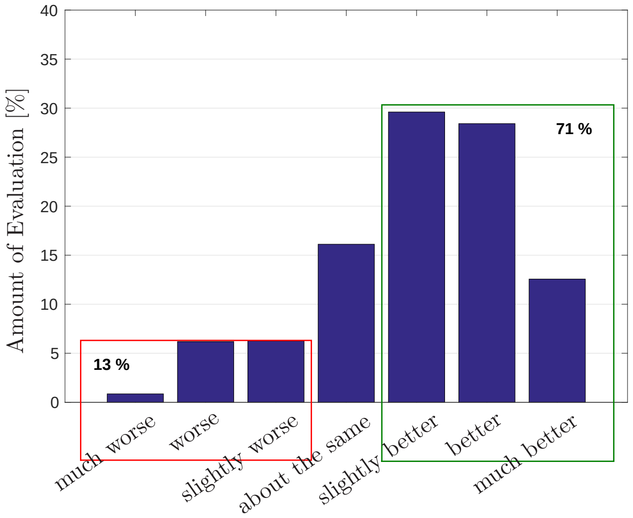


Figure 3.31: Results of all three parts of the listening tests, all speakers, the whole speech material and all listeners (1520 evaluations); “better” refers to the post-training recordings as being perceptually better than the pre-training recordings.

To evaluate the results in more detail Figure 3.32 shows the results separated into words, sentence and paragraphs. Words are evaluated worse than sentences or paragraphs. Due to the missing context words are more difficult to understand and pronounce even if the EL is working perfectly. Even one missing letter can already change the meaning of the word or make the word less intelligible. Moreover, detailed examination of the material shows that speaker F3 had more problems with the EL device than the other participants during training. Especially in pre-trainings, whispering speech could be heard. Some of the participants prefer this whispering in pre-trainings to the post-training stimuli because it increases intelligibility. Furthermore, we can not exclude that participants of the test confused the AB rating. However, this is a random effect and only effects the variance and should not bias our results.

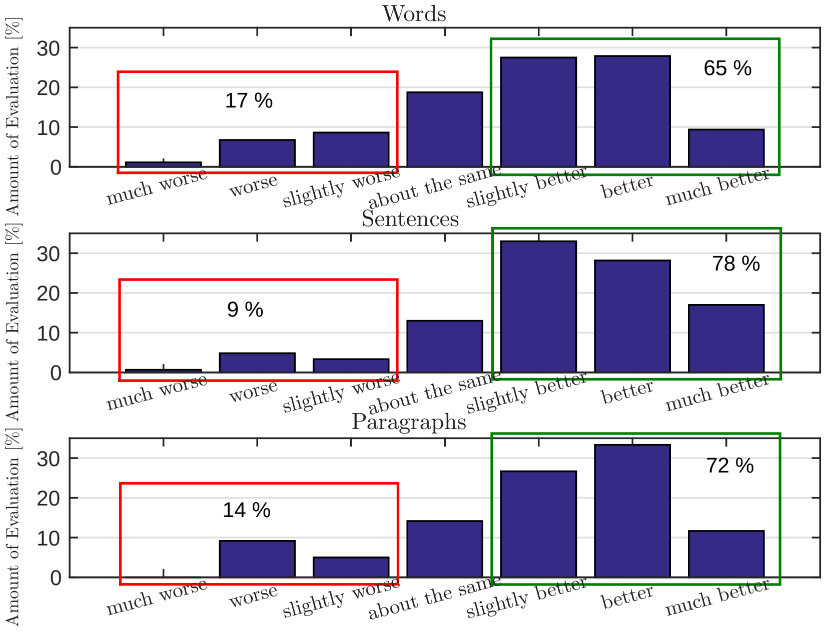


Figure 3.32: Results of the listening test for words, sentences and paragraph, averaged over all listeners and speakers. (800 evaluations for words, 600 evaluations for sentences and 120 evaluations for the paragraph); "better" refers to the post-training recordings as being perceptually better than the pre-training recordings.

3.4 Summary and Discussion

In this chapter we dealt with the control of the artificial excitation source, i.e., the EL device. From previous work of other authors we know the main problems of EL speech, which are: its unnaturalness and its inconvenience. The unnaturalness is mainly due to the constant patterns in the excitation, whereas the inconvenience originates from the physical appearance of the device and the need for one hand to operate it.

In Section 3.2 we investigated the artificial excitation signal itself, meaning the shape and its parameters. The final aim is to improve EL speech in general. Improving the f_0 contour is only a first step to reach this aim. Informal perceptual evaluations and earlier open-loop methods suggest that an artificial changing f_0 contour considerably improves the naturalness and the speech quality of EL speech. The strict periodicity of the excitation signal is clearly an aspect of the unnaturalness of EL speech. The excitation signal is created using a prototype oscillation which is repeated exactly with the f_0 period. More investigations need to be carried out on how and to which amount to include perturbations to increase the naturalness of the excitation signal. Furthermore, much more care needs to be placed onto the type of excitation signal. Although, we focused an the LF model as prototype of the excitation signal, this choice only resulted from an educated guess and was not based on extensive studies. This is one of the key issues not tackled in this thesis and will be re-addressed in the outlook section. So far, no one else investigated the influence of different excitation signal on the naturalness and intelligibility of EL speech.

A method to automatically learn the f_0 contour for speech produced using an artificial larynx device was presented which is inspired by voice conversion techniques. Based on the previously presented database, spectral features were extracted and a statistical model was trained. While the energy source, the larynx, is removed for laryngectomees, the vocal tract is still (fairly) unimpaired. The method is based on the assumption that spectrograms carry information about the speech prosody. In preliminary experiments we could reach high correlation coefficients between healthy f_0 contours and $\hat{f}_0$ contours based on all frames (voiced as well as unvoiced). These values decreased if only voiced frames are taken into account. The results demonstrate that fundamental frequency estimation based on a machine learning procedure is possible and, in terms of real-time application, preferable. Our proposed estimation algorithms are based on GMM models. Another proposed algorithm is based on HMMs where additionally the time structure is taken into account. This approach was rated high in a listening test, but ρ turned out to be very small. Further investigation need to be carried out to determine a possible superiority over the GMM based f_0 estimation method.

Evaluating f_0 estimation algorithms turned out to be very cumbersome and often not satisfying. Formal listening tests are the best method to evaluate the impact of the changing f_0. The evaluation was based on a listening test using comparison category rating. We could see that random f_0 was better than constant f_0 and that our proposed methods were rated about the same as healthy f_0 and significantly better than constant or random f_0. To sum this up: The artificial, changing f_0 contour improves electro-larynx speech and our proposed strategies are able to improve EL speech in our framework. This result confirms the findings in [Meltzner and Hillman, 2005] who state that EL speech can be improved by providing proper f_0 information.

The samples for the previous listening test were synthesized using PSOLA. In a later experiment we generated the excitation signal, output it via an EL speech system and spoke synchronously to the pre-determined f_0 patterns. In this case no synthesizing algorithms needed to be used and, therefore, also no synthesizing errors were generated. In this experiment we compared utterances spoken by two healthy subjects using the EL device. We provided two kinds of excitation signal: 1) f_0 taken from healthy speech and 2) $\hat{f}_0$ estimated based on a GMM voice conversion algorithm. The used excitation prototype was the *LF* pulse introduced earlier. Since the median of the listening test is -1 it can be concluded that the excitation signals with $\hat{f}_0$ are with a significance of 5 % "slightly worse" than with a healthy f_0 contour. Improving the estimation results and eliminating side issues we are convinced of the capability of our proposed algorithms. There are no standardized methods to evaluate changing f_0 contours for EL speech. Therefore, comparison between different approaches is not straightforward. The evaluation is also influenced by other factors, like SNR of the DREL noise, absolute mean fundamental frequency of the device and thus, of the gender and the used excitation signal. Nevertheless, the advantage of our proposed method is that it is automatic. No active control of the user, like arm movement, are needed. Our experiments showed that an automatic control of f_0 contours can improve EL speech.

The design of a listening test to evaluate EL speech is difficult. EL speech is very different from HE speech. Therefore, standardized methods to evaluate HE speech cannot be simply applied to EL speech. We suggest not to compare EL speech directly to HE speech because these types of speech are too different and different assessment criteria are applied by humans. In particular, naive listeners who are not used to the sound of EL speech are very critical. It is very important to communicate the upper bound of quality concerning EL speech.

An obvious point to consider is the introduced time delay. Recording the data, calculating the features, estimate the frequency and generate the excitation signal takes a certain period of time. Thus, the estimated healthy f_0 contour will be shifted in time. We did not tackle this problem and do not know the influence of the shifted f_0 contour on perception of EL speech. From informal listening tests we know that a time shift of the f_0 contour, created using PSOLA, of 20 ms is audible for healthy speech. However, a shift in EL speech is less critical because there are other artifacts which disturb EL speech much more.

In Section 3.3 we evaluated the possibility to use EMG signals for a hands-free control of the EL device. The advantage of an EMG based on/off control is the insensitivity of EMG to background noise. We developed a hardware to capture EMG signals which correlate with the activity of the muscles. We recorded EMG and speech samples from a male and a female speaker and performed on/off control using amplitude smoothing techniques and threshold approaches. In the evaluation we compared root mean square (RMS) envelope, Hilbert transform (HIL) and low-pass filtered (LP) signal together with single threshold (ST), double threshold (DT) and GMM based classification technique. We could reach very low total error rates with around 6 % for the male speaker and 10 % for the female speaker (LP-DT). The results turned out to depend on the SNR of the EMG signal. The thresholds were chosen speaker dependent, because they change with the SNR. The SNR was lower for the female speaker than for the male speaker. This corresponds to the findings in [Goldstein et al., 2004] which say that the amplitude of the EMG signal depends on the pitch; the lower the pitch, the larger the EMG amplitudes. The different envelope calculation methods differ in how smooth they become. The smoothing time is one of the limiting factors in real-time because we can only look a certain time into the future. The proposed Hilbert envelope method (HIL) has the advantage that it incorporates an amplitude follower and, therefore, includes a smoothing. In our experiments Hilbert envelope performed better than RMS and about the same than LP for low SNR values. For high SNR values LP outperformed RMS and Hilbert envelope. Furthermore, the Hilbert envelope incorporates a moving average filter which might be an advantage over the IIR low-pass filter (LP) because of the attack time of the 3-pole IIR filter. GMMs suffered from the fact that their might be differences in the conditions for the training utterances and for the test. We can avoid this problem when we apply an adaptation strategy in the real-time scenario [Reynolds et al., 2000]. In our experiments a low-pass filtered version of the EMG signals together with a double threshold detection or a GMM classifier outperformed other methods. In order to perform real-time smoothing, a time constant detection smoothing was implemented. It takes advantage of the preceding effect and smoothes fast variations in the detected signal. This especially improves *FEE* errors, because the activation is shifted in time.

For future work we have to consider several things: 1. Using EMG signals, non-speech related muscle movements will trigger the EL device. We assume that within a learning phase such unwanted events can be avoided. Exceptions are tracheostoma noise and noise from saliva. Furthermore, people have to gulp now and then which could trigger the EMG controlled device. We do not know yet, how to deal with that; 2. We did not investigate the difference of EMG signals for healthy subjects and laryngectomees; 3. The analysis was performed subject dependent. We can not talk about gender dependent differences because we only evaluated signals for 2 subjects, i.e., both gender.

Furthermore, we analyzed the training effects on an implemented EMG controlled EL speech. We used RMS envelope calculation method together with ST due to its simplicity. Three female and one male participant were recruited to perform nine training sessions. The training sessions over 6 days include vowel initiation, followed by vowel duration and vowel termination using the vowel [a]. Afterwards, articulation of words, sentences and paragraphs were trained. At the first and last day of training recordings were made for the evaluation. Measurements of the error (vowel initiation time, vowel termination time and vowel duration) showed that after the training sessions the variance of these errors were reduced. This leads to the conclusion that all participants could improve the handling of the device. Furthermore, 13 listeners attended an AB listening test with a comparison category rating scale where each pre- and post-training recording of the same speech sample was compared. The results show that 71 % of the post-training recorded speech samples were evaluated "slightly better", "better" or "much better" than the samples recorded before training. We revealed that there are measurable and audible learning effects in EMG-EL speech and that it is possible to improve handling, intelligibility and pleasantness. In this section we confirmed that EMG signals are practical and easy to use. Our tests showed that naive people can learn to handle the EMG controlled EL speech system within short time. A drawback of using EMG signals is the data acquisition in terms of electrode placement. Our tests are based only on healthy subjects with intact vocal tract. The anatomy of laryngectomees is, in some circumstances, dramatically different. Within this experiment we confirmed findings from [Goldstein et al., 2004]. Moreover, we prepared a system which works in real-time and can be included into our overall bionic EL speech system.

4 Transduction of Artificial Excitation Signal

4.1 Introduction

From previous chapters we can conclude that it is not enough to change some parameters of the excitation signal. In order to really bring forth fundamental changes we also need to have a look at the device itself. As mentioned in the literature review the removal of DREL can be managed through filtering methods but, to a certain amount, also by optimizing the mechanics of the device itself. Using a different transducer we plan to increase the SNR and make the EL speech system more pleasant to wear. Furthermore, we want to be able to feed different excitation signals to the device and be flexible in terms of shape and frequency.

In this chapter a new type of transducer is proposed and the advantages and disadvantages are described. Based on the transducer we built a prototype. We investigated the compensation of the transducer transfer function as well as the neck transfer function. Then, we chose an appropriate coupler disk. The coupler disk is mounted on top of the transducer to transfer the mechanical energy of the device through the neck tissue into the vocal tract to have a better impedance-matching to the neck tissue. Based on SNR calculations we also chose an excitation signal, e.g., the shape of the signal output from the speech system. A listening test compared the new speech system to the state-of-the-art product: the Servox Digital. We can give a proof of concept but also discuss the drawbacks and still open questions in terms of the design.

The following sections originate to a great amount from the joint work with and master's thesis of [Lüchtrath, 2015]. Fundamentals of electro-acoustics are summarized in [Zollner and Zwicker, 1993].

4.2 Electro-Larynx Devices

4.2.1 Conventional Electro-Larynx Device: Electro-Dynamic Transducer

The electro-larynx device consists of an oscillating transducer. Attached to the conventional transducer is a striker which conveys the oscillation to a coupler disk. The coupler disk is held against the neck tissue and the oscillation excites the air in the vocal tract. State-of-the-art EL devices are based on an electro-dynamic transducer. This kind of transducer is used in traditional loudspeakers. Figure 4.1(a) shows the structure of a conventional EL and Figure 4.1(b) shows the structure of the contained electro-dynamic transducer.

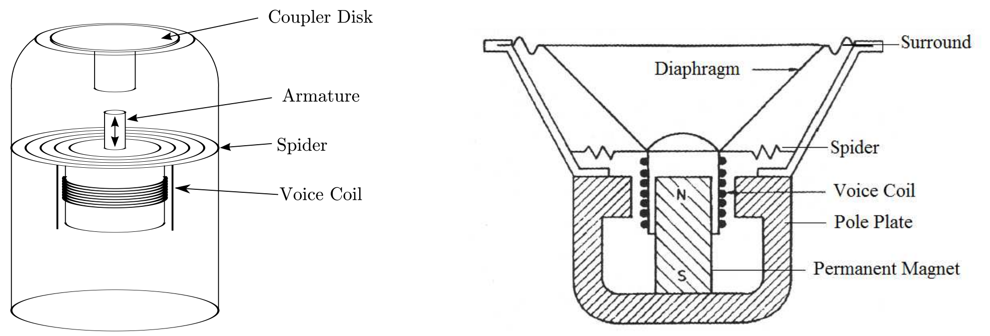
The electro-dynamic transducer is explained in the following: The driving force of the electro-dynamic transducer is the Lorentz force F_L. F_L is defined as the force acting on a current-carrying conductor of length l in a magnetic field [Zollner and Zwicker, 1993] and follows the equation:

$$F_L = B \cdot l \cdot I. \tag{4.1}$$

B is the induction of the magnetic field and I is the current. The force reaches its maximum when the direction of the current is orthogonal to the magnetic field lines. Therefore, the magnetic field should be constant which is realized using a permanent magnet. For the sake of completeness, we mention the further principle, which is that the current-carrying conductor is moving with v and induces a voltage U:

$$U = B \cdot l \cdot v. \tag{4.2}$$

Figure 4.1(b) shows the parts contained in the electro-dynamic transducer: The static permanent magnet and the moveable voice coil which is wound around cardboard, are the most important parts. The voice coil dives into the air gap of the permanent magnet. In order to prevent the movements of the magnet, it is attached to the housing of the conventional EL device. Due to the magnetic field of the permanent magnet, the F_L acts on the voice coil. The direction of I controls F_L and further on the direction of the voice coil movement and thus the diaphragm. The proportion between F_L and the I is linear for this transducer. In case of the conventional EL, a striker is connected to the voice coil which knocks due to the oscillations of the voice coil on a coupler disk. This impulse-like excitation makes the transducer non-linear. As a result we are very limited in the possible changes of the waveform of the excitation signal. In fact we can control the frequency of the device, but the parameters of the excitation signal are defined by the impulses.



(a) Electro-dynamic transducer of a conventional EL (inspired by [Houston et al., 1999]).

(b) Electro-dynamic transducer in a loudspeaker [Weselak and Graber, 2009].

Figure 4.1: Cross-section of conventional electro-dynamic EL device.

4.2.2 Proposed Electro-Larynx Device: Electro-Magnetic Transducer

An electro-dynamic transducer, as described in the previous section, is developed and optimized for loudspeakers. In that scenario electric energy is transformed into acoustic energy and radiated through the air. In contrast, electro-magnetic transducers are established in telephones, hearing aids and bone conductors. In such devices the coupling to the tissue and thus the power transmission is well defined. Therefore we propose an electro-magnetic transducer for a new EL speech system.

For the understanding of the electro-magnetic transducer Figure 4.2 is introduced. A detailed description including figures and demonstration can be found in the web: [Institut, 2015]. The electro-magnetic transducer consists of a coil which is wound around a permanent magnet and the diaphragm which contains metal. In the previous section we explained the electro-dynamic transducer, where the coil (current-carrying conductor) is moving. In case of the electro-magnetic transducer, the coil is fixed, but the magnetic field strength H is changed. H depends linearly on B:

$$B = \mu \cdot H, \tag{4.3}$$

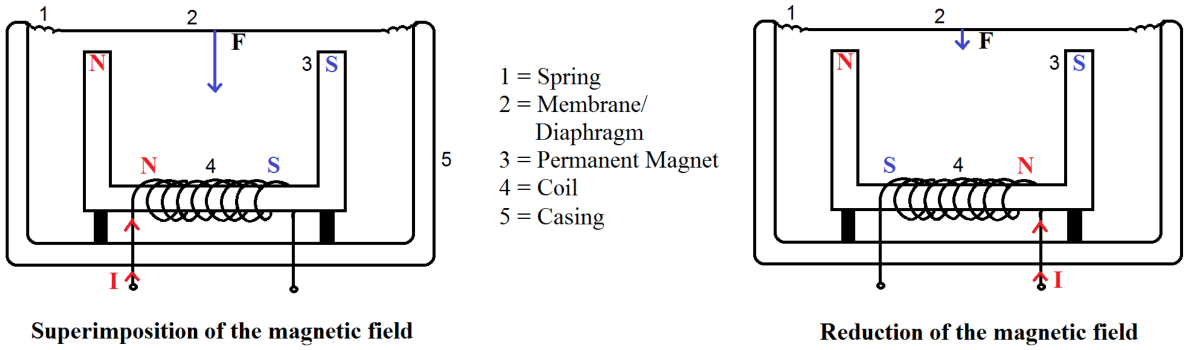


Figure 4.2: Cross-section of an electro-magnetic transducer; Left side: Superimposition of the magnetic field of the coil and magnet – force acting on the diaphragm is increased; right side: (current changes the direction) Reduction of the magnetic field of the magnet because magnetic field of the coil operates in the opposite direction – force acting on the diaphragm is reduced [Zollner and Zwicker, 1993].

with μ being the permeability. In other words, the diaphragm is moving and evokes a changing air gap between diaphragm and magnet and a changing magnet field. The assumptions to approximate the magnetic field are:

- The magnetic resistance depends only on the air gap between the magnet and the diaphragm, because the magnetic resistance of the diaphragm is, compared to the magnetic resistance of the air gap, negligibly.

- The relative alternation of the air gap remains small, otherwise the diaphragm would touch the magnet while getting displaced: $x_{\sim} \ll x_0$, where $x_{\sim}$ stands for the displacement and x_0 is the distance between the diaphragm and magnet.

The current-carrying coil in the electro-magnetic transducer generates a magnetic field which interacts with the magnetic field of the permanent magnet. The force F acting on the diaphragm depends on the magnetic field strength H in the gap between the permanent magnet and diaphragm and is given by

$$F = \mu \cdot S \cdot H^2, \tag{4.4}$$

with S being the cross-section of the air gap.

The polarization of the magnets depends on the magnetic flux Φ which is defined as:

$$\Phi = B \cdot S. \tag{4.5}$$

Inserting (4.3) and (4.5) into (4.4) results in

$$F = \frac{\Phi^2}{\mu \cdot S}. \tag{4.6}$$

In Figure 4.2 (left) current flows through the coil. The magnetic flux has the same direction in both magnets (permanent magnet and diaphragm). The force is in direction to the magnet because the polarization of the permanent magnet depends on the direction of the magnetic flux. If the current flows in the opposite direction (see Figure 4.2 – right) the polarization of both magnets change and the direction of the force stays in the same direction. This is also visible in (4.6): the force is independent from the direction of the magnetic flux. Due to the quadratic correlation between force and magnetic flux an alternating current (AC) results in oscillation of the diaphragm the double frequency.

In order to linearize this quadratic behavior, a direct current (DC) field H_0 with current I_0 is superimposed. From (4.6) we can expand to

$$F = \frac{(\Phi_= + \Phi_\sim)^2}{\mu \cdot S} = \frac{\Phi_=^2}{\mu \cdot S} + 2 \cdot \frac{\Phi_= \Phi_\sim}{\mu \cdot S} + \frac{\Phi_\sim^2}{\mu \cdot S}. \tag{4.7}$$

(4.7) consists of three parts: The first part is quadratic and is important for the maximum sound pressure, the second part is linear and depends on the DC and AC. DC amplifies the force on the diaphragm and, therefore, increases the efficiency. The third part is again quadratic and determines the distortion of the transducer.

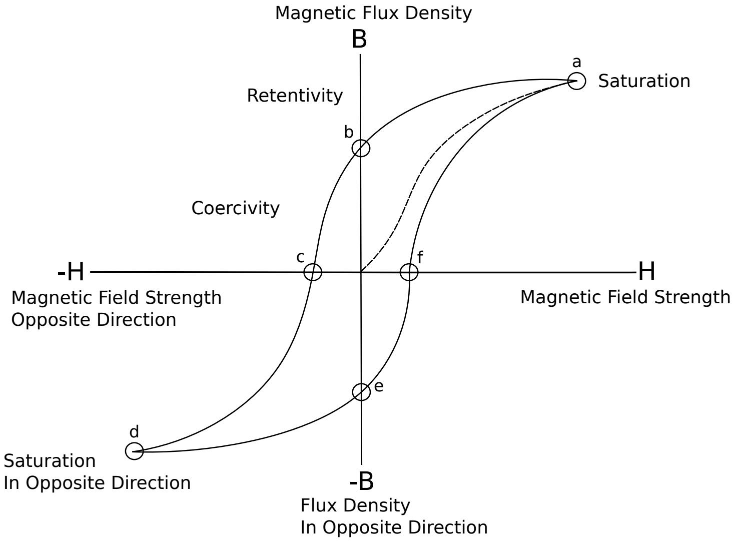


Figure 4.3: Hysteresis curve of a ferro magnet (inspired by [NDT, 2015]).

Due to its magnetic properties the diaphragm is always attracted to the magnet. This means that the air gap is crucial for this kind of transducer because the magnetic field depends on the air gap. Depending on the current flow, the diaphragm is more or less attracted to the permanent magnet. The air gap could be seen as a resistance, the bigger the gap, the higher the resistance. Additionally, the higher the magnetic flux, the higher is the force to close the magnetic circuit.

A further property is the polarization by the magnet which is characterized using the hysteresis curve in Figure 4.3. The hysteresis shows the relation between magnetic field strength and magnetic flux density. For yet un-magnetized materials, the curve starts in the origin (dashed line). In point (**a**) a saturation effect occurs. At this point an increase of the magnetic field strength will only slightly increase the magnetic flux density. A reduction of the magnetic field strength to zero will result in a remaining magnetic flux density in the material (**b**). This point is called *retentivity* or *remanence*. When the magnetic field strength is inverted the magnetic flux density is decreased to zero (**c**). This is the point of *coercivity*. The *coercivity* is needed to reduce the remaining magnetism of the material. Again, a saturation effect is visible when the magnetic field strength is further decreased (**d**). Now, increasing the magnetic field strength (in positive direction) leads to a lower increase than the previous decrease. The point where the magnetic flux density returns to zero is neither point **c**, nor the origin, but point **f**.

Every oscillation of the input current results in passing through the hysteresis curve which leads to power dissipation. The shape of the hysteresis curve and thus the power dissipation depends on the used material. The higher the frequency, the faster the oscillations and, therefore, the higher the power losses. This leads to a rapid decrease of efficiency of the electro-magnetic transducer for high frequencies.

To summarize: Technically an electro-magnetic transducer is an improvement in comparison to the normal electro-dynamic transducer explained above. The main advantage is the higher efficiency together with a lower power consumption due to its quadratic principle. The efficiency of the electro-dynamic transducer is about 1/7 to 1/10 times than that of an electro-magnetic transducer [Institut, 2015]. Another advantage is that the elements of an electro-magnetic transducer are quite small which results in an overall smaller construction. A small transducer leads to a small design of the EL which is necessary for a comfortable hands-free design. Disadvantages in comparison to the electro-dynamic transducer are the higher production of distortions which are a side effect of the quadratic principle and the rapid decrease of the efficiency towards higher frequencies.

Properties of the Electro-Larynx Transducers

The electro-magnetic transducer used in this thesis is a bone conductor BC2-E31 produced by the company BHM [BHM, 2015] which is normally used in bone conduction hearing aids. Figure 4.4 shows the dimensions of the transducer [BHM, 2004]. In western Europe one of the most common conventional EL devices is the Servox Digital by Servona. In this thesis we used this device for comparison with the proposed EL speech system.

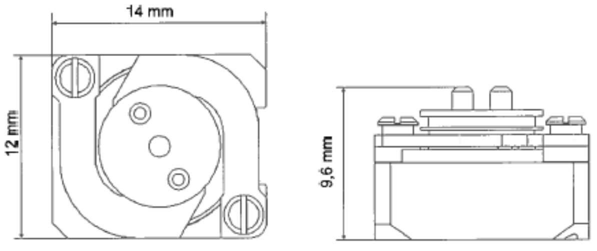


Figure 4.4: Cross-section of the electro-magnetic BHM BC2-E31 bone conductor [BHM, 2004].

We measured the impedance of both transducer according to the measurement setup in Figure 4.5. The ARTA software [Arta, 2015] includes a collection of programs for audio measurement and analysis and was installed on a computer. Test signals were generated within the software and sent to the Device under Test (DUT) via D/A converter (Audio Interface – RME Fireface UFX), QSC CX168 power amplifier and ARTA analyzer unit [Hiebl, 2014]. The analyzer unit was developed in our laboratory and is a connection between the hardware components of the measurement setup. The used measurement signal was a stepped sine signal. (DUT) were the electro-magnetic transducer and the Servox Digital device. The proposed transducer was mounted into a housing (see Appendix A, Figure A.3). In order to connect the Servox Digital to

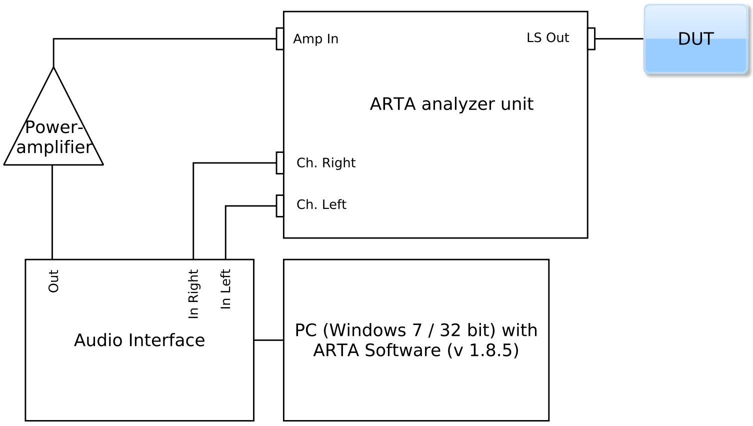


Figure 4.5: Impedance measurement setup (inspired by [Hiebl, 2014]).

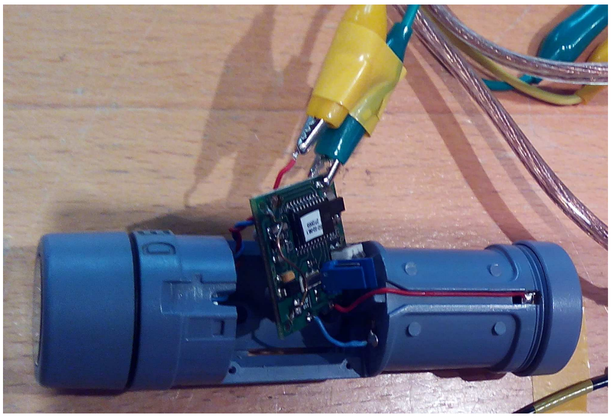


Figure 4.6: Conventional Servox Digital prepared for impedance measurement.

the measurement setup we bared the cables (see Figure 4.6). The results can be seen in Figure 4.7. The blue line is the measurement for the conventional electro-dynamic EL (Servox Digital) and the red line represents the proposed electro-magnetic transducer. The resonance in the blue line is a result of the coil-coupler disk structure of the conventional EL device. At a frequency of around 120 Hz the coil looses the connection to the coupler disk. This frequency depends on the EL device, the coupler disk and how tight the coupler disk is screwed. The increase of the impedance for frequencies higher than 300 Hz is visible for the proposed transducer. This results in damping with increasing frequencies. Our measurement matches to the impedance of the electro-magnetic transducer of the data sheet (Figure 4.8). The nominal impedance at 1 kHz is 60 Ohm ± 20 %. With increasing frequency also the impedance increases. The DC resistance (resistance at 0 Hz = real resistance of the coil without any frequency dependency) is 16 Ohm ± 10 % [BHM, 2004].

The conventional EL device has a length of 118 mm and a diameter of 35 mm. The weight without battery is about 60 g [Servona, 2015]. Unfortunately, no additional data for this transducer is available. The newly proposed electro-magnetic transducer has a maximum power consumption of 10 mW. The weight is only 8 g which makes the whole system with enclosure very light.

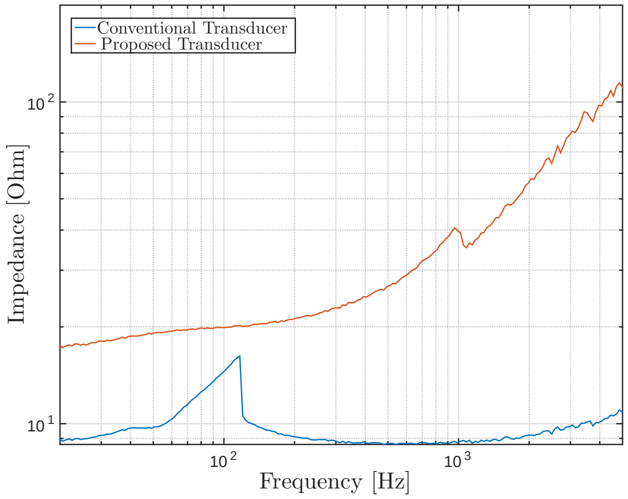


Figure 4.7: Impedance curve of the electro-magnetic transducer (proposed transducer) in red and the electro-dynamic Servox Digital EL (conventional transducer) in blue; Measurements were taken with neck tissue load.

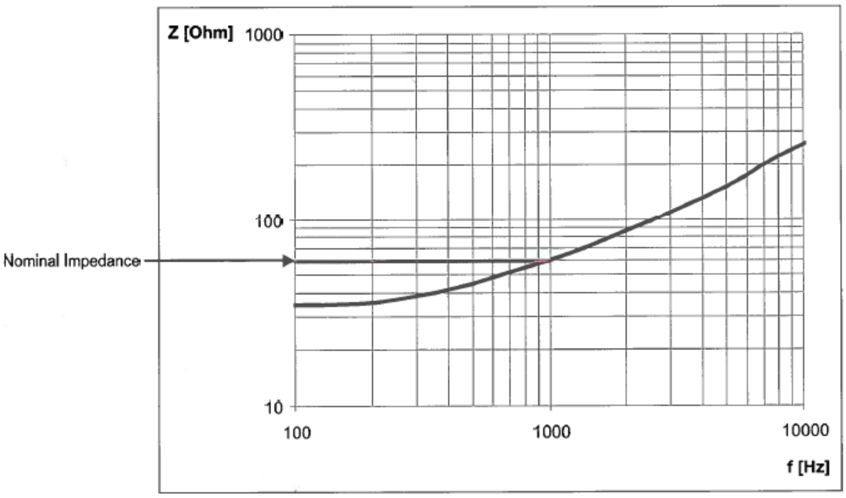


Figure 4.8: Impedance of the electro-magnetic transducer over the frequency according to the data sheet [BHM, 2004].

4.3 Experiment VI: Influence of Neck and Transducer Transfer Function

Following the signal chain, the excitation signal of the EL speech system is filtered by the transducer transfer function, by the neck transfer function and finally influenced by the vocal tract. In the following experiment, we wanted to investigate the effects of the transducer as well as the neck transfer function.

Transducer Transfer Function

The linear part of the transducer transfer function was measured with the setup shown in Figure 4.9. We measured the impulse response using the same measurement setup as in the previous section except that we record the output of the DUT using the omni-directional measurement

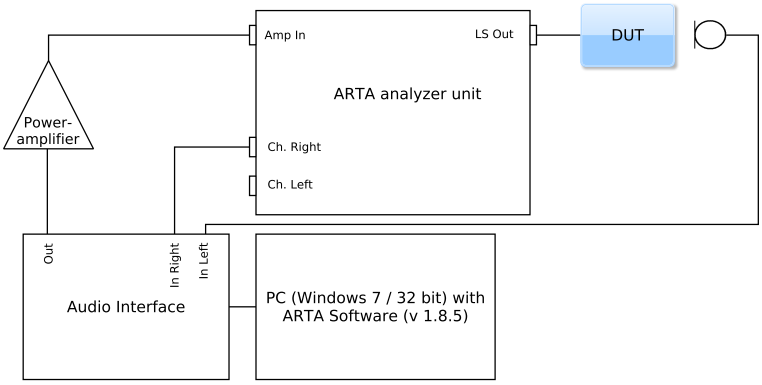


Figure 4.9: Transfer function measurement setup (inspired by [Hiebl, 2014]).

microphone DPA 4006-TL. Afterwards, the transfer function was calculated using the FFT. We used a maximum length sequence which is a pseudo-random, binary sequence as measurement signal. All measurements were repeated 3 times and averaged to obtain the transfer function results. The DUT was the proposed electro-magnetic transducer within a housing (see Appendix A, Figure A.4). All measurements were carried out using coupler disk 7 and 20 (for more information concerning coupler disks see next section). We did not measure the conventional device using this setup. The excitation signal of the conventional device is highly non-linear. The pseudo-random measurement signal is not able to excite the conventional device appropriately. Excitation output of the conventional EL device is only produced in a limited frequency range around 100 Hz. Alternative measurement procedures need to be applied for such non-linear transducers.

In the first measurement the proposed device was held directly in front of the measurement microphone (few mm). In order to prevent any damping we fixed the device at the cable (Figure 4.10).

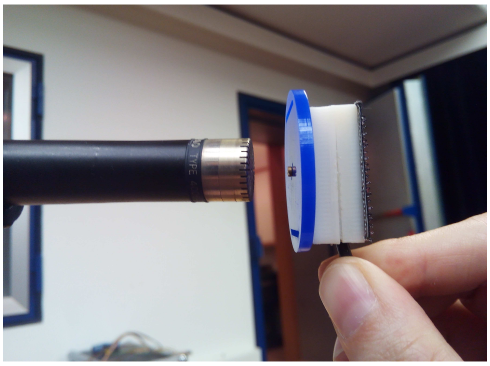


Figure 4.10: Measurement setup for impulse response.

We measured the impulse response one time without coupler disk, one time with coupler disk 7 and one time with coupler disk 20 (see Appendix A, Table A.1). Figure 4.11 shows the transfer functions of the proposed transducer. There is a strong resonance around 1150 Hz. The coupler disks shift the resonance towards a lower frequency. Additionally, they emphasize the frequency range between 100 Hz and the resonance frequency.

In the second measurement a person was holding the proposed transducer in the hand and simulated speaking. The microphone was directly in front of the persons lips. One time the transducer was pressed against the neck tissue (with tissue), one time the transducer was fixed

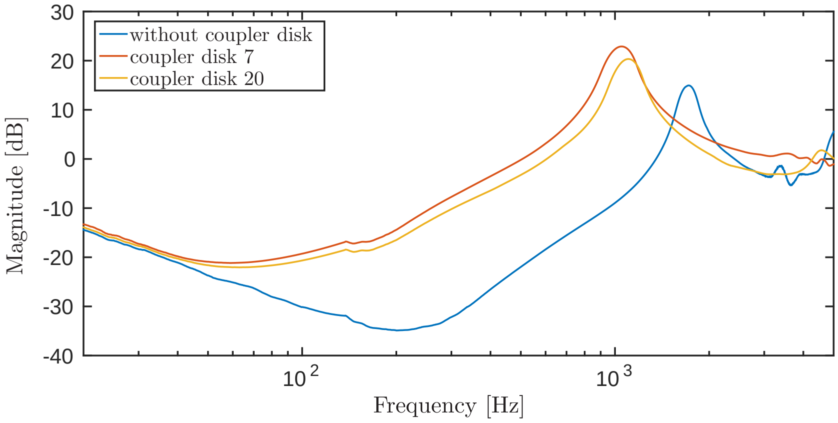


Figure 4.11: Measured transfer function of proposed transducer without and with two different coupler disks; transducer is always un-loaded.

without skin contact (without tissue). Figure 4.12 shows the results using coupler disk 7 and 20. In the blue line the transducer was loaded with the neck tissue. The mouth was closed, so no sound radiated from the lips. This measurement includes the effect of the transducer and the neck transfer function. In the upper as well as in the lower plot, a band-stop effect for the measurement without tissue can be seen.

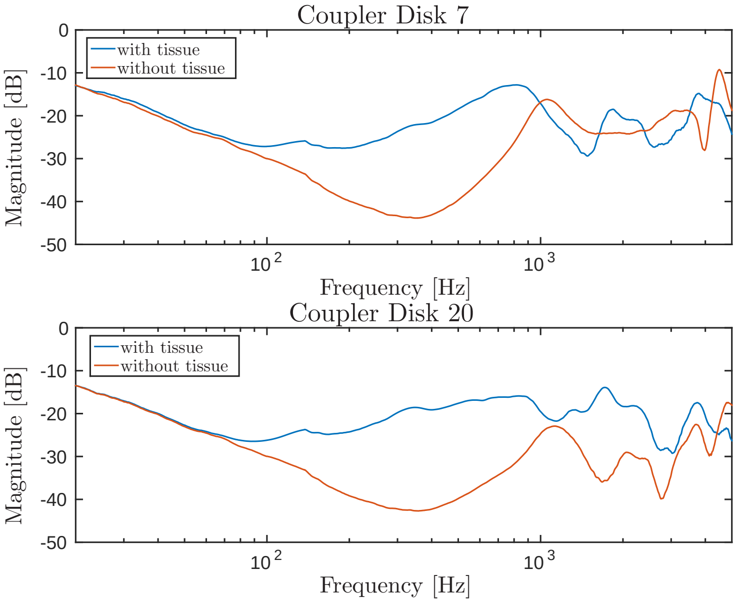


Figure 4.12: Measured transfer function of proposed transducer with two different coupler disks; blue line: transducer is loaded with neck tissue (closed mouth), red line: transducer is moving free (un-loaded).

In the third measurement the person who was holding the proposed transducer produced the two vowels [a] and [o]. In Figure 4.13 the transducer transfer functions loaded with the neck tissue and pronunciation of two different vowels are presented. The formant frequencies of the vowels are clearly visible.

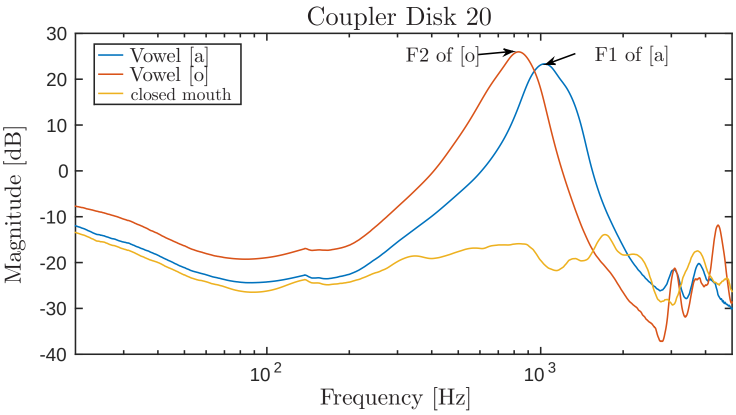


Figure 4.13: Measured transfer function of proposed transducer for coupler disk 20; blue line: vowel [a] is produced – first formant (F1) at around 1 kHz, red line: vowel [o] is produced – second formant (F2) at around 1 kHz, yellow line: closed mouth.

Neck Transfer Function

The neck transfer function was measured and analytically described by [Meltzner et al., 2003]. For the measurement a microphone recorded the sound pressure at the lips and an impedance head mounted on the exciter measured the acceleration. The neck transfer function is then defined as the ratio of the spectrum of the estimated volume velocity that excites the vocal tract to the spectrum of the acceleration delivered to the neck wall. The authors measured the neck transfer function of several people and reported a notable amount of inter-subject variability. Nevertheless, according to the author the neck transfer function can be accurately modeled as a low-pass filter due to the filtering effect of the tissue towards high frequencies. The analytical neck transfer function is suggested as two conjugate pairs of poles and one conjugate pair of zeros. Measurement results and analytical fit of the neck transfer functions for the groups laryngeal female, laryngectomized female, laryngeal male and laryngectomized male can be seen in Figure 4.14.

Results

Measuring the transfer function of the proposed transducer is not straightforward because the electro-acoustic properties are different when the transducer is under load. This means that the transfer functions of the neck as well as the transducer are measured.

[Meltzner et al., 2003] reported a difference between the implemented neck transfer function for laryngeal, female speakers and laryngeal, male speakers as well as the laryngectomized, females and males. Although the neck transfer functions vary from one person to another person it is not feasible to measure speaker individual transfer functions in real-world. The reported results suggest that there are differences of the neck transfer function between different groups, but the general behavior of the neck can be approximated using a low-pass filter. Nevertheless, individual measurements would be necessary for precise filtering.

In the sound pressure level measurements of Figure 4.12 the transducer shows band-stop characteristic when no load was attached. In case of the loaded transducer the neck transfer function, which is approximately a low-pass, amplifies the frequency range between 100 Hz and 1 kHz. This leads to a flatter combined transfer function. In future we have to establish a method for measuring the exact transfer function of the proposed transducer.

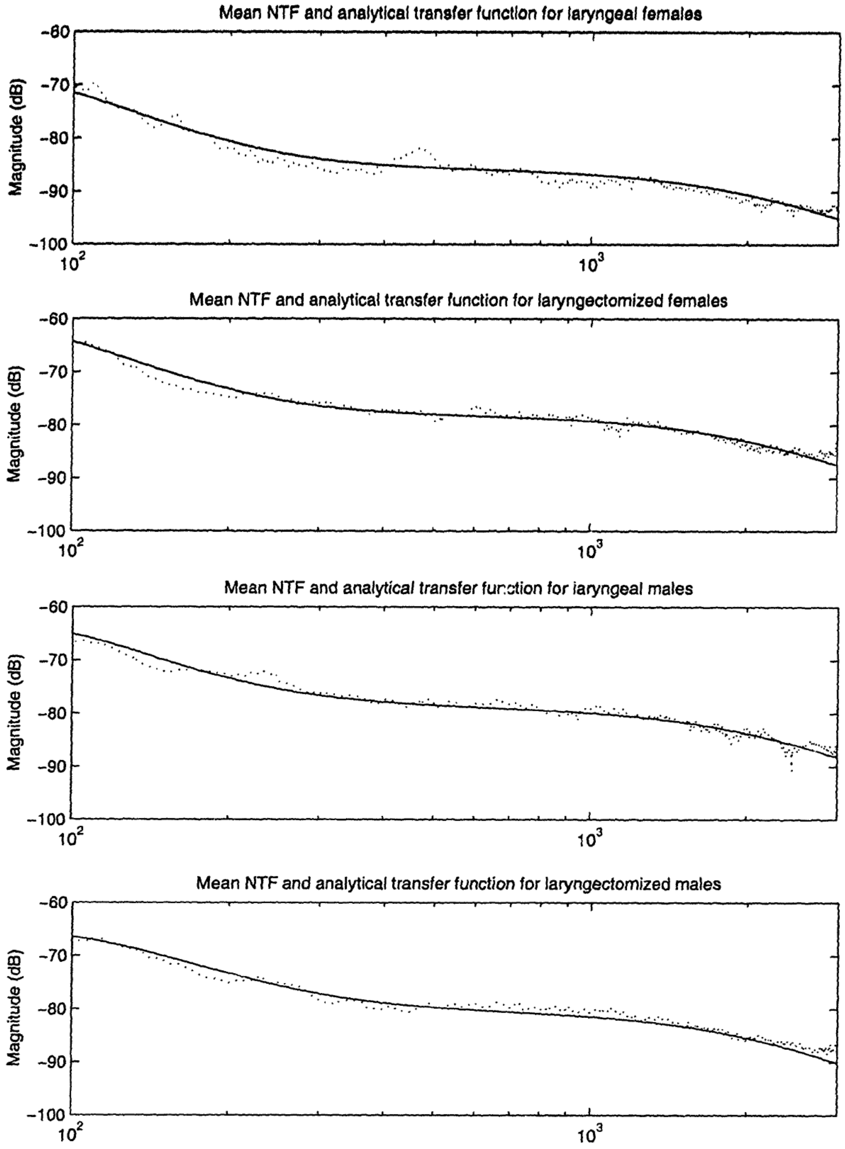


Figure 4.14: Mean of the measured neck transfer functions (dotted line) for laryngeal females and males as well as laryngectomized females and males with corresponding analytic transfer function fit (solid line) as proposed by [Meltzner, 1998].

4.4 Experiment VII: Choice of Appropriate Coupler Disk

The coupling element between the EL and the neck is a coupler disk and it plays an important role. Its tasks are to transmit as much energy as possible into the neck (in other words: to minimize the energy losses) and to be comfortable and pleasant to wear. In this section we investigate whether different coupler disks influence the proposed electro-magnetic transducer and if there is a way to find the optimal coupler disk for the transducer.

As said before, the coupler disk should be wearable. Therefore, we investigated the material of acrylic glass. The shapes varied from round to rectangular and flat to bent with different diameters. Additionally some coupler disks had slots. The slots reduce the weight of the coupler disk and have impact on the oscillation behavior. To investigate the effect of different coupler disks we chose to record short utterances and calculate the SNR as explained in Section 2.2.1. All coupler disks including their weight are shown in Table A.1 in the Appendix A. Each coupler disk was mounted onto the proposed electro-magnetic transducer (see Figure A.4). A healthy speaker uttered the German numbers (Eins, Zwei, Drei = One, two, three). The excitation signal of the transducer was the *Gauss Pulse* (see Section 3.2). The utterances were recorded with an omni-directional headset condenser microphone AKG HC 577 L at a distance of about 2 – 3 cm from the mouth corner as well as in 1 m distance with a condenser microphone AKG C414 with cardioid characteristic. These distances cover two setups: 1. The signal energy is highest and the disturbing DREL is lowest directly at the opening of the mouth, 2. In a typical speaking situation the dialogue partner is at a distance of about 1 m or more.

The results are shown in Table 4.1 for the recordings in 1 m distance as well as for the headset recordings. Analyzing the sum of the SNR values in 1 m distance and for the headset recordings, suggest that coupler disk 9 and 13 were working best. The SNR in 1 m distance are 14.91 dB and 14.07 dB. With headset the results are 24.30 dB and 24.13 dB with headset. The difference between the SNRs, however, is quite large (around 10 dB). Coupler disk 17 and 20 reach the highest values for the recordings in 1 m distance (17.16 dB and 17.08 dB). Similar values can be reached for the headset recordings. The worst coupler disks were found to be coupler disk 1 (14.45 dB), as well as coupler disk 3 (15.03 dB) for the recordings with the headset. For the 1 m recordings, coupler disk 15 (9.46 dB) and coupler disk 23 (7.86 dB) were working unsatisfactorily. For these coupler disk, the sound radiation to the back is large. As a result, much DREL noise reaches the microphone in 1 m distance.

Sometimes there is a huge difference between SNR values for 1 m distance and headset recordings (for example coupler disk 23). The sound radiation towards the back (and thus, towards the microphone in 1 m distance) depends on the size of the coupler disk, whether it is flat or not and on the slots. These parameters, together with the transducer build a complex mass-spring system. The position of the coupler disk at the neck, the angle of the bending and the contact pressure changes the sound radiation towards the back and the coupling to the tissue.

Since a listener is not standing directly next to the mouth, recordings in 1 m distance are of greater significance than the headset recordings. The coupler disks for the best SNR values in a distance of 1 m are rectangular with midsize dimensions (37×18 mm) and round with midsize to large dimension (40 mm to 50 mm). We suggest coupler disks with these dimensions and with similar SNR values to the headset recordings.

Interestingly, the optimal coupler disks have not the same shape as the round state-of-the-art Servox Digital coupler disk with a diameter of around 22 mm. The round, flat coupler disk with the dimensions 20×2 mm (coupler disk 1) is worst for headset recordings and amongst the worst in 1 m distance.

As described in Appendix A we designed a small prototype housing and attached the above chosen coupler disk on top of the housing. Furthermore, we attached an elastic band to the coupler disk in order to be able to wear it around the neck without need to hold it in our hands. This is necessary in order to develop a hands-free device. The drawback of fixing the coupler disk using an elastic band is that the contact pressure to the skin is reduced. As a result the coupling to the tissue might be less optimal and SNR will decrease. We recorded 6 utterances using the headset condenser microphone AKG HC 577 L: one time using the conventional EL device, one time using the proposed transducer with elastic band and one time using the proposed transducer without elastic band. The excitation signal was the *LF* (see Section 3.2). The calculated SNR values show that the conventional device reaches the highest value (10.67 dB). The SNR of the proposed transducer without elastic band is only slightly decreased (9.56 dB), whereas the proposed transducer with elastic band obtains a value of 7.12 dB. Note: the lower SNR values compared to the values in Table 4.1 result from the changed excitation signal.

C.d.	round	rect.	flat	bent	slot	Ø Size	SNR – 1 m	SNR – hs
1	×		×			20×2	10.94	14.45
2	×		×		×	20×3	12.53	21.97
3	×		×			20×5	10.50	15.03
4	×		×			30×2	12.62	19.73
5	×		×			40×2	12.68	20.78
6	×			×		40×2	15.21	20.64
7	×		×		×	40×3	12.88	23.02
8	×		×			40×5	14.75	21.23
9	×				×	40×5	14.91	24.30
10	×			×		50×2	10.78	15.62
11	×				×	50×2	13.40	22.87
12	×		×	×		50×3	16.03	20.65
13	×			×	×	50×3	14.07	24.13
14		×	×			18×8.5×2	11.39	16.88
15		×	×		×	18×8.5×3	9.46	19.13
16		×	×			18×8.5×5	15.24	21.02
17		×	×			37×18×2	17.16	20.96
18		×	×		×	37×18×3	11.32	17.92
19		×		×	×	37×18×3	16.71	17.54
20		×	×			37×18×5	17.08	21.05
21		×	×			75×36×2	12.51	23.88
22		×		×		75×36×2	12.17	19.82
23		×		×		75×36×5	7.86	23.27
24		×		×		75×36×3	12.42	20.37

Table 4.1: Dimension in mm and SNR in dB for different coupler disks (C.d.): 1 m and headset (hs).

4.5 Experiment VIII: Choice of Excitation Signal

The next step is to find the most appropriate excitation signal. We chose four different types of excitation signals: *Pulse*, *Gauss Pulse*, *LF* and *HGS*, which have been explained in Section 3.2. In Figure 4.15 we compare the spectra from the prototype excitation signal with recordings of this signal output from the transducer. For the recordings, the headset microphone AKG HC 577 L was used. The transducer was mounted into the housing (see Appendix A, Figure A.3) and coupler disk 7 was attached. The output from the transducer was oscillating in the air. The figures illustrate the frequency range which is transmitted to the microphone. The different excitation signals are dominated by two resonance frequencies: around 1000 Hz and around 4500Hz which originate from the resonances of the device housing and the coupler disk. For the *Gauss Pulse* low frequencies are more emphasized than for the other excitation signals, but the higher frequencies are more damped. Figure 4.16 shows the output of the transducer under neck tissue load. For these recordings the transducer was pressed at the skin and thus, the resonance frequencies are moved towards lower frequencies. Again, we used SNR calculations.

distance	*Pulse*	*Gauss Pulse*	*LF*	*HGS*
headset	9.08	11.09	14.35 *	13.01
1 m	6.88	5.72	8.14 *	6.74

*Table 4.2: SNR results for different excitation signals: headset microphone and 1 m distance; * indicates significant difference according to the Kruskal-Wallis test ($p<0.05$).*

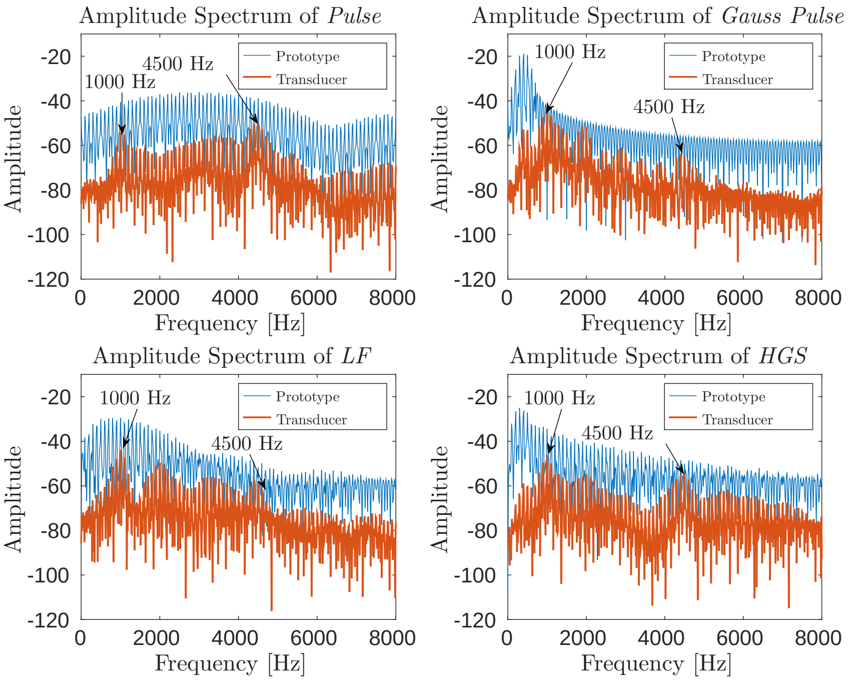


Figure 4.15: Different excitation signal: Prototype and output from transducer; spacing of the pulses corresponds to the prototype $f_0 = 100$ Hz.

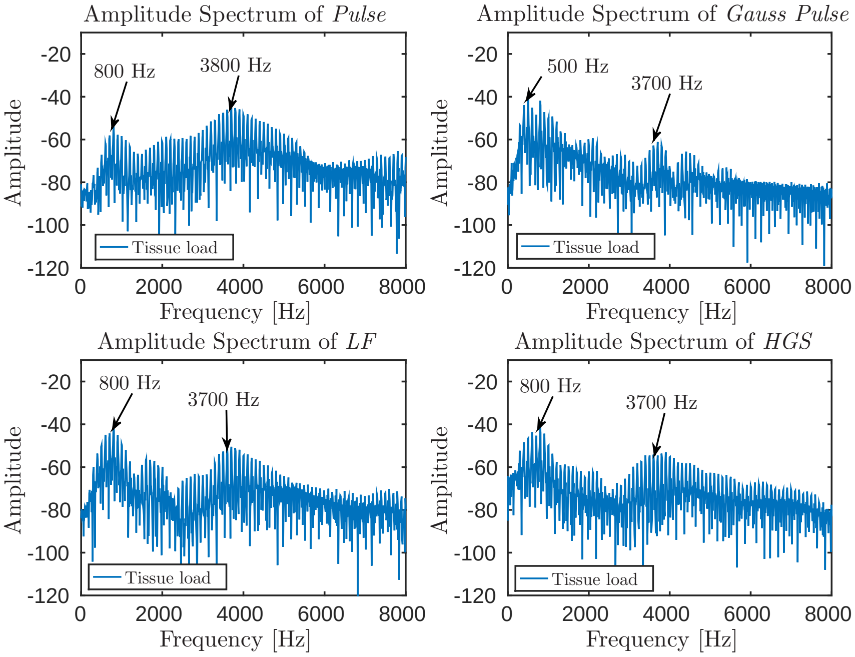


Figure 4.16: Different excitation signal: Output from transducer under neck tissue load; spacing of the pulses corresponds to the protytpe $f_0 = 100$ Hz.

The speech material is the German paragraph "Nordwind und Sonne" [Association, 1999] which consists of 6 separate sentences. We used coupler disk 20 (see Appendix A, Table A.1) which was mounted on the prototype and the proposed electro-magnetic transducer for the recordings. Also for this experiment, utterances were recorded with the headset and in 1 m distance. SNR calculations, as before, were conducted.

The results of this experiment are presented in Table 4.2. For both distances *LF* is ranked at first position. SNR values are significantly ($p<0.05$) better for *LF* excitation than all other excitation signals for both conditions, headset and 1 m recordings. We propose the *LF* model as best fitting excitation signal for our new bionic EL speech system. Please note, that this proposal is based on SNR measurements only and not confirmed by listening tests yet.

4.6 Experiment IX: Comparison of Conventional with Proposed Device

To compare the state-of-the-art EL device by Servona with the proposed speech system containing an electro-magnetic transducer, speech material was recorded where the sentences were uttered with both devices. Coupler disk 20 was attached to the proposed device. We used 24 phonetically rich utterances which were recorded with both devices using three different speakers (2 female and 1 male). For the proposed device we used an *LF* excitation signal. The devices were adjusted to a gender appropriate f_0 (100 Hz – male; 200 Hz – female). For the recordings we used the microphone AKG C414 and the recording software SpeechRecorder [Draxler and Jänsch, 2004]. We chose an office to have realistic environment including background noise and reverberation. The microphone was positioned at a distance of 1 m. Objective SNR calculations as well as an AB listening test were done. We chose the same utterance and the same speaker but with the two different EL devices for the AB pair. The order of the utterances as well as the assignment of the conventional and proposed speech system to A and B was randomized. To familiarize the listeners to the sound of EL speech, we presented four speech examples.

Objective Evaluation

For the speech material we calculated the SNR for every speaker, utterance and device (see Table 4.3). The mean SNR over all three speakers and all spoken utterances for the conventional device is 6.48 dB. Whereas the mean SNR over all three speakers and all spoken utterances for the proposed speech system is 5.97 dB. Figure 4.17 shows the boxplots with the SNR values for each device and speaker. Interestingly the SNR values for the male speaker are better for the proposed speech system.

Speaker	Servox Digital		proposed	
	$\overline{\text{SNR}}$ [dB]	$\overline{f_0}$ [Hz]	$\overline{\text{SNR}}$ [dB]	$\overline{f_0}$ [Hz]
F_1	6.83	202	4.75	203
F_2	6.94	202	5.03	202
M_1	5.67	101	8.13	100

Table 4.3: Mean SNR and mean f_0 of the conventional EL device and the proposed speech system.

Subjective Evaluation

The same 13 listeners as in experiment 3.2.3 were asked to rate how pleasant sample A is in comparison to B for all 72 sample pairs. As explained in Appendix B a comparison category rating scale was used.

The overall evaluation of the listening test for all speakers can be seen in Figure 4.18. The conventional device is assessed to be better than the proposed speech system with 70 %. Only 14 % rated the proposed speech system to be better than the conventional device. 16 % rated that both devices are about the same. Looking at Figure 4.19, where histograms are shown for

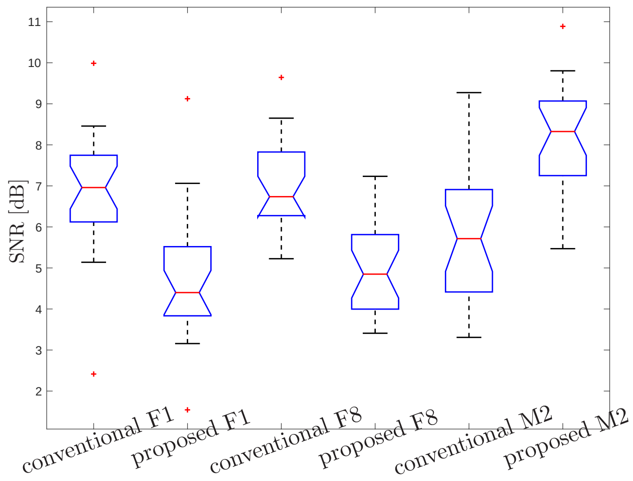


Figure 4.17: SNR results for every speaker and EL speech system; Averaged over each utterance, speaker and speech system (conventional electro-dynamic EL device and proposed electro-magnetic speech system).

each listener separately, we can see that all listeners voted in favor of the conventional device. However, the histograms in Figure 4.20 shows the ratings depending on the speaker of the utterances. In this case it is interesting that the male speaker was rated differently from the female speakers. For the male speaker a higher amount of ratings were in favor of the proposed speech system than for the female speakers. For the male speaker 30 % of the evaluations were in favor of the proposed speech system and 54 % for the conventional device. The results of the listening test correlate with the SNR calculations from Table 4.3 where the SNR of the utterances of the male speaker are better for the proposed speech system. It can be concluded that generally listeners prefer the conventional device, but the proposed speech system is better evaluated for the male speaker than for the female speakers.

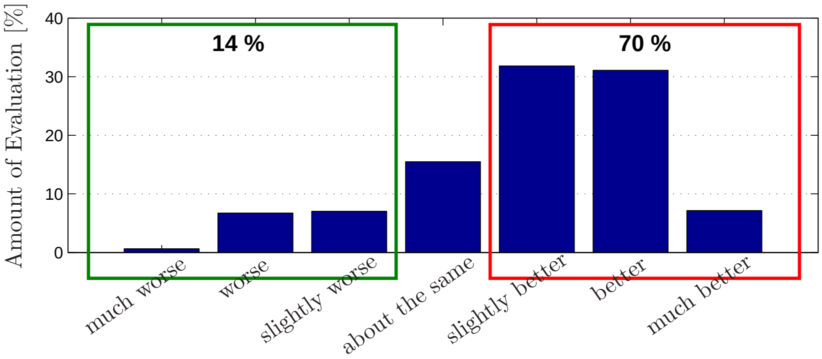


Figure 4.18: Averaged listening test results of all speakers and listeners for the comparison of the newly proposed electro-magnetic and the conventional electro-dynamic EL device; conventional Servox Digital EL device was assessed to be better than the proposed speech system with 70 %.

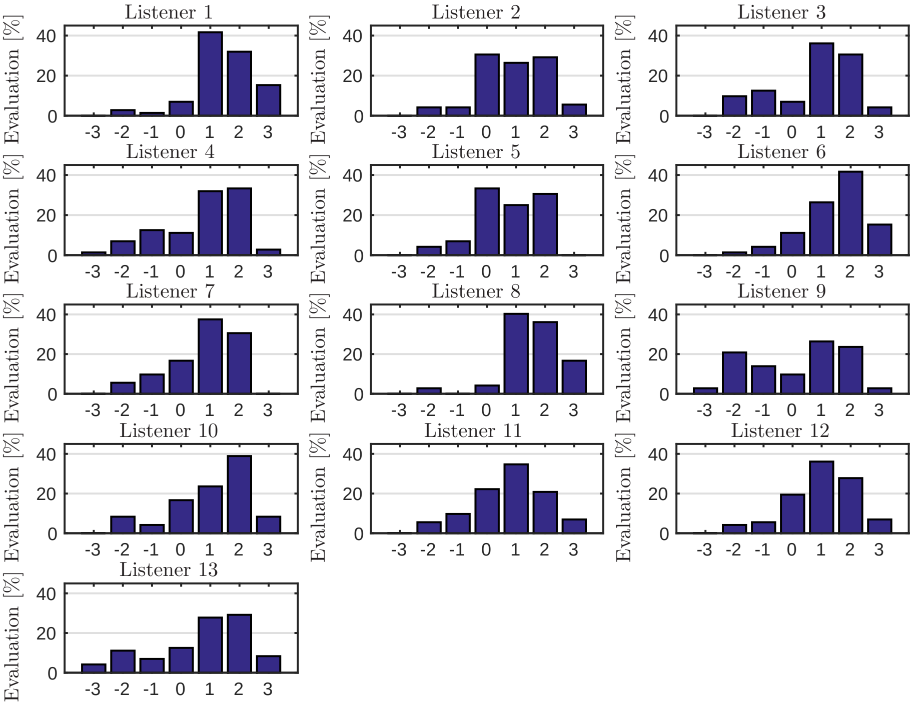
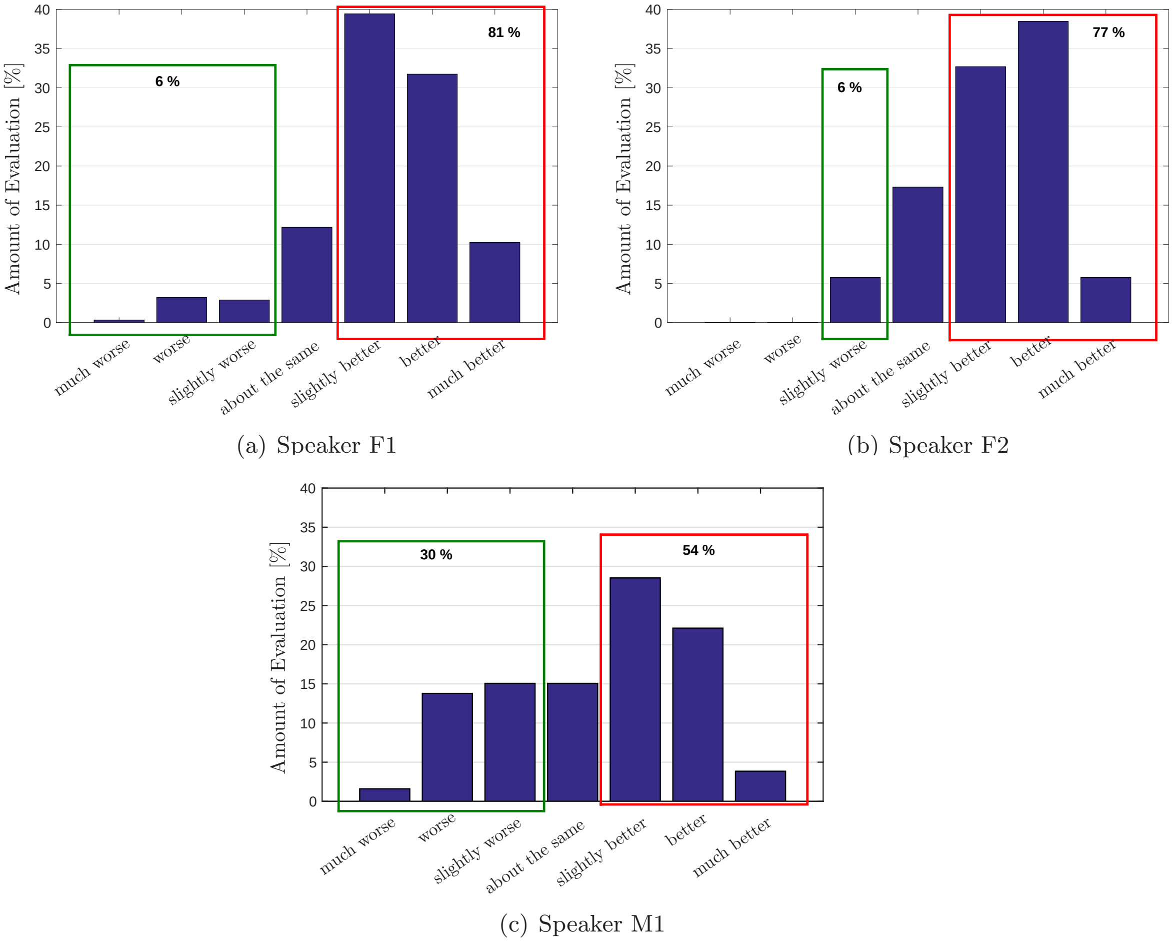


Figure 4.19: Detailed analysis of listening test results for each listener for the comparison of the two EL devices (proposed electro-magnetic speech system and conventional Servox Digital EL device); +3 means conventional system is much better, -3 means proposed system is much better.



(a) Speaker F1

(b) Speaker F2

(c) Speaker M1

Figure 4.20: Listening test results for the comparison of the two EL devices depending on the speaker; 'Better' refers to the conventional EL system as being perceptually better than the proposed EL system; Listeners rated female speakers more clearly than the male speaker (plot c).

4.7 Summary and Discussion

In this chapter a new type of transducer for the EL speech system was proposed. The advantages are:

- a higher efficiency, therefore, lower power consumption compared to the state-of-the-art systems
- smaller dimensions, this allows to design a new device which is also wearable in a hands-free manner
- possibility to playback an arbitrary excitation signal which can change the shape and frequency

Disadvantages are:

- rapid decrease of the efficiency for high frequencies
- lower displacement in comparison to an electro-dynamic transducer
- high production of distortions

The last point, high production of distortions, is not necessarily a drawback of the proposed transducer. Non-linear distortion can cause harmonics which can support the transduction through the neck tissue.

The newly proposed transducer was inserted into a small prototype EL housing and attached to an elastic band to be worn around the neck (see Appendix A). The transfer function of the neck and the transducer were investigated. The neck transfer function can be approximated using a low-pass filter, but it is difficult to adequately measure the transducer transfer function. The proposed transducer changed the characteristics of the transfer function when loaded with the neck tissue and damped by the holding hand. More investigations in terms of optimal pre-equalization of the excitation signal need to be carried out. Based on SNR measurements we chose the dimensions and shape of the coupler disk to be attached on top of this device. The optimal coupler disk turned out to be rectangular with midsize dimensions (around 4 cm). Again, the proposed transducer, the dimensions and weight of the housing and coupler disk and the contact pressure to the skin and the coupling built a highly complex mass-spring system. Finding the optimal parameters of each contributing part is an iterative process. Furthermore, we chose the type of excitation signal, also based on SNR measurements. It turned out that for the existing configuration the *LF* model outperforms other excitation signals in terms of SNR significantly. A listening test was realized to compare the conventional Servox Digital by Servona with the newly proposed device which contains an electro-magnetic transducer. The results of the test showed that the conventional device was “slightly better” than the proposed device. Also the SNR calculations showed that the SNR of conventional device was better but only for the female speakers. For the male speaker it was the other way around: SNR values for the proposed device were better than for the conventional device. This might be due to the lower fundamental frequency of the excitation signal. A lower frequency can result in a louder excitation signal. Therefore, the SNR for the male speaker and the conventional transducer decreased. Also the listening test showed that for the male speaker more evaluations were in favor for the proposed electro-magnetic device than for the female speakers. We do not think that different training levels of the individuals speaking in the recordings have an influence on the results.

Although the above mentioned listening results are in favor of the conventional device it must be mentioned that the conventional device is a commercially available product and on the market since decades. We claim that our proposed speech system is comparable in quality because the listeners are strongly distracted by DREL of the proposed device. This leads to the assumption that the votes of the listeners depended on the perceived intelligibility. During the design of the housing, no optimization in terms of damping DREL were considered. We claim that if we reduce the direct noise, the SNR will increase and, thus, the speech quality of the proposed speech system. This will cause a better speech intelligibility. The optimization of the proposed electro-magnetic transducer in terms of larger elements might improve the loudness and effectiveness of the proposed speech system. With further improvements the speech quality of the proposed EL speech system might be equivalent or superior to that of a state-of-the-art device besides its other listed advantages.

We conducted the recordings for the listening test in an office environment to reflect a situation as natural as possible. Looking back this might have been the wrong choice. Our prototype is not optimized in terms of noise suppression. Therefore, a comparison to a commercially available product might not be fair. A controlled situation, i.e., a recording studio environment might have led to a different outcome of the experiment. Furthermore, the participants of the listing tests revealed the importance of intelligibility. We suggest that at the moment our proposed prototype is not yet a mature competitor to the conventional EL device, which is optimized to be as intelligible as possible. More focus should be put on the excitation signal to optimize the intelligibility. In the future, an intelligibility test could confirm our claims.

5 Investigations on Statistical Voice Conversion

5.1 Fundamentals of Statistical Voice Conversion

As mentioned in the introduction's literature review, the formants of alaryngeal speech are different from the formants of laryngeal speech. The control of excitation shape and f_0 is not sufficient to correct these differences. In order to tackle this problem we want to investigate voice conversion techniques in this chapter. Statistical voice conversion (SVC) is a method which is capable to convert source speech of speaker A into target speech of speaker B. SVC is used in many applications such as text-to-speech synthesis, game and entertainment applications as well as speech-to-speech translation and dubbing. Our purpose of using SVC is voice restoration, i.e., we would like to restore healthy sounding speech for people without larynx. In our opinion full-fledged voice conversion will be needed in the future to achieve the highest quality EL speech.

There are various ways to modify a voice. Based on the source-filter theory of speech production the source as well as the filter can be modified. The modification in the source signal includes the modification of time, pitch and intensity. These kinds of transformation are usually referred to prosodic modifications. Most existing voice conversion systems convert spectral features (MFCC, line spectral frequencies, etc.). For filter modification usually the magnitude of the frequency response of the vocal tract is transformed. The modification of the magnitude can be performed without having an individual target (transformation from male to female voice, transformation from male to children voice, etc.) or with an individual target. Simple but efficient standard algorithms modify the filter to approximate the characteristics of the filter of the target speaker in the mean-square sense [Benesty et al., 2008]. From literature we know that both, the source signal and the transfer function of the vocal tract, carry cues for identification. Therefore, modification of both parts should be combined.

A typical conversion system consists of the training phase and the conversion phase (see Figure 5.1). During the training phase a conversion model is learned which captures the relationship between source and target speech features. The most common way to perform voice conversion is based on the usage of a parallel database (text-dependent), i.e., utterances from the source and target speakers are needed for training. It is also possible, though more challenging to use non-parallel (text-independent) data. However, it has been shown that better results could be obtained using parallel data [Mouchtaris et al., 2006].

[Stylianou et al., 1998] proposed a GMM based statistical voice conversion method. Within this method the continuous mapping is based on a soft clustering conversion function and converts the spectral as well as the prosodic features. It is one of the typical and efficient techniques.

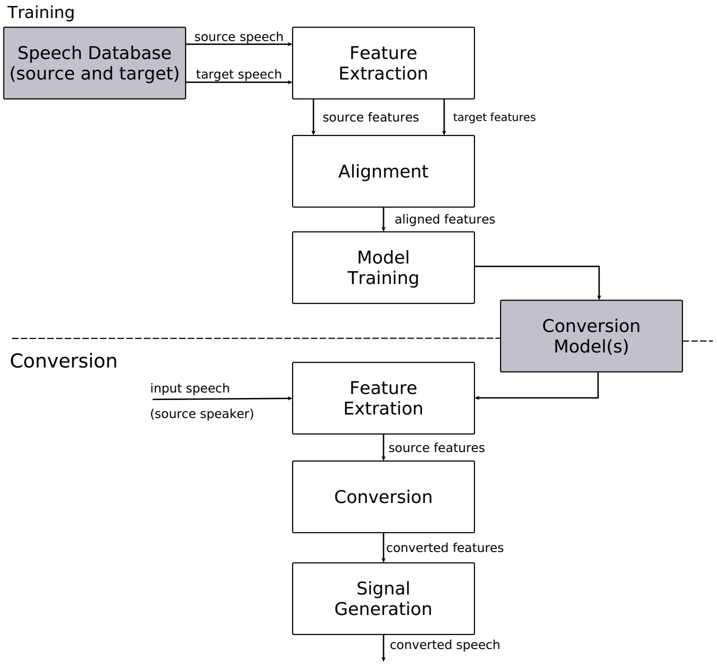


Figure 5.1: Block diagram of stand-alone voice conversion; the training phase generates conversion models based on training data (most common scenario: speech from both source and target speakers is included); conversion phase: the trained models can be used for converting unseen utterances of source speech [Nurminen et al., 2012].

[Toda et al., 2007a] proposed an SVC approach based on a GMM of the joint probability density of source and target features. The main contribution are the considered dynamic features and global variance (GV) in order to overcome two main problems: 1) the frame-based conversion process does not always cause appropriate spectral movements and 2) over-smoothing of the converted spectra. The conversion function is based on the assumptions that the source vectors follow a Gaussian mixture distribution and that the source and target vectors are jointly Gaussian. Instead of a minimum mean square error approach the authors used maximum likelihood estimation. The same authors introduced statistical voice conversion to improve naturalness of alaryngeal speech.

Latest work on the topic of voice conversion investigates different spectral features [Ghorbandoost et al., 2015]. They found out that cepstral features are more suitable for clustering and all-pole features are better for the analysis/synthesis stages. Thus, the authors propose a system which utilizes the properties of both kinds of features using feature combination together with classical GMMs as well as dynamic kernel partial least-squares regression methods. The authors could show that based on their proposed methods they outperform modern voice conversion methods in terms of speech quality and speaker individuality.

We tested the existing system developed by Toda et al. on our German parallel ELHE database described in Section 2.2. The aim of EL speech improvement is to improve the naturalness of EL speech as well as the perception of speaker identity. Converting spectral features has a huge influence on the naturalness of EL speech. However, within EL speech improvement such systems rely on the concept of open-loop processing. This means that a loudspeaker is needed to output the converted speech. In terms of speaker identity, [Helander and Nurminen, 2007] proposed that pure prosody alone can be used to recognize speakers, who are familiar to us, from HE speech. SVC is to a certain amount capable to convert prosodic features and, in further consequence, the quality of EL speech can be improved. It might be difficult to obtain many speech samples from laryngectomees and, therefore, findings of this chapter can be used to choose optimal data for the training and apply the models on various speakers.

Our database consists of extensive parallel speech material. In the following experiments we investigate the effects of the available speech material on the SVC system. We consider the number of utterances as well as the number of components used in training. We want to reveal the influence of the quality of EL and HE utterances used in training and test on the performance of the SVC system. As we assume that SNR is correlated to the quality of EL speech we investigate this property compared to SVC results. Not only SNR of EL speech, but also the efficiency of the time alignment algorithm have influence on the performance. Thus, we conduct an experiment based on speech material where HE speech is synchronized to EL during the speaking process. Within this chapter we show that conversion results in terms of objective measures depend on the used training data and try to identify parameters for good training speech material.

5.1.1 Algorithm

The framework used in this thesis is the same as in conversion from body-conducted unvoiced speech into healthy speech [Toda et al., 2012]. It extracts the features of the parallel training utterances, trains statistical models to convert them and synthesizes utterances based on the converted features. The most important definitions of the used algorithm are explained below.

Training Process

In the training of SVC, the GMM given in (5.1) models the joint probability density of the spectral features of EL speech $\mathbf{x}_t$ and the target feature vectors of HE speech $\mathbf{y}_t$ at frame t. The joint feature vector is $\mathbf{z}_t = [\mathbf{x}_t^T, \mathbf{y}_t^T]^T$. T denotes the transposition operation. (Note: $\mathbf{x}_t$ and $\mathbf{y}_t$ can have a different number of dimensions D, e.g., 39-dimensional MFCCs source feature vector and 1-dimensional f_0 target feature vector).

$$P(\mathbf{z}_t|\lambda) = \sum_{k=1}^{K} b_k \mathcal{N}\left(\mathbf{z}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \tag{5.1}$$

K is the number of mixture components, b_k are the weights for each component and $\mathcal{N}(\cdot)$ is a single, multivariate Gaussian distribution. For b_k the assumption $\sum_{k=1}^{K} b_k = 1$ holds. With $\lambda = \{(b_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k); k = 1, 2, \ldots, K\}$, the whole mixture model is described. The mean value vector $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ can be seen as a concatenation:

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_k^{(x)} \\ \boldsymbol{\mu}_k^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^{(xx)} & \boldsymbol{\Sigma}_k^{(xy)} \\ \boldsymbol{\Sigma}_k^{(yx)} & \boldsymbol{\Sigma}_k^{(yy)} \end{bmatrix}.$$

In the used framework static as well as dynamic (i.e., Δ) features are used for source and target: $\mathbf{X}_t = [\mathbf{x}_t^T, \Delta\mathbf{x}_t^T]^T$ and $\mathbf{Y}_t = [\mathbf{y}_t^T, \Delta\mathbf{y}_t^T]^T$. Source features of EL speech are segmental features. This means that successive feature vectors are stacked to capture the behavior in time. The construction, as explained in [Nakamura, 2010], is illustrated in Figure 5.2. The segmental source features $\mathbf{X}_t'$ at frame t is extracted using Principal Component Analysis (PCA). PCA reduces the dimensions of the concatenated feature vector $\mathbf{c}_t$ which is constructed from the static and dynamic feature vector $\mathbf{X}_t$ over ± L frames: $\mathbf{c}_t = [\mathbf{X}_{t-L}, \cdots, \mathbf{X}_t, \cdots, \mathbf{X}_{t+L}]^T$. Thus, $\mathbf{X}_t'$ and $\mathbf{Y}_t$ build the joint feature vector $\mathbf{Z}_t$ and are used to train the parameters for the GMM $\lambda^{(z)}$ in (5.1). In order to build the joint source and target feature vector the features need to be time aligned using DTW.

In the following experiments, three separate GMMs model the conversion functions. The source features $\mathbf{X}_t$ are represented in all three GMMs using segmental spectral features of the

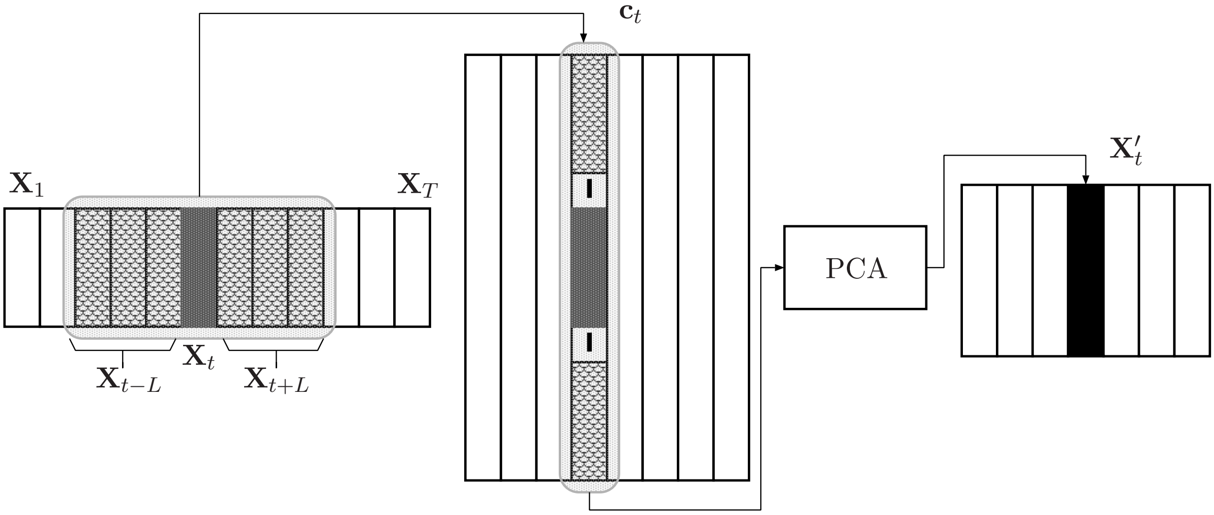


Figure 5.2: Construction of segmental feature vector from concatenated static feature vector (inspired by [Nakamura, 2010]).

source speech. The target features $\mathbf{Y}_t$ are 1) static and dynamic spectral features of the target speaker, 2) static and dynamic logarithmic f_0 of the target speaker and 3) static and dynamic aperiodic component (ac) of the target speaker.

Global Variance (GV)

As described in [Toda et al., 2007a] global variance (GV) is considered. GV models the variance of the target static feature vector of HE speech over an utterance and is represented with a Gaussian distribution as

$$P(\mathbf{v}(y)|\lambda^{(v)}) = \mathcal{N}\left(\mathbf{v}(y); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(v)}\right), \tag{5.2}$$

where $\mathbf{v}(y)$ is the GV of the target static feature vector $\mathbf{y}_t$: $\mathbf{v}(y) = [v(1), \cdots, v(d), \cdots, v(D)]^T$ is calculated as follows:

$$\mathbf{v}(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \frac{1}{T}\sum_{t=1}^{T} y_t(d)\right)^2. \tag{5.3}$$

The parameter set $\lambda^{(v)}$ consists of the mean vector $\boldsymbol{\mu}^{(v)}$ and the diagonal covariance matrix $\boldsymbol{\Sigma}^{(v)}$. The GV is calculated utterance by utterance.

Conversion Process

We want to maximize the (static and dynamic) target feature vector sequence $\mathbf{Y} = [\mathbf{Y}_1^T, \cdots, \mathbf{Y}_t^T, \cdots, \mathbf{Y}_T^T]^T$ given the (segmental) source feature vector sequence $\mathbf{X} = [\mathbf{X}_1^T, \cdots, \mathbf{X}_t^T, \cdots, \mathbf{X}_T^T]^T$. The static converted feature sequence $\hat{\mathbf{y}} = [\hat{\mathbf{y}}_1^T, \cdots, \hat{\mathbf{y}}_t^T, \cdots, \hat{\mathbf{y}}_T^T]^T$ is determined by maximizing the product of the conditional probability density of $\mathbf{Y}$ given $\mathbf{X}$ and the probability density of the GV. In the conversion process individual speech parameters of the target speech are independently converted using

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{Y}|\mathbf{X}, \lambda^{(z)}) P(\mathbf{v}(y)|\lambda^{(v)})^{\alpha} \qquad \text{subject to} \qquad \mathbf{Y} = \mathbf{W}\mathbf{y}. \tag{5.4}$$

The matrix $\mathbf{W}$ is used to extend the static feature vectors into the joint static and dynamic feature vectors [Toda et al., 2007a]. Input are the spectral segmental features extracted from EL speech and each of the separately trained GMMs. The result are the converted static features. α is the weight of the GV likelihood and GV is only considered for spectral conversion and not for f_0 conversion and aperiodic component conversion.

Feature extraction and the signal synthesis of the speech utterances are performed using the STRAIGHT analysis and synthesis methods [Kawahara et al., 1999]. In the signal synthesis step, the mixed excitation signal is constructed using the converted f_0 and converted aperiodic components. Finally, the generated excitation signal drives the filter controlled by the converted spectral parameters and creates the converted speech signal.

5.1.2 Error Measures

To compare results we calculated the following error measures:

Mel-Cepstrum Distortion (MCD)

MCD is a scaled Euclidean distance. First, the mel-cepstrum including Δ features is extracted for the target speech and the converted speech. The dimension D of coefficients is 48, no power information is used and only speech frames are considered for calculating the error measure:

$$\mathrm{MCD} = \frac{1}{T}\sum_{t=1}^{T} \frac{10 \cdot \sqrt{2 \cdot \sum_{d=1}^{D} (mc_{tar}[d,t] - mc_{conv}[d,t])^2}}{\ln 10} \tag{5.5}$$

with mc_{tar} and mc_{conv} being the d-th cepstral coefficient of the target and converted speech of frame t. Note that the coefficient $\frac{10}{\ln 10}$ is introduced to keep a consistency between MCD and spectral distortion in dB.

Aperiodic Component Distortion (ACD)

Aperiodic components (ac) [Kawahara et al., 2001] are used to create the excitation signal for speech synthesis. Instead of a hard voiced/unvoiced decision, the excitation is modeled as a combination of periodic and noise-like components with their relative contributions based on “voicing strengths” in separate bands across the frequency spectrum. Although most of a speaker’s prosodic features can be represented with f_0, aperiodicity still contains information that is perceptually significant [Shum, 2009]. Aperiodic components are calculated from upper and lower spectral envelopes for five frequency bands up to 8 kHz. The upper envelope is calculated by connecting spectral peaks and lower envelope is calculated connecting spectral valleys and the aperiodicity measure of a certain center frequency is defined as the lower envelope normalized using the upper envelope. The unit of aperiodic components is dB. Therefore, ACD is calculated as the root mean square error as follows:

$$\mathrm{ACD} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\sum_{d=1}^{D} (ac_{tar}[d,t] - ac_{conv}[d,t])^2\right)} \tag{5.6}$$

with ac_{tar} and ac_{conv} being the d-th aperiodic component of the target and converted speech of frame t.

(Pearson’s Linear) Correlation Coefficient (ρ)

ρ has already been explained in Section 3.2.2.

Voicing Error

The voicing error is classified into voiced-to-unvoiced (VUV) and unvoiced-to-voiced (UVV) errors. This kind of error is based on the f_0 contours of target and converted speech (see Figure 3.4, first plot) and is normalized to the length of the utterance. In the example of Figure 3.4 $VUV = 2.45\,\%$ and $UVV = 15.13\,\%$. Perceptually this error influences the breathiness of the converted speech. If VUV error is large, many frames are considered as unvoiced with a noisy excitation and vice versa, if UVV is large also unvoiced frames are converted with a periodic excitation signal which leads to a buzz-like distortion of the converted speech.

5.1.3 Experimental Setup

For training and evaluation we used the German parallel ELHE speech database described in Section 2.2. The database contains 7 speakers, 3 female and 4 male. Furthermore, we included speaker M08 who is a laryngectomee. The parameter of the feature extraction are as follows: 0th through 24th mel-cepstral coefficients were used as spectral features with a shift length of 5 ms. To extract segmental features ± 4 frames are concatenated and PCA reduces the dimensions to 50. For the target features, logarithmic f_0 values and aperiodic components of five frequency bands were used. At unvoiced frames, logarithmic f_0 values were set to 0. Number of utterances in training and test, and number of components for the GMMs are described separately for each experiment.

In most of the experiments the same 100 test utterances for each speaker are taken for evaluation (except experiment XIV because we used different material for this experiment). For the laryngectomee M08 all 35 utterances were used in test. The statistics of the test utterances are summarized in Table 5.1. We can see that the SNR$_{EL}$ strongly depends on the speaker, whereas SNR$_{HE}$ is approximately the same across speakers. The mean utterance length $\overline{len}$ and the total utterance length len show that the different speakers have different speaking styles in terms of duration. The difference in length between speakers is bigger for EL speech than for HE speech. The DTW cost is calculated between HE and EL utterances (without DREL suppression) according to the definitions explained in Section 3.2.3. Although the time differences between HE and EL is very small for speakers F03 and F07, who obtain the worst SNR values in EL speech, the DTW cost is not bigger compared to the other speakers. This raises the question whether DTW cost reflects the quality of DTW.

Speaker	SNR$_{EL}$ [dB]	SNR$_{HE}$ [dB]	$\overline{len_{EL}}$ [s]	$\overline{len_{HE}}$ [s]	len_{EL}	len_{HE}	DTW cost
F01	18.59	46.65	5.94	4.66	9 min 53 s	7 min 46 s	0.75±0.08
M02	21.84	46.76	6.33	4.28	10 min 32 s	7 min 7 s	0.70±0.07
F03	12.41	44.76	5.27	4.36	8 min 46 s	7 min 16 s	0.79±0.06
M04	20.27	47.85	6.51	4.69	10 min 51 s	7 min 48 s	0.85±0.09
M05	21.91	47.75	5.90	3.96	9 min 49 s	6 min 35 s	1.03±0.09
M06	19.11	47.00	5.59	4.16	9 min 18 s	6 min 55 s	0.81±0.07
F07	12.14	47.65	5.21	4.40	8 min 41 s	7 min 19 s	0.87±0.10
M08	17.79	-	3.90	-	2 min 16 s	-	-

Table 5.1: Database analysis for 100 test utterances; SNR values, averaged utterance length $\overline{len}$, total utterance length len and DTW cost (see Section 3.2.3) for EL and HE utterances.

Additionally, we compare our results to a reference framework with Japanese data. The number of GMMs for the spectrum, f_0 and ac was 32, 16 and 16. 40 utterances are used in training and 10 in the test. The Japanese speech data consists of one male laryngectomee source speaker and one female healthy target speaker. Statistics of the corresponding speech files are shown in Table 5.2.

		#	$\overline{f_0}$	σ_{f_0}	SNR [dB]	$\overline{len}$ [s]	*len*	DTW cost
Training	HE	40	228	36	45.54	6.64 s	4 min26 s	0.76±0.05
	EL		100	6	20.81	9.83 s	6 min33 s	
Test	HE	10	214	36	45.53	5.62 s	59 s	0.66±0.05
	EL		101	8	19.97	8.07 s	1 min25 s	

Table 5.2: Statistical analyses of Japanese speech material; Mean value and standard deviation of f_0 – $\overline{f_0}$,σ_{f_0}; SNR values, averaged utterance length $\overline{len}$, total utterance length len and DTW cost (see Section 3.2.3) for EL and HE utterances.

5.2 Experiment X: Overall Results with One-to-One and Many-to-One Conversion

In the overall results of the voice conversion technique we trained one-to-one and many-to-one models. For the one-to-one models we trained a model for each of the 8 speakers separately and tested on all test utterances from all speakers. The matched case corresponds to the case where the test utterances originate from the same speaker used in training. In the mismatched case the speaker model for a specific speaker is tested using the test utterances from a different speaker. In the many-to-one model we used a leave-one-out method, meaning that all utterances except for the utterances from one speaker are used to train the model. This leads to 7 many-to-one models, e.g., in the first model speaker one is left out, in the second speaker two, etc. Also for the many-to-one models we distinguish between matched and mismatched case as before. Experimental conditions are given in Table 5.3 and 5.4. Note that the number of components of the GMM for spectrum, f_0, and ac was adjusted speaker dependent depending on computational limits. In the implemented code the Cholesky decomposition is used to handle the matrices in the GMM. The Cholesky decomposition sometimes causes numerical problems when the values of the matrix are too small (e.g., too close to zero).

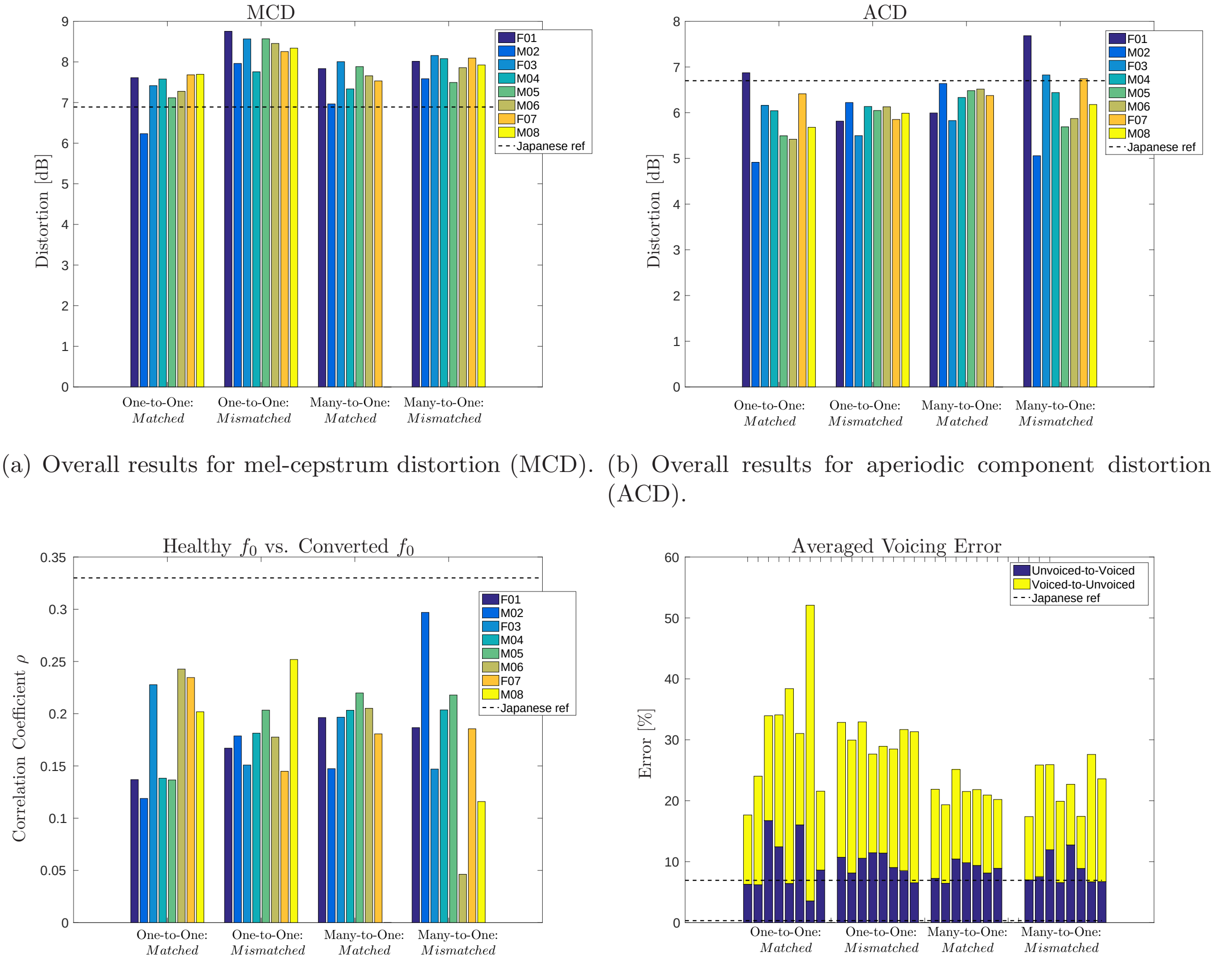
Speaker	F01	M02	F03	M04	M05	M06	F07	M08
# Training utterances	379	404	135	392	301	148	404	25
# Test utterances	735	735	735	735	735	735	735	10
#GMM (spectrum)	32	32	32	32	16	16	32	32
#GMM (f_0)	8	16	16	16	16	16	16	16
#GMM (ac)	16	16	16	16	16	16	16	16

Table 5.3: Experimental conditions for one-to-one conversion.

Speaker	F01	M02	F03	M04	M05	M06	F07
# Training utterances	1908	1908	902	1908	1908	899	1908
# Test utterances	735	735	735	735	735	735	735
#GMM (spectrum)	32	64	32	64	64	32	32
#GMM (f_0)	16	16	16	16	16	16	16
#GMM (ac)	16	16	16	16	16	16	16

Table 5.4: Experimental conditions for many-to-one conversion.

The results are shown in Figure 5.3 and Tables 5.5, 5.6, 5.7 and 5.8. Compared to the results using Japanese speech data MCD and ACD results are similar to our experiment. On average MCD reaches the lowest values for the one-to-one model in the matched case, followed by the many-to-one model in the matched case. Higher distortion can be observed for the



(a) Overall results for mel-cepstrum distortion (MCD).

(b) Overall results for aperiodic component distortion (ACD).

(c) Overall results for correlation coefficient ρ.

(d) Overall results for voicing error.

Figure 5.3: Overall results of experiment X (one-to-one and many-to-one models).

Speaker	o2o (matched)	o2o (mismatched)	m2o (matched)	m2o (mismatched)
F01	7.61	8.75	7.84	8.01
M02	6.23	7.96	6.97	7.58
F03	7.42	8.57	8.01	8.16
M04	7.58	7.76	7.34	8.08
M05	7.12	8.57	7.89	7.49
M06	7.28	8.45	7.66	7.86
F07	7.60	8.26	7.53	8.10
M08	7.69	8.34	-	7.92

Table 5.5: Mel-cepstrum distortion for all speakers; one-to-one (o2o) model and many-to-one (m2o) model in matched and mismatched case; grey: best results.

many-to-one model in the mismatched case and for the one-to-one in the mismatched case. This general result is as expected because the training of the conversion functions works well when speech material of a speaker is present in training and test. For the many-to-one model in the mismatched case an averaged speaker-independent model is available. This speaker-independent model can deal better with an unknown speaker than the speaker-dependent in one-to-one model (mismatched). The one-to-one is less realistic because in a realistic environment we have not

Speaker	o2o (matched)	o2o (mismatched)	m2o (matched)	m2o (mismatched)
F01	6.87	5.82	5.99	7.68
M02	4.91	6.22	6.64	5.06
F03	6.16	5.50	5.83	6.83
M04	6.04	6.14	6.33	6.44
M05	5.49	6.05	6.48	5.69
M06	5.42	6.13	6.52	5.87
F07	6.41	5.85	6.38	6.74
M08	5.68	5.99	-	6.18

Table 5.6: Aperiodic component distortion for all speakers; one-to-one (o2o) model and many-to-one (m2o) model in matched and mismatched case; grey: best results.

Speaker	o2o (matched)	o2o (mismatched)	m2o (matched)	m2o (mismatched)
F01	0.14	0.17	0.20	0.19
M02	0.12	0.18	0.15	0.30
F03	0.23	0.15	0.20	0.15
M04	0.14	0.18	0.20	0.20
M05	0.14	0.20	0.22	0.22
M06	0.24	0.18	0.21	0.05
F07	0.23	0.14	0.18	0.19
M08	0.20	0.25	-	0.12

Table 5.7: Correlation coefficient for all speakers; one-to-one (o2o) model and many-to-one (m2o) model in matched and mismatched case; grey: best results.

Speaker	o2o (matched)	o2o (mismatched)	m2o (matched)	m2o (mismatched)
F01	6.29/11.38	10.73/22.13	7.27/14.61	7.01/10.39
M02	6.22/17.81	8.15/21.81	6.46/12.88	7.57/18.30
F03	16.74/17.21	10.58/22.36	10.44/14.70	11.97/13.96
M04	12.45/21.64	11.47/16.20	9.84/11.66	6.57/13.35
M05	6.43/31.97	11.41/17.52	9.39/12.45	12.75/9.95
M06	16.04/15.00	9.05/19.44	8.15/12.79	8.89/8.54
F07	3.56/48.53	8.55/23.14	8.92/11.29	6.70/20.91
M08	8.63/12.95	6.54/24.79		6.73/16.86

Table 5.8: Voicing error (notation in table: VUV/UVV in [%]) for all speakers; one-to-one (o2o) model and many-to-one (m2o) model in matched and mismatched case; grey: best results.

enough parallel speech data to train a good conversion function. Our MCD results suggest that we can use a many-to-one model for statistical voice conversion. With further speaker adaptation the results can be improved. Comparing the MCD results and the SNR results of the EL test utterances in Table 5.1 reveal a indirect relation, e.g., speaker M02 reaches low MCD values together with high SNR values. Although the number of utterances for training was much smaller for the laryngectomee M08, the obtained results are similar to the other speaker in the database. Informal listening tests confirmed the conversion to more natural and intelligible speech. Figure 5.4 shows the waveform, spectra and f_0 contours before and after conversion for the laryngectomee M08. The spectrogram of the EL speech shows the constant harmonic structure resulting from the excitation signal. This can be removed which is shown in the converted spectrogram. Furthermore, the energy in the formants are shifted. ACD results do not correlate with any of the other objective measures. Therefore, we omit the measure in the

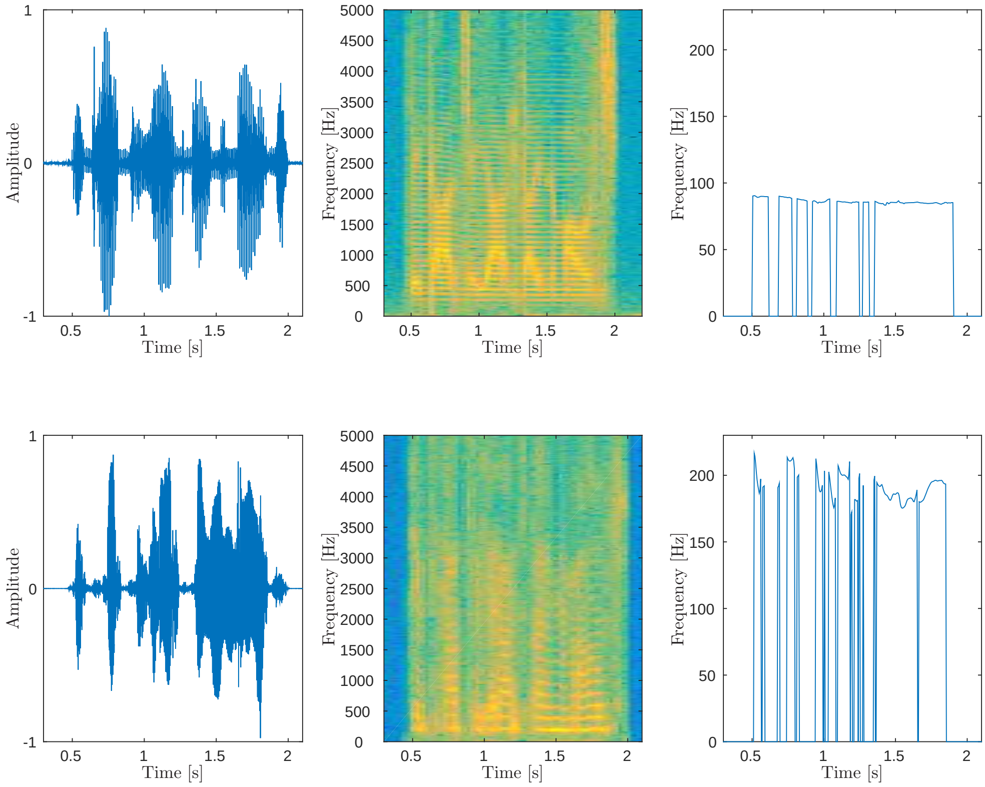


Figure 5.4: Waveform, spectra and f_0 contour for laryngectomee M08 before (upper plot) and after (lower plot) conversion.

next experiments. ρ reaches low values between 0.1 and 0.3. A trend can be seen for the variance of the voicing error between speakers: the voicing error increases for one-to-one matched case but decreases for many-to-one models. Furthermore, it can be observed that the variance in many-to-one case (matched) decreases between speakers for all error measures. In terms of ρ and voicing error Japanese data clearly outperforms the results of the German parallel ELHE database. The calculation of ρ is based on the comparison between HE speech and converted EL speech. Therefore, DTW needs to be applied in order to obtain aligned f_0 values. It seems that ρ is very sensitive to DTW errors. The temporal structure of the parallel Japanese utterances is carefully designed. The utterances contain separate phrases with pauses in between. The pauses in Japanese EL utterances are the same as in the Japanese HE utterances. Therefore, we expect that DTW performs better for the Japanese speech than for the German utterances. The voicing error is calculated based on the same DTW results and therefore, this errors shows the same behavior.

It is not possible to conclude intelligibility from objective measures. Informal listening tests have shown that the converted speech is improved in terms of naturalness while being intelligible in many cases. To gain more insight in what is important for training the conversion function more experiments are carried out in the following. We want to investigate the perceptual meaning of the objective measures and find a way how to optimize the conversion results.

5.3 Experiment XI: Influence of Utterance Number

In this experiment we investigate the influence of the number of utterances in training. We trained different models with 50, 100, 150, 200, 250, 300, 350 and 400 utterances in training. Data from speaker M02 is used. The number of components for the GMM for the spectrum, f_0

and ac are 8, 8 and 16, respectively. When 50 utterances are taken for training, the number of components for the aperiodic components needed to be reduced to 8 due to numerical problems of matrix inversion.

Looking at Figure 5.5 one can see that the MCD decreases slightly when more utterances are used in training. However, the values do not change very much anymore using more than 200 utterances in training which leads to the conclusion that the algorithm does not benefit from a bigger database.

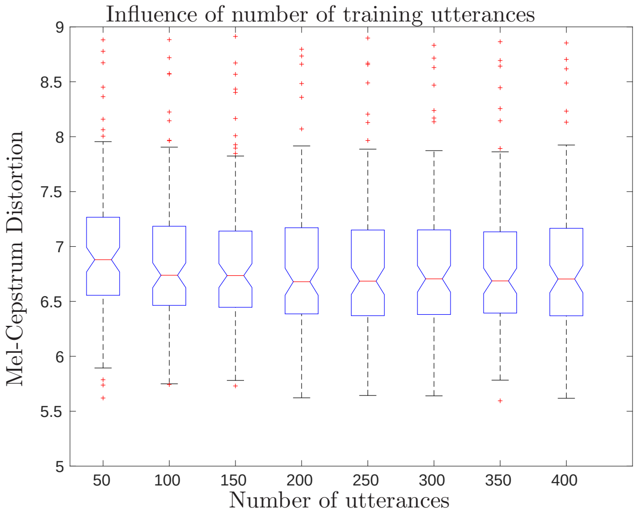


Figure 5.5: Mel-cepstrum distortion of experiment XI (utterance number in training).

Looking at the other objective measures in Table 5.9 suggests that the ACD is increasing with the number of training utterances. Also ρ tends to be higher for a small number of training utterances. For the voicing error: UVV is smaller for a small training set, but VUV decreases using more training utterances.

# Training utterances	MCD	ACD	ρ	UVV [%]	VUV [%]
50	6.97 ± 0.67	4.78	0.37 ± 0.36	4.38 ± 2.51	25.68 ± 12.97
100	6.87 ± 0.65	4.81	0.35 ± 0.36	4.90 ± 2.89	22.88 ± 12.50
150	6.85 ± 0.66	4.84	0.32 ± 0.35	4.83 ± 2.91	23.08 ± 11.89
200	6.82 ± 0.67	4.85	0.33 ± 0.36	5.59 ± 3.00	21.49 ± 12.03
250	6.81 ± 0.66	4.84	0.36 ± 0.36	4.73 ± 2.80	23.30 ± 12.04
300	6.82 ± 0.66	4.84	0.32 ± 0.33	5.95 ± 3.25	20.93 ± 11.79
350	6.85 ± 0.76	4.83	0.32 ± 0.32	6.21 ± 3.35	20.07 ± 11.64
400	6.81 ± 0.66	4.84	0.32 ± 0.33	6.00 ± 3.24	20.84 ± 11.71

Table 5.9: Objective measures for experiment XI (dependency on number of utterance).

5.4 Experiment XII: Influence of Number of Mixtures

In this experiment we investigate the number of mixtures used in training. Again speech material of speaker M02 is investigated. 200 utterances are used in training, 100 in test. The number of mixtures are changed for the spectrum between 2, 4, 8 and 16, for f_0 between 4, 8, and 16, and for ac between 2, 4, 8, 16 and 32.

Results are shown in Table 5.10. MCD decreases with increasing number of components in the GMM for the spectrum. Though, the number should be optimized individually for each

#GMM (spectrum)	#GMM (f_0)	#GMM (ac)	MCD	ACD	ρ	UVV [%]	VUV [%]
2	16	16	7.16	4.78	0.33	7.84	18.96
4	16	16	6.94	4.78	0.34	7.75	18.85
8	16	16	6.84	4.87	0.35	6.21	18.31
16	4	16	6.71	4.87	0.32	7.40	17.03
16	8	16	6.71	4.87	0.34	5.22	21.09
16	16	2	6.71	4.81	0.34	6.21	18.05
16	16	4	6.71	4.82	0.34	6.21	18.05
16	16	8	6.71	4.84	0.34	6.21	18.05
16	16	16	6.71	4.87	0.34	6.21	18.05
16	16	32	6.71	4.87	0.34	6.21	18.05

Table 5.10: Objective measures for experiment XII (dependency on number of mixtures).

speaker. The changing number of components for spectral features have influence on ρ and the voicing error, too, and of course the number of GMMs for f_0 also influences the voicing error but not the MCD. Also in this experiment ACD increases with increasing number of components. This experiment leads to the conclusion that the number of components should be around 16 for 200 utterances in training. The values will be influenced by the quality of the source and target speech. Furthermore, the number of mixtures strongly depends on the number of data available for training the parameters in the GMM.

5.5 Experiment XIII: Choosing Best Utterances for Training

In this experiment we investigate the influence of the chosen utterances for training. For the baseline setup again speaker M02 was chosen. 100 utterances were taken for training, 100 for test. The number of mixtures was 8, 8, 16 for mel-cepstrum, f_0 and aperiodic components, respectively.

The criterion for choosing the utterances were: 1. randomly from the training set, 2. SNR calculations using equation 2.1, then the utterances were chosen where the SNR for HE as well as for EL was best/worst and 3. MCD between HE and EL utterances (best and worst). Simulations with random utterances was carried out 5 times and averaged. Results for this experiment are given in Table 5.11 and Figure 5.6. The results show that MCD decreases for *worst SNR*, but there is no difference between choosing the utterances randomly or based on the SNR or MCD criterion. Voicing error reaches the best value in the *rand* case (low UVV together with low VUV).

#GMM (spectrum)	MCD	UVV [%]	VUV [%]
rand	6.86 ± 0.66	5.10 ± 0.75	23.17 ± 11.51
best SNR	6.89 ± 0.70	6.22 ± 3.45	19.60 ± 11.73
worst SNR	6.92 ± 0.65	4.35 ± 2.57	24.98 ± 12.20
best MCD	6.90 ± 0.66	3.86 ± 2.71	26.07 ± 12.78
worst MCD	6.95 ± 0.66	5.14 ± 3.15	21.05 ± 11.80

Table 5.11: Objective measures for experiment XIII (criterion for choosing utterances in training).

5.6 Experiment XIV: Synchronized Data

Generating a good and appropriate speech database is challenging, especially for parallel voice conversion tasks. The major part of our database is not synchronized, e.g., one kind of speech

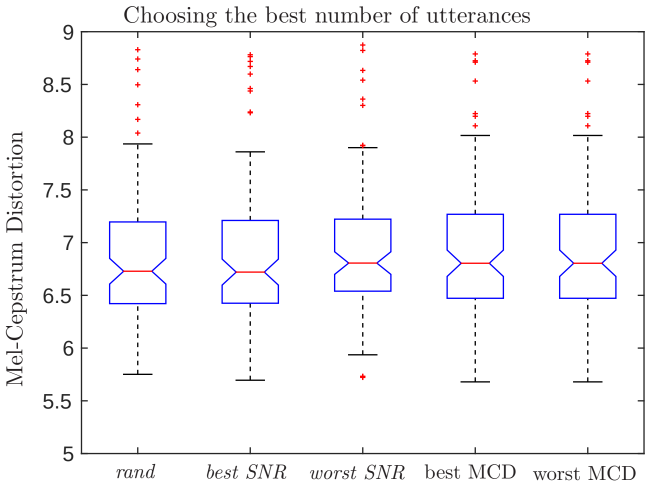


Figure 5.6: Mel-cepstrum distortion of experiment XIII (choosing good utterances for training).

is not intentionally manipulated to match the other kind of speech, because we instructed the speakers to behave and act in their healthy speaking style.

Nevertheless, we recorded synchronized data, i.e., the natural utterances are spoken synchronized to the EL utterances(two speakers – F01 and M02). In order to do so we first recorded EL utterances. Then, we put the recorded EL utterances to the headphones and recorded HE utterances while speaking synchronized to the EL utterances which were presented at the headphones. This situation is unnatural because healthy utterances are spoken too slowly. These utterances were taken to train the conversion models. Moreover, we evaluate these utterances based on different recording sessions. This simulates the real-world scenario where the conditions for the training and test are different. For this experiments 2 models are trained: 1. based on synchronized data, 2. based on non-synchronized data, and evaluated on 2 conditions: A. test utterances from the same session – matched case and B. test utterances from a different session – mismatched-case. For both speaker we used 49 utterances in training and 18 in test. The number of mixtures for the male speaker were 16, 8 and 16 for spectrum, f_0 and ac, respectively. For the female speaker the number of mixtures for the spectrum were set to 8.

	Name	MCD	UVV [%]	VUV [%]
M02	sync: matched	6.22 ± 0.49	4.09 ± 2.76	6.77 ± 1.71
	sync: mismatched	7.09 ± 0.79	4.46 ± 3.04	11.86 ± 3.97
	not sync: matched	6.68 ± 1.24	3.71 ± 2.90	16.01 ± 10.39
	not sync: mismatched	7.03 ± 0.79	3.44 ± 3.01	18.62 ± 11.22
F01	sync: matched	7.45 ± 0.83	6.88 ± 2.86	8.57 ± 3.15
	sync: mismatched	7.78 ± 0.74	11.67 ± 4.68	13.44 ± 4.49
	not sync: matched	7.87 ± 1.56	6.11 ± 3.44	9.86 ± 3.25
	not sync: mismatched	8.44 ± 0.42	5.32 ± 4.40	20.64 ± 6.80

Table 5.12: Objective measures for experiment XIV (synchronization experiment: speakers M02 and F01).

The results are listed in Table 5.12 with a detailed analysis in terms of the MCD, the most representative measure, in Figure 5.7(a) for the male speaker and Figure 5.7(b) for the female speaker. In general, error measures are lower for the male speaker compared to the female speaker. The reason is the SNR of the source EL speech and the absolute f_0 of the device. In terms of MCD, results are better for synchronized data than for not synchronized data. For the

SNR	sync train		not-sync train	
	F	M	F	M
	17.28	21.17	21.65	19.49

SNR	matched test		not-matched test	
	F	M	F	M
	17.36	22.25	18.98	20.31

Table 5.13: SNR analysis for synchronization experiment for training and test utterances; sync and not-sync refer whether HE speech utterances are spoken synchronized (same temporal structure) to EL speech utterances or not; matched and not-matched refer to different recording sessions.

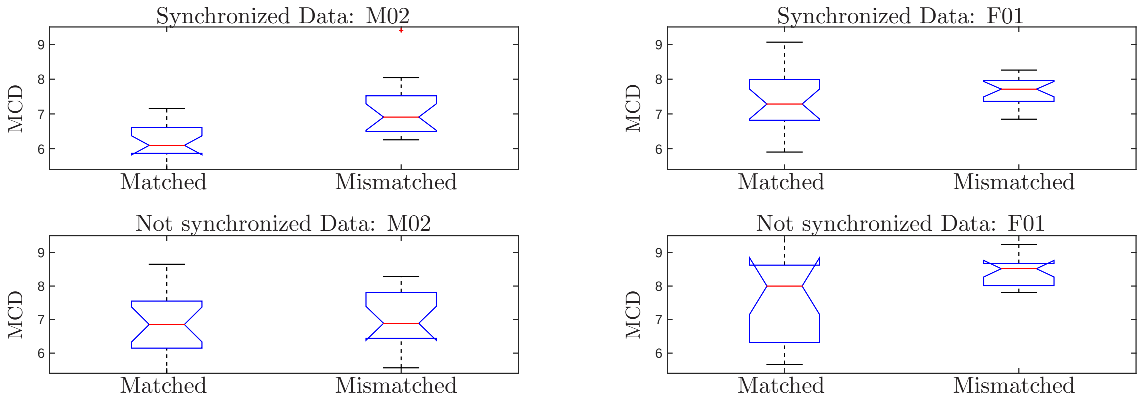
male speaker there is only one significant difference: synchronized data between matched and mismatched case.

If the data is not synchronized results for the matched case are decent but suffer from the huge variance. The variance decreases in the mismatched case due to the more stable quality of EL speech, but the distortion increases because of the different training conditions. As can be seen in 5.7(a), there is no significant improvement from not synchronized data to synchronized data if mismatched speech data is taken for the male speaker. The results correspond to an informal listening test where the utterances were blindly ordered according to their pleasantness.

A strong evidence for better conversion results using synchronized data is the decreased variance for the synchronized matched case. For the female speaker results for the mismatched case are better in terms of variance. The reason for this is that the speech utterances from the mismatched case were recorded a few months after the first session. Due to the improved handling of the device, the quality of EL speech for the female subject increased. This is also visible in the SNR calculations in Table 5.13. The quality of the male speaker did not change over time.

Besides the effects on MCD this experiment shows the influence on the VUV error. The UVV error is about the same for all experiment setups, but the VUV error is much better for the synchronized experiment scenario (for both matched and mismatched data). If this error is big, the converted speech sounds more breathy, e.g., like whispered speech.

However, the experiment shows that the modeling of the joint probability is better for synchronized data and the conversion results in terms of the above mentioned error measure is better in that case.



(a) Results of the synchronization experiment: Male speaker M02.

(b) Results of the synchronization experiment: Female speaker F01.

Figure 5.7: Mel-cepstrum distortion (MCD) of experiment XIV (synchronization experiment).

5.7 Summary and Discussion

In this chapter we evaluated the German parallel ELHE database using the statistical voice conversion technique proposed by [Toda et al., 2007a]. We evaluated the method in terms of speaker dependency and were able to obtain more natural EL speech compared to the original (not converted) EL speech. Informal listening tests showed that although the naturalness of the converted speech has been improved the intelligibility is insufficient in some cases. It is still an open question how intelligibility can be improved in the context of statistical voice conversion.

We could show that the conversion results are strongly speaker dependent. In terms of the objective measure MCD, which turned out to approximate the perceived quality, the one-to-one conversion model showed the best results. This behavior is not surprising. For the many-to-one model MCD values were increased, but the difference to the one-to-one model was not significant. In order to improve the results we should investigate conversion based on eigenvoices as suggested by [Toda et al., 2007b].

To gain more detailed insight on how to train the best models for the voice conversion task we checked the influence of the number of utterances used in training and number of mixtures used for training the spectral features, f_0 and aperiodic components. It turned out that for our database more utterances in training did not improve the conversion results. In fact, taking more than 200 utterances did not have any influence on the quality of the converted speech in terms of objective measures. The number of mixtures for the models depended on the number of utterances in the first place.

We tested different methods to choose the best number of utterances, but neither an SNR based criterion nor an MCD based criterion showed significantly better conversion results. Although we investigated the number of utterances, the SNR, and the error for MCD as indicators to choose the best utterances no clear evidence could be found that one of these parameters are responsible for the success of voice conversion.

However, we could show that a careful design of the database improves the quality of the trained conversion function. An experiment with synchronized and non-synchronized data showed that a perfect time alignment leads to significantly better results. This means that a database with mismatch between source and target speech leads to conversion functions with reduced quality. Then, dynamic time warping (DTW) is important to align source and target utterances and DTW performance is crucial to obtain good conversion results.

To conclude: It is better, but less realistic, to use synchronized data for training the conversion function. The number of training utterances and parameters for the training are also crucial but need to be chosen in a speaker dependent way. The training of the conversion model does not benefit from a huge database with insufficient quality.

Without doubt, SVC is a good method to obtain natural voice conversion results, but effects in terms of reverberation and quality mismatch in training and test still need to be investigated. Moreover, this method is limited to telecommunication applications (open-loop approaches) and it is difficult to include a fall-back mechanism. This means that if speech cannot be converted intelligibly, there is no way to understand the meaning of the utterance.

6 Discussion and Conclusion

6.1 Summary of this Thesis

To reach the goal of a new bionic EL speech system we dealt with control mechanisms of the artificial excitation signal as well as the artificial excitation source itself. We recorded a German parallel ELHE database using the conventional EL device and analyzed its properties. ASR results show that EL speech can be applied to ASR without significant drawbacks. Our proposed strategy to generate a changing, and thus more natural f_0 contour is based on statistical methods using GMMs. Listening tests suggested that a changing f_0 improves naturalness of EL speech and that a GMM approach outperforms a strategy which is based on randomly changing f_0 values. We compared the healthy fundamental frequency to the estimated one. The natural fundamental frequency is the best possible reference. Therefore, it is not surprising that the listeners in this evaluation preferred this optimum. To tackle a hands-free approach for EL speech we studied EMG bio-signals. We built a hardware and performed objective as well as subjective tests. EMG is a very promising method to control the on/off signal for the EL device. There is also a strong learning effect which means that users are able to improve the hands-free control using EMG signals after a short time. Furthermore, subjective evaluation confirm the improvement of naturalness compared to pre-training. We proposed a new electromagnetic transducer as artificial excitation source and provided a proof-of-concept for this kind of transducer. Furthermore, SNR measurements were used to characterize the shape of an appropriate coupler disk and the used excitation signal. In a listening test the conventional device still outperforms the proposed speech system in terms of pleasantness, but the main differences turned out to be SNR and intelligibility. Changing the electro-mechanical transducer is essential because it is not possible to change the waveform shape of the excitation signal using the conventional device.

The overall system, which combines all discussed steps into a closed-loop approach, is depicted in Figure 6.1. For capturing the speech sound we use an AKG HC 577 L headset microphone. EMG signals are acquired using Skintact34 electrodes and pre-processed on the bio-signal shield attached to the ARDUINO DUE micro-controller board. Microphone signals as well as pre-processed EMG signals are amplified using the Allen&Heath ZED-10. The mixing console is connected to a personal PC with Linux operating system and Matlab. We use Playrec [Humphrey, 2014], a Matlab utility that provides real-time in- and output of audio signals. The programs to extract features and estimate a natural f_0 contour, as well as the amplitude smoothing and thresholding of the EMG signals, are implemented in Matlab. Furthermore, the excitation signal, with the parameters f_0 and on/off message of the EMG signals is generated and sent to the microphone amplifier of the mixing console using Playrec. The excitation signal drives the electro-magnetic transducer and thus, closes the loop. Informal tests prove the feasibility of the system. EMG can control on/off signals and a changing excitation signal is produced without becoming unstable. The computational power of the PC is sufficient to execute the algorithms in real-time. To make the system portable, the current supply of the transducer as well as the implementation of the algorithms on a DSP or FPGA need to be considered. Voice conversion using statistical methods is a promising strategy for a telecommunication approach and can lead to very natural results. However, we did not yet confirm the naturalness and intelligibility in listening tests.

6.2 Lessons Learned

In this thesis we push the frontiers of investigations and implementations of a new bionic electro-larynx speech system. Although dead ends are inevitable in such a process, we gain insight into the problems related to electro-larynx speech. We question the conventional system and think of ways how to improve it. We tackle different problems concerning EL speech and try to connect solution strategies to form an overall answer to the problem. Although we analyze and investigate the problems, the combination of our solutions is still not satisfying and needs to be improved.

After conducting all these experiments and consideration we suggest the following system: The closed-loop approach is the best choice because then we can use the still remaining anatomy of a laryngectomee and the technical impact is as small as possible. We think people prefer to use their own body to produce speech instead of the need to depend altogether on technical gadgets. We know that the modern world consists of and lives from technical achievements but still, the nature out there is older than we are, and it took the nature billions of years to optimize in the way it is now. Engineering is important and in some cases absolutely essential but not exclusively. This is also the motivation to attach the adjective **bionic** to our new proposed speech system. We would like to bring the human anatomy in synergy with the EL speech system instead of separating the two parts. The first responsibility of electro-larynx users lies with themselves. They need to learn how to use an EL device in a proper way. Only if the basics are working, the speech system can work at its optimum. EMG is a good possibility to control on/off signals for the EL speech system. Also here the user has to cooperate for optimal working. The preliminary choice of a new transducer turned out to be promising, but here many improvements must be exploited. The excitation signal needs to be optimized inherently as well as in combination with the used transducer. In our opinion, changing the electro-mechanical transducer is essential for a successful improvement of the speech system. Although we put much effort into the design there are still open questions and challenges. Furthermore, the optimal design of the speech system needs to be investigated in order to minimize the directly radiated noise. Telecommunication is a technical achievement. Therefore, we suggest open-loop approaches in that case. SVC can lead to very natural speech quality, but questions in terms of training, implementation, and adaption still need to be investigated.

Based on our experiments, we believe in a major revision of the conventional system. Although the conventional system is quite simple and problems related to this system are quite obvious, the solution is not straightforward at all. We believe that a closed-loop approach is the best solution for laryngectomees because the technical part is kept as small as possible. The statistical voice conversion, which we investigated, would be optimal for telecommunication applications. Such algorithms could be implemented in a telephone device and amazing results could be expected.

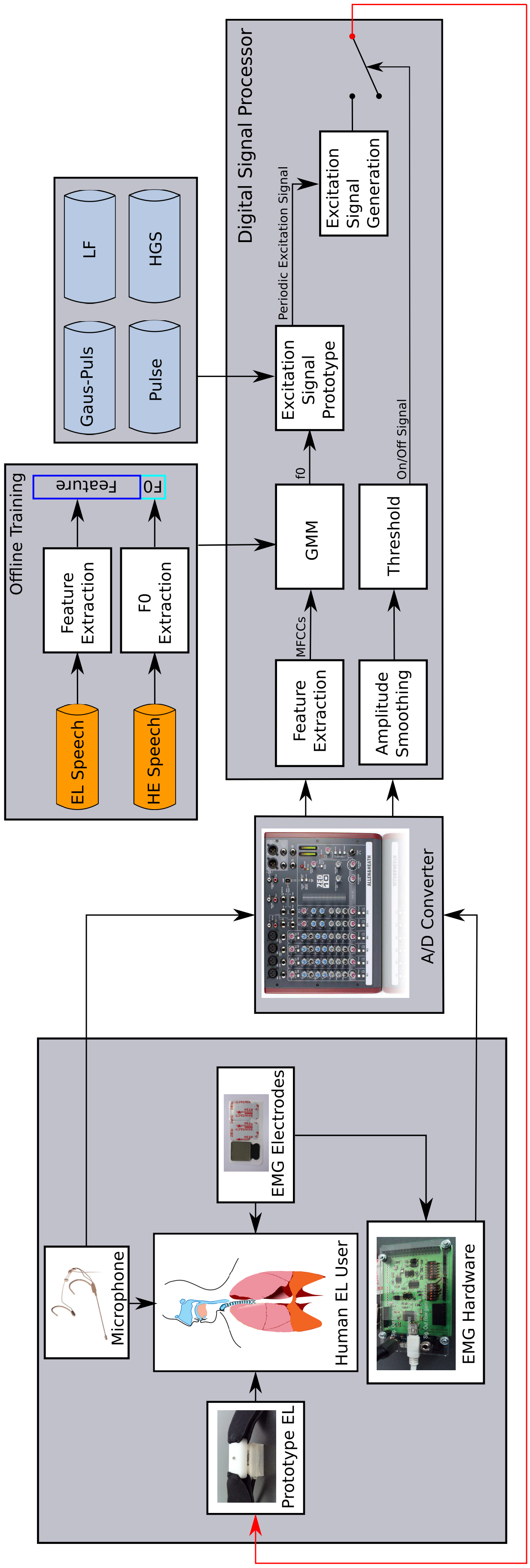


Figure 6.1: Block diagram of overall system

6.3 Future Work

EL speech improvement includes three main goals: 1. naturalness, 2. intelligibility and 3. identity (gender, speaker). These goals influence and depend on each other. After developing basic strategies for EL speech improvement, further focus needs to be put on their coherence and interaction.

For our experiments, we suggest the following improvements in future work:

1. The relation between ASR results and subjective rating could be determined in [Schuster et al., 2006] for tracheo-esophageal speech. An ASR system trained with laryngeal speech was used to obtain recognition results for alaryngeal speech. A lower syllable and lower word accuracy could be shown compared to laryngeal speech. Expert listeners confirmed the objective results. In future work we could confirm the relation between intelligibility and ASR results for EL speech and use ASR as objective measure to evaluate intelligibility of EL speech.

2. A critical point within this thesis is the evaluation of the algorithms. Comparison of f_0 contours, for example, is a difficult task. Comparing f_0 contours is not straightforward because the perception can not be mapped adequately. It is not important that the exact f_0 is the same, a certain variation in absolute value and time is acceptable and does not change the naturalness of the resulting speech. Objective and subjective evaluation are not standardized for EL speech and, therefore, benchmarks need to be developed and formulated.

3. Many algorithms, e.g., DREL suppression algorithms, are evaluated in noise free conditions. There might be an influence on the performance of the algorithms depending on the scenario. This has not been investigated in this or many of the related works. In this thesis we used two different algorithms for DREL suppression: 1. spectral subtraction and 2. modulation filtering. Both algorithms work well for clean speech but fail in noisy conditions (e.g., recordings in office environment). This behavior has not been tackled so far.

4. In the listening test conducted in Section 4.6 the conventional device, the Servox Digital, outperforms our proposed speech system. We suggest that a careful prototyping in terms of directly radiated noise suppression can increase the performance of our proposed system. In this experiment it turned out that intelligibility is the main property a system should fulfill, i.e., a more natural system can not overcome the problem of reduced intelligibility. Intelligibility tests would unveil the potential of our proposed system.

5. Although [Arifin et al., 2014] reported that the correlation between EMG and loudness is larger than EMG and fundamental frequency, the importance of loudness for EL speech has not been investigated in great detail. In case of healthy speech production loudness is important, although to a lesser amount to intonation compared to fundamental frequency. The loudness for EL speech is mainly influenced by the vocal tract and the opening of the mouth (loudness for [a] is larger than for [i]). Although loudness will have little influence on naturalness it must be considered in future realizations of the EL speech system due to completeness.

6. One major drawback of occupation of a hand and to press the EL against the neck tissue to turn on and off the EL 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. Furthermore, an adaptive algorithm for threshold detection would also help to improve the EMG controlled EL speech system. For daily usage a portable and reliable working hardware must be designed which

provides an easy handling and comfortable wearing of the electrodes. Concerning on/off control: In healthy speech voiced and unvoiced phonemes are constantly alternating. We claim that a rapidly, frequently changing on/off signal will not increase acceptance of EL speech. Such fast fluctuations will be disturbing and will result in reduced naturalness, probably also in reduced intelligibility. This should be confirmed using listening tests.

7. Concerning speaker dependent voice conversion, an important point is the subjective evaluation. In our experiments, we could improve naturalness a lot, but for untrained listeners it is impossible to understand the meaning of the spoken sentences, i.e., converted speech is often not intelligible. However, with pre-training of the listening effort intelligibility is improved a lot, e.g., if a person knows what will be said, it is understandable. This leads to the conclusion that one can improve the naturalness of EL speech using statistical voice conversion, but the resulting artifacts will decrease intelligibility.

8. Furthermore, there is a strong need to improve the gender variation, which means that women are actually perceived as women. Further work is necessary to improve the sound quality, while preserving the identity of the voice and to give the user an individual voice characteristic. We want to investigate different excitation signals and their power to fulfill above mentioned characteristics.

A
Device Design

A.1 Prototyping of the Device

An electro-magnetic transducer was proposed in Chapter 4. In order to use the transducer we modeled a housing using AutoCAD and created a 3D model which we printed.

The transducer (Figure A.1) was connected to a spring which was fixed to the spring suspension in the housing (Figure A.2). The ends of the spring (also called feet) were glued into the spring suspension. We used a special glue in order to enable the spring to move flexibly and that no obstruction appears at any side. The electro-magnetic transducer was mounted into the notch of the spring with the 3 holes. The spring-suspended mounting of the transducer is needed to allow a higher force transmission. After the feet of the spring were glued into the spring suspension and the cables were connected to the transducer (Figure A.3), the housing could be closed. The cables are connected inside of the housing to allow for cable relief.

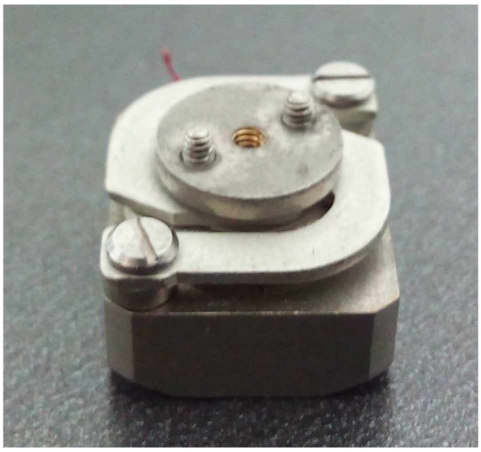


Figure A.1: Proposed electro-magnetic transducer (BHM BC2-E31).

Figure A.4 shows the new, electro-magnetic transducer with the spring mounted into the housing and a coupler disk attached on top of it. Screwing the coupler disk onto the transducer combines transducer and coupler disk with the spring. Since this design is just a prototype, other improvements of the housing and dimensioning are required to lead to further enhancement of the new system. It turned out that the main drawback of this prototype is the missing damping

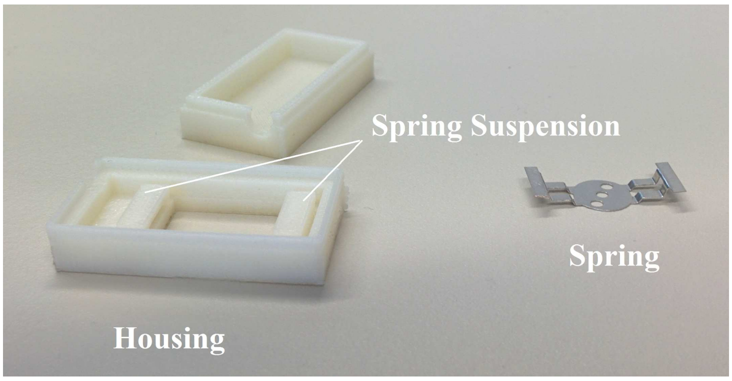


Figure A.2: Proposed housing with spring suspension and spring for the proposed electro-magnetic transducer (BHM BC2-E31); designed using AutoCAD and 3D printed.

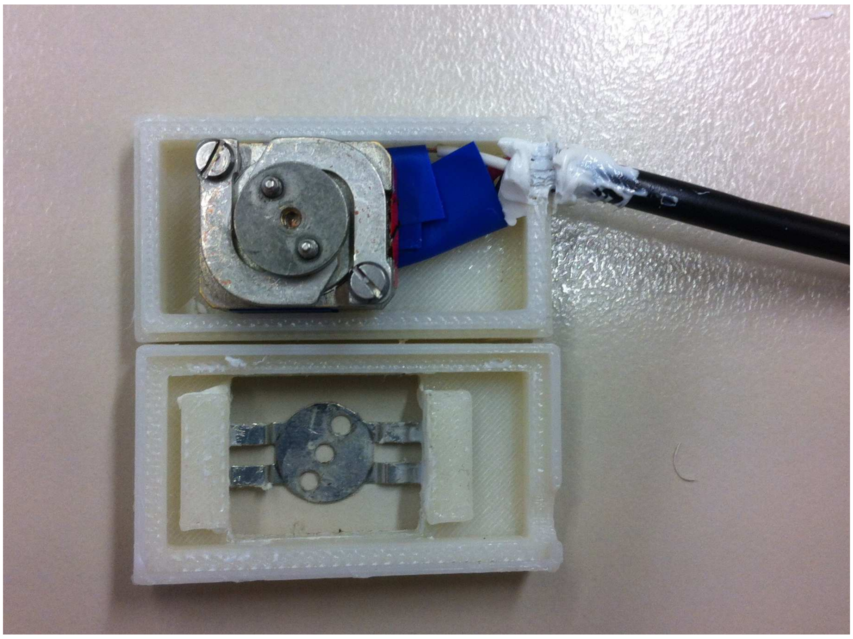
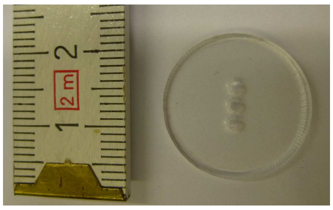
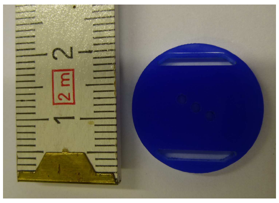
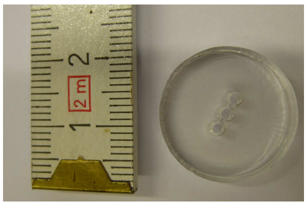
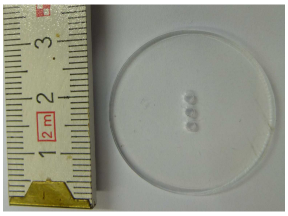
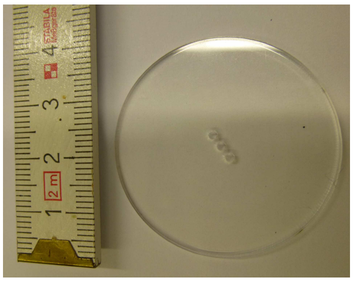
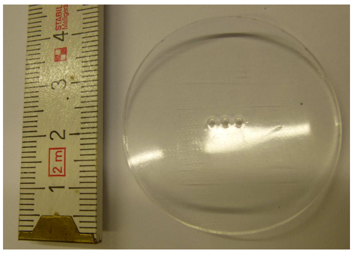
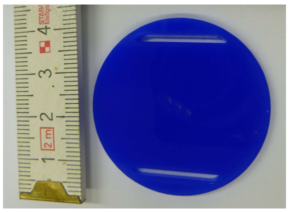
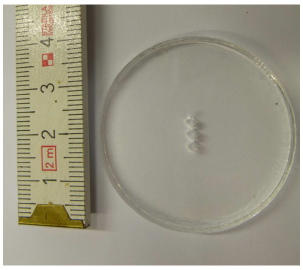
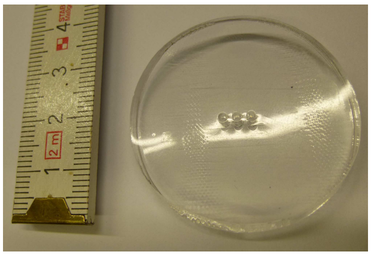
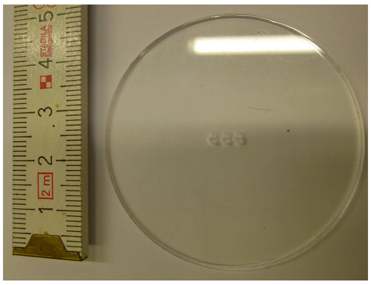
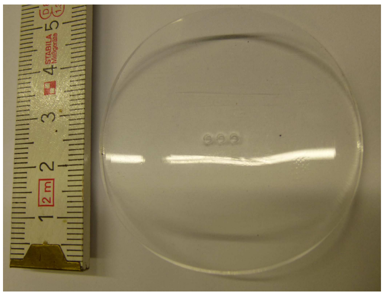
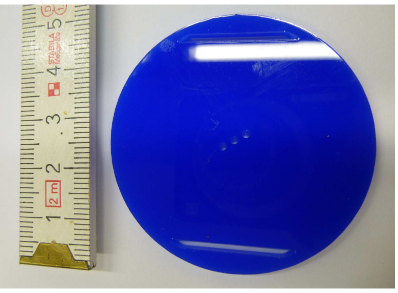
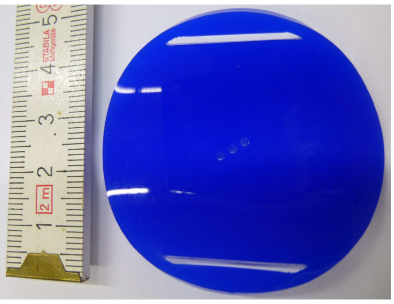
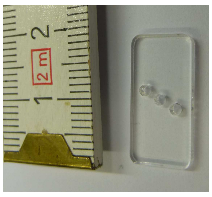
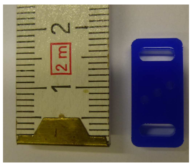
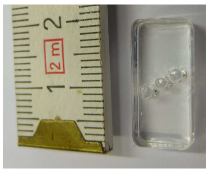


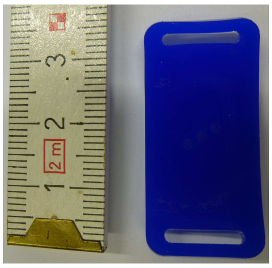
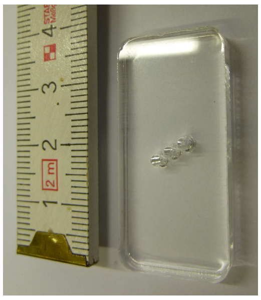
Figure A.3: Proposed electro-magnetic transducer (BHM BC2-E31) fixed and glued inside the housing.

of the DREL noise. The coupler disks with dimensions, weight and pictures are shown in Table A.1.

C.d.	round	rect.	flat	bent	Ø Size [mm]	Weigth [g]	Picture
1	×		×		20×2	0.74	
2	×		×		20×3	0.93	

C.d.	round	rect.	flat	bent	Ø Size [mm]	Weigth [g]	Picture
3	×		×		20×5	1.64	
4	×		×		30×2	1.69	
5	×		×		40×2	2.95	
6	×			×	40×2	3.00	
7	×		×		40×3	4.10	
8	×		×		40×5	6.81	
9	×			×	40×5	6.80	

C.d.	round	rect.	flat	bent	Ø Size [mm]	Weigth [g]	Picture
10	×		×		50×2	4.78	
11	×			×	50×2	4.80	
12	×		×		50×3	6.44	
13	×			×	50×3	6.37	
14		×	×		18×8.5×2	0.37	
15		×	×		18×8.5×3	0.46	
16		×	×		18×8.5×5	0.83	

C.d.	round	rect.	flat	bent	Ø Size [mm]	Weigth [g]	Picture
17		×	×		37×18×2	1.53	
18		×	×		37×18×3	2.12	
19		×		×	37×18×3	2.06	
20		×	×		37×18×5	3.55	
21		×	×		75×36×2	6.47	

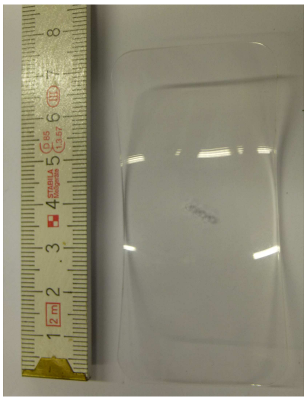
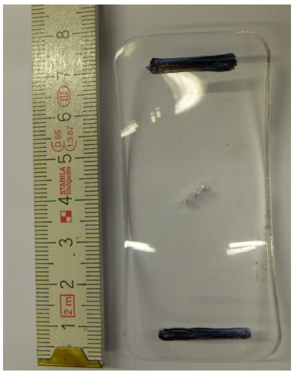
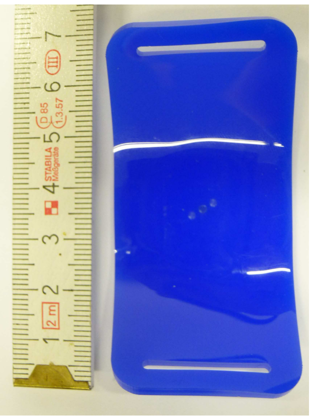
C.d.	round	rect.	flat	bent	Ø Size [mm]	Weigth [g]	Picture
22		×		×	75×36×2	6.46	
23		×		×	75×36×5	5.58	
24		×		×	75×36×3	8.66	

Table A.1: Dimensions, weight and pictures of different coupler disks (C.d.).

A.2 Mounting of the Device

The device should be wearable not only portable. Therefore, we choose a Velcro tape to attach the device to the neck. This design makes a hands-free design possible. Our first choice was to fix the device from the back to guarantee a high contact pressure. We found out that this causes distortions. Therefore, we fixed the coupler disk and not the whole device (see Figure A.5).

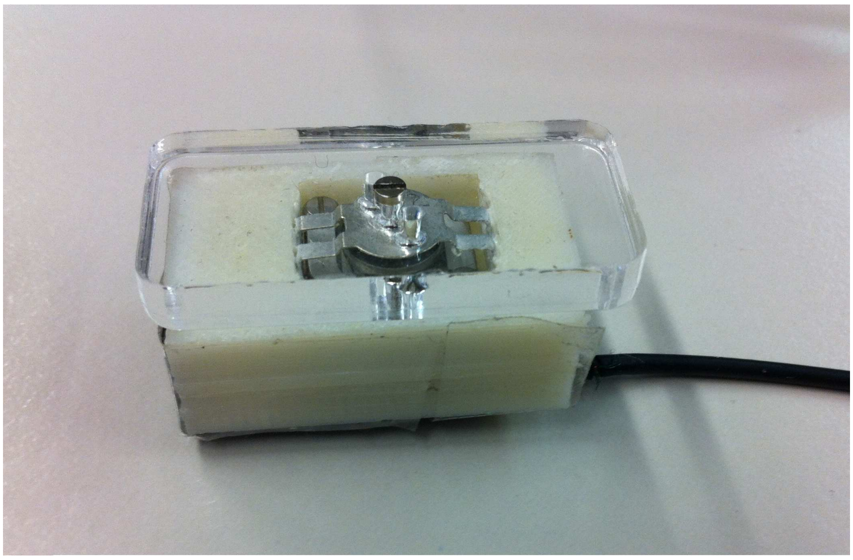
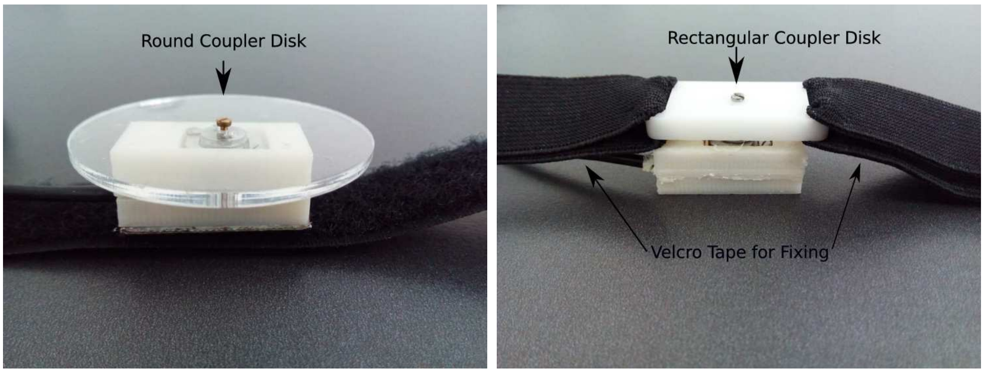


Figure A.4: Lateral view of the proposed EL device with electro-magnetic transducer (BHM BC2-E31) in the 3D printed housing with a mounted coupler disk on top.



(a) Previous design of proposed device, housing and mounting.

(b) Revised design of proposed device, housing and mounting.

Figure A.5: Previous and revised design of proposed device, housing and mounting.

B

Listening Tests

B.1 Audiometry

To verify that participants in our listening tests are normal hearing people, we measured the audiogram using the Audio-Ton CAS 1000. Audiometry is used to determine the threshold of hearing. The standards to perform audiometry is described in DIN EN ISO 8253. The audiometry software generates tones with different frequency and amplitude. The generated tones are presented to a listener via appropriate amplifiers, bone conductors and headphones or loudspeakers. The headphones are ear-embracing to be able to eliminate noise and disturbing directly radiated signals. We neglect the use of bone conductors. The investigated frequency range is between 125 Hz and 8 kHz. Within this frequency range we examine octaves (125 Hz, 250 Hz, 500 Hz) until 500 Hz and halve octaves (750 Hz, 1000 Hz, 1500 Hz, 200 Hz, 300 Hz, 4500 Hz, 6000 Hz, 8000 Hz) above 500 Hz.

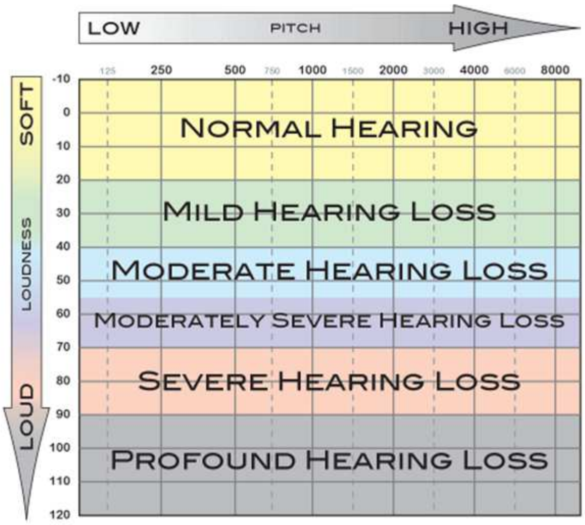
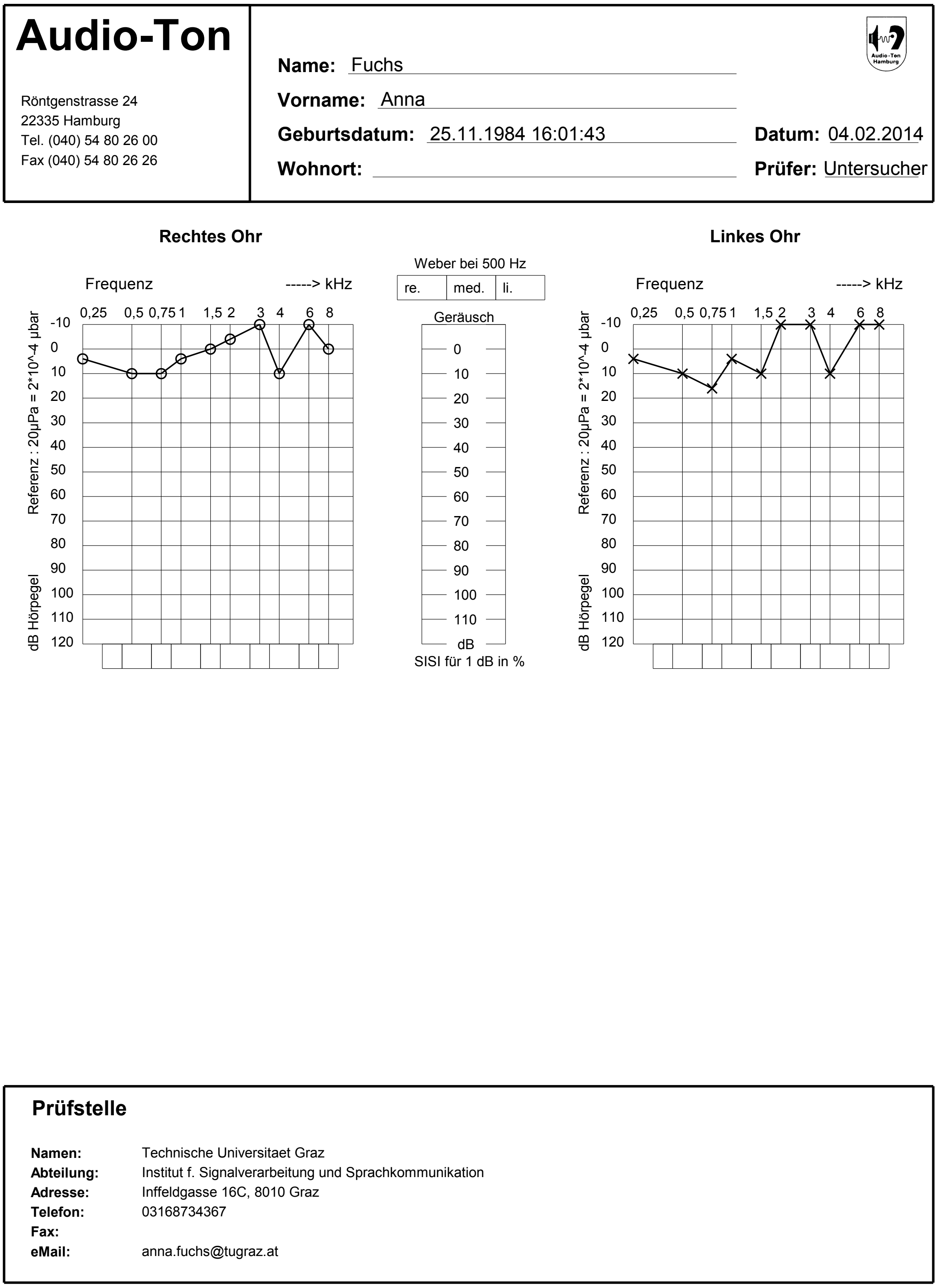


Figure B.1: Different hearing loss thresholds [Sensiple, 2015].

During the measurement of the threshold of hearing several requirements need to be considered:

- The subject should remove glasses or disturbing accessories during the measurement.
- The subject should avoid any disturbing movements.
- Contemporary presentation of test tone and optical or acoustic attraction need to be avoided.
- Start the test with superior ear.

The test stimuli are presented to the test person and changed in steps of 5 dB. When the test stimulus is audible to the test person, the value is marked in the audiogram. The frequency value in [Hz] is on the x-axis and the hearing level in [dB] is on the y-axis. The values for "Normal hearing" are considered up to 15 dB (see Figure B.1). The measurements for one person can be seen in Figure B.2 for the right and left ear.



Audio-Ton

Röntgenstrasse 24
22335 Hamburg
Tel. (040) 54 80 26 00
Fax (040) 54 80 26 26

Name: Fuchs

Vorname: Anna

Geburtsdatum: 25.11.1984 16:01:43

Wohnort:

Datum: 04.02.2014

Prüfer: Untersucher

Prüfstelle

Namen:	Technische Universitaet Graz
Abteilung:	Institut f. Signalverarbeitung und Sprachkommunikation
Adresse:	Inffeldgasse 16C, 8010 Graz
Telefon:	03168734367
Fax:	
eMail:	anna.fuchs@tugraz.at

Figure B.2: Audiogram of a normal hearing test person.

B.2 Listening Test

Subjective listening tests are the gold standard to evaluate speech quality. A number of standards and recommendations exist which have been developed by experts in the field from industry and academia due to a large-scale need to address a problem. The listening tests conducted by us are inspired by the ITU-R recommendation which is used for broad bandwidth (20 Hz – 20 kHz) signals. The main focus of this listening test is upon applications related to audio for radio communication.

Listening tests are capable to provide following information, according to [Bech and Zacharov, 2006]:

- Identify whether or not audio stimuli are perceptually identical.
- Establish whether a sample is perceptually equivalent, superior or inferior to another sample with regard to audio quality.
- Define to what degree a sample is superior to another in terms of audio quality.

...

It is also important to note that listening tests are not directly able to:

- Identify and locate the problem parameter of an audio algorithm.

...

- Define how developers should improve their systems to obtain significant audio quality improvement.

In most of our listening tests we wanted to compare two systems or two algorithms. Therefore, we decided to use an AB listening test with comparison category rating (CCR) which is described in the following:

Test paradigm Paired rating hidden reference.

Description Subjects are presented with a pair of samples for each item. The hidden reference is identified and subjects are asked to rate the test items against each other on a 7-point categorical comparison category rating scale.

Primary Application Applied to evaluations of systems that may improve or degrade speech quality compared to the reference.

Scale 7-point comparison category rating scale. Subject is instructed to rate the second sample against the first.

Number of subjects 32 (minimum of 12 subjects from either gender).

Stimuli 8 samples per talker, 2 male, 2 female.

Analysis approach The first step is to record the data relative to the reference. Apply ANOVA and report means (CMOS).

Speech stimuli are presented in a random order unique to each listener. Figure B.3 shows the CCR assessment categories ranging from -3 to 3 where the values indicate whether stimulus B sounds better and by how much better, than stimulus A. Please note that we reversed the question in our listening tests by mistake (A compared to B, not B compared to A). We implemented a graphical user interface in MATLAB for conducting the listening tests (see Figure B.4).

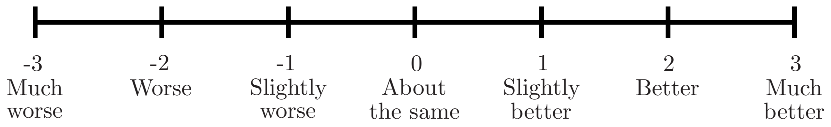
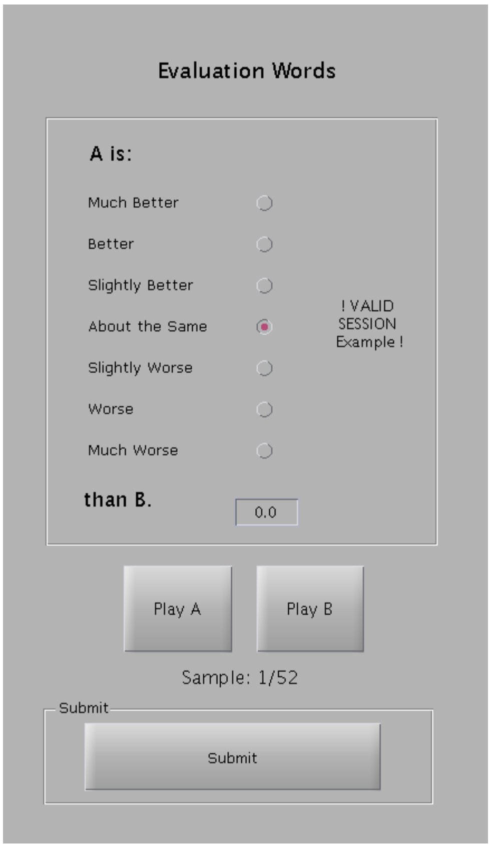


Figure B.3: Degradation rating scale.



Evaluation Words

A is:

Much Better ○
Better ○
Slightly Better ○
About the Same ●
Slightly Worse ○
Worse ○
Much Worse ○

! VALID SESSION Example !

than B. 0.0

Sample: 1/52

Figure B.4: Graphical user interface implemented in MATLAB *for evaluation with two playback buttons, 7-point comparison category rating scale, a sample counter and a submit button.*

We decided against the Mean Opinion Score (MOS) which is used to test human user's perception of the quality of telephone speech and similar. This kind of test is reference free where people judge the quality of a presented stimulus on a scale between 1 and 5, 1 representing bad quality with very annoying impairment and 5 being excellent quality with imperceptible impairment. In our opinion judging EL speech without reference is critical when naive listeners are chosen for the listening test because these people are not used to EL speech and can only produce reliable scores relative to other stimuli.

Bibliography

[Ahmadi et al., 2014] Ahmadi, F., Araujo Ribeiro, M., and Halaki, M. (2014). Surface electromyography of neck strap muscles for estimating the intended pitch of a bionic voice source. In *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 37–40.

[Amon, 2014] Amon, C. (2014). Electrolarynx Control using Electromyographic Signals. Master's thesis, Graz University of Technology, Signal Processing and Speech Communication Laboratory.

[Arifin et al., 2014] Arifin, F., Sardjono, T. A., and Purnomo, M. H. (2014). The relationship between electromyography signal of neck muscle and human voice signal for controlling loudness of electrolarynx. *Biomedical Engineering: Applications, Basis and Communications*, 26(05):1450054–1–1450054–7.

[Arta, 2015] Arta (2015). Audio measurement and analysis software. `http://www.artalabs.hr/`. Accessed: 2015-04-17.

[Association, 1999] Association, I. P. (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. A Regents publication. Cambridge University Press.

[Atkinson, 1978] Atkinson, J. E. (1978). Correlation analysis of the physiological factors controlling fundamental voice frequency. *The Journal of the Acoustical Society of America*, 63(1):211–222.

[Azarnoush et al., 2007] Azarnoush, H., Mir, F., Agaian, S., Jamshidi, M., and Shadaram, M. (2007). Alaryngeal speech enhancement using minimum statistics approach to spectral subtraction. In *IEEE International Conference on System of Systems Engineering (SoSE 2007)*, pages 1–5.

[Basha, 2011] Basha, S. (2011). Enhancement of Electrolaryngeal Speech. Master's thesis, Electrical Engineering, Indian Institute of Technology Bombay.

[Basha and Pandey, 2012] Basha, S. and Pandey, P. (2012). Real-time enhancement of electrolaryngeal speech by spectral subtraction. In *National Conference on Communications (NCC)*, pages 1–5.

[Bech and Zacharov, 2006] Bech, S. and Zacharov, N. (2006). *Perceptual Audio Evaluation : Theory, Method and Application*. John Wiley & Sons, Chichester.

[Benesty et al., 2008] Benesty, J., Sondhi, M. M., and Huang, Y., editors (2008). *Springer Handbook of Speech Processing*. Springer, Berlin.

[Bergauer and Janknecht, 2011] Bergauer, U. and Janknecht, S. (2011). *Praxis der Stimmtherapie*. Springer Berlin Heidelberg.

[BHM, 2004] BHM (2004). Bc2-e31 technical data sheed.

[BHM, 2015] BHM (2015). BHM - Innovative hearing systems made in Austria. `http://www.bhm-tech.at/index.php?id=2`. Accessed: 2015-04-17.

[Bishop, 2007] Bishop, C. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.

[Boersma and Weenink, 2007] Boersma, P. and Weenink, D. (2007). Praat ver. 4.06, software. `http://www.praat.org`. Accessed: 2015-08-02.

[Brown and Feinstein, 1977] Brown, W. and Feinstein, S. (1977). Speaker Sex Identification Utilizing a Constant Laryngeal Source. *Folia Phoniatrica et Logopaedica*, 29(3):240–248.

[Cheah et al., 2015] Cheah, L., Bain, J., Gonzalez, J., Ell, S., Gilbert, J., Moore, R., and Green, P. (2015). A user-centric design of permanent magnetic articulography based assistive speech technology. In *8th International Conference on Bio-inspired Systems and Signal Processing, Lisbon*.

[Christensen et al., 2012] Christensen, H., Cunningham, S., Fox, C., Green, P., and Hain, T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. In *13th Annual Conference of the International Speech Communication Association (InterSpeech 2012)*, pages 1776–1779.

[Cole et al., 1997] Cole, D., Sridharan, S., Moody, M., and Geva, S. (1997). Application of noise reduction techniques for alaryngeal speech enhancement. In *Proceedings of IEEE Conference. Speech and Image Technologies for Computing and Telecommunications. TENCON'97*, pages 491–494, Brisbane, Australia.

[Coleman, 1976] Coleman, R. O. (1976). A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech, Language & Hearing Research*, 19(1):168–80.

[Cox and Doyle, 2014] Cox, S. and Doyle, P. (2014). The influence of electrolarynx use on postlaryngectomy voice-related quality of life. *Otolaryngology–Head and Neck Surgery (OTO-HNS)*, 150(6):1005–1009.

[De Armas et al., 2014] De Armas, W., Khondaker, A. M., and Tom, C. (2014). Vocal frequency estimation and voicing state prediction with surface EMG pattern recognition. *Speech Communication*, 63–64(0):15 – 26.

[Doi et al., 2014] Doi, H., Toda, T., Nakamura, K., Saruwatari, H., and Shikano, K. (2014). Alaryngeal speech enhancement based on one-to-many eigenvoice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):172–183.

[Draxler and Jänsch, 2004] Draxler, C. and Jänsch, K. (2004). Speechrecorder – a universal platform independent multi-channel audio recording software. In *Proceedings of the IV. International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal.

[Ellis, 2003] Ellis, D. (2003). Dynamic time warp (DTW) in Matlab. `http://labrosa.ee.columbia.edu/matlab/dtw/`. Accessed: 2015-07-25.

[Espy-Wilson et al., 1998] Espy-Wilson, C. Y., Chari, V., MacAuslan, J., Huang, C., and Walsh, M. (1998). Enhancement of electrolaryngeal speech by adaptive filtering. *Journal of Speech, Language & Hearing Research*, 41:1253–1264.

[Fant et al., 1985] Fant, G., Liljencrants, J., and Lin, Q.-G. (1985). A four parameter model of glottal flow. Technical Report STL-QPSR Nos. 2-3, Royal Institute of Technology, Stockholm, Sweden.

[Freeman et al., 1989] Freeman, D., Cosier, G., Southcott, C., and Boyd, I. (1989). The voice activity detector for the pan-european digital cellular mobile telephone service. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1989)*, pages 369–372, Glasgow.

[Ghorbandoost et al., 2015] Ghorbandoost, M., Sayadiyan, A., Ahangar, M., Sheikhzadeh, H., Shahrebabaki, A. S., and Amini, J. (2015). Voice conversion based on feature combination with limited training data. *Speech Communication*, 67(03):113 – 128.

[Giet, 1956] Giet, F. (1955–1956). Kann man in einer Tonsprache flüstern? *Lingua*, 5(0):372 – 381.

[Girbes and Elbers, 2014] Girbes, A. R. and Elbers, P. W. (2014). Speech in an orally intubated patient. *New England Journal of Medicine*, 370(12):1172–1173.

[Goldstein et al., 2004] Goldstein, E., Heaton, J., Kobler, J., Stanley, G., and Hillman, R. (2004). Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity. *IEEE Transactions on Biomedical Engineering*, 51(2):325–332.

[Goldstein et al., 2007] Goldstein, E., Heaton, J., Stepp, C. E., and Hillman, R. (2007). Training effects on speech production using a hands-free electromyographically controlled electrolarynx. *Journal of Speech, Language & Hearing Research*, 50(2):335–351.

[Griffin, 2015] Griffin (2015). http://www.griffinlab.com/. `http://www.griffinlab.com/`. Accessed: 2015-04-13.

[Griffin, 1998] Griffin, C. J. (1998). Artificial larynx with frequency control. US Patent 5.812.681.

[Hagmüller, 2009] Hagmüller, M. (2009). *Speech Enhancement for Disordered and Substitution Voices*. PhD thesis, Graz University of Technology.

[Hanquinet et al., 2006] Hanquinet, J., Grenez, F., and Schoentgen, J. (2006). Synthesis of disordered voices. *Nonlinear Analyses and Algorithms for Speech Processing*, pages 231–241.

[Hansen and Pellom, 1998] Hansen, J. H. L. and Pellom, B. L. (1998). An effective quality evaluation protocol for speech enhancement algorithms. In *Proceedings of the international conference on Speech and Language Processing*, pages 2819–2822.

[Hänsler and Schmidt, 2004] Hänsler, E. and Schmidt, G. (2004). *Acoustic Echo and Noise Control: A Practical Approach*. John Wiley & Sons.

[Hart, 1981] Hart, t. J. (1981). Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America*, 69(3):811–821.

[Hashiba et al., 2007] Hashiba, M., Hashiba, M., Sugai, Y., Izumi, T., Ino, S., and Ifukube, T. (2007). Development of a wearable electro-larynx for laryngectomees and its evaluation. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 5267–5270.

[Heaton et al., 2011] Heaton, J., Robertson, M., and Griffin, C. (2011). Development of a wireless electromyographically controlled electrolarynx voice prosthesis. In *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5352–5355, Boston, MA, USA.

[Heimomed, 2015] Heimomed (2015). http://www.heimomed.de. `http://www.heimomed.de`. Accessed: 2015-04-13.

[Helander and Nurminen, 2007] Helander, E. and Nurminen, J. (2007). On the importance of pure prosody in the perception of speaker identity. In *8th Annual Conference of the International Speech Communication Association (InterSpeech 2007)*, pages 2665–2668. ISCA.

[Hermansky and Morgan, 1994] Hermansky, H. and Morgan, N. (1994). RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589.

[Hiebl, 2014] Hiebl, H. (2014). Elektroakustik, Labor. Laboratory Tutorial.

[Hirsch, 2002] Hirsch, G. (2002). Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task, version 2.0. ETSI STQ-AURORA DSR working group.

[Hofe et al., 2013] Hofe, R., Ell, S. R., Fagan, M. J., Gilbert, J. M., Green, P. D., Moore, R. K., and Rybchenko, S. I. (2013). Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Communication*, 55(1):22 – 32.

[Houston et al., 1999] Houston, K., Hillman, R., Kobler, J., and Meltzner, G. (1999). Development of sound source components for a new electrolarynx speech prosthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999)*, volume 4, pages 2347–2350, Phoenix, Arizona, USA.

[Hülser and Rothmund, 2015] Hülser, G. and Rothmund, F. (2015). Automatic Speech Recognition for Electro-Larynx Speech. Technical report, Graz University of Technology, Signal Processing and Speech Communication Laboratory.

[Humphrey, 2014] Humphrey, R. (2008-2014). Playrec - Multi-channel Matlab Audio. `http://www.playrec.co.uk/contact.html`. Accessed: 2015-08-01.

[Institut, 2015] Institut, D. H. (2015). Der elektromagnetische Hörer. `http://www.dhi-online.de/DhiNeu/12_Fachtec/FtHgTec/06_Wandler/Fthgtec_0604_2.html`. Accessed: 2015-04-13.

[ITU, 2001] ITU (2001). ITU-T recommendation P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

[Jochum and Reiner, 2008] Jochum, C. and Reiner, P. (2008). Comparison of Excitation Signals for an Electronic Larynx. Master's thesis, Graz University of Technology, Signal Processing and Speech Communication Laboratory.

[Kabir et al., 2008] Kabir, R., Greenblatt, A., Panetta, K., and Agaian, S. (2008). Enhancement of alaryngeal speech utilizing spectral subtraction and minimum statistics. In *International Conference on Machine Learning and Cybernetics*, volume 7, pages 3704–3709.

[Kawahara et al., 2001] Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *Models and analysis of vocal emissions for biomedical applications*, pages 59–64, Firenze, Italy.

[Kawahara et al., 1999] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3-4):187–207.

[Kikuchi and Kasuya, 2004] Kikuchi, Y. and Kasuya, H. (2004). Development and evaluation of pitch adjustable electrolarynx. *Speech Prosody*, pages 1–4.

[Klammer, 2015] Klammer, H. (2015). Learning Effects for Electromyographically controlled Electrolarynx Speech. Master's thesis, Graz University of Technology, Signal Processing and Speech Communication Laboratory.

[Kleijn et al., 2015] Kleijn, W., Crespo, J., Hendriks, R., Petkov, P., Sauert, B., and Vary, P. (2015). Optimizing speech intelligibility in a noisy environment: A unified view. *IEEE Signal Processing Magazine*, 32(2):43–54.

[Kubert et al., 2009] Kubert, H., Stepp, C., Zeitels, S.M. anad Gooey, J., Walsh, M., Prakash, S., Hillman, R., and Heaton, J. (2009). Electromyographic control of a hands-free electrolarynx using neck strap muscles. *Journal of Communication Disorders*, 42(3):211–225.

[Lakhtakia and Martín-Palma, 2013] Lakhtakia, A. and Martín-Palma, R. J., editors (2013). *Engineered Biomimicry*. Elsevier, Boston.

[Lasarcyk, 2005] Lasarcyk, E. (2005). Wie deutsch klingt Englisches im Deutschen – und warum? Master's thesis, Universität Bonn, Institut für Kommunikationsforschung und Phonetik.

[Law et al., 2009] Law, I., Ma, E., and Yiu, E. (2009). Speech intelligibility, acceptability, and communication-related quality of life in chinese alaryngeal speakers. *Archives of Otolaryngology–Head and Neck Surgery (OTO-HNS)*, 135(7):704–711.

[Leggetter and Woodland, 1995] Leggetter, C. and Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171 – 185.

[Liu et al., 2006] Liu, H., Zhao, Q., Wan, M., and Wang, S. (2006). Enhancement of electrolarynx speech based on auditory masking. *IEEE Transactions on Biomedical Engineering*, 53(5):865–874.

[Lohscheller, 2003] Lohscheller, J. (2003). *Dynamics of the Laryngectomee Substitute Voice Production*. PhD thesis, Shaker-Verlag, Aachen, Germany.

[Loscos and Bonada, 2006] Loscos, A. and Bonada, J. (2006). Esophageal voice enhancement by modeling radiated pulses in frequency domain. In *Proceedings of 121st Convention of the Audio Engineering Society*, pages 1064–1067, San Francisco, CA, USA.

[Lüchtrath, 2015] Lüchtrath, L. I. (2015). Parametrization of a Transducer for Electro-Larynx Speech Production. Master's thesis, Graz University of Technology, Signal Processing and Speech Communication Laboratory.

[Madden, 2013] Madden, B. (2013). *Augmented Control of a hands-free Electrolarynx*. PhD thesis, Dublin Institute of Technology.

[Madden et al., 2010] Madden, B., Nolany, M., Burkez, T., Condron, J., and Coyle, E. (2010). Intelligibility of electrolarynx speech using a novel actuator. In *The Irish Signals and Systems Conference*, pages 159–162, Cork, Ireland.

[Madhushankara et al., 2015] Madhushankara, M., Prasad, K., Chaitanya, C., and Bhat, S. (2015). A low power low frequency oscillator for driving electrolarynx. In *International Conference on VLSI Systems, Architecture, Technology and Applications (VLSI-SATA)*, pages 1–3, Daejon, Korea.

[Maier et al., 2010] Maier, A., Haderlein, T., Stelzle, F., Nöth, E., Nkenke, E., Rosanowski, F., Schützenberger, A., and Schuster, M. (2010). Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. *EURASIP Journal on Audio Speech Music Processing*, 2010:926–951.

[Matsui et al., 2014] Matsui, K., Kimura, K., Pérez, A., Rodríguez, S., and Corchado, J. M. (2014). Development of electrolarynx by multi-agent technology and mobile devices for prosody control. In *Highlights of Practical Applications of Heterogeneous Multi-Agent Systems. The PAAMS Collection Communications in Computer and Information Science*, volume 430, pages 54–65. Springer Science + Business Media.

[McLoughlin, 2014] McLoughlin, I. (2014). Super-audible voice activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(9):1424–1433.

[McLoughlin, 2015] McLoughlin, I. (2015). The bionic voice project. `http://www.lintech.org/bionicvoice/`. Accessed: 2015-04-13.

[Meltzner, 2003] Meltzner, G. (2003). *Perceptual and Acoustic Impacts of Aberrant Properties of Electrolaryngeal Speech*. PhD thesis, Massachusetts Institute of Technology.

[Meltzner et al., 2011] Meltzner, G., Colby, G., Deng, Y., and Heaton, J. (2011). Signal acquisition and processing techniques for sEMG based silent speech recognition. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, volume 2011, pages 4848–4851, Boston, MA, USA.

[Meltzner, 1998] Meltzner, G. S. (1998). Measuring the Neck Transfer Function of Laryngectomy Patients. Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

[Meltzner and Hillman, 2005] Meltzner, G. S. and Hillman, R. E. (2005). Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *Journal of Speech, Language & Hearing Research*, 48(4):766–79.

[Meltzner et al., 2003] Meltzner, G. S., Kobler, J. B., and Hillman, R. E. (2003). Measuring the neck frequency response function of laryngectomy patients: Implications for the design of electrolarynx devices. *The Journal of the Acoustical Society of America*, 114(2):1035–1047.

[Merlo et al., 2008] Merlo, M., Li, G., and Bachman, M. (2008). A remotely powered and wirelessly controlled intraoral electrolarynx. In *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3459–3462, Vancouver, Canada.

[Merriam-Webster, 2015] Merriam-Webster (2015). http://www.merriam-webster.com/dictionary/bionic. `http://www.merriam-webster.com/dictionary/bionic`. Accessed: 2015-04-22.

[Meyer-Eppler, 1957] Meyer-Eppler, W. (1957). Realization of prosodic features in whispered speech. *The Journal of the Acoustical Society of America*, 29(1):104–106.

[Milner and Shao, 2007] Milner, B. and Shao, X. (2007). Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction. *IEEE Transactions on Audio, Speech & Language Processing*, 15(1):24–33.

[MIT, 2015] MIT, T.-R. (2015). http://www.technologyreview.com/lists/breakthrough-technologies/2013/. `http://www.technologyreview.com/lists/breakthrough-technologies/2013/`. Accessed: 2015-04-22.

[Mouchtaris et al., 2006] Mouchtaris, A., der Spiegel, J. V., and Mueller, P. (2006). Nonparallel training for voice conversion based on a parameter adaptation approach. *IEEE Transactions on Audio, Speech & Language Processing*, 14(3):952–963.

[Moulines and Charpentier, 1990] Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467.

[Nachtigall, 2002] Nachtigall, W. (2002). *Bionik: Grundlagen und Beispiele für Ingenieure und Naturwissenschaftler*. Springer Berlin Heidelberg.

[Nagle et al., 2012] Nagle, K., Eadie, T., Wright, D., and Sumida, Y. (2012). Effect of fundamental frequency on judgments of electrolaryngeal speech. *American Journal of Speech-Language Pathology*, 21(2):154–66.

[Nakajima et al., 2006] Nakajima, Y., Kashioka, H., Campbell, N., and Shikano, K. (2006). Non-Audible Murmur (NAM) Recognition. *IEICE Transactions on Information and Systems*, E89-D(1):1–4.

[Nakamura, 2010] Nakamura, K. (2010). *Speaking-Aid Systems using Statistical Voice Conversion for Electrolaryngeal Speech*. PhD thesis, Nara Institute of Science and Technology.

[Nakamura et al., 2011] Nakamura, K., Toda, T., Saruwatari, H., and Shikano, K. (2011). Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech. *Speech Communication*, 54(1):134–146.

[Nakatani and Irino, 2004] Nakatani, T. and Irino, T. (2004). Robust and accurate fundamental frequency estimation based on dominant harmonic components. *The Journal of the Acoustical Society of America*, 116(6):3690–3700.

[NDT, 2015] NDT (2015). The hysteresis loop and magnetic properties. `https://www.nde-ed.org/EducationResources/CommunityCollege/MagParticle/Physics/HysteresisLoop.htm`. Accessed: 2015-04-13.

[Ng, 1996] Ng, M. L. (1996). *A perceptual and acoustic study of alaryngeal speech in adult Cantonese-speaking males*. PhD thesis, University of Connecticut.

[Norton and Bernstein, 1993] Norton, R. L. and Bernstein, R. S. (1993). Improved laboratory prototype electrolarynx (lapel): Using inverse filtering of the frequency response function of the human throat. *Annals of Biomedical Engineering*, 21(2):163–174.

[Nurminen et al., 2012] Nurminen, J., Silen, H., Popa, V., Helander, E., and Gabbouj, M. (2012). *Speech Enhancement, Modeling and Recognition- Algorithms and Applications*. InTech.

[Ooe, 2012] Ooe, K. (2012). Development of controllable artificial larynx by neck myoelectric signal. In *26th European Conference on Solid-State Transducers*, volume 47, pages 869 – 872, Krakow, Poland.

[Ooe et al., 2000] Ooe, K., Fukuda, T., and Arai, F. (2000). A new type of artificial larynx using a PZT ceramics vibrator as a sound source. *IEEE/ASME Transactions on Mechatronics*, 5(2):221–225.

[Pandey et al., 2002] Pandey, P., Bhandarkar, S., Bachher, G., and Lehana, P. (2002). Enhancement of alaryngeal speech using spectral subtraction. In *14th International Conference on Digital Signal Processing (DSP), 2002*, volume 2, pages 591–594, Santorini, Hellas, Greece.

[Pandey and Basha, 2010] Pandey, P. C. and Basha, S. K. (2010). Enhancement of electrolaryngeal speech by spectral subtraction, spectral compensation, and introduction of jitter and shimmer. In *Proceedings of 20th International Congress on Acoustics (ICA)*, pages 3851–3854, Australia, Sydney.

[Perrachione et al., 2014] Perrachione, T. K., Stepp, C. E., Hillman, R. E., and Wong, P. C. M. (2014). Talker identification across source mechanisms: Experiments with laryngeal and electrolarynx speech. *Journal of Speech, Language & Hearing Research*, 57:1651–1665.

[Pineda-Rico et al., 2008] Pineda-Rico, Z., Dieck-Assad, G., Martinez-Chapa, S., and Avila-Ortega, A. (2008). A switching capacitor CMOS based device for hands-free electrolarynx activation using electromyographic signals. In *IEEE Conference on Electronics, Robotics and Automotive Mechanics*, pages 8–13, Cuernavaca, Morelos, Mexico.

[Pratapwar et al., 2003] Pratapwar, S. S., Pandey, P. C., and Lehana, P. K. (2003). Reduction of background noise in alaryngeal speech using spectral subtraction with quantile based noise estimation. In *Proc. of 7th World Multiconference on Systemics, Cybernetics and Informatics (SCI)*, pages 408–413, Orlando, Florida.

[Qi and Weinberg, 1991] Qi, Y. and Weinberg, B. (1991). Low-frequency energy deficit in electrolaryngeal speech. *Journal of Speech, Language & Hearing Research*, 34(6):1250–1256.

[Rabiner et al., 1976] Rabiner, L. R., Cheng, M. J., Osenberg, A. E., and McGonegal, C. A. (1976). A comparative performance study of several pitch detection algorithms. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1976*, pages 399–418.

[Reynolds et al., 2000] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41.

[Rosso et al., 2012] Rosso, M., Siric, L., Ticac, R., Starcevic, R., Segec, I., and Kraljik, N. (2012). Perceptual evaluation of alaryngeal speech. *Coll. Antropol.*, 36(2):115–118.

[Saikachi, 2009] Saikachi, Y. (2009). *Development, Perceptual Evaluation, and Acoustic Analysis of Amplitude-based F0 control in Electrolarynx Speech.* PhD thesis, Harvard-MIT Division of Health Sciences and Technology.

[Saikachi et al., 2009] Saikachi, Y., Stevens, K. N., and Hillman, R. E. (2009). Development and perceptual evaluation of amplitude-based f0 control in electrolarynx speech. *Journal of Speech, Language & Hearing Research*, 52(5):1360–1369.

[Sauert and Vary, 2006] Sauert, B. and Vary, P. (2006). Near end listening enhancement: Speech intelligibility improvement in noisy environments. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 493 – 496, Toulouse, France.

[Schiefer and Hagen, 2000] Schiefer, J. and Hagen, R. (2000). Rehabilitation laryngektomierter Karzinompatienten. *Der Onkologe*, 6(1):36–43.

[Schiel and Baumann, 2006] Schiel, F. and Baumann, A. (2006). Phondat 1, corpus version 3.4. `http://www.bas.uni-muenchen.de/Bas/BasFormatseng.html`.

[Schultz and Wand, 2010] Schultz, T. and Wand, M. (2010). Modeling coarticulation in emg-based continuous speech recognition. *Speech Communication*, 52(4):341–353.

[Schuster et al., 2006] Schuster, M., Haderlein, T., Nöth, E., Lohscheller, J., Eysholdt, U., and Rosanowski, F. (2006). Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. *European Archives of Oto-Rhino-Laryngology and Head & Neck*, 263(2):188–193.

[SECOM, 2015] SECOM (2015). http://www.secom.co.jp/personal/medical/myvoice.html. `http://www.secom.co.jp/personal/medical/myvoice.html`. Accessed: 2015-04-13.

[Sensiple, 2015] Sensiple (2015). The national heraing test. `http://www.nationalhearingtest.org/wordpress/?p=786`. Accessed: 2015-04-13.

[Servona, 2015] Servona (2015). http://www.servona.de. `http://www.servona.de`. Accessed: 2015-04-13.

[Shanxi Datong University, 2015] Shanxi Datong University (2015). American spoken english – for chinese speakers. `http://americanspokenenglish.weebly.com/chapter-2-pronunciation.html`. Accessed: 2015-06-01.

[Sharifzadeh, 2011] Sharifzadeh, H. R. (2011). *Reconstruction of Natural Sounding Speech from Whispers*. PhD thesis, Nanyang Technological University, Singapore, School of Computer Engineering.

[Sharifzadeh et al., 2010] Sharifzadeh, H. R., McLoughlin, I. V., and Ahmadi, F. (2010). Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec. *IEEE Transactions on Biomedical Engineering*, 57(10):2448–2458.

[Shum, 2009] Shum, S. (2009). A gmm-straight approach to voice conversion. Technical report, Berkeley University of California, Berkeley, California, USA.

[Shute, 2003] Shute, B. (2003). *Perceptions of Artificial Larynx Reliability According to Laryngectomees and Speech-Language Pathologists*. PhD thesis, Gonzaga University, Spokane, Wash.

[Stepp, 2008] Stepp, C. E. (2008). Electromyographic Control of Prosthetic Voice after Total Laryngectomy. Master's thesis, Massachusetts Institute of Technology.

[Stylianou et al., 1998] Stylianou, Y., Cappé, O., and Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2):131–142.

[Sugio et al., 2007] Sugio, Y., Kanetake, R., Tanaka, A., and Ooe, K. (2007). Work of PZT ceramics sounder for sound source artificial larynx. In *International Symposium on Micro-NanoMechatronics and Human Science*, pages 237–242, Nagoya, Japan.

[Takahashi et al., 2005] Takahashi, H., Nakao, M., Kikuchi, Y., and Kaga, K. (2005). Alaryngeal speech aid using an intra-oral electrolarynx and a miniature fingertip switch. *Auris Nasus Larynx*, 32:157–62.

[Takahashi et al., 2008] Takahashi, H., Nakao, M., Kikuchi, Y., and Kaga, K. (2008). Intra-oral pressure–based voicing control of electrolaryngeal speech with intra-oral vibrator. *Journal of Voice*, 22(4):420–429.

[Tanaka et al., 2014a] Tanaka, K., Toda, T., Neubig, G., Sakti, S., and Nakamura, S. (2014a). Direct f0 control of an electrolarynx based on statistical excitation feature prediction and its evaluation through simulation. In *15th Annual Conference of the International Speech Communication Association (InterSpeech 2014)*, pages 31–35, Singapore.

[Tanaka et al., 2014b] Tanaka, K., Toda, T., Neubig, G., Sakti, S., and Nakamura, S. (2014b). A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation. *IEICE Transactions on Information and Systems*, E97-D(6):1429–1437.

[Tanaka et al., 2014c] Tanaka, K., Toda, T., Neubig, G., Sakti, S., and Nakamura, S. (2014c). An inter-speaker evaluation through simulation of electrolarynx control based on statistical f0 prediction. In *Asia Pacific Signal and Information Processing Association (APSIPA)*, Siem Reap, Cambodia.

[Toda et al., 2007a] Toda, T., Black, A. W., and Tokuda, K. (2007a). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech & Language Processing*, 15(8):2222–2235.

[Toda et al., 2012] Toda, T., Nakagiri, M., and Shikano, K. (2012). Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *IEEE Transactions on Audio, Speech & Language Processing*, 20(9):2505–2517.

[Toda et al., 2007b] Toda, T., Ohtani, Y., and Shikano, K. (2007b). One-to-many and many-to-one voice conversion based on eigenvoices. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, volume 4, pages IV–1249–IV–1252.

[Tribolet et al., 1978] Tribolet, J., Noll, P., McDermott, B., and Crochiere, R. (1978). A study of complexity and quality of speech waveform coders. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1978)*, volume 3, pages 586–590.

[Uemi et al., 1994] Uemi, N., Ifukube, T., Takahashi, M., and Matsushima, J. (1994). Design of a new electrolarynx having a pitch control function. In *3rd IEEE Proceeding of International Workshop on Robot and Human Communication (RO-MAN)*, pages 198–203, Nagoya.

[Valbret et al., 1992] Valbret, H., Moulines, E., and Tubach, J. P. (1992). Voice transformation using PSOLA technique. *Speech Communication*, 11(2-3):175–187.

[van Rossum et al., 2002] van Rossum, M. A., de Krom, G., Nooteboom, S. G., and Quené, H. (2002). Pitch accent in alaryngeal speech. *Journal of Speech, Language & Hearing Research*, 45:1106–1118.

[Vary and Martin, 2006] Vary, P. and Martin, R. (2006). *Digital Speech Transmission: Enhancement, Coding and Error Concealment.* John Wiley & Sons.

[Veldhuis, 1998] Veldhuis, R. (1998). A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. *The Journal of the Acoustical Society of America*, 103(1):566–571.

[Vieregge et al., 1996] Vieregge, W., Pahn, J., and Schutte, H. (1996). *Patho-Symbolphonetik: auditive Deskription pathologischer Sprache.* Zeitschrift Für Dialektologie und Linguistik. Franz Steiner Verlag.

[Wan et al., 2012] Wan, C., Wu, L., Wu, H., Wang, S., and Wan, M. (2012). Assessment of a method for the automatic on/off control of an electrolarynx via lip deformation. *Journal of Voice*, 26(5):674.e21 – 674.e30.

[Weiss et al., 1979] Weiss, M. S., Yeni-Komshian, G. H., and Heinz, J. M. (1979). Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx. *The Journal of the Acoustical Society of America*, 65(5):1298–1308.

[Welch, 2015] Welch, M. (2015). http://www.limcosolutions.com/about `http://www.limcosolutions.com/About%20us/About%20us.htm`. Accessed: 2015-07-27.

[Wells, 1995] Wells, J. (1995). Computer-coding the IPA: A proposed extension of SAMPA. Unpublished notes. Department of Phonetics and Linguistics, University College London.

[Weselak and Graber, 2009] Weselak, W. and Graber, G. (2009). Elektroakustik.

[White, 1994] White, J. P. (1994). *Acoustic and Perceptual Characteristics of Electronic Artificial Larynges.* PhD thesis, Northwestern University, Evanston, Illinois.

[Wohlmayr and Pernkopf, 2010] Wohlmayr, M. Stark, M. and Pernkopf, F. (2010). A mixture maximization approach to multipitch tracking with factorial Hidden Markov Models. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, pages 5070 – 5073, Dallas, Texas, USA.

[Wu et al., 2013] Wu, L., Wan, C., Wang, S., and Wan, M. (2013). Improvement of electrolaryngeal speech quality using a supraglottal voice source with compensation of vocal tract characteristics. *IEEE Transactions on Biomedical Engineering*, 60(7):1965–1974.

[Yan et al., 2014] Yan, N., Ng, M., and Lee, T. (2014). Improving the sound quality of an electronic voice box. In *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, Siem Reap, Cambodia.

[Young et al., 2006] Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006). *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.

[YourTone, 2015] YourTone (2015). http://www.dencom.co.jp/product/yourtone/yt_eng.html. `http://www.dencom.co.jp/product/yourtone/yt_eng.html`. Accessed: 2015-07-16.

[Zollner and Zwicker, 1993] Zollner, M. and Zwicker, E. (1993). *Elektroakustik*. Springer.