Graz University of Technology

Doctoral Thesis

Speech Enhancement for Disordered and Substitution Voices

Martin Hagmüller

Signal Processing and Speech Communication Laboratory
Faculty of Electrical and Information Engineering
Graz University of Technology, Austria

Advisors:
Prof. Dr. Gernot Kubin, Graz University of Technology, Austria
Dr.habil. Jean Schoentgen (FNRS), Univ. Libre de Bruxelles, Belgium

Graz, September 2009

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 entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ....
(Unterschrift)

Englische Fassung:

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)

Abstract

This thesis presents methods to enhance the speech of patients with voice disorders or with substitution voices. The first method enhances speech of patients with laryngeal neoplasm. The enhancement enables a reduction of pitch and a strengthening of the harmonics of voiced segments as well as decreasing the perceived speaking effort. The need for reliable pitch mark determination on disordered and substitution voices led to the implementation of a state-space based algorithm. Its performance is comparable to a state-of-the art pitch detection algorithm but does not require post processing.

A subsequent part of the thesis deals with alaryngeal speech, with a focus on Electro-Larynx (EL) speech. After investigating an EL speech production model, which takes into account the common source of the speech signal and the directly radiated EL (DREL) sound, a solution to suppress the direct sound is based on the different temporal properties of the propagation paths. Time-invariant signal components, which can be attributed to the DREL sound are filtered out in the modulation frequency domain. Another issue with EL speech production has been addressed, namely the flat F0 contour. Based on the observation that prosodic information is conveyed in whispered speech, we have assumed that formants can be used as substitute intonation cues. We therefore derive an artificial F0 contour from the speech formants and impose it on the EL speech signal. The artificial intonation contour was preferred in a subjective listening test.

Kurzfassung

Diese Arbeit präsentiert Methoden zur Sprachverbesserung von Patienten, die an einer Stimmstörung leiden oder mit einer Ersatzstimme sprechen. Die erste Methode verbessert Stimmen von Patienten, die an einem Neoplasma laryngis leiden. Die Verbesserung besteht aus einer Verringerung der Stimmgrundfrequenz und der Stärkung der harmonischen Komponenten der stimmhaften Sprachsegmente. Ein Ergebnis ist die Verringerung der wahrgenommenen Sprechanstrengung. Aufgrund des Bedarf von einem zuverlässigen System zur Bestimmung der einzelnen Stimmzyklen für Stimmstörungen und Ersatzstimmen wurde ein Algorithmus, der im Zustandsraum arbeitet, implementiert. Die Ergebnisse sind vergleichbar mit Grundfrequenzbestimmungssystemen am Stand der Technik, jedoch ohne die Notwendigkeit einer Nachbearbeitung der Ergebnisse.

Ein weiterer Teil der Arbeit beschäftigt sich mit kehlkopfloser Sprache. Hauptaugenmerk wird auf Elektro-Larynx (EL) Sprache gelegt. Wir präsentieren ein Model der EL Spracherzeugung unter Berücksichtigung der gemeinsamen Quelle von Sprachsignal und dem direkt abgestrahltem EL (DREL) Geräusch. Wir schlagen eine Lösung zur Unterdrückung dieses DREL Geräusches vor, die auf den zeitlich unterschiedlichen Eigenschaften der Ausbreitungspfade basiert. Zeitinvariante Signalanteile, die dem DREL Geräusch zugeschrieben werden können, werden im Modulationsfrequenzbereich herausgefiltert. Ein weiteres Problem von EL Sprache ist die flache Intonationskontur. Aufgrund der Beobachtung, dass in Flüstersprache prosodische Information übermittelt werden kann, haben wir angenommen, dass Formanten als Ausgangssignal für die Berechnung einer künstlichen Intonationskontur verwendet werden können und diese dann dem EL Sprachsignal aufgeprägt wird. Diese künstliche Intonationskontur wurde in subjektiven Hörtests bevorzugt.

Acknowledgment

This thesis would not have been possible without the support of many people.

First and foremost, I want to thank my wife Daniela who has been at my side for all the time I was working on the thesis. She joined me in celebrating successes and encouraged me in times of frustration. I thank my children Sara Kristin and Kilian Lukas for allowing me go to work every morning.

I am indebted to my supervisor Prof. Gernot Kubin, whose expertise has guided me through the endeavor of creating this thesis. His bright mind in combination with a kind spirit makes him a very special person. What I admire most is his ability to always focus on the positive.

I am grateful to Prof. Gerhard Friedrich and his team at the Phoniatric Department at the ENT Hospital in Graz for supporting my work. Especially Jutta Chibidziura-Priesching, a speech therapist and Dr. Markus Gugatschka have been of crucial support in providing their expertise from a therapeutic and medical point of view and organizing recording sessions at the clinic. Nothing would have been possible without the patients, who offered their time and were willing to have their voice recorded.

The evaluation of the algorithms would not have been possible without the support of many people who took the time to listen to some rather strange speech sounds and even tried to express an opinion about it.

A very special thank goes to Dr. Jean Schoentgen, my second advisor, who has taken the time to come to Graz and spend a week discussing the thesis with me and – from a different point of view – helping to improve the text considerably.

I am also thankful to Servona for providing an electro-larynx, that made my research much easier.

I am thankful to all the members of the Signal Processing and Speech Communication Laboratory, who make it a great place to work.

Finally, I thank God for giving me a mind able to research.

Graz, September 2009 Martin Hagmüller

To Daniela, Sara and Kilian

Contents

List of Figures

List of Tables

Notation

$e(t)$	EL signal
$F1$	First Formant Frequency
f_0	Fundamental Frequency
$H_d(t)$	EL direct path transfer function
$h_d(t)$	EL direct path impulse response
$h_v(\tau, t)$	Time varying vocal tract impulse response
$h_{nv}(\tau, t)$	Time varying combined neck and vocal tract impulse response
$H_n(t)$	Neck transfer function
$h_n(t)$	Neck impulse response
$n(t)$	Environmental noise
$s(t)$	Speech signal
$s_d(t)$	Directly radiated sound of EL
$s_u(t)$	Noise-like speech signal
$s_v(t)$	Voiced speech signal
$u(t)$	Noise-like sound source
$x[n]$	Discrete-time signal

Acronyms

ANOVA	Analysis of variance . . . 69
ASR	Automatic speech recognition . . . 59
BAS	Bavarian archive for speech signals . . . 108
CCR	Comparison category rating . . . 94
CMOS	Comparison mean opinion score . . . 97
CI	Cochlear implant . . . 109
DREL	Directly radiated electro-larynx . . . 87
EGG	Electro-glottogram . . . 32
EL	Electro-Larynx . . . 107
ES	Esophageal . . . 78
FIR	Finite impulse response . . . 63
GRBAS	Grade, Roughness, Breathiness, Aesthenia, Strain
GMM	Gaussian mixture model . . . 12
HMM	Hidden Markov Model . . . 13
HNR	Harmonics-to-Noise Ratio . . . 25
IIR	Infinite impulse response . . . 63
ITU-T	International Telecommunications Union - Telecommunications standardization sector . . . 67
LF	Liljencrants-Fant . . . 9
LSF	Line Spectral frequency . . . 13
LP	Linear prediction . . . 86
LPC	Linear Prediction Coefficient . . . 89
LMR	Linear Multivariate Regression . . . 12
LTI	Linear time-invariant . . . 52

MOS Mean Opinion Score . . . 56
NFRF Neck frequency response function . . . 14
PSOLA Pitch synchronous overlap add . . . 89
RBH Roughness, Breathiness, Hoarseness . . . 20
SLT speech and language therapist . . . 19
SNR Signal-to-noise ratio . . . 62
SVD Singular value decomposition . . . 37
STFT Short-time Fourier-transform
TD-PSOLA Time-domain pitch-synchronous overlap-and-add . . . 101
TE Tracheo-esophageal . . . 89
VQ Vector quantization . . . 12
XML eXtensible markup language . . . 108

Chapter 1

Introduction

1.1 Motivation

Voice problems are a very common issue among humans. They can be broadly classified into functional disorders due to a misuse of the voice apparatus and organic disordersdue to an organic change of the voice organ. Whatever the reason, voice disorders can greatly influence the life of affected people. For example, a teacher is not able to teach, because his or her voice is not capable to meet the requirements of the job. According to a recent study [SKNBF+06], 69 % of all teachers suffer of some kind of vocal symptom over a lifetime, compared to 39 % of the non-teaching population.

Even more, some people, after suffering voice problems over a longer period of time, are confronted with the diagnosis of laryngeal cancer. While at an early stage there is a good chance of healing and being able to continue the previous live, sometimes the last chance is to remove the entire larynx. Vocal communication as usual is not possible anymore. The person has to learn to use a substitution voice, which sounds very different compared to a natural voice. The social stigma, which can go along with the medical situation, poses the danger to lead the person into social isolation. The estimated number of total laryngectomees is about 600.000 people worldwide [LN07] and 21000 laryngectomees in Germany with about 3000 additional laryngectomy operations performed every year [SH00].

Specially for alaryngeal speech, one direction of research and practice is to find ways to replace the larynx with something that can take over the task that the larynx fullfilled, namely the production of voice and the switch between trachea and esophagus. While the possibility to reach this goal is still far away, this thesis looks at this issue from a signal processing point of view. While there has been a lot of progress in natural speech signal processing, only little improvent of disordered voices can be attributed to signal processing methods. Most research has been done on assessment of voice disorders. Enhancement of disordered and substitution voices still is a challenging task, in the framework of which only little progress has been made over the last decades. This work aims to add another step toward providing

patients with tools that could possibly help to improve their quality of life.

1.2 Overview

The thesis starts with a short overview of voice disorders with focus on substitution voices and presents the state-of-the-art in alaryngeal voice enhancement. The remainder of the thesis is divided into two parts, laryngeal versus alaryngeal voice enhancement.

Part I: Laryngeal voice enhancement

Chapter 3 introduces a method to improve the voice of subjects with laryngeal neoplasm. For the evaluation the term of perceived speaking effort was introduced.
Related publication:

- Martin Hagmüller and Gernot Kubin. Voice enhancement of male speakers with laryngeal neoplasm. In *Proceedings of the International Conference on Spoken Language Processing*, pages 541–544, Jeju Island, South Korea, October 4–8 2004.

Chapter 4 presents the analysis and improvement of the Poincaré pitch marking method as proposed by [Kub97]. Several improvements to increase reliability were investigated, among which are the automatic level control of the time-domain signal, choosing the first pitch mark at a position, where trajectories are the most parallel, and clustering of candiate points to exclude points from a wrong trajectory. The method was compared to a state-of-the-art pitch detection algorithm.
Related publications:

- Martin Hagmüller and Gernot Kubin. Poincaré sections for pitch mark determination in dysphonic speech. In *Proceedings of 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, pages 281–284, Firenze, Italy, December 10–12 2003.

- Martin Hagmüller and Gernot Kubin. Poincaré sections for pitch mark determination. In *Proceedings of Nonlinear Signal Processing (NOLISP)*, pages 107–113, Barcelona, Spain, April 19-22 2005.

- Martin Hagmüller and Gernot Kubin. Poincaré pitch marks. *Speech Communication*, 48(12):1650–1665, December 2006.

Part II: Alaryngeal voice enhancement

Chapter 5 is dedicated to the proposal of a method to effectively reduce the direct Electro-Larynx (EL) sound.
Related publication:

- Martin Hagmüller and Gernot Kubin. Methode zur Trennung von Signalpfaden und Anwendung auf die Verbesserung von Sprache mit Elektro-Larynx. Austrian Patent application, February 4 2009

Chapter 6 deals with the use of speech formants to generate an artificial pitch contour for alaryngeal speech.

Related publication:

- Martin Hagmüller. Pitch contour from formants for alaryngeal speech. In *Proceedings of 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, pages 205–208, Firenze, Italy, December 13–15 2007.

Chapter 2

Disordered Voice Enhancement

A pathological voice can be a major handicap for social interaction. Problems can arise due to bad intelligibility especially in situations with high vocal demands, such as a classroom situation or other places with high background noise. Speakers can also suffer from a too short speaking time owing to a quickly fatigued voice.

Effects of voice disorders on speech coders used in telecommunication devices have been reported to be a reduction of speech intelligibility and naturalness. [JPP02a, JPP02b]. More and more telephone-based information services rely on Automatic speech recognition (ASR). Experiments using isolated-phone ASR trained with laryngeal speech, yielded a mean word accuracy of 28.7 $\pm$ 12.1 % for tracheo-esophageal speech compared to 57.6 $\pm$ 6.1 % for a healthy control group [SHN$^+$06]. Besides the application of ASR error rate as intelligibility measure, those results suggest that alaryngeal speakers are excluded from using ASR-based information services.

Laryngeal disorders and substitution voices differ from each other. We first briefly discuss some characteristics of laryngeal disorders and then cover the special features of alaryngeal voices.

2.1 Laryngeal Disorders

Laryngeal disorders can be functional, i.e., due to wrong usage of the voice production mechanism, or organic, i.e., due to changes in the anatomy of the voice production system, e.g., vocal nodules (see, e.g., [WSKE96] for an indepth medical coverage of vocal disorders).

In a clinical setting, there exists a widely used perceptual classification system of voice disorders (GRBAS) [Hir81], which assigns the disorder a general grade (G) and then specifies the extent of roughness (R) and breathiness (B) (additive noise due to excessive amount of airflow). Asthenia (A) and strain (S) may also be rated but results have shown to be inconsistent among raters [BWVdHC97]. In the German speaking countries a compact version of the GRBAS scale is common: Roughness,

Breathiness, Hoarseness (RBH) [NAW94], in which the term hoarseness is used to designate the overall degree of disorder.

Roughness is the perceptual cue of irregularity of the vocal fold oscillation. Perceived breathiness is due to an excessive amount of turbulent airflow in voiced speech, e.g., because of insufficient closure of the vocal folds. This is a simplification, but an indepth analysis would be beyond the scope of this thesis. Numerous attempts have been made to quantify perceptual impressions of disordered voices by means of signal analysis (see e.g. [Tit94, Sch06]).

2.2 Substitution Voices

In case of laryngeal cancer at an advanced stage, the only possibility to stop the further advance of the cancer and, thereby, saving a patient's life is to remove the entire larynx or parts of it, possibly including the vocal folds. This results in the loss of the usual voice production mechanism, based on a vibration of the vocal folds. In addition, in the case of total laryngectomy the trachea opening is surgically moved to the neck. The new opening is called the tracheostoma, so the airflow from the lungs is not passing through the vocal tract anymore. The patients have then to rely on a substitution voice production mechanism [BHIB03]. There are three major mechanisms available:

Esophageal voice (ES): Air is swallowed into the esophagus and is than burped back up again (Fig. 2.1, middle image). Substitute folds – also called neoglottis – are then vibrating to produce the source of the speech sound. The substitution voice production can be enhanced by surgical modification of the neoglottis [HKS+99].

Tracheo-esophageal voice (TE): A valve between the trachea and the esophoagus is surgically inserted. The valve allows to let the air from the lungs flow through the vocal tract and excite the neoglottis, as with the esophageal voice, but with a more continous, extended air stream. (Fig. 2.1, right image).

Electro-Larynx (EL): A hand held device, which produces a buzz-like sound is held against the neck. The vocal tract is excited by this sound and is used to shape the speech sound. For people using the first two substitution voice production methods, the Electro-Larynx can be used as a fallback method, when the above approaches do not work.

If the larynx is still working as a switch between trachea and esophagus, but parts are removed, so that the vocal folds cannot be used for voice production anymore, the patient has to rely on a substitution voice. In some cases a surgical remodelling of the larynx is necessary to enable voice production.

All of those voices have major shortcomings, which are discussed in the following section 2.3.

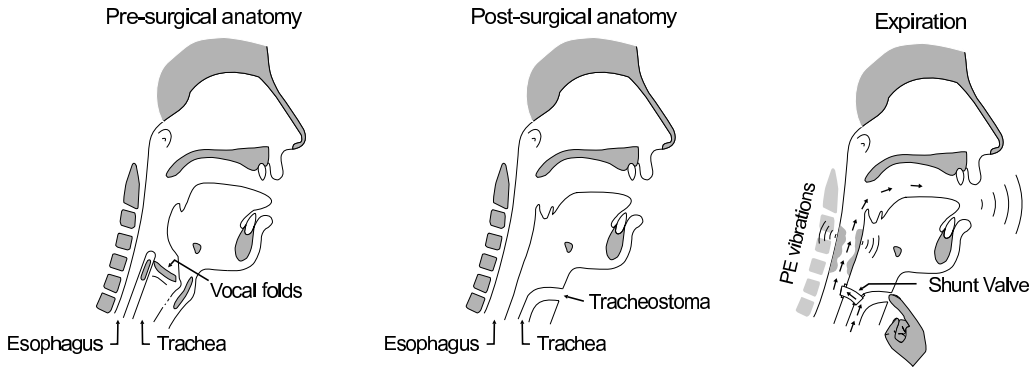


Figure 2.1: *Schematics of pre-surgical anatomy (left) , post surgical anatomy (middle) and tracheo-esophageal voice production (right). (from [Loh03]).*

2.3 Acoustical Characteristics of Alaryngeal Speech

A number of studies have been performed to evaluate the characteristics of substitution voices. While some key issues of the acoustical characteristics of substitution voices have been identified, research is still going on to find objective methods to classify substitution voices. Since substitution voices have very different properties compared to laryngeal speech, most voices cannot be classified by classification systems designed for laryngeal voices. Several approaches to classify substitution voices have been proposed, so far, e.g. [Tit94, vABKvBPH06, MMVdB$^+$06].

Voicing Source

The change of the voicing mechanism has the largest impact on the quality of alaryngeal speech. This has effects on different acoustic features of the voice. First the fundamental frequency of a substitution voice is usually at about 100 Hz or lower and the HNR is significantly lower than in normal speech. Significant differences exist between EL speech on the one hand and ES and TE speech on the other. Therefore, we look at the voicing characteristics separately.

Electro-Larynx (EL): The electro-larynx voice sounds very mechanical, due to the monotonuos excitation signal, which is strictly periodic at a constant pitch. Another serious problem of EL speech is the presence of a constant background noise due to the direct sound radiated from the device to the listener [EWCM$^+$98].

It has been shown previously that providing EL Speech with a natural pitch contour in combination with other enhancements is perceived as beneficial [MH05]. Earlier studies have shown that the flattening of the F_0 contour decreases intelligibility of sentences spoken by healthy subjects [LW99, LB03]. For those studies, sentences with normal F_0 countour were recorded and resynthesized with a flat F_0 at the mean F_0 of the sentence. Intelligibility tests were carried out in different listening environments with the natural pitch contour and the flat pitch.

Another problem is the unnatural source spectrum, which makes the speech sound less human. One deviation from healthy human speech is the lack of low-frequency components below 500 Hz [QW91], due to the mechanical properties of the EL device. In [MKH03] a model has been presented which helps to calculate the transfer function of the neck, which transmits the sound from the EL to the vocal tract.

Esophageal and Tracheo-esophageal: The excitation signal produced by the substitute folds is often irregular, which results in a very rough voice. Therefore, F_0 cannot be reliably extracted by means of most digital signal processing methods. Further, the oscillation of the substitute folds cannot be controlled very well.

Both TE and ES speech exibit very low average fundamental frequency, which is a problem especially for female speakers [KKP+06]. A study of the gender perception on TE speech [SS02] has shown that, while female TE speaker run the risk of being identified as male or gender neutral, most of the time the gender of the speaker is identified correctly. A study which compared female laryngeal, Esophageal (ES) and Tracheo-esophageal (TE) speech showed that concerning fundamental frequency a significant difference exists between laryngeal and alaryngeal speech, but not between ES and TE speech [BLG01]. The mean fundamental frequency for female alaryngeal speakers was approximately 110 Hz for ES speakers and 130 Hz for TE speakers. Similar results have been obtained for speakers of variable proficiency [MTM00].

Substitute folds do not always allow to produce a controlled pseudo-periodic voicing. High irregularities may characterize these vibrations, which result in a very rough voice. Therefore, standard methods to exctract pitch or evaluate the quality of voice often fail [Tit94].

Formants

While the articulatory organs are usually not effected, the configuration of the vocal tract is changed due to the removal of the larynx. The most important change is the shortening of the vocal tract, which changes the position of the formants. Several studies have been carried out to investigate the effects of the altered anatomy of alaryngeal speakers. Generally, higher formant frequencies have been reported [WHS80, vABRKvBH97, Mel03, KPK+07].

Energy supply

While TE and EL voices have a steady energy supply, this is not the case for the ES voice. Furthermore, it is not or only in a very limited range possible to modulate the energy supply of the ES voice. While excellent TE speakers are able to modulate the intensity of their voice satisfactorily, less proficient TE speakers have difficulties. The EL only offers a fixed energy level, which for some models can be switched to a higher level by pressing a button.

Speaking rate/ rhythm

A study with good female ES and TE speech has shown that while the speaking rate is comparable for TE and ES speech, the latter has a significantly increased pause length [BLG01]. Speaking rate is lower and pauses are longer compared to laryngeal speech. The esophageal voice also suffers from sentence prosody problems, because the amount of air that can be swallowed limits the duration of speech phrases considerably. Well trained EL speakers do not have any limitation concerning speaking rate and rhythm.

2.4 State-of-the-Art of Disordered Voice Enhancement Approaches

Clinical methods for disordered voice restoration include voice therapy, surgical removeal of excess tissue on the vocal folds, vocal fold medialization, etc. In case of total laryngectomy, the most popular surgical procedure is the use of the tracheo-esophageal shunt, which enables the patient to use pulmonary air for substitution voice production (see Section 2.2). An overview of the state-of-the-art of voice rehabilitation after laryngectomy can be found in [BHIB03], overviews of EL speech enhancement can be found in [CM00], and more recently in [LN07].

Possible applications for voice enhancement based on digital signal processing are, e.g., a portable electronic device which helps the patient to cope with acoustically difficult situations in everyday life. For example, for voice telephony, there is almost no other possibility for augmentative communication. It could also support speech therapy, to ease voice use until the patient has learned new patterns of voice production. Another possible application could be motivating for the patient by giving a preview of what the voice could sound like after successful speech therapy or surgery.

So far, only few attempts have been made to enhance laryngeal dysphonic speech. Approaches include reducing breathiness using singular value decomposition [MDB01] or state-space approaches to noise reduction [MM02a]. Most of the work concerning disordered voice enhancement has focused on alaryngeal speech.

2.4.1 Voice Replacement

In section 2.3, we have explored the shortcomings of alaryngeal voice. Because, it is one of the major problems of alayngeal speech, several approaches have been presented to enhance the disordered voicing of neoglottis-excited speech.

One possible method is to replace human voicing altogether by some artificial excitation signal. Source models exist that describe the glottal source signals by a set of parameters. The most popular model, which has been used before to simulate the glottal excitation signal is the Liljencrants-Fant (LF)model [QWBH95, BQ97, AJ06,

PY06]. The glottal flow is fixed by 4 independent parameters, which determine voice timbre [Fan97].

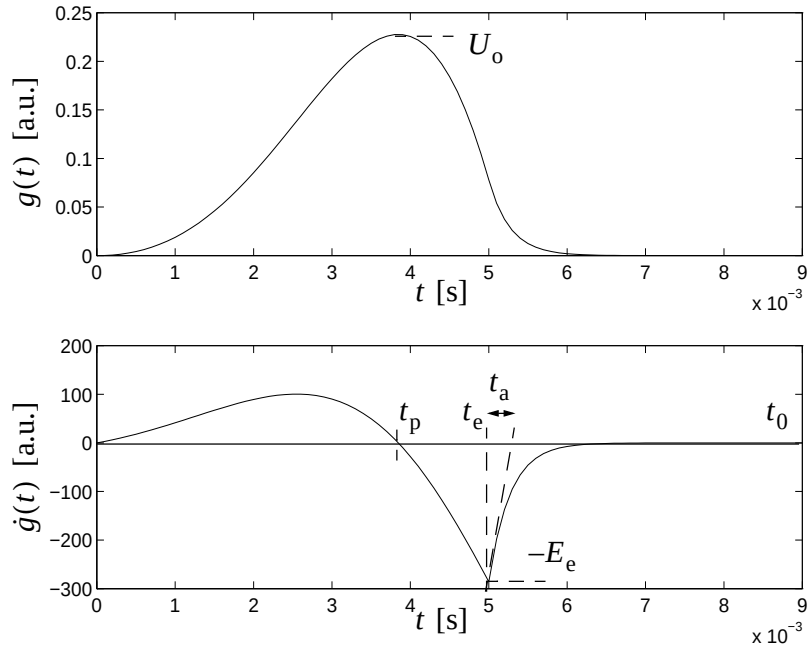


Figure 2.2: *Liljencrants-Fant model.*

Using a model that defines one glottal cycle only has a major disadvantage as pointed out in [HHKM99]:

> The waveform is defined over a single cycle and repeated, and as a consequence, all harmonics are in lock-step with the fundamental. Other glottal models with more sophisticated parameterization of a single cycle suffer from the same problem, even if noise is added or the 'arrival times' of the impulses are dithered.

They, therefore, propose to use a sample of a healthy human excitation signal. Sample based synthesis is widely used in musical instrument synthesizers and – at the current state-of-the-art – offers the best available sound quality. This source signal sample can than be used to excite an Linear Prediction Coefficient (LPC) based filter [MH99]. The signal has to be of a resonable length, i.e., many periods, to avoid a perceived strict periodicity.

Intonation Contour

A problem for alaryngeal voicing substitution is that there is either no pitch contour available – in case of EL – or it is not possible to get a reliable measurement of the F_0 contour – for TE and ES speech. A possibility is to create an artificial pitch contour from the speech energy envelope [LB06]. The energy envelope is one possibility

to convey accentuation if no F_0 is available [vRdKNQ02]. The energy envelope has also been used to provide whispered speech with an F_0 contour [MC02]. The energy contour has been scaled and offset so that the transformed contour matches the average pitch and the dynamic range of the speaker. This does not produce a linguistically correct intonation, but it does give a useful pitch, which can be influenced by the speaker by modulating whisper intensity.

The generation of a fundamental frequency contour has been introduced to EL devices earlier. Several approaches have been proposed so far. They either require manual interaction to change the pitch or only provide some predefined pitch contour. There are some commercially available approaches such as using a switch from the standard pitch level to a different level to mark accentuation. A more flexible approach is to use a pressure sensitive button, which allows a continuous pitch contour to be produced [Gri98]. Those two approaches require a manual control of the pitch contour, which is hardly used by the patients in practice. An automatic approach to generating a more natural pitch contour is applying a declination curve with predefined time constants to the constant pitch. The pitch level is reset every time the EL sound button is released [tHCC90, p.171f]. Another method is to use the expiration air pressure as a means to control the pitch contour. It uses a pressure sensor that is put on the stoma, so pitch can be controlled by lung pressure [UITM94].

Those options are commercially available, but none of them offers a satisfactory solution to the problem of unnatural pitch contour for EL speech. Therefore the standard EL speaker uses an EL device with constant pitch.

2.4.2 Voice Conversion

We have learned that the formant frequencies of alaryngeal speech are changed due to the changed anatomy of laryngectomized people. Because of the removal of the larynx the vocal tract is shortened and the pharynx has a different shape without the larynx. To improve voice intelligibility voice conversion techniques have been applied to move the formant frequency closer to modal speech formants. Voice conversion is often used for corpus-based speech synthesis to avoid the long recording times of additional speakers [SCM98]. The timbre of a speaker from an existing speech corpus is changed by modifying the formant frequencies to match another speech timbre. There are essentially two approaches.

One is to use a rule-based approach, which transforms the parameters. The other method uses a statistical learning approach, where a model of the mapping between the source and the target speaker characteristics is fitted.

Rule-based voice conversion

Some of the differences between alaryngeal speech and normal speech can be modeled by a few modifications of the formant properties. In the literature different rules are used. [AJ06] proposes to make ES speech more intelligible by expanding the

formant bandwidths. As mentioned above, alaryngeal speakers shift their formants to higher frequencies due to the shortened vocal tract length. Therefore, Loscos *et al.* [LB06] propose to move the formants down by a frequency dependent mapping function:

$$|X_s(f)| = |X_a(f^\alpha)|, \tag{2.1}$$

where $|X_s(f)|$ is the target spectral envelope and $|X_a(f)|$ is the spectrum envelope of the analysed signal. In this work, α has been chosen as 1.02. In addition, Pozo *et al.* [PY06] have reported a spectral tilt in TE speech, which favors the high frequenzy band. The auhtors have used a 6bB/octave roll-off filter (see Fig. 2.3) to de-emphasize this high frequency band.

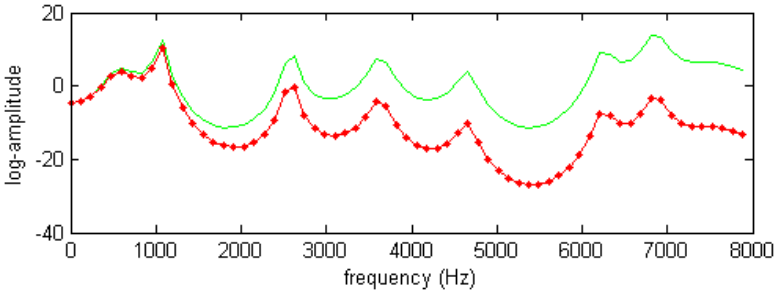


Figure 2.3: *De-emphasis of the high frequencies for alaryngeal speakers (from [PY06]).*

Statistical voice conversion

Statistical voice conversion consists of a learning stage, during which a model of the mapping between the source and target speakers is trained. This model is then used for the mapping of spectral characteristics of one speaker to the other (see Fig. 2.4). One of the earliest proposed method was mapping of Vector quantization (VQ) codebooks [ANSK88]. Later approaches have been based on Linear Multivariate Regression (LMR) [VMT92] and Gaussian mixture models (GMMs)) [SCM98], among others.

Voice Conversion has been also applied to the enhancement of alaryngeal speech. In [BQ97] a VQ and an LMR-based voice conversion system were used. The VQ method was modified by a chirp-z transform and cepstral weighting to decrease formant bandwith and the LMR method was modfied by overlapped subset training to reduce spectral discontinuities. The alaryngeal voicing source was replaced by an LF model. The two methods were applied to alaryngeal speech and evaluated by perceptual tests. Results showed that the listeners preferred the modified speech over the original alaryngeal speech, while no clear preference between VQ and LMR was shown.

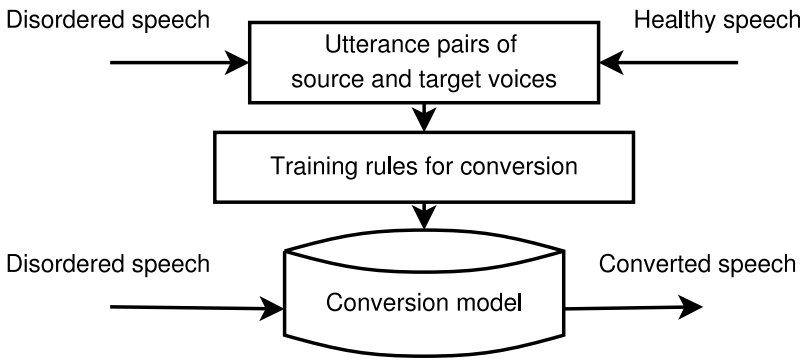


Figure 2.4: *Voice conversion.*

[ATNMPM06] used a VQ-mapping to enhance EL speech and [NTSS06] used GMM mapping to transform very low-volume EL speech to whisper, based on a non-audible-murmur enhancement approach.

2.4.3 Spectral Noise Reduction

Alaryngeal speech suffers not only from a distorted spectrum, but also from spectral frame-to-frame variability due to the instable voicing. Replacing the alaryngeal voice source may remove instabilities but inverse-filtering related distortions cannot be fully removed.

Spectral cue smoothing

Some attempts have been made to smooth the formant trajectories of alaryngeal speech. Matsui *et al.* [MH99] have proposed a LPC based source-filter approach, where a 3pt median smoothing of both formant frequencies and bandwidths is performed. A similar approach has been reported by Pozo *et al.* [PY06], in the framework of which a 10pt median filter is applied on the Line Spectral frequencys (LSFs), which have been derived via LPC analysis (see Fig. 2.5).

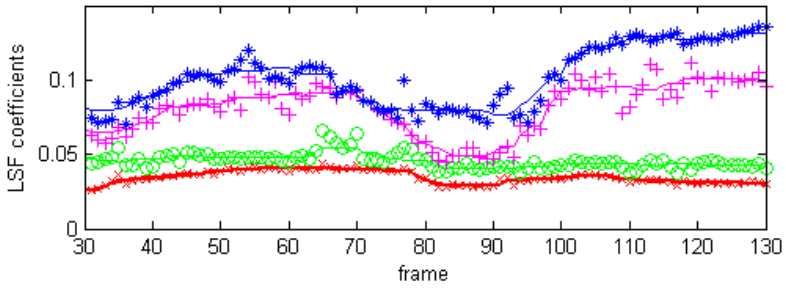


Figure 2.5: *Smoothing of LSF parameters (from [PY06]).*

Another problem is the injection noise of ES speech, which is heard prior to a speech utterance [JGN97]. To reduce this noise, a Hidden Markov Model (HMM)

speech recognition system has been trained including the injection sound as a unit to be recognized, which when recognized is suppressed. The authors also presented another method, based on a morphological filter. Using the mean and the frame-to-frame derivatives of the spectrum of the signal, the injection noise can be identified. They report promising error rates for both the HMM and the morphological filter based method.

Suppression of directly radiated electro-larynx sound

Attempts to suppress the Directly radiated electro-larynx (DREL) sound exist. One approach was to use adaptive filters to remove the DREL sound, but due to the shared source of EL speech and DREL sound, a heuristic had to be used, which turns off the adaptation process during voiced speech [EWCM+98, NWWL03].

Other approaches include several variants of spectral substraction methods, which are widely used for speech enhancement [CSMG97, PBBL02, LZWW06b]. Spectral substraction suffers from the problem that the direct noise is synchronized with the tract excitation and additionally, that environmental background noise and the directly radiated EL noise have completely different properties.

The high correlation of the DREL sound with EL speech, makes that conventional noise supression methods are unsuited for processing DREL sound. Heuristics have been used to take care of this problem.

2.4.4 E-Larynx Design Improvement

While the above discussion focusses on enhancing the EL signal by means of digital signal processing, the following section presents work which tries to improve the EL signal generation itself. The design of the EL has not been changed fundamentally since 1959 [BHD59].

In Houston *et al.* [HHKM99], a new design of the EL with a linear electro-acoustic transducer is proposed. Conventional ELs are nonlinear transducers similar to an old fashioned door bell with a mallet. They have has very limited possibilities of changing the sound of the device. The spectral characteristics of the signal is determined by the mechanical properties of the device. The novel design of the linear transducer is very similar to a loudspeaker, with a moving coil. So in principle an arbitrary sound can be generated (see Fig. 2.6)

To reduce the directly radiated sound leaking from the EL device they also considered the impedance of the neck to which the EL has to be matched with, so that most of the energy is transmitted through the neck into the vocal tract. They measured the mechanical load of the neck with an electro-dynamic shaker driven with white noise. The results were used for the design of a linear transducer prototype.

Norton *et al.* [NB93] and Meltzner *et al.* [MKH03] have worked on improving the EL by measuring the Neck frequency response function (NFRF). They have hoped to improve EL speech by incorporating the NFRF in the EL spectrum design. They have used a Brüel & Kjær mini-shaker to exite the neck with white gaussian noise.

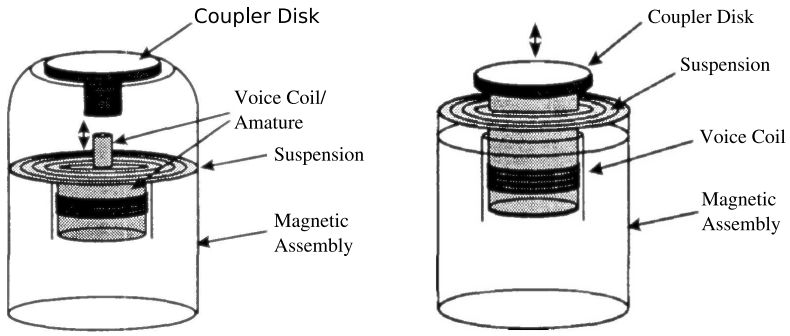


Figure 2.6: *Linear moving coil EL vs. conventional nonlinear EL (from [HHKM99]).*

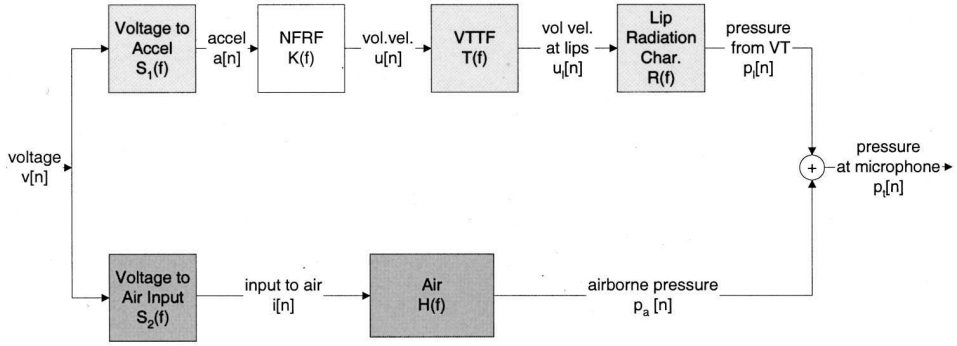


Figure 2.7: *EL noise model with the directly radiated path and the vocal tract path. NFRF . . . Neck frequency response function. VTTF: Vocal tract transfer function (from [MKH03]).*

They have designed an excitation signal that takes into account the NFRF. Figure 2.7 shows the model used in [MKH03]. While Norton *et al.* used healthy test subjects who held their breath, Meltzner used both alaryngeal and laryngeal speakers. In addition different places of excitation at the neck were tested. [NB93] reported an improved speech sound when using the proposed excitation signal. They also put a 1“ thick foam shielding around the EL and reported a suppression of the directly radiated noise by 20 dB.

2.4.5 Enhanced Voice Production System with E-Larynx

A few authors have tried not only to redesign the EL , but to fundamentally change either the control of the EL or the sound production.

E-Larynx with Non-Audible-Murmur Microphone

This approach [NTSS06] uses a low volume sound emitting device (EL), which is placed at the neck. A non-audible-murmur microphone behind the ears picks up the structure-borne sound modulated by the vocal tract. Due to the different transfer

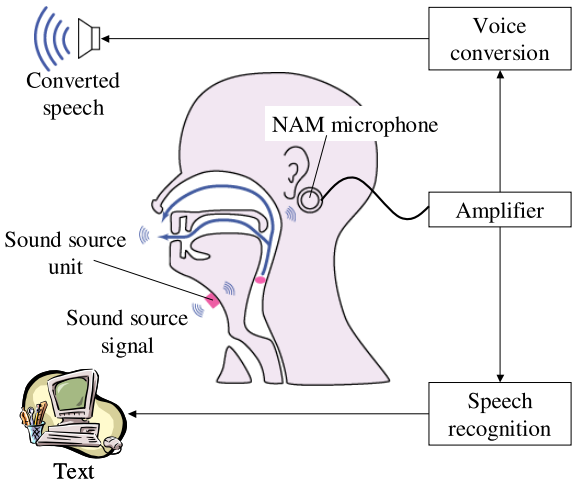


Figure 2.8: *ELarynx with Non-audible-murmur microphone (from [NTSS06]).*

function from the neck to behind the ears in comparison to in front of the mouth, a voice conversion step is necessary to convert the signal to intelligible speech. The authors decided to convert the picked-up sound to whispered speech (Fig. 2.8). The advantage is that due to the small excitation volume the directly radiated sound is small.

Electromyographic E-Larynx Control

The following method, which is clearly not only a signal processing approach, draws heavily on biomedical research. One major problem is the control of the conventional EL, which enables only very limited control of prosody and voicing, which is usually done by hand. By using the electromyographic signals of the neck strap muscle picked up by surface electrodes plus some signal processing it is possible to control the on/off signal and the values of the fundamental frequency of the EL (See Fig. 2.9, [GHK+04]). Experiments with three larygectomized and four healthy subjects showed that after a few hours of training they were able to initiate, sustain and terminate the EL signal. A majority was also able to produce an intonation contour [GHSH07].

2.5 Discussion

While the discussed methods document some research to enhance the communication abilities of people with alaryngeal voices, the only widely used technical aid is still the electro-larynx. Furthermore, the EL design has not changed fundamentally since the 1960s.

For alaryngeal speakers, only little technical progress has been made to provide them with a voice that is close to normal. To make a human voice sound natural many aspects have to be considered. Important features are the correct temporal

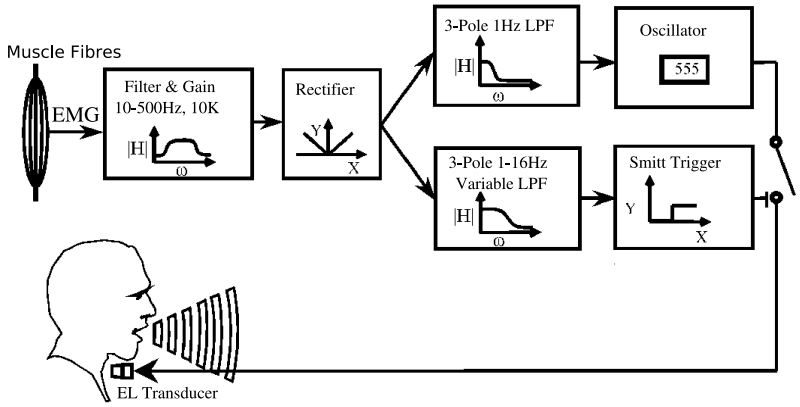


Figure 2.9: *Electromyographic E-Larynx Control (from [GHK+04]).*

evolution of the speech spectrum, the speech rhythm, and the fundamental frequency contour. Alaryngeal speech lacks naturalness owing to a lack of most of the above mentioned features. The difficulty of the task for signal processing approaches lies in the fact that only the acoustic speech output is accessed and not the linguistic content. So only small progress has been made, such as the reduction of some artefact noises that occur with substitution voices, and minor improvements of the voice quality and intelligibillity.

Chapter 3

Voice Enhancement for Laryngeal Neoplasm*

3.1 Introduction

Within the framework of this small study, we investigate a type of pathology, i.e., laryngeal neoplasm. Laryngeal Neoplasm is excess tissue in the larynx, due to cancer. This tissue has to be surgically removed. Such a voice has two major audible characteristics. First, the voice is too high (mean pitch $>$ 200 Hz for male speakers), and second, it is breathy. The overall impression of the speech of speakers with this kind of pathology is a voice under strain. We want to appy signal processing methods to improve the voice quality. Reducing breathiness and pitch may make the voice sound more natural and less strained. We are aware of the limited practical applications, since patients with laryngeal neoplasm have to undergo surgery as soon as possible, so one could call this a purely academic exercise. Still, the insights that might be gained by the attempt to enhance such voices make it worth the effort.

3.2 Speech Data

The speech tokens for this study were recordings made at the phoniatric department at the Graz Ear, Nose & Throat University Clinic. The recordings have been part of a dissertation on the evaluation of disordered voices. The recordings were performed in a quiet room and sampled at 44.1kHz with a quantization of 16bit. A detailed description of the database and recording procedure can be found in [Wre98].

We selected four male subjects with a diagnosis of laryngeal neoplasm. The average age was 57 years. Two speech and language therapists (SLTs) have agreed

*This chapter is an extended version of a previously published paper:
Martin Hagmüller and Gernot Kubin. Voice enhancement of male speakers with laryngeal neoplasm. In *Proceedings of the International Conference on Spoken Language Processing*, pages 541–544, Jeju Island, South Korea, October 4–8 2004.

Table 3.1: *Unprocessed patient data: evaluation using RBH rating (0 no disorder ... 3 severe disorder).*

ID	age	R	B	H
A	57	0	2	2
B	52	1	2	2
C	49	1	3	3
D	71	1	2	2

on the perceptual rating of the four subjects before processing the signals (table 3.1). The rating is based on the Roughness, Breathiness, Hoarseness (RBH) scale [NAW94].

Very often voice quality is evaluated using sustained vowels. Since this is not a natural way of speaking and, therefore, not the best choice to judge voice quality, a variety of German test phrases was used as well. The utterances were:

- sustained vowels *('i e a o u')*
- CV repetitions (*'dada nini soso lolo'*)
- isolated words (*'Maat Wanne Miete Minne Rose Wolle'*)
- numbers (*'eins zwei drei vier fünf sechs sieben acht neun zehn'*)
- days of the week (*'Montag Dienstag Mittwoch Donnerstag Freitag Samstag Sonntag'*)
- short sentences (*'Nie und nimmer nimmt er Vernunft an', 'Eine Maus saust aus dem Haus'*)

3.3 Speech Enhancement

Figure 3.1 shows an overview of the steps of the algorithm. First the pitch marks are determined, then a periodicity enhancement algorithm is applied, the pitch is lowered and as post-processing the periodicity enhancement algorithm is applied again. Lack of perfect periodicity is a common feature of dysphonic speech. The

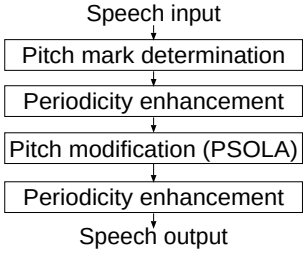


Figure 3.1: *Schematic diagram of voice enhancement.*

periodicity enhancement step, which performs cycle regularisation, is computed to reduce breathiness. Since breathiness is usually perceived as noise, one would think conventional noise reduction algorithms would be suitable to enhance breathy voice disorders. Breathiness in a hoarse voice is, however, usually correlated with the speech signal. The required condition of independence of signal and noise is therefore not satisfied.

Pitch modification is carried out to lower the voice of the subjects, because male neoplasm voices are too high-pitched. After the Time-domain pitch-synchronous overlap-and-add (TD-PSOLA) pitch modification, the periodicity enhancement is performed a second time to reduce processing artefacts owing to PSOLA.

3.3.1 Pitch Mark Determination

Both pitch modification (TD-PSOLA) and periodicity enhancement operate pitch-synchronously. So, the first step is to detect every pitch cycle correctly. For pitch modification using TD-PSOLA, the best performance is obtained when the pitch marks are set at the energy maximum of each cycle. This step is carried out using the 'Praat' (a software package for speech processing [BW07]) pitch mark determination function. For speech segments considered 'unvoiced' by 'Praat', pitch marks are set at a fixed period. To avoid artificial periodicity in unvoiced segments, a jitter of 10% has been superimposed on the calculated pitch marks in unvoiced speech fragments. Problems observed with pitch mark detection algorithms such as 'Praat', when applied to disordered voices, have been a motivation for exploring alternative pitch mark determination methods. The results of this exploration are presented in chapter 4. While this improved performance in some respect, for the current application in this chapter the 'Praat' algorithm has been the method of choice. For further discussion see section 4.6.

3.3.2 Pitch-Synchronous Overlap-Add

The TD-PSOLA algorithm [ML95b] separates the signal into overlapping windowed frames of pairs of pitch cycles and then rearranges the cycles according to a desired new pitch contour (Fig. 3.2). The newly arranged signal pieces are then added to obtain the modified signal.

The new intonation contour is calculated by multiplying the analysis pitch contour with a factor α:

$$\mathbf{P}_s(t) = \mathbf{P}_a(t)\alpha, \tag{3.1}$$

where the vectors $\mathbf{P}_a$ and $\mathbf{P}_s$ contain the durations of the analysis (a) and synthesis (s) periods, i.e. $P_a(i) = t_a[i+1]-t_a[i]$, where $t_a(i)$ are the time markers of the analysis cycles. α is the pitch modification factor. The time markers of the synthesis pitch cycles are determined recursively: $t_s[j+1] = t_s[j] + \mathbf{P}_a(i)\alpha$. Resynthesis is carried out by choosing the pitch cycle with a time marker close to the target time marker. In case of lowering the fundamental frequency some pitch cycles drop out, in case of

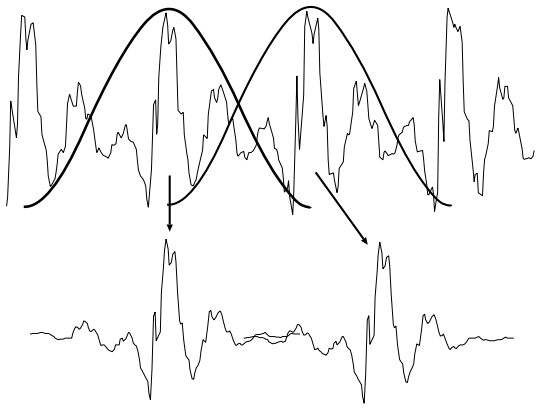


Figure 3.2: *Principle of pitch modification using TD-PSOLA.*

raising the pitch some have to be used twice. Since this algorithm operates at the speech-cycle level, the correct identification of the energy maximum of each cycle is crucial for the performance of the algorithm.

3.3.3 Periodicity Enhancement

Kleijn [Kle02] proposed a method for periodicity enhancement for voiced speech at the output of a digital speech coder. The enhancement algorithm should preserve the signal energy and keep low the changes between the original and the modified signal. This should avoid artefacts, which occur, e.g., when noise-like signal parts are made more periodic by the enhancer or sudden energy changes appear in the signal. The periodicity enhancement is performed by averaging neighboring frames that are shifted by multiples of the current period. What this algorithm does in the time domain is similar to what in [HKM01] is done in the phase space, but with less computational effort. While the algorithm was designed as a post filter to enhance speech coding algorithms, the property of enhancing periodicity is also interesting to reduce breathiness in disordered voices. In the case of breathiness small random components are present in the speech signal.

Let $\mathbf{s}[j]$ be a speech frame with length K, at time instant j, and $\tilde{\mathbf{s}}[j]$ the corresponding enhanced speech frame that replaces signal frame $\mathbf{s}[j]$. We introduce a signal-norm (energy) preservation condition to avoid frame-to-frame energy jumps:

$$\|\tilde{\mathbf{s}}[j]\| = \|\mathbf{s}[j]\|, \tag{3.2}$$

where $\|\cdot\|$ denotes the Euclidean norm.

The second constraint is to keep the difference between $\tilde{\mathbf{s}}[j]$ and $\mathbf{s}[j]$ small (modification constraint):

$$\|\mathbf{s}[j] - \tilde{\mathbf{s}}[j]\|^2 \leq \beta \|\mathbf{s}[j]\|^2, \tag{3.3}$$

with $\beta \in [0, 1]$.

The degree of periodicity of the enhanced signal can be measured by

$$\eta_{\mathcal{J}} = \sum_{j \in \mathcal{J}} \sum_{m \in \mathcal{I} - \{0\}} \alpha_m \langle \tilde{\mathbf{s}}[j], \tilde{\mathbf{s}}[j, m] \rangle, \tag{3.4}$$

where α_m is a window function, $\mathcal{I}$ is a set of integers that describes the support of this window (e.g. $\mathcal{I} = -3, -2, \ldots, 3$) and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product and $\mathcal{J}$ is a set of frame indices, which are centered around consecutive pitch marks. The aim is to maximize the periodicity criterion $\eta_{\mathcal{J}}$, which can be achieved by maximizing every

$$\eta[j] = \sum_{m \in \mathcal{I} - \{0\}} \alpha_m \langle \tilde{\mathbf{s}}[j], \mathbf{s}[j, m] \rangle, \tag{3.5}$$

The local periodicity criterion (3.5) under the constraints (3.2) and (3.3) with the Langrange multipiers λ_1 and λ_2 can be written as

$$\eta[j]' = \sum_{m \in \mathcal{I} - \{0\}} \alpha_m \langle \tilde{\mathbf{s}}[j], \mathbf{s}[j, m] \rangle + \lambda_1 \|\tilde{\mathbf{s}}[j]\|^2 + \lambda_2 \|\mathbf{s}[j] - \tilde{\mathbf{s}}[j]\|^2. \tag{3.6}$$

After some algebra, we can write the result of the optimization as:

$$\tilde{\mathbf{s}}[j] = A\mathbf{y}[j] + (B + 1)\mathbf{s}[j] \tag{3.7}$$

where $\mathbf{y}[j]$ is given as

$$\mathbf{y}[j] = \sum_{m \in \mathcal{I} - \{0\}} \alpha_m \mathbf{s}[j, m]. \tag{3.8}$$

and

$$A = \sqrt{\frac{(\beta - \frac{\beta^2}{4} \|\mathbf{s}[j]\|^2}{\|\mathbf{y}[j]\|^2 - \frac{\langle \mathbf{y}[j], \mathbf{s}[j] \rangle}{\|\mathbf{s}[j]\|^2}}}, \tag{3.9}$$

and

$$B = -\frac{\beta}{2} - A \frac{\langle \mathbf{y}[j], \mathbf{s}[j] \rangle}{\|\mathbf{s}[j]\|^2} \tag{3.10}$$

In case the modification constraint is not active, i.e., $\lambda_2 = 0$ in Equ. 3.6, the modification results in

$$\tilde{\mathbf{s}}[j] = C\mathbf{y}[j], \tag{3.11}$$

where

$$C = \sqrt{\frac{\|\mathbf{s}[j]\|^2}{\|\mathbf{y}[j]\|^2}}. \tag{3.12}$$

Kleijn [Kle02] summarizes the algorithm as:

1. compute $y[j]$, A, B, and C,

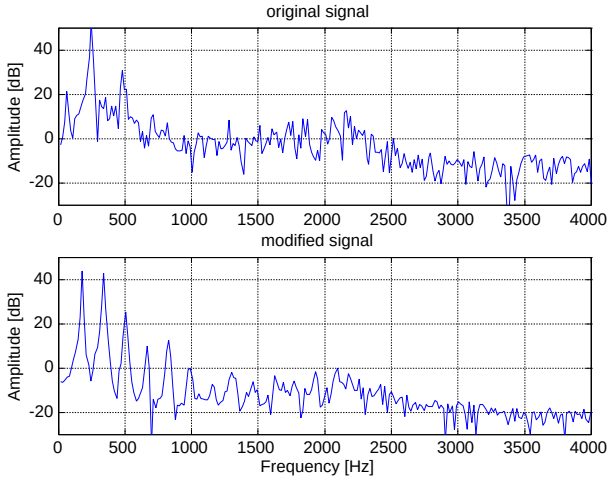


Figure 3.3: *Upper plot: spectrum of a disordered sustained vowel. lower plot: spectrum of modified signal.*

2. if $\|\mathbf{s}[j] - C\mathbf{y}[j]\|^2 \leq \beta\|\mathbf{s}[j]\|^2$
   (if modification constraint is fulfilled for scaled version of $y[j]$)

   then $\tilde{\mathbf{s}}[j] = C\mathbf{y}[j]$
   (then use scaled version of $y[j]$ as output, calculation of A and B is not necessary)

   else $\tilde{\mathbf{s}}[j] = A\mathbf{y}[j] + (B+1)\mathbf{s}[j]$
   (else incorporate modification constraint for calculating the output)

3.3.4 Results

The spectrum and the waveform of the original disordered vowel (Speaker A) and the modified signal are compared in Fig. 3.3 and Fig. 3.4, respectively. One can see the influence of the period modification in the distance between the signal harmonics and the increased length of the signal period by a factor of approximately 1.4. The periodicity enhancement can be seen in the increase of the spectral harmonics and the reduction of the additive noise in the signal waveform. Fig. 3.5 shows the original and the lowered intonation contour. It has to be noted, that the main impact on the sound quality comes from the pitch modification. The final periodicity enhancement step is used to reduce processing artefacts caused by the pitch modification.

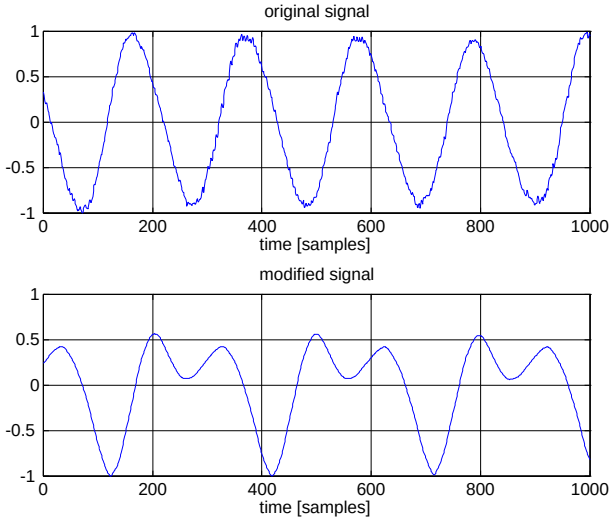


Figure 3.4: *Upper plot: waveform of original signal. lower plot: waveform of modified signal (the different waveform shape is due to the PSOLA period modification).*

3.4 Evaluation

3.4.1 Objective Evaluation

Mean pitch, jitter, shimmer and Harmonics-to-Noise Ratio (HNR) of the original and processed speech tokens were analysed using 'Praat'. The difference of the acoustic features between the processed and original speech tokens is shown in Tab. 3.2.

The most prominent difference is, of course, the mean pitch value, which has been lowered systematically in all speech samples. The other acoustic features stay more or less the same with only small changes. At least for the HNR ratio we would have expected increased values for the modified speech tokens.

3.4.2 Subjective Evaluation

The evaluation of this kind of disordered speech enhancement may be done in three different ways.

Table 3.2: *Calculation of acoustic features for every speaker. The difference between the processed and the original speech tokens is shown.*

Speaker	Mean Pitch [Hz]	Local Jitter (%)	Local Shimmer (%)	HNR (dB)
A	-60.7	0.13	-1.43	0.90
B	-64.5	0.19	0.03	-0.48
C	-86.3	0.13	-0.80	0.55
D	-66.4	-0.97	-1.07	-1.37

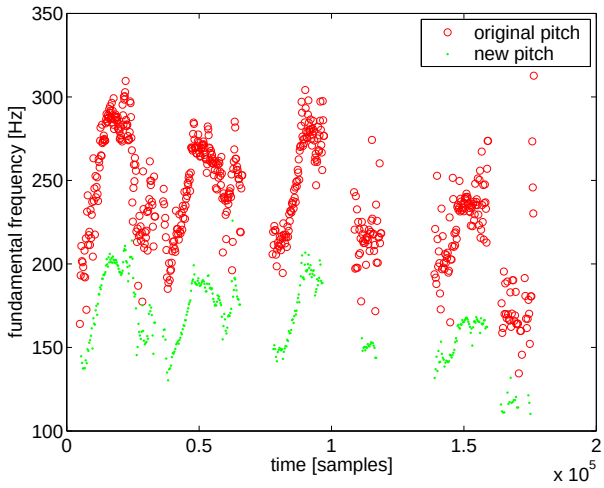


Figure 3.5: *Global lowering of the pitch by a factor of* $\alpha = 0.7$. *Original Pitch Contour (o). Modified Pitch Contour (·).*

1. Professional SLTs can judge the original versus the modified voice to assess the modification from a medical point of view.
2. Naive listeners should be included because the results may not necessarily be of medical relevance, but have to be seen in the context of everyday life.
3. The patients should express their opinion about their own modified voice, whether they feel comfortable with this new voice and could imagine to use a device which modifies their own voice.

For the present study, the subjective evaluation was performed by trained SLTs and naive listeners only.

Since the speech utterances are modified by signal processing algorithms the usual evaluation methods for dysphonic speech are not sufficient to describe the effect of the enhancement algorithms adequately. Therefore, in addition to the RBH rating scale, other evaluation attributes (including those recommended for telephone transmission quality or synthetic speech [ITU96, VHH98]) are used.

For the comparison of the modified speech tokens with the original ones the following features were assessed:

Naturalness: How would you judge the naturalness of the utterance?

Listening effort: How much effort does it take to listen to and understand the utterance?

Speaking effort: The term speaking effort has previously been used in the literature. In the area of affective disorders, i.e., depression, speaking effort is defined as gesticulating and looking at the interviewer during patients' own speech [GBB97]. This is a psychologic definition, which is not used here. Other definitions relate to speaking effort as the effort to increase the intensity of

the voice, which is accompanied by an changed spectral balance, emphasizing higher frequencies [SvH96].

Svec *et al.* [STP03] used a questionnaire, which included the self-evaluation of the speaking effort level (1-10 scale; 1 for no effort, 10 for an extreme effort to speak). It was used to assess vocal fatigue, which involves an increase in the speaker's perception of effort when speaking [CK04].

The latter definition is very close to what is asked here. The difference is that we evaluate the speaking effort as perceived by the listeners. So the question is: How much effort do you assume does the speaker need for this utterance? A high speaking effort can for example be perceived as a pressed voice or strain. Velsik Bele [Bel07] describes hyperfunctional/pressed voice production as follows:

> The voice sounds strained, as if the vocal folds are compressed during phonation and produced with great laryngeal effort. The air pulses through the glottis have low amplitude, and the time interval of minimum flow is long.

Noise: Please judge noise or artefacts either due to the voice disorder or signal processing?

Acceptability: Is the sound of the voice acceptable to you?

General opinion: What is your general opinion of the tokens regarding their overall impression?

3.4.3 Results

In table 3.3 the results for the SLT evaluation of the processed speech samples are presented (the two SLTs agreed on the results). The overall hoarseness index and in particular the breathiness of the voices was reduced. The roughness of the voice was not influenced by the processing.

There were eleven naive listeners who had no prior exposure to severely disordered voices. All of them were native Austrian German speakers. The utterances were presented in the original and the modified version. The subjects could listen to the test examples as often as they wanted. They were asked to rate the difference

Table 3.3: *Evaluation by SLT using RBH rating (0 no disorder ... 3 severe disorder). Left: Original data. Right: Processed patient data.*

ID	age	R	B	H
A	57	0	2	2
B	52	1	2	2
C	49	1	3	3
D	71	1	2	2

ID	age	R	B	H
A	57	0	1	1
B	52	1	1	1
C	49	1	2	2
D	71	1	1	1

between the modified and original utterance regarding the before mentioned features according to the following differential scale:

Rating :	Score
much better :	3
better :	2
slightly better :	1
about the same:	0
slightly worse :	– 1
worse :	– 2
much worse :	– 3

Assuming a normal distribution of the comparison opinion scores, estimates of the Comparison mean opinion score (CMOS), i.e., the mean $\hat{\mu}$, and the standard deviation $\hat{\sigma}$ were calculated. For every mean, a confidence interval CI_{95}, is also determined. The CI_{95} shows the range where the true value of the true mean is expected to lie, with a confidence of 95%. The results are presented in Fig. 5.15.

Results averaged over all four speakers are presented in table 3.4 and figure 3.6. The features *speaking effort* and *acceptability* have the lowest variance and a similar rating. The improvement of both features is significantly different from zero at a 95% confidence level. Both might be related to each other since one feels more comfortable listening when a speaker does not seem to need a lot of effort to speak.Noise and Naturalness are still signficantly different from zero at a 95% confidence level, but not as clear anymore as the previous two features. The other features, i.e., Listening effort and General Opinion, show rather small changes between the processed and unprocessed tokens, but a high standard deviation and do not achieve a 95% confidence level of being significantly different from zero.

An interesting observation was that all results for isolated vowels had high standard deviations, which means that the subjects did not agree on the characteristics. This supports the above mentioned concerns about the use of isolated vowels only. Those attribute might be meaningless on sustained vowels for naive listeners.

Table 3.4: *Subjective evaluation of the pathologic speech enhancement. Comparison category rating (-3 . . . 0 . . . 3) averaged over all speakers before and after processing by naive listeners.*

Feature	Mean	Std. dev.
Naturalness	0.4	1.3
Listening effort	0.1	1.3
Speaking effort	0.8	0.8
Noise	-0.3	1.2
Acceptability	0.7	1.1
General opinion	-0.1	1.48

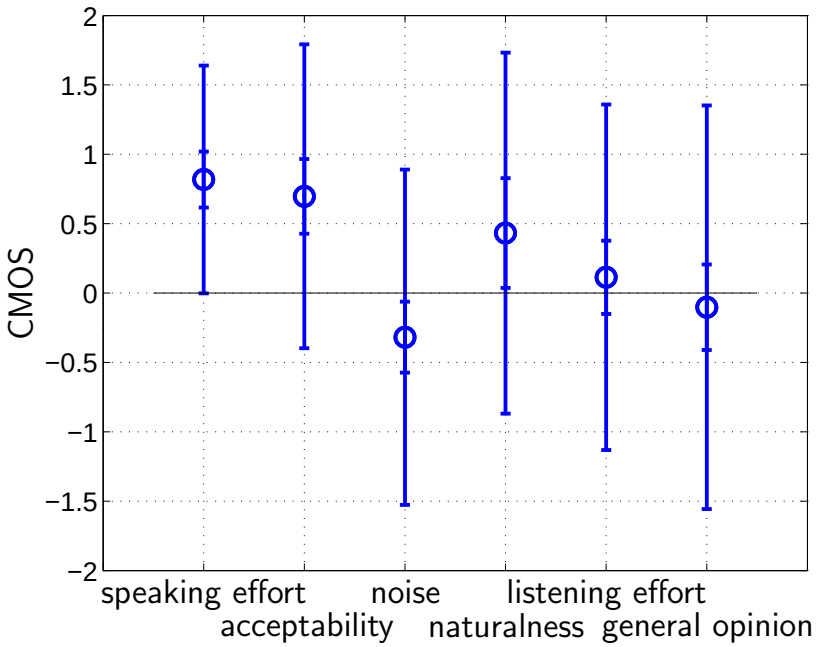


Figure 3.6: *Results for the subjective evaluation of disordered voice enhancement. Shown are the estimated mean $\hat{\mu}$, the 95% confidence interval CI_{95} and the estimated standard deviation $\hat{\sigma}$.*

The *noise* feature received rather negative ratings, with high variance, though. The negative noise rating is probably due to the processing artefacts. This points to the fact, that humans are very sensitive concerning processing artefacts. This is an important point for future work.

An analysis of variance shows that the results are speaker dependent. Speaker A has received the best, whereas Speaker D the worst ratings.

3.5 Conclusions

Even though the processing does not turn a disordered voice into a healthy voice, some characteristics of the voices has been improved. The breathiness rating is lower in all of the processed voices compared to the original speech samples. The most consistent result was that the modified speech was found to be more lax than the original.

The performance of the algorithm is speaker-dependent. Voice disorders have different acoustic outcomes depending on the type and degree of the illness, so the development of a fit-all algorithm is challenging.

The lowering of the fundamental frequency of a speaker results different pitch characteristics depending on the use of absents of such a modification algorithm. The acceptance of such a modification of the voice has still to be determined. Such

an algorithm could be made part of a telecommunication system, improving the speech of the dysphonic speaker, before it is transmitted to the receiver. In digital telephones such as mobile phones this can be done by using a software plugin.

The correct determination of pitch marks is crucial for the successful performance of the algorithm. For dysphonic speech, state-of-the-art pitch determination algorithms only offer limited success. Therefore, improvement of the pitch mark determination is an important future work. The following chapter introduces a pitch mark determination method using state-space embedding and Poincaré sections.

Chapter 4

Poincaré Pitch Marking†

4.1 Introduction

This work is motivated by the need for a pitch marking system, which can be applied on running speech and gives results in real-time. The application which it is needed for is a speech enhancement system for speakers with disordered voices as proposed in the previous chapter 3. For practical use in everyday life the processing delay has to be as short as possible. For example, if a potential voice augmentation device is used to enhance telephone conversations, the processing delay of the device adds up with the delay already introduced by the speech coders for the telephone channel. With this in mind the aim is a reliable algorithm, which only needs a short frame buffer to keep the processing delay as low as possible.

With Poincaré sections we choose an approach that originates in nonlinear system theory. Since this field is rather new to the speech processing community this chapter presents the theoretical background and provides a step-by-step description of the algorithms.

The resulting pitch marking system is compared to the pitch marking system provided by the speech processing software package 'Praat' [BW07].

4.1.1 Applications of Pitch Marks

Pitch marks are essential for several speech processing methods. For speech modification, [ML95b] proposed the Time-domain pitch-synchronous overlap-and-add (TD-PSOLA) technique. It enables pitch modification of a given speech signal without changing the time duration and vice versa. Single speech cycles are manipulated and, therefore, they have to be determined reliably. This method is widely used in concatenative speech synthesis to modify the stored speech segments according to the desired prosody (i.e., intonation (F_0) and syllable duration).

†This chapter is an edited version of the previously published paper:
Martin Hagmüller and Gernot Kubin. Poincaré pitch marks. *Speech Communication*, 48(12):1650–1665, December 2006.

Pitch-synchronous speech enhancement is another application, for which pitch-marks are needed. [Kle02] has proposed a procedure based on averaging over neighboring speech cycles and has applied it to coded speech. The method enhances deterministic harmonic components and reduces stochastic noise. In chapter 3 we apply this speech enhancement method for disordered voice augmentation.

Pitch marks can also be used to determine a local frequency contour and to calculate cycle-based cues such as jitter [Sch03]. Based on such analysis, other applications such as intonation recognition [NBW+02] for dialogue systems or voice disorder detection for health care are possible [Tit94].

If a system has to be applied in a live communication setting, then a major requirement for the algorithm is that the results can be obtained in real time with minimal delay.

4.1.2 Nonlinear Processing of Speech

Linear methods have been applied successfully to speech processing problems and are widely accepted in the speech processing community. Not all phenomena occurring in human speech can be explained by linear models. In the particular case of disordered voices the limitations of linear models are clearly observable. Therefore, non-linear approaches for speech processing have received a wider attention for just over a decade.

With the increasing popularity of non-linear dynamical systems analysis, researchers have started to apply low dimensional dynamical models to speech processing [Tis90]. Specifically, vocal fold oscillation has received a considerable amount of attention from the viewpoint of non-linear dynamics, see, e.g., [HBTS95, GOG+99]. Phenomena like bifurcations, subharmonics or period-doubling – as occur in diplophonic voice – and chaotic behavior have all been observed in the human voice. Human speech has been examined in terms of Lyapunov exponents and correlation dimensions, among others ([BM96, KM96, KM05]). For an overview of nonlinear speech processing see [Kub95]. From meetings dealing specially with nonlinear speech processing several publications resulted, which also provide an overview of the state-of-the-art in the field [Bim03, CEFZM05, FZJE+06].

For disordered voices, non-linear approaches have received considerable attention for analysis, in particular as an objective alternative to auditory voice evaluation methods (e.g., [GOT99, Tit94, JZM06, LMMR06]). More recently, state-space approaches have been used for noise reduction (e.g., [HKM01, HKM00, MM02b, JLPY03]) and to improve automatic speech recognition, [IPJ05].

4.1.3 Pitch Marks - Perception - Harmonicity - Periodicity

Depending on the application, one wants to analyze either the temporal periodicity or spectral harmonicity of the signal (see a speech signal and its corresponding Electro-glottogram (EGG) signal in Fig. 4.1). This is of interest, when investigating irregularities of the vocal frequency. If signal modification is implemented in the

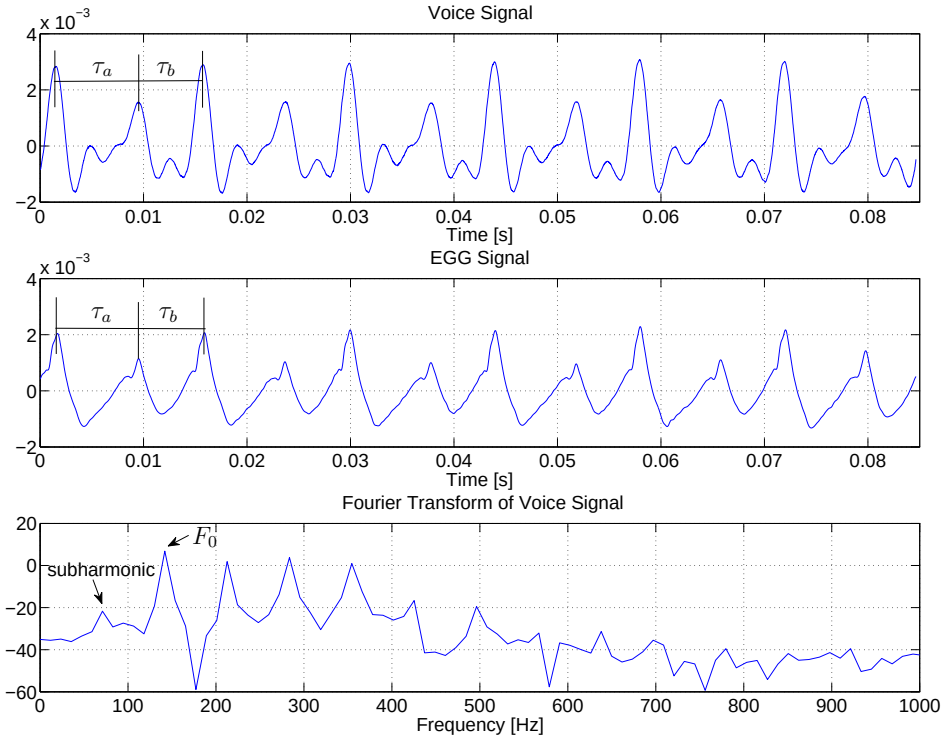


Figure 4.1: *Top: Wave-form view of a signal fragment (vowel 'i'). The two alternating fundamental periods, τ_a and τ_b can be analysed. Middle: EGG signal of the speech fragment. Bottom: Fourier transform of the voice signal. The subharmonic of the signal can be seen.*

frequency domain, such as harmonic plus noise modeling [SLM95], the necessary infomation is only captured if the lowest harmonic is considered. For time-domain based signal modification such as TD-PSOLA [ML95b] or analysis of the vocal fold movement, the glottal cycle length has to be captured in the time domain to reconstuct the irregularities of the voice.

The perception of a subharmonic depends on its relative energy. In case of a weak subharmonic, the perceived pitch agrees with the glottal cycle length, but a change in vocal quality such as roughness may occur [Tit94]. When the subharmonic is strong, a half pitch may be perceived.

4.2 Background and Related Work

Here, we present the background needed to understand the proposed algorithm. A more indepth coverage can be found in [KS04].

4.2.1 Embedding of Dynamical Systems

To analyze a dynamical system, the most efficient representation is the state space. In continuous time, the dynamics of a system is defined by m first-order ordinary differential equations:

$$\frac{d}{dt}\mathbf{s}(t) = \mathbf{f}(\mathbf{s}(t)), \qquad t \in R \tag{4.1}$$

where $\mathbf{f}$ is a set of m functions, $\mathbf{f}(\mathbf{s}(t)) = (f_1(\mathbf{s}(t)), f_2(\mathbf{s}(t)), \ldots, f_m(\mathbf{s}(t)))$ and $\mathbf{s}(t)$ is the m-dimensional state vector of the system. Equation 4.1 is also called the *flow* of a system.

In discrete time the dynamics are described by a map $\mathbf{F}$:

$$\mathbf{s}[n+1] = \mathbf{F}(\mathbf{s}[n]), \qquad n \in \mathbb{Z}. \tag{4.2}$$

In both cases the state-space trajectory is specified by the temporal evolution of the state vector $\mathbf{s} \in \mathbb{R}^m$.

An attractor is a bounded subset of the state space onto which, after some transient, the trajectories converge. Trajectories initially outside the attractor but within its 'basin of attraction' will evolve towards the attractor. The attractor can be a point, a curve, or a more complicated topological structure.

Delay Embedding. Human speech is usually available only as a one-dimensional signal, $s[n]$. Therefore, one has to convert this scalar signal into a state space representation. It has been shown that a non-linear dynamical system can be embedded in a reconstructed state space by the method of delays [KS04]. The state space of a dynamical system can be topologically equivalently reconstructed from a single observed one-dimensional system variable [Tak81].

A trajectory $\mathbf{s}(n)$ in an M-dimensional state-space can be formed by delayed versions of the speech signal $s[n]$,

$$\mathbf{s}(n) = \{s[n], s[n-\tau_d], ..., s[n-(M-1)\tau_d]\}, \tag{4.3}$$

where τ_d is the delay, which has to be chosen so as to optimally unfold the hypothetical attractor.

Embedding Dimension. The optimal choice of the dimension for state space embedding is an important issue. If the dimension is too small the reconstructed trajectories may intersect. On the other hand, if the dimension is too large, the computational effort rises. For a D-dimensional attractor, it is sufficient to reconstruct an $M \geq 2D+1$ state space vector [Tak81]. Later this result has been generalized by [SYC91] to $M > 2D_F$, where D_F is the (fractal) box counting dimension of the attractor. In practice, though, values of $M > D_F$ may be sufficient. The method of detecting false nearest neighbors can be used to determine the minimal necessary embedding dimension in practice [KS04].

Embedding Delay. From a practical point of view, the goal is to unfold the attractor optimally. That means that the extension of the attractor should be roughly the same in all dimensions. This is the case when the reconstructed state vector components have minimal statistical dependence on each other. The most natural approach would be to use the autocorrelation function of the signal, which is linear statistics. Both linear and non-linear dependencies can be calculated by the auto-mutual information, also known as the time-delayed mutual information. This is an information theoretic concept, which is based on the Shannon entropy. It reports the statistical dependencies between two signals, one of which is delayed. The auto-mutual information for a time delay τ is defined as:

$$I_\epsilon(\tau) = \sum_{i,j} p_{i,j}(\tau) \ln p_{i,j}(\tau) - 2 \sum_i p_i \ln p_i, \tag{4.4}$$

where ϵ is the bin width of the histogram estimate of the probability distribution of the data, p_i is the probability that the signal has a value which lies in the ith bin of the histogram and $p_{i,j}$ is the probability that $s(t)$ is in bin i and $s(t+\tau)$ is in bin j. The time lag τ of the first minimum of the auto-mutual information is the optimal delay time for the embedding. The width ϵ can be set rather coarse, because only the dependence of I_ϵ on τ is of interest and not the value of $I_\epsilon(\tau)$ itself [KS04].

Poincaré Plane. If one chooses an arbitrary point on an attractor in an M-dimensional space, one can define a hyperplane orthogonal to the flow of the trajectories at that point. This hyperplane is called the Poincaré plane (Fig. 4.2). All trajectories that return to a certain neighborhood of the selected point, cross the hyperplane and can be represented by their intersection with this plane of $M-1$ dimensions.

4.2.2 Pitch Mark Detection in State Space

[Kub97] first suggested to use Poincaré planes for the determination of pitch marks and mentioned special applications for signals with irregular pitch. Experiments showed promising results for an example of vocal fry, owing to period doubling. The cycle length was recovered correctly.

Later [MM98] applied Poincaré maps to epoch marking for speech signals, with glottal closure instants set as initial point. They obtained promising results, but reported failure to resynchronize after stochastic fragments of speech.

More recently, [Ter02] proposed another state space approach to pitch determination, using space-time separation histograms. Each pair of points on the trajectory in reconstructed state space is separated by a spatial distance r and a time distance Δt. One can draw a scatter plot of Δt versus r or, for every time distance Δt, count the $(r, \Delta t)$ pairs within a certain neighborhood r. This count can then be normalized to 100% to yield a histogram (Fig. 4.3).

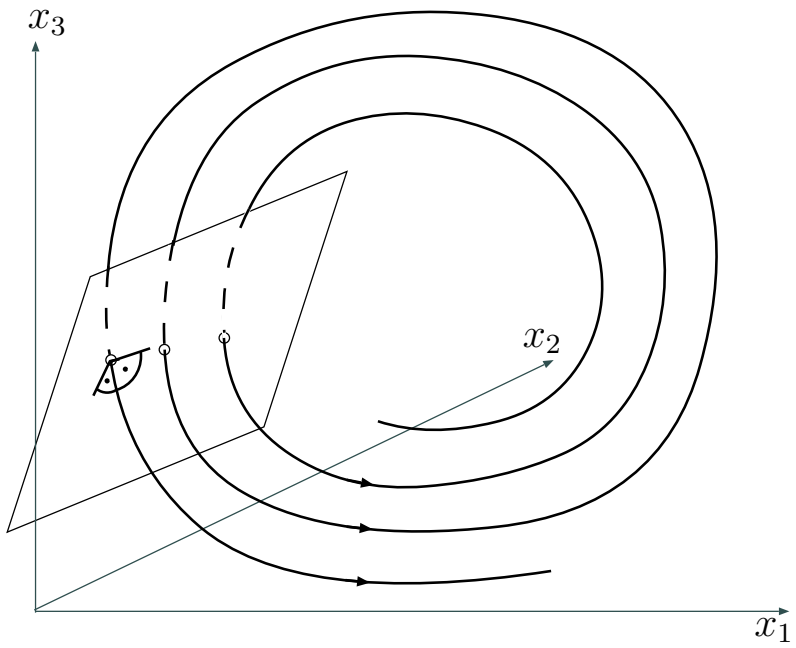


Figure 4.2: *Placement of the Poincaré plane orthogonal to the flow of the trajectories.*

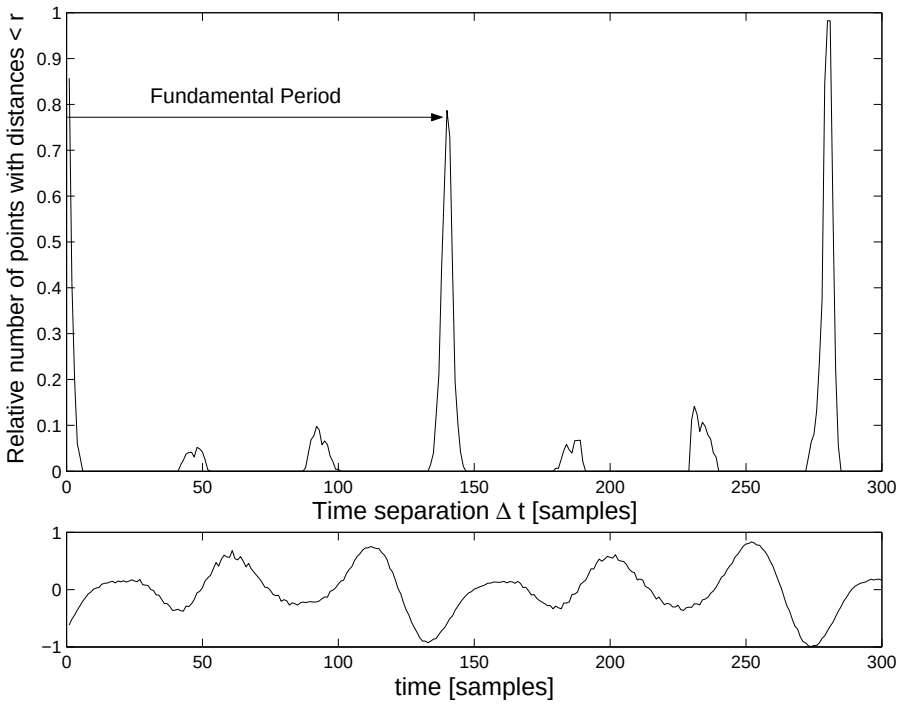


Figure 4.3: *Histogram of space-time separation. Top: The normalized number of points within a certain state-space neighborhood r for every time distance Δt is plotted. Bottom: Time-domain waveform plot.*

In case of periodicity, the histogram is large at certain Δt values, whereas for others it is low. The first maximum of the histogram indicates the fundamental period. Compared to the auto-correlation function the peak is sharper and, therefore, the author claims, offers improved performance. In case of noise-like signals, the histogram is more spread out over all time distances Δt. Since histograms are based on several cycles, pitch marks cannot be determined reliably with this approach. The computed fundamental period is an average over the frame length.

4.3 Algorithm

This work builds on the aforementioned approaches and is an improved version of the work previously described in [HK03] and [HK05a]. A step-by-step guide is included in Matlab-like notation.

4.3.1 Pre-Processing

The algorithm works on a frame-by-frame basis to handle the slowly changing parameters of the speech production system. For pitch mark detection, noise has to be removed, otherwise, e.g., for hoarse voices, the attractor is hardly visible using 3-dimensional embedding (Fig. 4.4). If the embedding dimension is high enough, intersections with the Poincaré plane would still correspond to the pitch period, but with less reliability. However, at a high enough noise level the algorithm would break down.

To reduce noise in the attractor, a Singular value decomposition (SVD) embedding approach has been proposed [BK86], but similar results can be achieved by linear-phase low-pass filtering (Fig. 4.5). The latter is computationally less demanding.

To remove the influence of the changing amplitude, automatic gain control is applied for every frame of the input signal $s_0[n]$. First, the signal envelope $\nu(n)$ is calculated:

$$\nu[n] = \frac{1}{K} \sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} |s_0[n+k]| \tag{4.5}$$

$$s[n] = \frac{s_0[n]}{\nu[n]}, \qquad \forall \nu[n] > \nu_{th} \tag{4.6}$$

where $s[n]$ is the speech signal, K is the length of the moving average filter, which is set to the maximum expected fundamental period and ν_{th} is a threshold to avoid overamplification of low-energy non-speech sections. This moves the trajectories of pseudo-periodic signals closer together, which means that the attractor is contracted, if it was spread out owing to amplitude variations (Fig. 4.6).

Then the signal is upsampled to $f_s = 48$ kHz to increase the resolution of the pitch marks. The embedding in state space is implemented by the method of delays, the embedding dimension is chosen to be $M = 8$. Experiments have shown that this

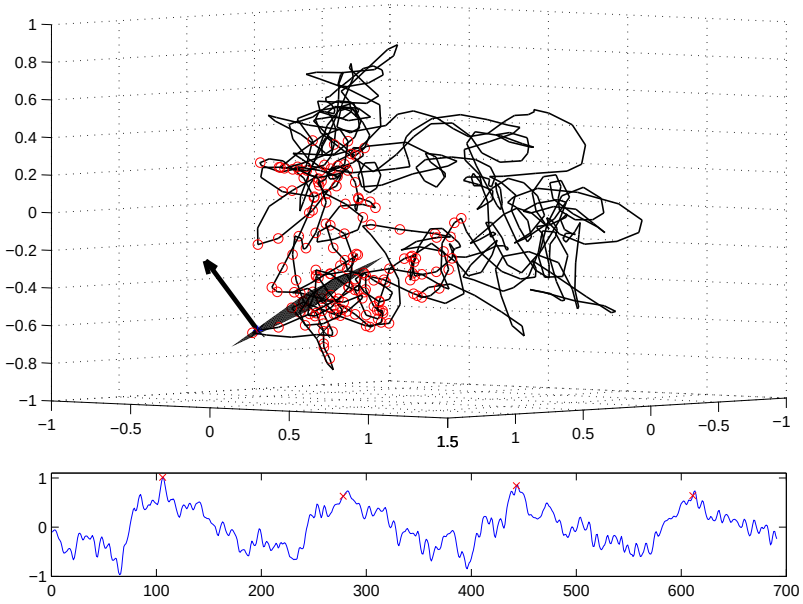


Figure 4.4: *Vowel 'a' from a dysphonic speaker. Top: Projection of state-space embedding on 3 dimensions and Poincaré plane. Circles are neighbors. Bottom: Time-domain waveform plot. No low-pass filter. Crosses correspond to the pitch marks, i.e., the point where the trajectories go through the Poincaré plane.*

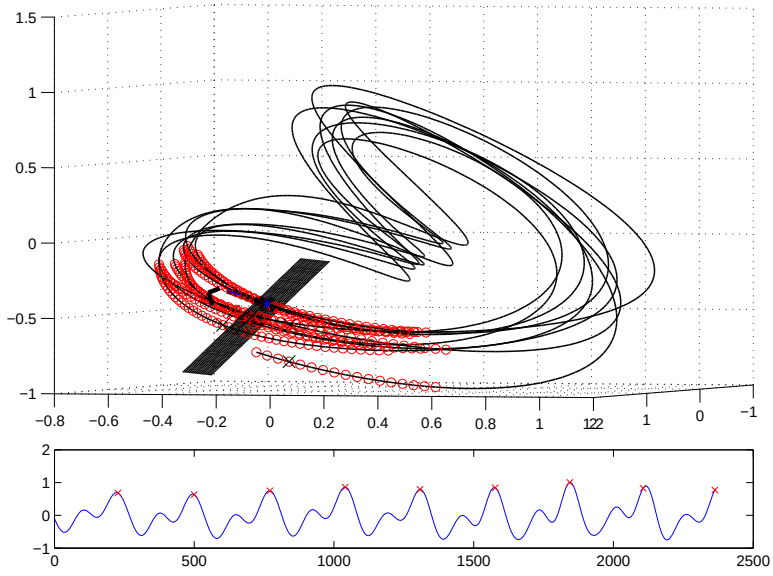


Figure 4.5: *Vowel 'a' from a dysphonic speaker. Top: Projection of state-space embedding on 3 dimensions and Poincaré plane. Circles are neighbors. Bottom: Time-domain waveform plot. Low-pass filter. Crosses correspond to the pitch marks, i.e., the point where the trajectories go through the Poincaré plane.*

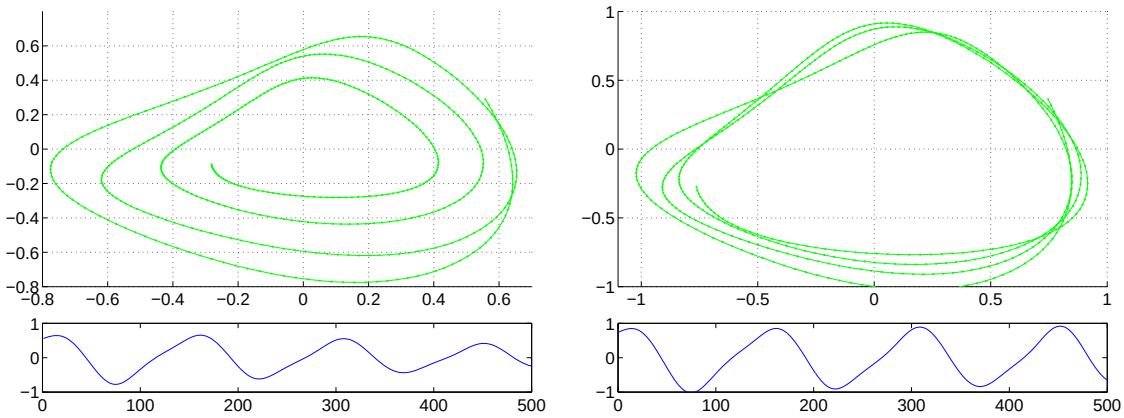


Figure 4.6: *Top: Projection of state-space embedding on two dimensions . Bottom: Time-domain waveform plot. Left: No automatic gain control applied. Right: Automatic gain control applied.*

number gives the most reliable results over different kinds of speech tokens, though the algorithm may generally work with embedding dimensions $M \geq 3$. For every frame a state-space matrix $\mathbf{S}$ is formed:

$$\mathbf{S} = \begin{pmatrix} s[0] & s[-\tau_d] & \dots & s[-(M-1)\tau_d] \\ s[1] & s[1-\tau_d] & \dots & s[1-(M-1)\tau_d] \\ \vdots & \vdots & \ddots & \vdots \\ s[N] & s[N-\tau_d] & \dots & s[N-(M-1)\tau_d] \end{pmatrix}, \tag{4.7}$$

where N is the frame length, M is the embedding dimension and τ_d the embedding delay. Each row represents a point $\mathbf{s}(n)$ in state space.

4.3.2 Poincaré Plane

At the heart of the algorithm is the calculation of the Poincaré hyperplane. First a point at time n_0 has to be chosen on the trajectory. This can be a peak or trough in the time-domain waveform. A possible source of error is the choice of peaks in the signal that are not the maximum (see Fig. 4.13). An alternative is the selection of the initial point in state space.

This alternative tries to find an area in state space, where the local bundle of trajectories is the least diverging. The initial point is placed inside this area. This is optimal for finding pitch marks using the Poincaré plane, and the time-domain waveform is not considered anymore.

Around this chosen query point $\mathbf{s}(n_0) = \mathbf{S}(n_0, :)$, the state space is searched for the k closest points according to the Euclidean distance measure. These k point form a neighborhood $\mathcal{N}(n_0)$. This can be done by calculating the Euclidean distance between $\mathbf{s}(n_0)$ and all other points $\mathbf{s}(n)$ of the state-space matrix $\mathbf{S}$:

$$d_{eucl}(n) = \sum_{m=1}^{M} \left(S(n,m) - S(n_0,m)\right)^2 . \tag{4.8}$$

More computationally efficient methods exist to search for the neighbors in state space (e.g., [Sch95]).

Then a mean flow direction $\mathbf{f}(n_0)$ of the trajectories in this neighborhood $\mathcal{N}(n_0)$ is calculated,

$$\mathbf{f}(n_0) = mean(\mathbf{s}(n+1) - \mathbf{s}(n)) \qquad \forall n \in \mathcal{N}(n_0), \tag{4.9}$$

For which only trajectories are considered that point roughly in the same direction as the initial flow vector ($\mathbf{f}_0^T(n)\mathbf{f}_0(n_0) > 0.6$, where $\mathbf{f}_0$ is a unit-length vector, for $f_0(n)$ no flow averaging is applied), i.e., orthogonal flow vectors or flow vectors in the opposite direction will not be considered.

So for every frame the Poincaré hyperplane is defined as the hyperplane through $\mathbf{s}(n_0)$ that is perpendicular to $\mathbf{f}(n_0)$ (Fig. 4.7 (b)).

$$(\mathbf{P} - \mathbf{s}(n_0))^T \cdot \mathbf{f}(n_0) = 0, \tag{4.10}$$

where $\mathbf{P}$ is any point on the Poincaré plane.

To calculate the intersections with the plane, the points before and after the passing of the trajectory through the plane have to be found. If

$$(\mathbf{s}(n+1) - \mathbf{s}(n_0))^T \cdot \mathbf{f}(n_0) > 0 \;\&\; (\mathbf{s}(n) - \mathbf{s}(n_0))^T \cdot \mathbf{f}(n_0) < 0 \tag{4.11}$$

then the points $\mathbf{s}(n)$, $\mathbf{s}(n+1)$ are just before and after the plane (Eq. 4.11) and

$$(\mathbf{s}(n) - \mathbf{s}(n_0))^T \cdot \mathbf{f}(n_0) = 0 \tag{4.12}$$

if point $\mathbf{s}(n)$ lies exactly on the Poincaré plane (Eq. 4.12).

The exact location of the intersection points is calculated by linear interpolation between the two points before and after the intersection, i.e., between $\mathbf{s}(n)$ and $\mathbf{s}(n+1)$. The intersection points with the Poincaré plane and their corresponding time indices are considered to be pitch mark positions.

The length of one frame is chosen so that at least two cycles at the expected minimum frequency fit into the frame. If the signal is pseudo-periodic, the trajectory then returns at least once into the chosen neighborhood and intersects the Poincaré hyperplane and a pitch mark can be detected. The frame hop size depends on the pitch mark in the current frame. The beginning of the next frame is set to the current pitch mark.

4.3.3 Post-Processing

The voiced/unvoiced decision is based on several criteria. First, a frame is considered unvoiced if the energy of the low-pass filtered signal is below a threshold. If the

energy criterion misses an unvoiced frame, the frame is further analyzed in the state space in case of observed fluctuations of the fundamental cycle length by considering several cues and calculating a cost for each candidate pitch mark. First, the Euclidean distance of the pitch mark candidate points to the query point is considered. Second, the distances between all candidate points are considered and a grouping of the candidates in two clusters is performed. If the result includes a cluster that is clearly removed from the points around the query point, those points are attributed a higher cost. A third cue is the time-domain energy in the vicinity of a candidate pitch mark; a low energy is associated with a higher cost. Candidate points with a cost above a threshold are discarded. In addition to the detection of unvoiced sections, this also reduces the occasional hit of the first formant (Fig. 4.7).

Fig. 4.8 shows a flow diagram of the pitch marking system.

4.3.4 Pseudo Code

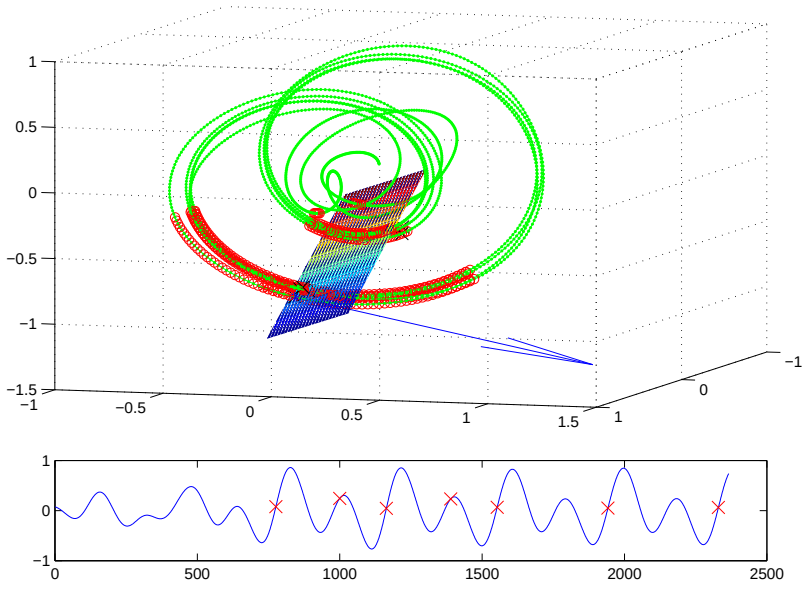
· Input speech signal $x(n)$

· Low-pass filter

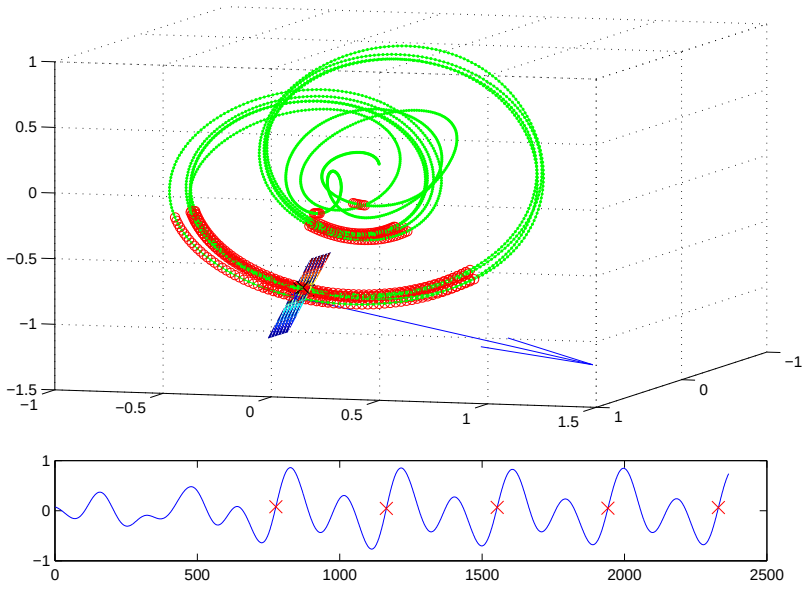
· Upsample (if necessary)

· WHILE index $<$ lastindex - framelength

- Get frame with framelength at index
- IF energy(frame) < threshold,
  - · set frame unvoiced
  - · take next frame
- END
- Apply automatic gain control
- Normalize frame
- Choose initial point, $x(n_0)$, in time-domain
- Embed segment in pseudo state space (dimension M, delay τ_d)
  $\mathbf{x}(n) = [x(n), x(n-\tau_d), ..., x(n-(M-1)\tau_d)]$
- Select k neighbors in state space neighborhood $\mathcal{N}(n_0)$ of $\mathbf{x}(n_0)$
- Compute estimate of average vector flow $\mathbf{f}(n_0)$
  $\mathbf{f}(n_0) = mean(\mathbf{x}(n+1) - \mathbf{x}(n)) \qquad \forall n \in \mathcal{N}(n_0)$
- Define Poincaré hyperplane perpendicular to $\mathbf{f}(n_0)$ going through $\mathbf{x}(n_0)$
  $(\mathbf{x}(n_0) - \mathbf{P})^T \cdot \mathbf{f}(n_0) = 0$
- Calculate intersection of trajectories through Poincaré plane by interpolation between samples neighboring Poincaré plane
- IF std(T0) > 0.15 · median(T0)
  - · Discard points far away from $\mathbf{x}(n_0)$
- END
- Set index to beginning of next frame



(a)



(b)

Figure 4.7: *Vowel 'a' from a dysphonic speaker. Top: Projection of state-space embedding in 3 dimensions and Poincaré plane. Circles are neighbors. Bottom: Waveform plot. (a) Wrong parts of the trajectories chosen. (b) Correct placement of Poincaré plane and pitch marks shown as crosses.*

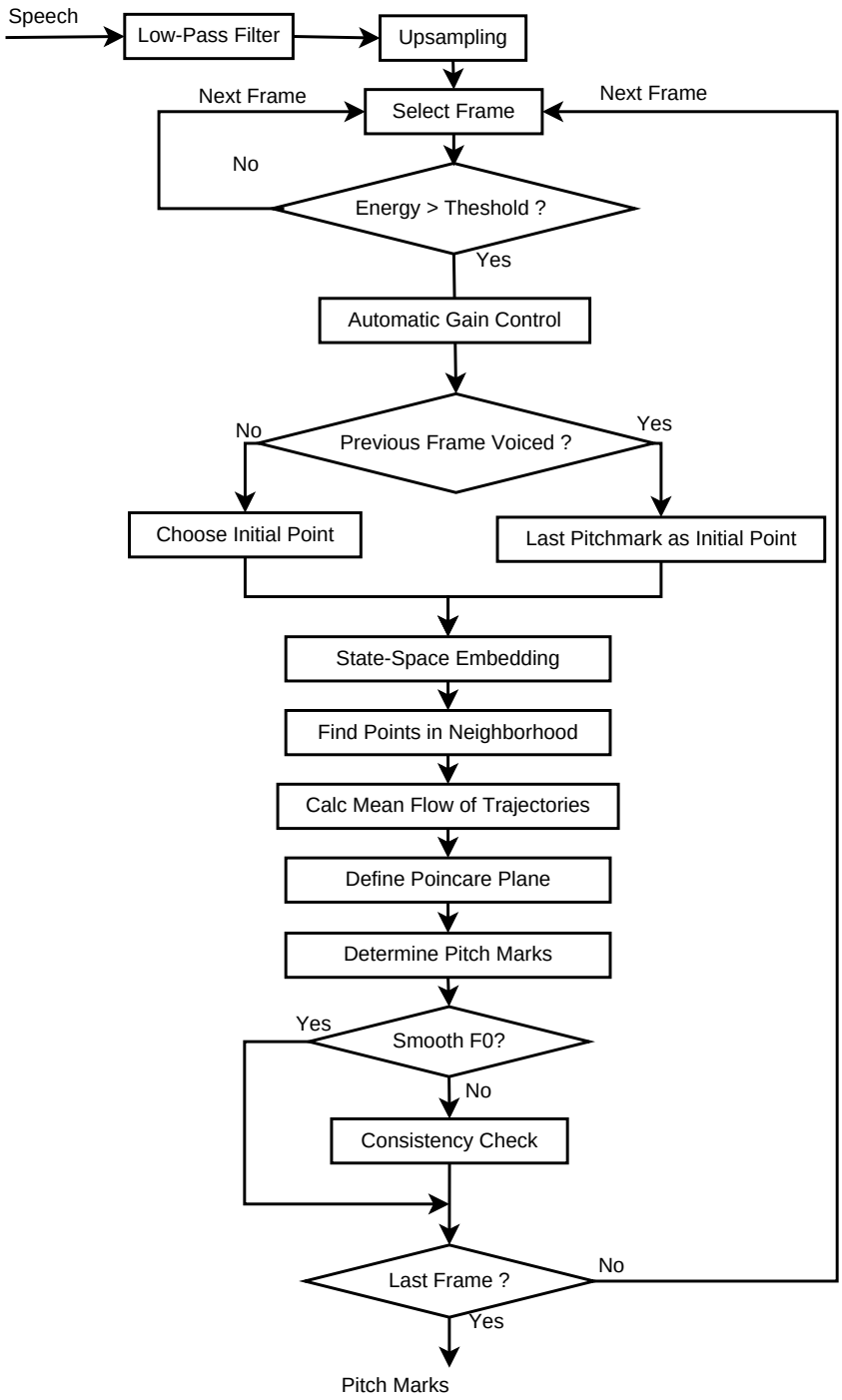


Figure 4.8: *Flow diagram of pitch marking system.*

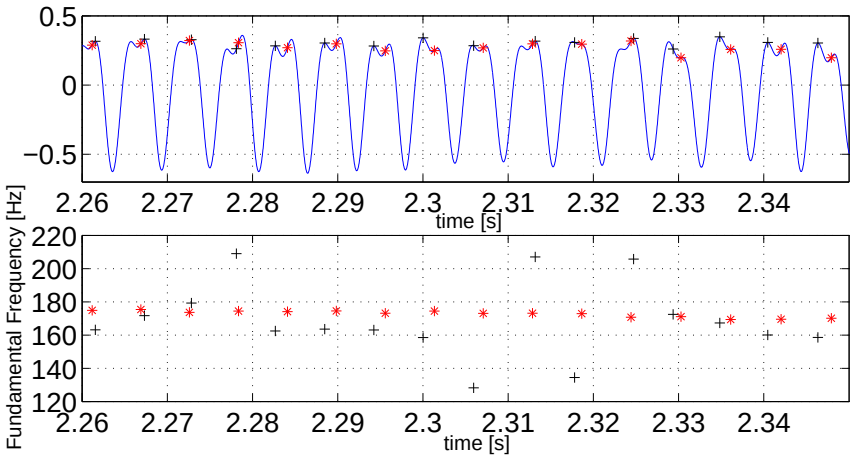


Figure 4.9: *Changing peaks. Comparison of results from the Poincaré method (∗) and the peak-picking method from 'Praat' (+). Top: Waveform with pitch marks. Bottom: Fundamental frequency estimates computed from pitch marks.*

· `END WHILE`

· `Output pitch marks`

4.4 Discussion

There are some interesting results and open problems, which we discuss in this section.

Jitter in Peak Detection Algorithm. Fig. 4.9 shows a signal section with a temporal evolution of the prominence of its positive peaks. The positive signal changes from a dual peak with its maximum on the right into a dual peak with its maximum on the left and back again. Peak picking pitch marking algorithms such as 'Praat' (see appendix A) switch the pitch mark back and forth between the left and the right peak. This introduces spurious jitter, which is an artefact of the algorithm. Since the Poincaré method is not based on time domain signal properties, the pitch cycles are followed correctly, staying at the same position over the whole signal fragment.

To compare the two methods quantitatively, we calculate the period perturbation factor (PPF):

$$PPF = \frac{100\%}{N-1} \sum_{i=1}^{N-1} \frac{u(i) - u(i-1)}{u(i)}, \tag{4.13}$$

where $u(i)$ is the cycle length sequence, corresponding to the time between pitch marks shown in figure 4.9. For given speech segment, the Poincaré method yields PPF = 0.14 % and the Praat peak-picking method: PPF = 15.1 %, which is over 100 times larger as the Poincaré based value.

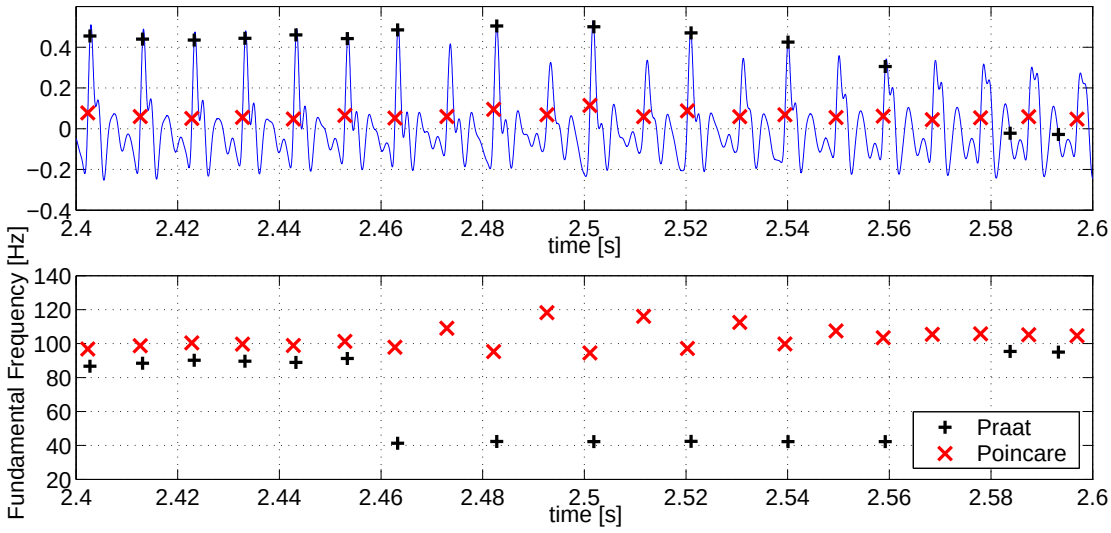


Figure 4.10: *Pitch marks of transient glottalization ('her o(n)'). Comparison of the results from the Poincaré method and 'Praat'.*

Biphonation. In figure 4.10, a fragment ('her o(n)',) is shown of a recording (sentence *'rl040 – Judith found the manuscripts waiting for her on the piano'* from the Bagshaw database [Bag94]). A transient biphonation due to glottalization is observed.

Other algorithms like Praat by [BW07] either completely fail for such events or detect a period doubling if the fixed minimum pitch value allows for such a long period. The Poincaré method recognizes the rapidly alternating pitch cycles correctly. Of course in this case it is a matter of definition whether alternating cycles or period doubling is the correct interpretation. Still, we consider our approach more useful, since the subpartials can be derived from our result in a second step, if desired. This is not possible the other way round. If only the subpartial is known, the cycle alternation in the time domain cannot be recovered anymore.

In Fig. 4.11 the alternating size of the time-domain waveform peaks can be seen in the state-space plot as two different bundles of the trajectory.

Phase Drift. In Fig. 4.12, we see that the phase of the Poincaré pitch marks does not remain constant over the whole signal fragment in the plot. One sees that the marks slowly evolve from the positive to the negative peaks. This is a problem, which occurs only occasionally and a change of analysis parameters usually removes the problem for one signal fragment, but may introduce a drift at another place.

However, no general design rule could be derived so far to remedy this problem.

In applications for which high accuracy is an important issue or for which the pitch marks should be synchronized with a glottal event, such as an excitation maximum or glottal closure, this may be a disadvantage of this method. The reason for this drift may lie in the changing dynamics of the speech production system or small

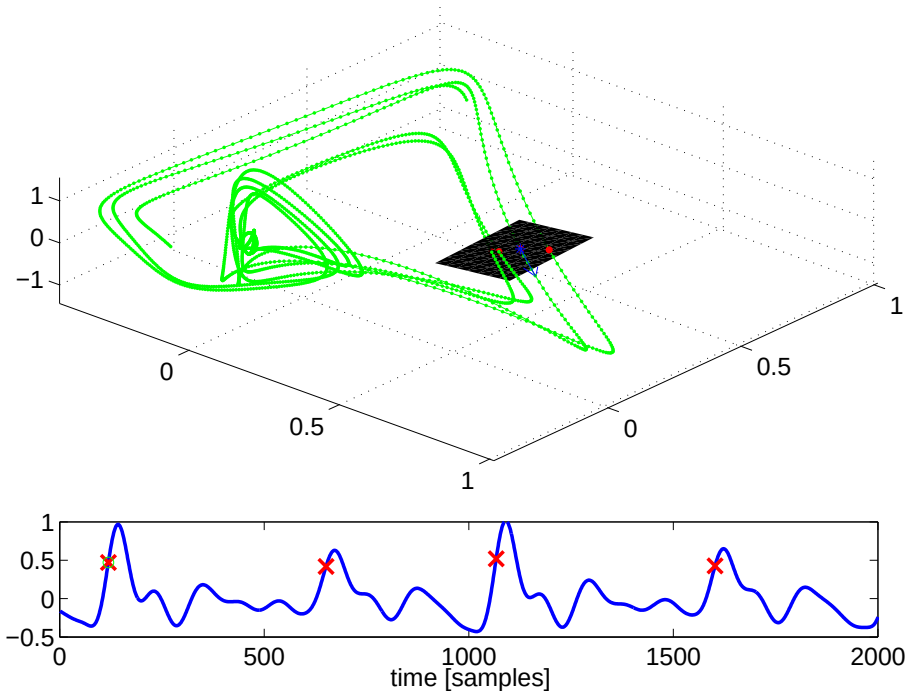


Figure 4.11: *State space representation of a segment of voice with biphonation. The two different trajectories of the speech production system can be clearly seen.*

errors in the calculation of the mean flow vector.

A phase drift was also reported by [Man99], but in comparison to his approach we interpolate between two sampling points to determine the intersection of the trajectories with the Poincaré plane more precisely (see Sec. 4.3.2).

Resynchronization. In Fig. 4.13, the same signal fragment is shown twice after low-pass filtering with two different cut-off frequencies. After each unvoiced segment, the initial point $\mathbf{s}(n_0)$ has to be set again. The search frame for the initial point is set to the first half of the current possibly voiced frame. As mentioned in section 4.3.2, the position of the current frame depends on the last pitch mark of the previous frame (if voiced).

In the upper plot of Fig. 4.13, the initial point was set to the smaller peak at the onset of the voiced segment, consequently the following pitch marks are in the corresponding positions in the following cycles. The lower plot shows a situation, for which the search frame was positioned to mark the maximum peak of the cycle, which is followed perfectly throughout the rest of the frame.

A possible solution would be to introduce a more sophisticated peak search algorithm, with some look-ahead to determine, whether a better choice for the initial point is available. This comes at the cost of reduced real-time capabilities of the algorithm. If the initial points are chosen based on the bundling of the trajectories in state space, this is, of course, not an issue since time-domain peaks are no criterion

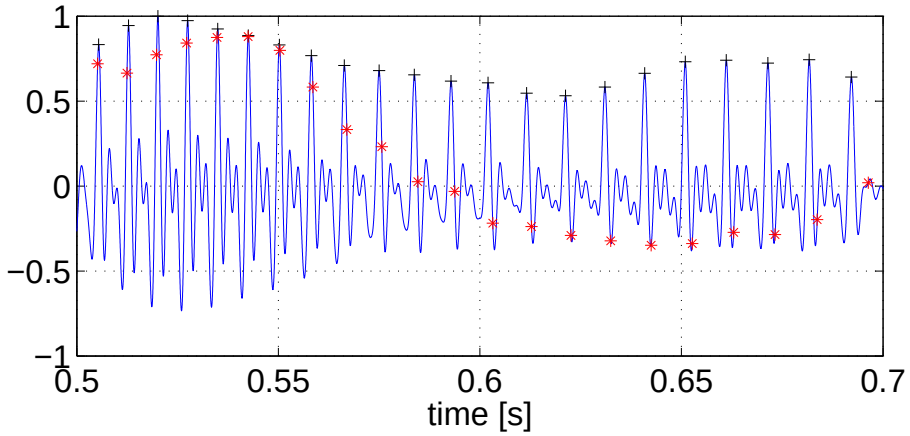


Figure 4.12: *Phasedrift of the pitch marks. The marks evolve from positive to negative peaks. Results from 'Praat' (+) and Poincaré (∗).*

for this choice, anyway.

4.5 Evaluation

To evaluate the performance of the pitch marking algorithm in a quantitative way, it is compared with a state-of-the-art pitch marking software, that is *Praat*, which is a software for speech processing and computer linguistics [BW07]. For pitch mark determination, *Praat* first does pitch determination using auto-correlation and a least-cost path search through the analyzed frames with several pitch candidates per frame. The resulting pitch contour provides a narrow search range for the pitch marking algorithm. It starts from an absolute extremum in a frame and determines the pitch marks by finding the cross-correlation maximum in the range given by the previously calculated pitch contour.

The database used for the evaluation of the pitch mark algorithm is freely available. It includes speech and EGG signals of a male subject (117 seconds of speech) and a female subject (147 seconds of speech) uttering 50 sentences each and the corresponding pitch marks obtained by a pulse location algorithm applied to the EGG signal [Bag94].

4.5.1 Formal Evaluation

Due to several issues discussed in Section 4.4, the evaluation of the algorithm cannot be performed assuming a fixed phase of the pitch marks. Since evaluation of pitch marks which are not associated with a fixed temporal event (such as a peak in the time domain) is difficult, a conversion to vocal frequency (F_0) values for a fixed time step of 10 ms is performed. In case less than two pitch marks per 10 ms are found the window is extended to 20 ms. If still less than two pitch marks are found, the

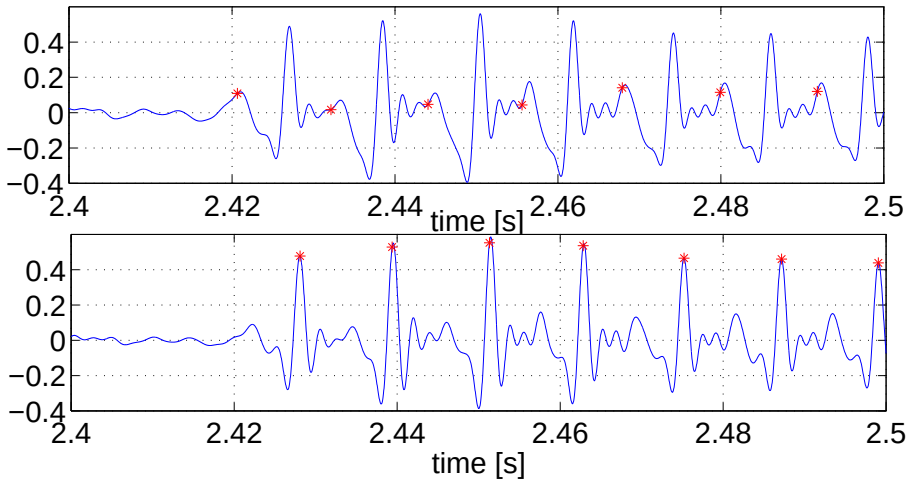


Figure 4.13: *Resynchronization after unvoiced period. Depending on the frame position different initial points are chosen, but synchronization is kept independent of initial phase throughout voiced interval.*

fragment is considered to be unvoiced. For every window, an average value for the fundamental frequency is calculated. For the male files the range for the expected frequencies was set to 50-300 Hz, for the female files to 150-400 Hz. The procedure is rather an evaluation of the fundamental frequency determination than the pitch marking capabilities of the algorithm. With this restriction, the performance of the algorithm can be compared to the results from 'Praat'.

Comparisons with other afore mentioned related algorithms from the literature (section 4.2.2) are difficult, since none of the referenced works deal with running speech, using full sentences [Kub97, MM98]. In addition they do not provide any formal evaluation results themselves.

4.5.2 Results

In tables 4.1 and 4.2, the results are shown for the database. The results for the Poincaré method and the 'Praat' pitch marking algorithm [BW07] are presented. The voicing/devoiced errors are at the top. In the first column the total ratio of voiced frames for the speech material can be seen, then the voicing errors and finally the devoicing errors are shown. A voicing error is set if an unvoiced frame is falsely considered as voiced by the pitch detection algorithm. A devoicing error is set if a voiced frame is falsely considered as unvoiced. Voiced speech covers roughly half of the total length of the recordings.

In the lower part of the table the accuracy of the results for the voiced frames is shown. Fundamental frequency errors are differentiated into the percentage of errors <1 %, <5 % and <10 % of the reference fundamental frequency. It is desired to have a high percentage of the errors in the 'errors <1 %' column.

Discussion

The performance of the two algorithms is comparable. While the Poincaré method has slightly better results for the voicing errors, Praat has a lower error rate for devoicing errors. Both algorithms can be tuned to favor either voicing or devoicing errors, depending on the application. On the one hand, it may be better to capture all pitch marks while running the risk of having too many. On the other hand, it may also be desirable to ensure that the pitch marks that are found are all correct, while running the risk of not finding all of them. In general, a balance between voicing and devoicing errors at a low level is desired.

The accuracy of the voiced frames found is also similar. For both algorithms, the errors of roughly 70 % of the F_0 values are <1 %, while more than 90 % of all errors are below 5 %. For female speech, the results of the two algorithms are very close, while for male speech the Poincaré method outperforms Praat by about 5 % in its relative accuracy.

4.6 Conclusion

In this chapter, we have presented an approach to pitch mark determination, which applies methods from dynamical systems analysis to speech signal processing. The algorithm has been presented in a way, that a reader new to dynamical systems is able to understand the principles and implement the algorithm.

While the results are promising, the algorithm does not outperform state-of-the-art pitch detection algorithms such as 'Praat' in all situations. A clear advantage of the new algorithm has been shown for biphonic voices. The rapidly alternating cycles are recognized correctly, where in contrast, 'Praat' only detects subharmonics. It is known that peak-picking algorithms can introduce jitter artefacts if the analyzed signal has dual competing peaks, which vary in amplitude. In that case, the Poincaré method tracks the cycles correctly.

One has to keep in mind, though, that the performance achieved by 'Praat' is

Table 4.1: *Results for a female speaker, 147 s of speech, 46.7 % voiced speech. The voiced column shows the percentage of detected voiced frames. The voicing errors are the percentage of frames falsely considered as voiced, and vice versa for the devoicing errors. Errors < 1 %, 5 % and 10 % show the percentage of pitch errors which are below 1 %, 5 % and 10 % relative error, respectively.*

	voiced	voicing errors	devoicing errors
Poincaré	50.3 %	3.0 %	2.2 %
Praat	51.6 %	4.4 %	0.6 %
	errors <1 %	errors <5 %	errors <10 %
Poincaré	71.7 %	91.7 %	95.8 %
Praat	70.7 %	91.2 %	95.7 %

only possible by looking at a whole speech fragment. In a first stage a algorithm determines a pitch contour by searching for an optimal path through a number of pitch candidates. The contour is then used as a constraint for the pitch mark determination. The results are further improved by a sophisticated post-processing algorithms, which consider previous and following frames next to the current one. This higher-layer processing is not used by the Poincaré method, since it has been designed for real-time applications, and it is intended to keep the delay strictly of the order of a single frame.

The intention has been to develop a reliable method for pitch mark determination for pitch synchronous voice modification as presented in chapter 3. Due to the phasedrift discussed above (Sec. 4.4) the algorithm is not able to surpass the performance achieved with Praat for the PSOLA task at the current stage of the work. This problem is inherent to the algorithm, since in state-space time-domain amplitude peaks cannot be detected. Since the main goal of the thesis are signal processing methods for disordered voices, the pitch marking topic has been left at this stage to be able to further explore the main topic.

Table 4.2: *Results for a male speaker, 117 s of speech, 51.82 % voiced speech. The voiced column shows the percentage of detected voiced frames. The voicing errors are the percentage of frames falsely considered as voiced, and vice versa for the devoicing errors . Errors < 1 %, 5 % and 10 % show the percentage of pitch errors which are below 1 %, 5 % and 10 % relative error, respectively.*

	voiced	voicing errors	devoicing errors
Poincaré	54.6 %	2.7 %	8.7 %
Praat	57.7 %	5.8 %	4.8 %
	errors <1 %	errors <5 %	errors <10 %
Poincaré	74.8 %	96.1 %	98.3 %
Praat	69.1 %	93.2 %	97.6 %

Chapter 5

Multipath Signal Separation for Electro-Larynx Speech

5.1 Introduction

This section focuses on a problem of the Electro-Larynx (EL) method for substitution voice production. This problem is due to the Directly radiated electro-larynx (DREL) sound, which in addition to the speech sound, reaches the ear of the listener unmodulated and interferes with the modulated speech sound (see Fig. 5.1). If only the acoustic channel is available (e.g., over the telephone) this disturbance makes communication difficult. We have investigated methods to reduce the DREL sound to improve speech communication in situations other than face-to-face communication.

We look at how a speech sound is formed in EL speech and how unwanted components affect the sound received by the listener.

5.1.1 Electro-Larynx Speech Production Model

Speech production with an EL is different from laryngeal speech production in several aspects. To clarify the EL speech production we present an EL speech production model, which extends a simpler model [NWWL03]. Since we later point out the difference between a time-variant and a time-invariant signal path, we shortly introduce the necessary representations and notations for time varying filters [CM82].

5.1.2 Time-Variant Linear Filtering

The output $y(t)$ produced by the input signal $x(t)$ and time-variant impulse response $g(\tau, t)$ is

$$y(t) = \int_{-\infty}^{\infty} g(\tau, t)x(\tau)d\tau, \tag{5.1}$$

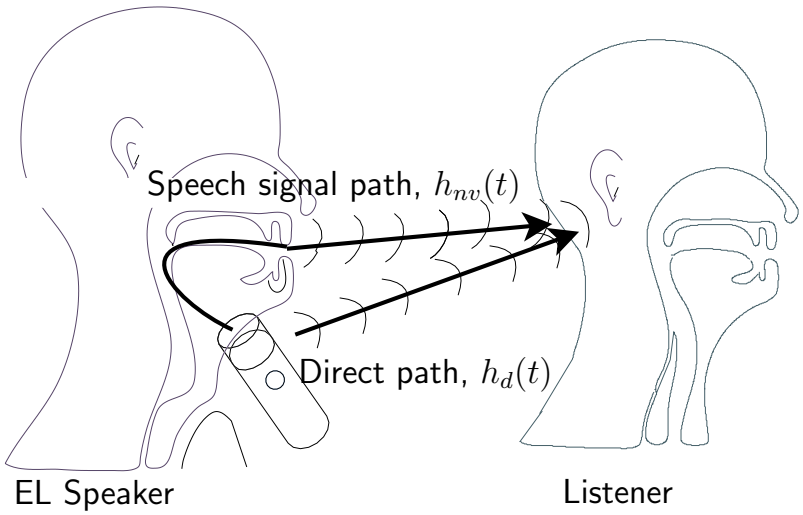


Figure 5.1: *Electro-Larynx (EL) speech. The EL signal is transmitted to the listener over two paths with different impulse responses. One path is through the time-variant vocal tract, $h_{nv}(\tau, t)$. The other path is through the time-invariant direct path, $h_d(t)$, and gives rise to the DREL sound.*

where $g(\tau, t)$ is the impulse response at time t for an excitation with an impulse at time τ. If we substitute $h(\tau, t) = g(t - \tau, t)$, we can write the above equation like the convolution of a time-varying impulse response with the input signal:

$$y(t) = \int_{-\infty}^{\infty} h(\tau, t)x(t - \tau)d\tau, \tag{5.2}$$

where $h(\tau, t)$ is the time-varying impulse response at time t caused by an impulse applied τ time units earlier. This is a convenient notation, since in case of a time-invariant impulse response, Equ. 5.2 becomes the convolution integral known for Linear time-invariant (LTI) systems. The commutative property of the convolution integral is also valid in the time-varying case, so:

$$y(t) = \int_{-\infty}^{\infty} h(t - \tau, t)x(\tau)d\tau, \tag{5.3}$$

Zadeh [Zad50] showed that one can apply a Fourier transform to the impulse response $h(t, \tau)$ with respect to τ and obtain a time-varying transfer function:

$$H(\omega, t) = \int_{-\infty}^{\infty} h(\tau, t)e^{-j\omega\tau}d\tau \tag{5.4}$$

So the input-output equation can be written in the frequency domain as:

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(w)H(\omega, t)e^{j\omega t}d\omega \tag{5.5}$$

5.1.3 Electro-Larynx Speech Production

We look at the speech production mechanism used in EL speech. See Fig. 5.2 for a model, which is closely related to the source-filter model of laryngeal speech.

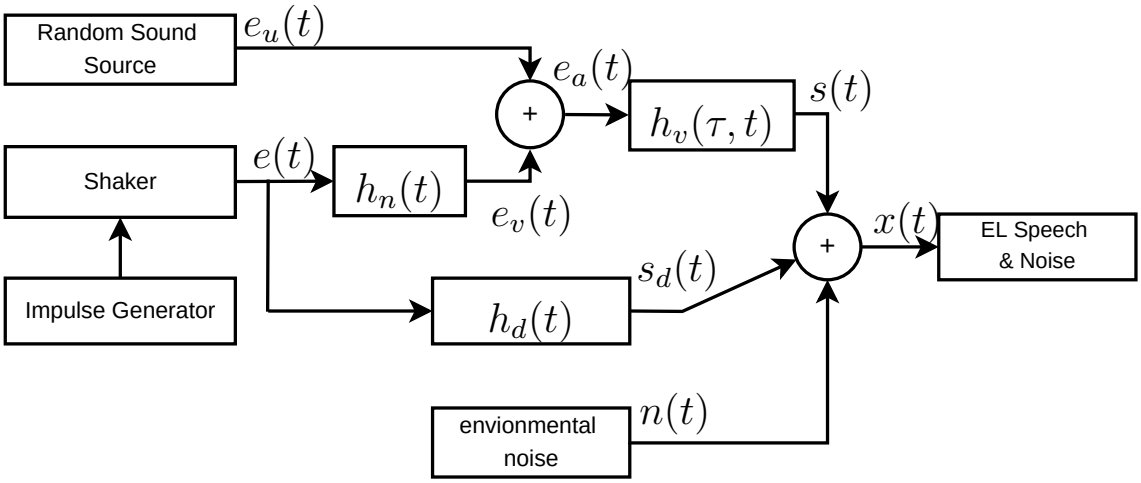


Figure 5.2: *EL speech model. The speech signal $s(t)$ is a superposition of the EL source signal $e(t)$ filtered by the time-invariant neck impulse response $h_n(t)$ and the time-varying vocal tract impulse response $h_v(\tau, t)$ and a random sound source $e_u(t)$ convolved with $h_v(\tau, t)$. In addition to the speech sound, the EL sound $e(t)$ convolved with the time-invariant EL direct path impulse response $h_d(t)$ and environmental noise $n(t)$, which is not related to speech production, arrive at the listener's ears.*

The difference to normal speech production is that the voicing source is produced outside the human body by the EL device. Due to imperfect coupling between EL and neck tissue, only a fraction of the EL energy is transferred through the neck tissue with impulse response $h_n(t)$ into the vocal tract. This yields the signal, $e_v(t)$, which excites the vocal tract. Although the neck impulse response depends on several parameters, such as the exact placement of the EL or the pressure of the EL on the neck, it is safe to assume that during a speech fragment it does not change much, so that $h_n(t)$ can be regarded as time-invariant,

$$e_v(t) = \int_{-\infty}^{\infty} e(t-\tau) h_n(\tau) d\tau. \tag{5.6}$$

Unvoiced sounds $s_u(t)$ are produced by the patient with the usual mechanisms of healthy subjects by using the limited air volume available in their mouth $e_u(t)$. Though, limitations exist for glottal sounds, such as [h]. The positions of the sound sources, $e_u(t)$, $e_v(t)$ are assumed to be fixed and the same. The sum of $e_v(t)$ and $e_u(t)$ is filtered by the vocal tract impulse response $h_v(\tau, t)$, which is modulated by articulatory movement. An unvoiced/voiced switch, which is often seen in the widely used source-filter model for speech production, is not included, because only excellent EL speakers manage to turn off the EL for unvoiced sounds. For reasons of

simplicity, the lip radiation is included in the vocal tract impulse response, $h_v(\tau, t)$. This results in the speech signal $s(t)$.

$$s(t) = \int_{-\infty}^{\infty} \left[e_v(t-\tau) + e_u(t-\tau) \right] h_v(\tau, t) d\tau. \tag{5.7}$$

The part of the energy of the EL signal $e(t)$, which is not transferred into the vocal tract, is directly radiated into the environment and reaches the listener on a direct path with impulse response $h_d(t)$. The amount of the leaked energy depends on the position of the EL on the neck and the contact pressure. As discussed above, we assume those factors to be changing slowly compared to the articulatory modulation. In addition, the room impulse response is also changing the directly radiated EL source signal. It is save to assume that the room impulse response is also changing slowly compared to the speed of articulatory movement. The filtered leakage sound then is $s_d(t)$. This direct sound disturbs the perception of EL speech.

$$s_d(t) = \int_{-\infty}^{\infty} e(t-\tau) h_d(\tau) d\tau \tag{5.8}$$

In addition to the directly radiated sound of the EL device, an environmental noise component, $n(t)$, has also to be considered. It includes every additional noise source not correlated with the speech signal. For EL speakers, the stoma noise is part of this signal. Because breathing and speaking is decoupled, the stoma noise is not correlated with the EL speech signal.

The acoustic signal, $x(t)$, which is picked up by a microphone or perceived by a human listener can be summarized as:

$$x(t) = s(t) + s_d(t) + n(t) \tag{5.9}$$

We can also re-organize the EL-speech model such that the voiced and noise-like speech components have separate signal paths (Fig. 5.3), i.e.,

$$x(t) = s_u(t) + s_v(t) + s_d(t) + n(t). \tag{5.10}$$

The noise-like speech generation is straight-forward and can be calculated by solving the convolution integral of $e_u(t)$ and $h_v(t, \tau)$:

$$s_u(t) = \int_{-\infty}^{\infty} h_v(t-\tau, t) e_u(\tau) d\tau \tag{5.11}$$

For the voiced signal part, it is convenient to merge the impulse response of the neck, $h_n(t)$, and the vocal tract, $h_v(\tau, t)$, into one neck-vocal tract impulse response, $h_{nv}(\tau, t)$. The voiced speech part can then be calculated as:

$$s_v(t) = \int_{-\infty}^{\infty} h_{nv}(t-\tau, t) e(\tau) d\tau \tag{5.12}$$

In this model we see that the EL signal is transmitted to the listener or microphone over two different paths. In one path the EL signal is filtered in a time-varying

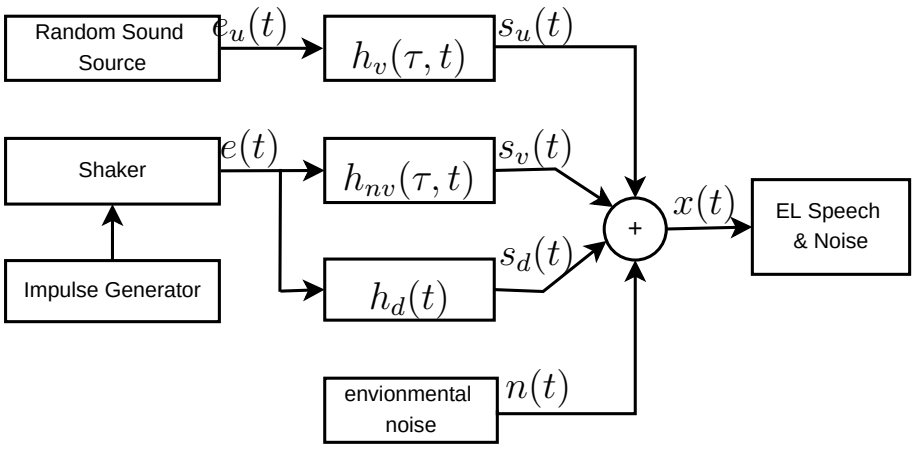


Figure 5.3: *EL speech model. The EL signal $e(t)$ produces a time-varying voiced speech signal $s_v(t)$ and a time-invariant directly radiated EL sound $s_d(t)$.*

manner to produce voiced speech sounds, $s_v(t)$ and through the other one the EL sound $s_d(t)$ is directly radiated to the listener.

5.2 Previous Approaches

Since the DREL sound $s_d(t)$ is a major issue with EL speech, there have been some attempts to reduce this signal part. We present previous approaches to reduce the DREL sound.

5.2.1 Adaptive Filter

It has been proposed to be reduce the EL direct sound by an adaptive filter using an LMS algorithm [EWCM+98] and independent component analysis (ICA) [NWWL03]. This adaptive filter is based on a signal and noise model. A primary microphone picks up the signal, $y(t)$, assuming that

$$y(t) = s_u(t) + s_v(t) + s_d(t), \tag{5.13}$$

where $s_v(t) + s_u(t)$ is the speech signal and $s_d(t)$ is the DREL sound which is the EL sound $e(t)$ convolved with the impulse response $h_d(t)$ as in Equ. 5.8. A reference microphone picks up the DREL signal, $s_d(t)$. The adaptive filter tries to minimize the error signal $a(t) = y(t) - z(t)$, where $z(t) = e(t) * h_a(t)$ by modelling the impulse response $h_d(t)$. If the adaptation is successful then $a(t) = s_v(t) + s_u(t)$ (see Fig. 5.4)

It was stressed in both papers that the adaptation has to be controlled because of the correlation between the noise signal $e(t)$ and the speech + noise signal $y(t)$. Indeed the adaptive filter is based on a signal model for which the speech signal and the interferer are statistically independent.

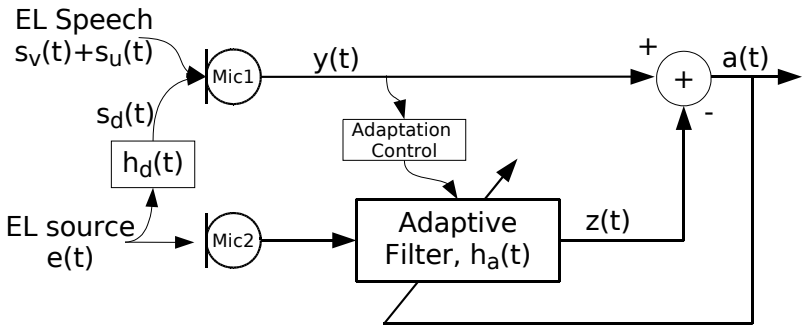


Figure 5.4: *Adaptive filter noise reduction. The adaptive filter coefficients $h_a(t)$ are modified to minimize the error signal $a(t)$.*

In the following we discuss the problem in more detail considering the input signals to the two microphones in the framework of the EL speech production model (Fig. 5.5).

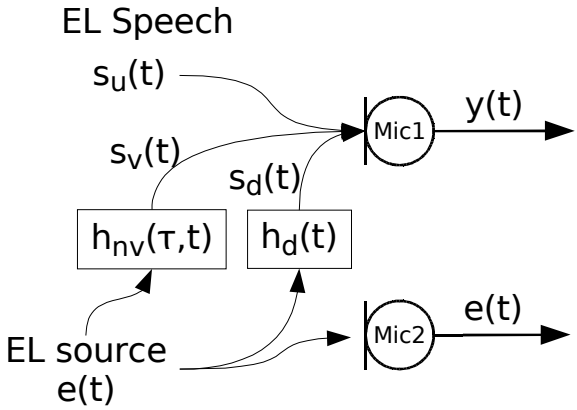


Figure 5.5: *Adaptive filter noise reduction input considering the EL speech production model.*

The EL speech sound $s_v(t)$ and the directly radiated noise of the EL $s_d(t)$ are correlated. The adaptive filter does not only approximate $h_d(t)$ but $h_d(t) + h_n v(t)$. To overcome this problem, the authors introduce an adaptation control that stops the adaptation of the filter in case of sonorant sounds (i.e. sounds produced without turbulent airflow or bursts). In case of unvoiced sounds and strong background noise the authors found out that the correlation between the directly radiated EL sound and the EL speech signal is very low; therefore the adaptation is meaningful. Both papers report a significant improvement of the filtered speech over the original. Niu *et al.* also report a significant increase of acceptability on a Mean Opinion Score (MOS) scale.

5.2.2 Spectral Subtraction

The directly radiated noise of an EL is only slowly varying and, therefore, Cole *et al.* [CSMG97] have applied spectral subtraction and root cepstral subtraction noise suppression to EL speech. The spectral subtraction method is based on estimating the noise power spectrum and then subtracting this spectrum from the signal power spectrum. The noise is estimated during non-speech intervals. They report an improvement of the processed speech compared to the original EL speech.

Subsequent approaches have used several variants of the spectral subtraction method [PBBL02, LZWW06b, LZWW06a]. One of the proposals is based on the following.

$$|\hat{S}(\omega)|^2 = \begin{cases} |Y(\omega)|^2 - \alpha \left|\hat{L}(\omega)\right|^2, & \text{if} \frac{\left|\hat{L}(\omega)\right|^2}{|Y(\omega)|^2} \leq \frac{1}{\alpha+\beta} \\ \beta \left|\hat{L}(\omega)\right|^2, & \text{otherwise} \end{cases}, \tag{5.14}$$

where $\hat{S}(\omega)$ is the enhanced speech spectrum, $\hat{L}(\omega)$, the noise power spectrum estimate and $Y(\omega)$ the noisy speech spectrum. $\alpha(\alpha \geq 1)$ is the subtraction factor and $\beta(0 \leq \beta \leq 1)$ the spectral noise floor. To further enhance the speech sound, Liu *et al.* [LZWW06b, LZWW06a] include auditory masking, which does a weighting of the noise by means of a Linear Prediction Coefficient (LPC) based weighting filter. The noise estimation is based on minimum statistics.

The authors report improved MOS rates for EL speech without and with background noise consisting of Gaussian white noise or babble noise. A follow-up paper to the spectral subtraction by Pandey *et al.* [PBBL02] introduced a quantile based noise estimation for EL speech [PPL03]. The noise estimation is similar to minimum statistics, but does not use the minimum of the spectral bins over the observation period, but a certain quantile over the observation period.

5.2.3 Cepstrum Based Processing

The unnatural sound of the EL speech is due to the unnatural excitation signal of the EL device. One proposal was to replace the EL excitation signal with a natural excitation signal obtained via cepstral deconvolution [MDEWM99]. The DREL and EL speech excitation is removed via cepstral deconvolution and replaced by the natural one. A drawback is that there is always need for a reference sentence spoken by a healthy speaker, which is then used to substitute the excitation signal of the EL speaker. A real-world application is, therefore, not feasible.

5.2.4 Notch Comb Filter

Since the frequency of the EL is usually constant, the most obvious approach would be to use a notch filter to filter out the DREL sound. In practice this has to be a notch comb filter since, not only the fundamental frequency, but the harmonics

have to be cancelled, too. The first problem that arises, is that the fundamental frequency has to be determined exactly, otherwise the higher harmonics do not get cancelled exactly. An adaptive interference canceler or an adaptive notch comb filter could follow the actual fundamental frequency in case of any drift. If the notch filter sits exactly on the harmonics of the EL sound, then not only the DREL sound, but also the speech signal is degraded, because the EL harmonics are the carriers for the modulation which is performed by the articulatory organs. Because of the remaining sidebands due to the modulation, the speech signal is still intelligible, but since the harmonics (carriers) are missing it sounds like whispered speech, which is not necessarily desired. See Fig. 5.6 for a spectrogram of an original speech utterance with EL and the same utterance processed with a notch filter.

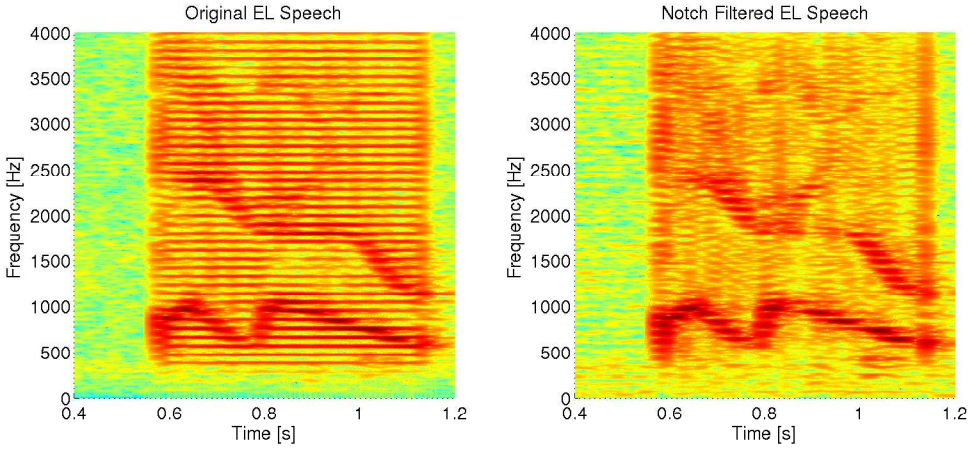


Figure 5.6: *Top: Spectrogram of the EL speech phrase, 'Hello'. Bottom: Spectrogram of notch comb filtered EL speech.*

5.2.5 Discussion

The state-of-the art approaches for DREL sound suppression suffer from a main problem. The underlying signal model assumes independence of the speech signal and the disturbing signal. In case of the DREL sound and voiced EL speech this assumption cannot hold, since both have the same physical signal source. To make the above approaches work, heuristics are necessary to obtain useful results. Another problem is the on-and-off switching of the EL, which results in an abrupt change of the DREL noise condition. The results of perceptual tests are difficult to compare, since no standards exist. Different testing protocols are used in different publications. We therefore develop a method to reduce the DREL sound that takes into account the model introduced in figures 5.2 and 5.3. Furthermore, we propose a testing protocol, which might serve as a basis for future standardization in this domain.

5.3 Multi-Path Separation

Our approach takes advantage of the different properties of the EL speech sound and the DREL sound. To simplify the problem, we only consider the EL sound source and omit unvoiced sounds and environmental noise. The EL sound is transmitted to the listener via two separate paths. In Fig. 5.7, the previously introduced EL speech model is shown for only those signal paths which are taken by the EL sound source. As 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, this signal is only modulated at a very low frequency 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.

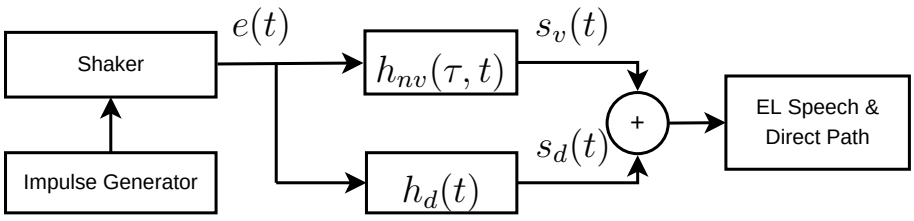


Figure 5.7: *Simplified EL speech model. EL signal $e(t)$, neck-vocal tract impulse response $h_{nv}(\tau, t)$, EL direct path impulse response $h_d(t)$, EL speech sound $s_v(t)$, directly radiated EL signal $s_d(t)$.*

5.3.1 Modulation Filtering

Time-varying path:

First, we consider the path through the vocal tract, where the EL sound source is modulated and filtered by a time-varying vocal tract. Previous research on human speech has shown that signal components with a low modulation frequency can be suppressed without loss of intelligibility. While Drullman *et al.* [DFP94] have shown that modulation frequencies above 4 Hz are important for speech intelligibility, later studies have come to the conclusion that there is no loss of intelligibility if all the modulation frequencies are suppressed below 1 Hz [APHA96]. Experiments with the evaluation of the error rate of an automatic speech recognition system have confirmed this value [KAHP99]. Band-pass filtering in the modulation spectrum domain has previously been proposed as a means of channel normalization for acoustic pre-processing in Automatic speech recognition (ASR) and noise reduction [HM94, HWA95].

Time-invariant path

To suppress a time-invariant signal in the modulation frequency domain, we can place a notch filter at a modulation frequency $f_n = 0\,\text{Hz}$. In practice, it is safe to assume that the modulation frequency of the DREL noise, f_m, is below 1 Hz. Therefore, to suppress the very slowly varying direct EL signal path, we apply a high-pass filter with $f_c = 1\,\text{Hz}$ on the modulation spectrum of the speech signal corrupted by the DREL sound.

Modulation filtering framework

The filter, that suppresses the slowly varying components, is applied in the modulation frequency domain (see [Sch07] for an-indepth study of modulation domain filtering). The modulation frequency domain filter is, in principle, a filterbank with a detector that separates the modulating signal $m[n]$ and carrier $c[n]$. The modulating signal can then be processed with a filter for every frequency band. After filtering, the modulating signal and carrier are multiplied again and the reconstruction filter bank yields the overall modified speech signal (Fig. 5.8)

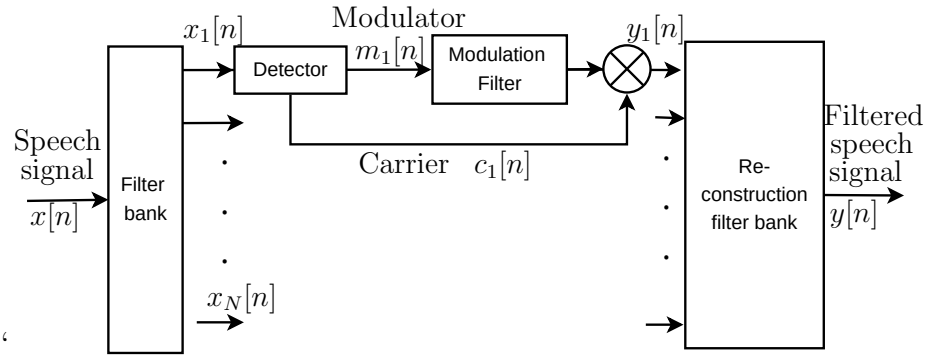


Figure 5.8: *Temporal filtering of the modulation spectrum. A filter bank, e.g., based on the STFT, is used for the conversion of the speech signal into multiple bandpass channels. The modulator signal, which has to be detected from a specific bandpass channel of the the filterbank output can then be filtered to get the desired signal.*

5.3.2 Implementation of the Modulation Filtering

The speech signal is processed at a sampling rate, $f_s = 16kHz$, and is first high-pass filtered with a cutoff frequency, $f_c = 80Hz$, to remove additive low-frequency noise. The next step is an STFT of the speech signal $x[n]$:

$$X[l,k] = \sum_{n=-\infty}^{\infty} x[n]w[lR-n]e^{-j\frac{2\pi kn}{N}}, \tag{5.15}$$

for $k = 0, \ldots, N-1$, $w[n]$ is an analysis window of length N and l is a frame index, R the hopsize. We then represent the signal $X[l,k]$ in terms of magnitude and phase:

$$X[l,k] = A[l,k]e^{j\phi[l,k]}, \tag{5.16}$$

where $A[l,k] = |X[l,k]|$ and $\phi[l,k] = \arg(X[l,k])$. Then we assign the magnitude to the modulating signal, $m[l,k] = A[l,k]$ and $c[l,k] = e^{j\phi[l,k]}$ to the carrier. The filter is then applied to the time trajectories of the modulator $m[l,k]$, for every frequency bin k. The carrier is not modified and used directly for resynthesis. While the application of the high-pass filter to magnitude $A[l,k]$ itself reduces the DREL sound somewhat, better suppression is achieved when the modulating signal is compressed with a static nonlinearity before filtering. Previously [HWA95], $A_c[l,k] = A[l,k]^{2/3}$ has been proposed. Since our application is different, we have performed informal listening tests based on different compression laws. The tests and the results are described later in section 5.3.3.

To reduce the DREL sound, all compressed spectral bins are filtered in time with a 1^{st} order Butterworth high-pass filter with a cutoff frequency $f_c = 1\ Hz$ (Fig. 5.9). The output of the filter may be negative. Therefore, negative values of the filtered modulating signal are replaced by small positive random values, since such negative values only appear for small magnitudes. After filtering, the modulator $A_{cm}[l,k]$ is again expanded $A_m[l,k] = A_{cm}[l,k]^{3/2}$ and, using the original phase, $Y_m[l,k] = A_m[l,k]e^{j\phi[l,k]}$ is transformed back into a time domain signal $y[n]$ by the inverse STFT using the overlap-add method [AR77]. We see this operation in the following equation.

$$y[n] = \sum_{l=-\infty}^{\infty} \delta[n - lR] \frac{1}{N} \sum_{k=0}^{N-1} Y[l,k] e^{j\frac{2\pi kn}{N}} \tag{5.17}$$

The inverse fourier transform of every $Y[l,:]$ is calculated and placed at the position determined by frame index l and hopsize R. The time-domain signal $y[n]$ is determined by adding the overlapping frames.

We have observed high power level changes in EL speech owing to the on/off switching of the EL. These are sometimes amplified by the multipath separation method. Dynamic range compression at the output reduces this problem considerably.

To simplify the problem, we have above reduced the signal representation to the two paths driven directly by the EL device as shown in Fig. 5.7. Now we revert to the complete signal model and analyze how the method (Fig. 5.9) affects other signal components. The actual speech signal is better described by the more complete model in Fig. 5.3. Unvoiced speech $s_u(t)$ has the same temporal modulation properties as voiced speech $s_v(t)$, so it is processed like voiced speech. For the environmental noise $n(t)$, we have not made any assumptions. The noise is treated according to its temporal properties, slowly varying noise will be suppressed and noise with a modulation frequency similar to the speech signal will be treated like

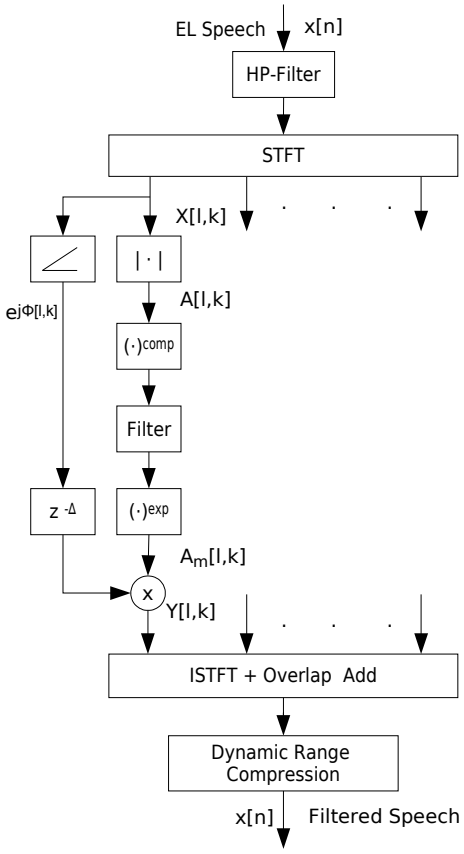


Figure 5.9: *Detailed signal flow graph for modulation filtering of the speech signal including compression and expansion of the modulator and final dynamic range compression.*

speech. So for many environmental noise types, our approach serves as implicit noise suppression system. Having said this, noise suppression is not discussed further in the remainder of this thesis, since the focus is on the suppression of the DREL sound.

5.3.3 Parameter Optimization

In this subsection, we investigate the influence of some key parameters on system performance. Given the impracticality of performing formal listening tests for every potential parameter combination, a simple two-stage evaluation was used. A mean segmental Signal-to-noise ratio (SNR) value was calculated as the ratio of the signal including the DREL sound and an estimation of the DREL sound averaged over all measured frames in the log domain (Equ. 5.18). The final decision was made by informal listening tests. For this evaluation, a single sentence from the database described in the next section was used. Non-speech parts were removed manually. The DREL sound power level is estimated with the minimum statistics approach

proposed in [Mar01]. The measurement uses a signal token with the EL turned on all the time, i.e., the noise level is at the level of the DREL sound.

$$SNR_{seg} = \frac{1}{K}\sum_{i=1}^{K}\left(10\log\left(\frac{\sum_{\kappa=1}^{N}(s(\kappa+iN)+s_d(\kappa+iN))^2}{\sum_{\kappa=1}^{N}s_d^2(\kappa+iN)}\right)\right), \tag{5.18}$$

where K is the number of frames with length N.

Modulation filter impulse response length. To get an insight into the modulation filter impulse response length, we replaced the first-order high-pass Infinite impulse response (IIR) filter with an Finite impulse response (FIR) filter. The truncated impulse response of the IIR filter was taken as the FIR filter coefficients. The shortest impulse response length sufficient to achieve a value of the SNR above which it could not be improved further was determined.

We have calculated the mean segmental SNR depending on the modulation filter impulse response length from 0 ms to 0.4 ms. A impulse response length of $0ms$ is equivalent to not applying any modulation filter. In addition, the length of the analysis window has been varied. The resulting plot is shown in Fig. 5.10. The figure shows that increasing the filter length above 0.25 seconds does not improve the SNR. This is qualitatively the same for different frame sizes. Informal listening tests have confirmed that an impulse response length of 0.25 s is sufficient and larger values do not significantly suppress the DREL sound more.

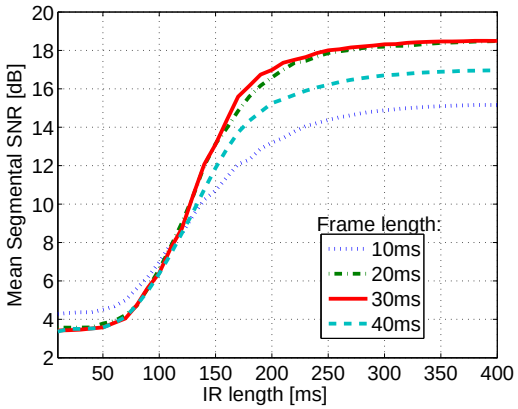


Figure 5.10: *Mean segmental SNR depending on the modulation filter length. We observe that an increase of the modulation filter length above 250 ms does not yield a significant increase of SNR anymore.*

Analysis frame and hop size. In addition, segmental SNR evaluation has been performed to determine the optimal analysis frame length and hop rate. The analysis

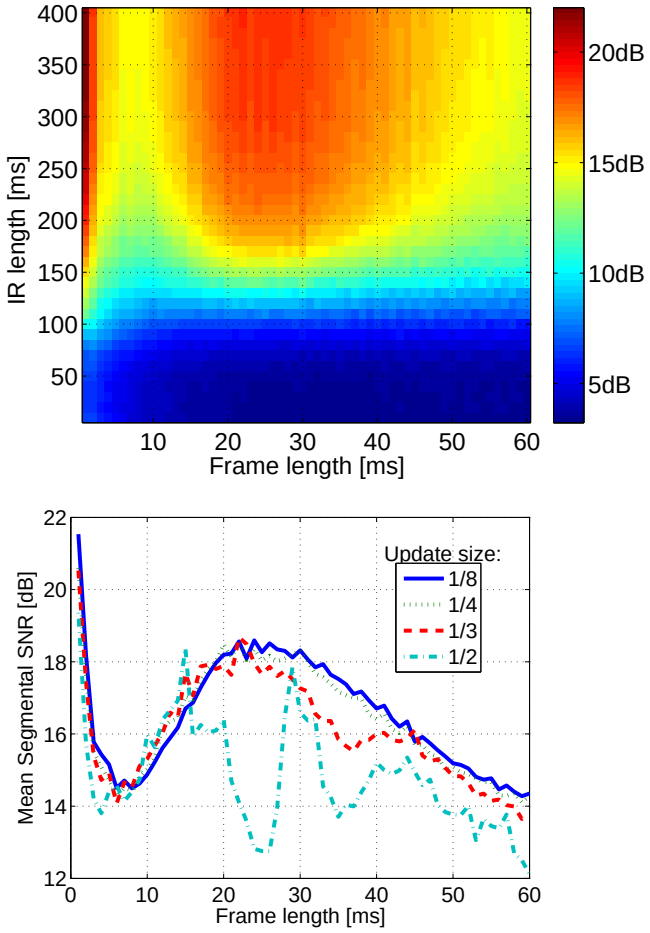


Figure 5.11: *Top: Mean segmental SNR depending on the analysis frame length and impulse response (IR) length. One sees the optimal area of operation at frame lengths of* $20 - 30ms$ *and IR lengths above 200ms.*
Bottom: Mean segmental SNR depending on the analysis frame length and hop size. One sees a flat peak at frame lengths of 20-30 ms and the best suppression of the DREL sound at hopsizes of 1/4 and 1/8 of the frame length. Framelengths up to 10 ms provide no meaningful results in terms of informal listening.

frame length varied between 1 ms and 60 ms and the hop size was varied from 1/8, 1/4, 1/3 to 1/2 of the frame length.

The mean segmental SNR values can be seen in Fig. 5.11. The SNR peak at a frame length of 1ms is an artifact only of the measurement method. Listening tests do not show any improvement in the signal quality for such small frames. This is due to the fact that such small frame length are in the order of the fundamental period of the signal. We see a smaller peak at a frame length of 15ms, which is identical to the fundamental period of the EL signal. This peak is only visible for large hop sizes. For frame lengths between 20ms and 30ms, the mean segmental SNR has a flat

maximum at hop sizes of 1/8 and 1/4 of the frame length. For hop sizes of 1/2 and 1/3 of the frame length, a considerable decrease of the segmental SNR is seen. For these large hop sizes, the observation of SNR peaks at multiples of the fundamental period suggests using a frame lengths related to the EL pitch. However, we have observed that this pitch is not stable for some devices. We, therefore, recommend to use a frame length of approximately 25ms and a hop size of 1/4 of the frame length.

Modulator Signal Compression. As mentioned above, the suppression of the DREL sound is better, when the modulating signal is compressed prior to filtering. We have varied the compression exponent from 0.01 to 1 (no compression). We applied the same segmental SNR measure as described above, see Fig. 5.12 for results. In opposition to previous evaluation results, the calculated optimal segmental SNR value does not predict the best compression exponent suggested by informal listening tests. While with increasing compression of the modulating signal (exponent decreasing down from 1.0 to 0.23), the suppression of the DREL sound does increase, but at the same time also the speech signal gets suppressed as well. Based on informal listening, a compression exponent, comp = 0.4, is chosen as the best value to be formally assessed in later listening tests.

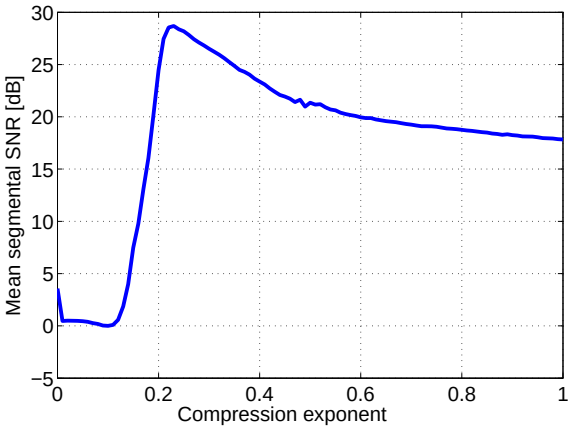


Figure 5.12: *Mean segmental SNR depending on the compression of the modulator signal. When the compression exponent equals* 0*, no modulation domain filtering has been performed.*

5.4 Formal Performance Evaluation

Finally, we evaluate whether the proposed suppression of the direct EL path offers any improvement for EL speech communication.

We perform this evaluation either using objective measures of quanitative signal properties, or subjective listening tests, in the framework of which human listeners are asked to judge a given speech token.

5.4.1 Speech Database

Quality assessment in speech processing requires speech data for the given problem area. Since no EL speakers database is publicly available, we collected our own.

Four laryngectomized and ten trained healthy subjects were recorded while using an Electro-Larynx. The recorded material was edited, i.e., silent periods at the beginning and end of an utterance were removed and recordings that were unfit for further processing, e.g., because the subjects were not able to produce the desired utterance, were discarded. A set of 450 utterances were kept for the listening test. This adds up to about 3 minutes of speech for each of the speakers. For more detailed information on the recording conditions, refer to appendix B.

5.4.2 Objective Evaluation

In the previous sections, we have presented mean segmental SNR values for the original and enhanced speech utterance selected for parameter optimization. For these parameters, we observe an improvement, when evaluating the complete database, of approximately 15-20dB, depending on the quality of the original sound file. In a spectrogram we observe a significant visible reduction of the DREL sound (Fig. 5.13).

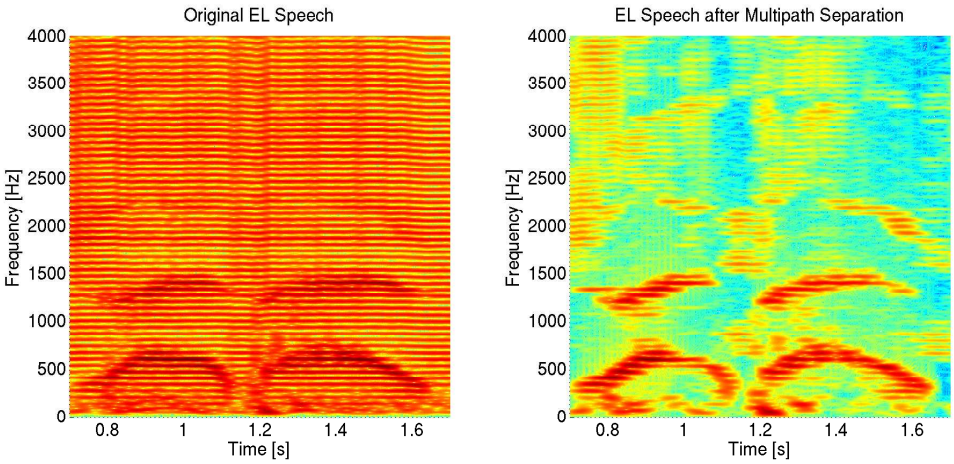


Figure 5.13: *Spectrogram of EL speech utterance in German 'zwei drei'. Left: Original. Right: Signal after multipath separation. One can see the strong DREL sound characterized by time-invariant harmonics in the left plot and a significant reduction of the direct path from the EL in the right plot.*

SNR measures are standard in the literature, so this measure was included to allow comparison with other publications. Since the receiver of the speech output is a human listener, SNR values and other so-called objective measures must always be complemented by subjective assessment.

5.4.3 Subjective Evaluation

Even for unprocessed alaryngeal voices, no widely accepted method exists to evaluate the quality of speech, though there is work going on in that direction (e.g. [MMVdB+06, MMCB+06]). Specific protocols are necessary because the usual assessment methods are valid for laryngeal speech only. Therefore, we have designed a test, which fits the special requirements of evaluating alaryngeal speech.

Subjective speech quality evaluation has great importance for the rating of algorithms and transmission systems used in speech processing. Depending on the application, different procedures are useful and several standards have evolved. For example, the International Telecommunications Union - Telecommunications standardization sector (ITU-T) standard P.800 [ITU96] covers subjective determination of speech transmission quality. The Speech Quality MOS scale is not usable for EL speech, because compared to laryngeal speech, it will always be rated in the lowest MOS categories.

EL speech is in some respect similar to (early) speech synthesis systems, so it makes sense to also consider the ITU-T recommendation P.85 for speech output systems [ITU94]). While this recommendation uses absolute category rating (ACR) the suggested questions and answers are crafted explicitly for artificial speech. The most relevant questions in our context concern the *listening effort*, which is also defined in the P.800 recommendation, and *voice pleasantness*.

5.4.4 Design of the Listening Test

Due to the above mentioned problems, we decided to use a comparison category rating (CCR) instead of an ACR test protocol, which compares a modified EL speech token with a reference unprocessed EL speech token.

Spectral Subtraction. To have a reference, we not only compare the output of the multipath separation with the original EL speech, but also with a state-of-the-art general noise suppression algorithm. We choose an algorithm that has achieved the best results in conventional speech enhancement tasks and is well documented in the literature. We choose the nonlinear log-spectral amplitude minimum mean square estimator of the noise as proposed in [EM85], with a decision-directed estimation of the a priori SNR [EM84] combined with a quantile-based approach [SFB00] and a time-varying smoothing factor [Mar01].

Test questions. The unprocessed EL speech is compared to an EL sound token after multipath separation. Three questions are asked. First, we ask for a judgment of the overall impression of the processed EL speech utterance compared to the original. Second, since the DREL sound makes the EL speech harder to understand, we ask whether the listening effort is affected by the multipath separation. If processed EL speech is easier to understand, this would ease communication with EL speakers. Finally, we want to know whether the suppression of the DREL sound

out-weights a possible introduction of processing artefacts, which occur frequently in speech enhancement. This leads to the following sets of questions and opinion scores:

General Impression: How is the overall quality of the second sample compared to the quality of the first?

Speech Quality:	Score
much better :	3
better :	2
slightly better :	1
about the same:	0
slightly worse :	– 1
worse :	– 2
much worse :	– 3

Listening Effort: What is the effort to listen to and understand the second speech sample compared to the first?

Listening Effort:	Score
much less :	3
less :	2
slightly less :	1
about the same:	0
slightly more :	-1
more :	-2
much more :	-3

Background Noise: How do you perceive the background noise of the second speech sample compared to the first?

Background Noise :	Score
much better :	3
better :	2
slightly better :	1
about the same :	0
slightly more annoying:	-1
more annoying :	-2
much more annoying :	-3

The subjects hear an original token and a processed token or vice versa, in random order. The three questions are presented simultaneously and the subjects can listen to the tokens three times at most (Fig. 5.14).

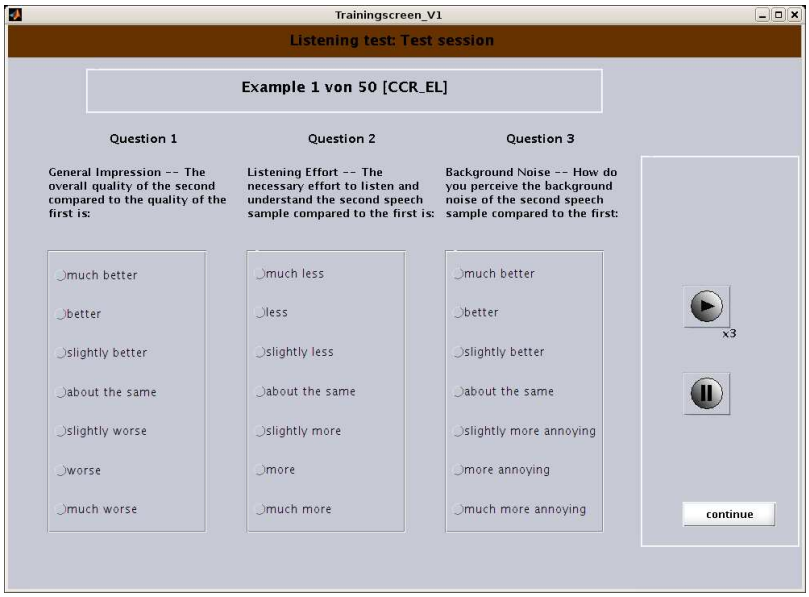


Figure 5.14: *User interface for the listening test.*

Listeners. Most of the listeners had no previous or only little experience with EL speech. There were 4 female and 10 male listeners. The mean age was 32 years, all of them were native German speakers and had normal hearing abilities. The reliability of the listeners was checked by introducing null-pairs, where the same sample was presented twice within a comparison pair. Two subjects did not identify the null-pairs, so they were excluded from the evaluation, leaving 12 listeners, 3 female and 9 male. The listeners used high quality studio headphones (AKG K-271) and were asked to adjust the volume to a comfortable level.

5.4.5 Results

The results of the listening tests were analysed in several ways. Assuming a normal distribution of the comparison opinion scores, estimates of the comparison mean opinion scores (CMOS), i.e., the mean $\hat{\mu}$, and the standard deviation $\hat{\sigma}$ were calculated. For every mean, a confidence interval CI_{95}, was also determined. The results are presented in Fig. 5.15.

While a small improvement concerning the background noise is visible for both systems, the perceived overall quality is lower for the modified tokens when compared to the original EL speech utterances. This is true for both the multipath and the spectral subtraction method. A significant difference concerns the listening effort, for which the spectral subtraction method performs significantly worse than the multipath separation method. Furthermore, a high standard deviation of the test results is noted. Analysis of variance Analysis of variance (ANOVA) [Lin74] revealed that a major contribution to the high variance is a high inter-listener variability. For additional analysis, the results were split into two listener groups of equal size. One

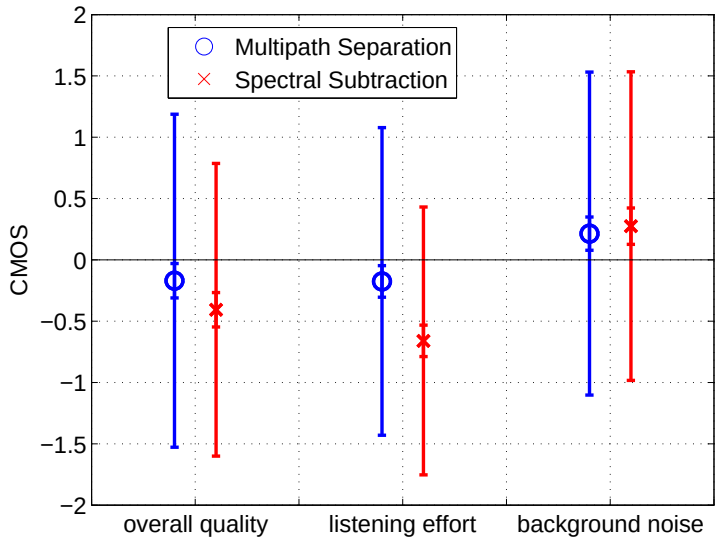


Figure 5.15: *Comparison Mean Opinion Score (CMOS) for Multipath Separation (∘) and Spectral Subtraction (×) compared each to the unprocessed signal for three evaluation qualities. Shown are the estimated mean $\hat{\mu}$, the 95% confidence interval CI_{95} and the estimated standard deviation $\hat{\sigma}$.*

group consisted of listeners who rated the modified tokens rather positively, the other group preferred the original tokens. The results for the three test questions are shown in Fig. 5.16.

If we assume that two groups of listeners exist having different preferences, the results may be interpreted differently. For the multipath separation method, one group perceives the reduction of the directly radiated EL sound positively and express the opinion that the listening effort is slightly less than for the original toekn. This is also reflected in the overall perceived quality. Spectral subtraction and multipath separation are rated equally with regard to background noise.

The sound of EL speech is unusual for most of the listeners. To investigate whether the familiarity of the listeners with EL speech affects their judgement, they were partitioned into two groups, the speech and language therapists (SLT, 3 subjects) and the rest. Results (see Fig. 5.17) show that the SLTs judge the modified tokens to be worse compared to those listeners who were not previously exposed to EL speech. They even judge the background noise of the processed signals worse than the background noise of the original signals. This may be due to their long experience with EL speech, which they do not perceive as an uncomfortable sound anymore.

We have used both laryngeal and alaryngeal speakers in our database, so the question arises about the relevance of the results of the laryngeal speakers with respect to the alaryngeal cases. The results for both groups are presented individually and compared in Fig. 5.18. All qualities are judged similar for both speaker groups, though one has to take into account the larger confidence interval, especially for the

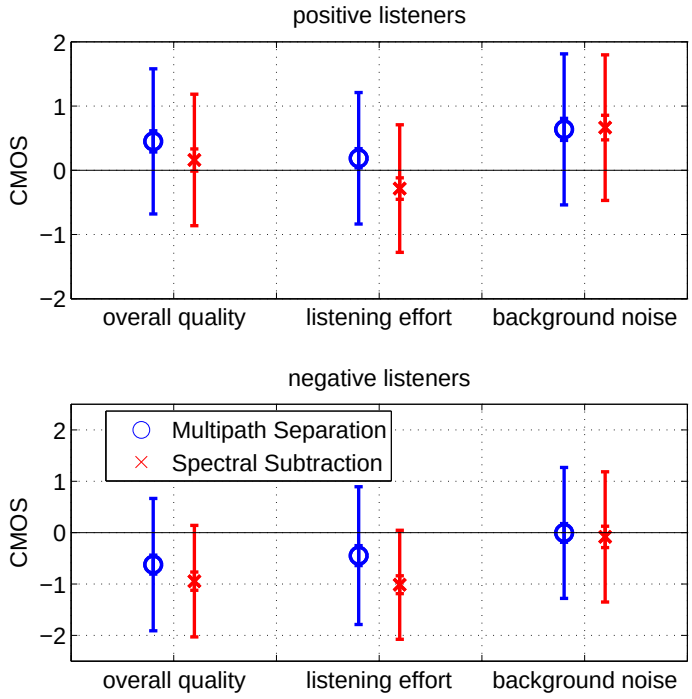


Figure 5.16: *Comparison Mean Opinion Score (CMOS) for Multipath Separation (∘) and Spectral Subtraction (×) compared to the unprocessed signal for three evaluation qualities. Top: Listeners who prefer the modifications. Bottom: Listeners who prefer the original. Shown are the estimated mean $\hat{\mu}$, the 95% confidence interval CI_{95} and the estimated standard deviation $\hat{\sigma}$.*

alaryngeal results, due to the lower sample size.

5.4.6 Discussion

The analysis of the listening test gave mixed results with a very high standard deviation. Reasons may be the following. First, it is difficult for listeners to evaluate voices that are not part of their usual listening experience. So far, not even the medical community has agreed on an evaluation procedure for substitution voices [MMVdB+06]. In current evaluation methods, inter- and intra-judge variations are a well-known problem. Second, while we believe that the modulation filtering approach is a helpful method to remove the DREL sound, the current implementation introduces some distortion of the EL speech signal itself, which is rated negatively by some of the listeners.

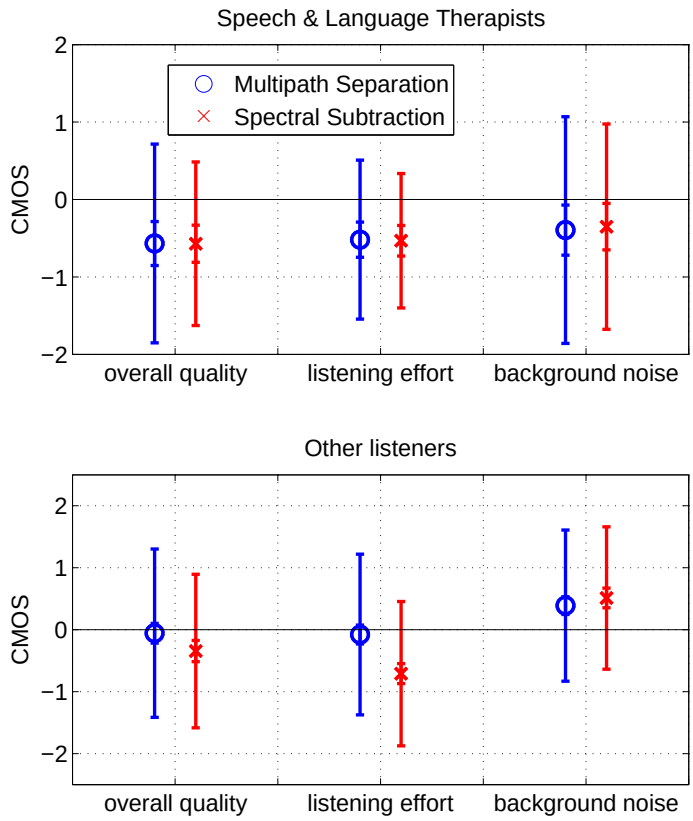


Figure 5.17: *Comparison Mean Opinion Score (CMOS) for Multipath Separation (∘) and Spectral Subtraction (×) compared to the unprocessed signal for three evaluation qualities. Top: Speech and language therapists. Bottom: Other listeners. Shown are the estimated mean $\hat{\mu}$, the 95% confidence interval CI_{95} and the estimated standard deviation $\hat{\sigma}$.*

5.5 Conclusion

We have presented a model for the production of Electro-Larynx speech. Based on this model, we have presented an approach to separate the different signal paths of EL speech. The principle is to take advantage of the different temporal properties of the two main signal paths. The time-invariant path can be suppressed to obtain an improved EL speech signal.

We observe that only some of the listeners appreciate the modifications of the signal positively. For those listeners, the perceived overall quality rating reflected the positive comparison ratings of both listening effort and background noise.

We do not know yet why other listeners preferred the original to the modified EL speech token. Most listeners were not familiar with EL speech, which is very different from modal speech. This may have influenced the rating of the listeners.

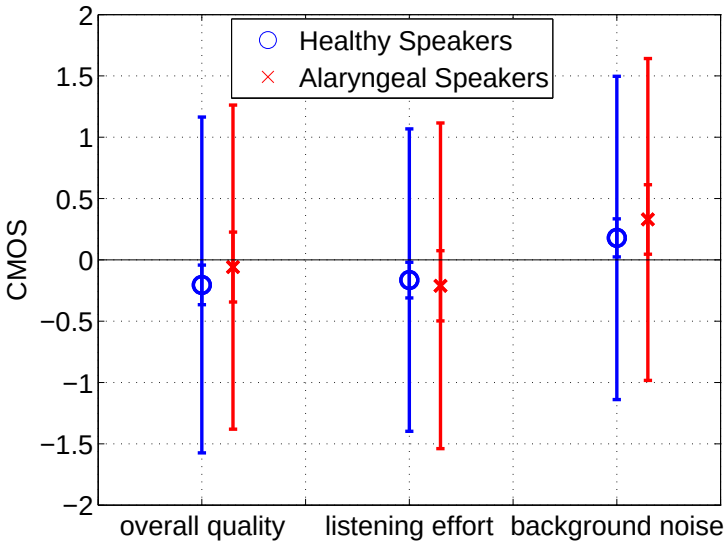


Figure 5.18: *Comparison Mean Opinion Score (CMOS) for healthy (∘) and alaryngeal (×) speakers for three evaluation qualities for the multipath separation method. Shown are the estimated mean $\hat{\mu}$, the 95% confidence interval CI_{95} and the estimated standard deviation $\hat{\sigma}$.*

The similar rating of the spectral subtraction method with respect to the background noise at the prize of increased listening effort and decreased overall quality can be explained with the fact that this method uses a signal model that assumes statistical independence or at least linear uncorrelatedness of the speech and noise signals. This condition is clearly violated in the EL case for which both the desired signal and the DREL sound are driven by the same physical signal source. Therefore, the spectral subtraction algorithm not only modifies the noise signal but tends to modify the speech signal more than our method.

We believe that using a signal model that best represents the underlying speech production mechanism is a requirement for a satisfactory solution of a speech processing problem. A weakness of our current approach is that the magnitude spectrum or modulating signal is always positive, i.e., it has a DC component. Therefore, a highpass filter in the modulation frequency domain also decreases all the values of the magnitudes of any spectral component. Some may therefore become negative, which has unwanted consequences.

Recently, a new approach to modulation filtering has been proposed by [Sch07]. The author argues that in case of using the magnitude and phase of the filterbank outputs as modulators and carriers, signals are not band-limited anymore. Any modification in one sub-band then also affects the neighboring bands, which can introduce artifacts. Therefore, one has to apply a new envelope detection approach that limits the bandwidth of the resulting signals. The application of those findings could potentially improve the quality of the presented multipath separation method.

For a broader outlook, one has to consider that all discussed signal enhancement approaches have a limited scope in daily life applications. Both the proposed approach and the state-of-the-art methods have in common that a reasonable application can only be found in the context of telecommunications, where speaker and listener are not connected acoustically or where the direct acoustic signal path is negligible. This would apply to telephone conversations and also when using public address systems. The suppression of the directly radiated EL sound for face-to-face communication is not possible by signal enhancement methods alone, it would involve a redesign of the EL device, as previously proposed by [HHKM99] (see Section 2.4.4).

Chapter 6

Prosody for Alaryngeal Speech

Alaryngeal speech – and in particular – Electro-Larynx (EL) speech suffers from inadequate prosody. This chapter explores a possibility for superimposing an (a posteriori) intonation contour on a speech signal, when the fundamental frequency is either not or only partially measurable as in voices using the substitute folds, or is constant as for EL speech. We will see that the formants can be used to calculate an artificial F_0 contour that sounds natural. In addition, the formants can also be used to accentuate words in a sentence, which can be translated into an F_0 peak at that word.

From an application point of view, a pitch contour indirectly derived from the speech signal of an alaryngeal speaker may improve conversation over a telephone channel by inserting the pitch modification into the signal path. For EL speakers, a system for generating an artificial fundamental frequency contour, can be used to control the EL excitation frequency directly. This may be even made more effective, because with this direct feedback the speaker could be enabled to learn to use the formants as a means to convey prosodic information. This application is not limited to telecommunication situations.

In the introductory chapter, we have already discussed earlier approaches to prosody reconstruction (see section 2.4.1, p. 10). We first give an introduction to speech prosody and discuss the relations of prosody and speech formants, before we evaluate other possibilities of generating pitch for alaryngeal voices.

6.1 Introduction to Speech Prosody

Scientists of different backgrounds have carried out research on prosody for a long time. This subject has been of interest for linguistics, speech technology, speech pathology and foreign language acquisition, among others. Unfortunately, no commonly accepted terminology has been developed. Therefore, we start with a short introduction and definition of terms related to prosody (mostly according to [NP92] and [Hag01], if not noted differently).

Prosody organizes supra-segmental features, i.e., features which are not related to a single location in a speech utterance (e.g., specific phones). Its domain of interpretation is well beyond phone boundaries, concerning syllables, words, phrases and sentences. Prosody describes the temporal evolution of the relationships of amplitude, duration and fundamental frequency of speech. It provides cues for syntactic information (segmentation, resolving ambiguity, conversational structure), and paralinguistic information about the speaker. Various suprasegmental features are parts of prosody, such as intonation, rhythm (duration, pauses, tempo), rate, accentuation and timbre. Here, we only cover the most important acoustic features.

6.1.1 Acoustic Correlates of Prosody

Fundamental frequency. Intonation is the variation of the fundamental frequency F_0 in a speech utterance. It is the most important part of prosody. Fundamental frequency is the best scientifically covered property of prosody, possibly because it is easy to measure compared to some of the other features mentioned above. In almost every language, most utterances have a downward trend after the first stressed word, called declination. It is normally reset at major syntactic boundaries. This effect is assumed to be correlated with the declining air pressure in the lungs [tHCC90, p. 121ff]. Lower F_0 peaks at the end of a phrase are, therefore, perceived as strong as higher peaks at the beginning of a phrase. This is because F_0 peaks are perceived relative to the declination contour.

Intensity. Intensity is related to the loudness of a speech signal. Accentuated syllables are usually perceived as louder. Whereas in Germanic languages loudness is important for placing accents, in Romance languages there are almost no changes in intensity in accentuated syllables.

Rhythm. Rhythm integrates such effects as speaking rate, phone and syllable duration and pauses. It is difficult to measure with automatic approaches, although it is an important feature.

6.1.2 Role of Prosody in Speech

The information conveyed by prosodic features can be subdivided into three groups, linguistic, para-linguistic and extra-linguistic. The former two are communicative, whereas the latter is not communicative but informative, see Tab. 6.1 (from [Mix98]).

The linguistic features refer to the way a message is formally coded and organized into units of a language. They correspond to the surface structure of the message on a still rather abstract level. The actual meaning of the message can often not be decoded without also interpreting the paralinguistic communication. An ironical undertone, for example, can change the meaning of an utterance completely.

communicative		informative
linguistic (lexical, syntactic, semantic)	para-linguistic	extra-linguistic
segmentation, sentence mode, disambiguation, discourse organization	speakers intention, attitude	gender, age, regional and social background, voice disorder

Table 6.1: *Information conveyed by intonation (from[Mix98]).*

Syntax: One basic function of prosody is to organize speech utterances into phrases and sentences, which help the listener to process speech in smaller units than the whole speech flow. It also signals the function of a phrase or sentence such as a question or imperative and also the syntactic structure as main or subordinate clause. An important role of prosody is solving ambiguity.

1. 'Vielleicht. Am Montag bei mir. Paßt das?'
   'Maybe. On Monday, at my place. Is that OK?'

2. 'Vielleicht am Montag. Bei mir paßt das.'
   'Maybe on Monday. That's possible for me.'

Here two equal sets of words get a different meaning through prosodic variation, i.e. it is prosody that dissolves the ambiguity (expample taken from [NBK+97]).

Accentuation - Stress: Another role of prosody is the placement of accents on words and phrases. Thus a sentence with the same words and different prosody placing the focus on different words can get completely different meanings. Syllable accent can even have lexical importance. In German there are a few minimal pairs of words which are segmentally equivalent and only distinguished by the position of their syllable accent (e.g. 'umgehen' (to handle) vs. 'umgehen' (to avoid) [Mix98]). Stressed syllables are longer, louder, and/or have F0 patterns that cause them to stand out against unstressed syllables.

Speaking style: Intonation is heavily influenced by the speaking style. Main categories are read, narrative and spontaneous speech as well as formal versus familiar.

Personality: Prosody is a very personal characteristic; gender, age-group, health and sometimes even vocational cues can be communicated via prosody [NP92]. For instance, in earlier research, we have studied the prosodic differences of regional variants of German [BDH+04, Hag01].

6.1.3 Conclusion for Alaryngeal Speech

The total lack of or the lack of control over some prosodic features is a major shortcoming of alaryngeal speech. It has already been mentioned in the introduction (see. Sec. 2.3) that a natural pitch contour is perceived as most beneficial to the perceived quality of the EL substitution voice [MH05]. Earlier studies have shown that the flattening of the F_0 contour yields a lower intelligibility of sentences spoken by healthy subjects [LW99, LB03]. The limited control over the speech rhythm, due to the necessary inhalation of air in the esophagus makes Esophageal (ES) speech less natural. For a real-time enhancement of alaryngeal speech prosody the improvement of rhythm is out of reach. This would mean a massive intrusion into the timing of the speech, which is not possible for real-time processing.

We focus on the introduction of a fundamental frequency contour as a means for improving the prosodic quality of alaryngeal speech.

6.2 Relations Between Perceived Pitch and the Speech Spectrum in Unvoiced Speech

The question is, can an alaryngeal speaker use speech features other than fundamental frequency to convey prosodic information? If so, can we use those features to synthesize a fundamental frequency contour to reinforce prosodic information?

In this section we investigate the relations between pitch and the speech spectrum, when no fundamental frequency is present. We consider previous research and own experiments.

6.2.1 Whispered Pitch

As early as 1956, research on the perception of pitch in whispered speech has found, that even without voicing, in a tone language like Chinese, there is no significant loss in intelligibility when whispering instead of using voiced speech [ME56]. In tone languages lexical distinctions may be based on the pitch contour.

Meyer-Eppler [ME57] measured the formant frequencies of vowels, which were whispered with different pitches. He summarizes his research on the realization of vowels in whispered speech as follows:

> Spectrographic analysis of whispered vowels and words shows that there exist two substitutes for pitch movements, which in voiced speech are used to indicate different prosodic features. The whispered vowels [e], [i], and [o] substitute spectral noise for pitch, whereas [a] and [u] possess some formants whose position changes with the intended "pitch".

Thomas [Tho69] performed a different experiment regarding pitch in whispered vowels. Listeners used a sinusoidal tone generator to match the perceived pitch of a whispered vowel. Results showed that the subjects always choose the perceived

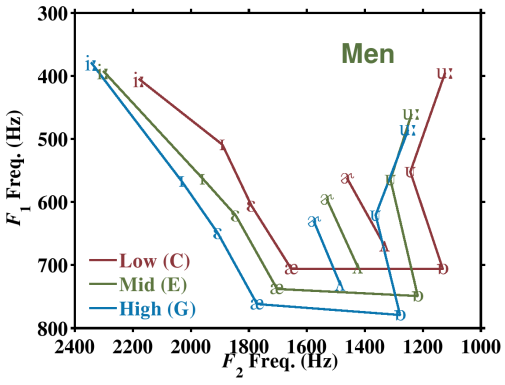


Figure 6.1: *Formant frequencies for male vowels at intended pitch (from [Kie04]).*

formant frequency of F2 as the perceived pitch, with a maximum error of 4%. The intra-subject variability of the chosen frequency was only 2.2 %.

Holmes *et al.* [HS83] suggested that this relation between pitch and formants also exists in voiced speech. To get more natural sounding speech synthesis he proposed to adjust the formant frequencies according to the fundamental frequency.

Another study with pitch perception in whispered speech [HM99] showed that an up/down change of F1 and/or F2 induces a corresponding pitch perception change. While there is a high correlation of pitch perception and formant frequency changes by $\geq$ 40 Hz of either F1 or F2, the correlation raises to even higher values when changing F1 and F2 simultaneously.

A recent study [Kie04] has shown that formants F1 and F2 move depending on the intended pitch of sung whispered vowels (see Fig. 6.1).

6.2.2 Alaryngeal Pitch

A previous study on alaryngeal pitch accent revealed that alaryngeal speakers are able to convey prosody [vRdKNQ02]. The study included esophageal and tracheo-esophageal speakers, who placed the emphasis in a sentence according to the surrounding context, e.g., 'The ball did not fly over the wall; the ball flew over the FENCE' vs. 'The shoe did not fly over the fence; the BALL flew over the fence'. Listeners had to repeat the utterance placing the emphasis at the perceived location. While the study showed that alaryngeal speakers are able to convey emphatic stress, the analysis of the accoustic features showed inconclusive results. One feature which was not included was the speech formants.

Similar to whispered speech, in EL speech there is no pitch contour that can be modulated. While in whispered speech no harmonic signal exists at all, in EL speech F_0 is constant. This of course evokes a primary pitch perception, which is at a constant level. Any pitch-like sensation which should for example place an accent on words would have to override this primary pitch. In opposition to whispered speech, there have been no studies concerning the relations between formants and

perceived pitch. We therefore investigate if pitch accent can be conveyed in EL speech.

Alaryngeal Singing. We have carried out an informal experiment to explore whether the findings for whispered speech are reproducible for EL speech. To simplify the problem and avoid the variability of the formants in running speech, a sung passage of a single syllable has been chosen. A speech and language therapist has sung a short musical scale – a fifth up and down – on the syllable 'la' with her normal voice, whispering and using an EL (Fig. 6.2). In addition, a male subject was asked to sing a scale up an octave. For whispered and EL speech, both subjects were asked to produce the musical scale as good as possible. The scale was to be sung at a comfortable pitch, when applicable.

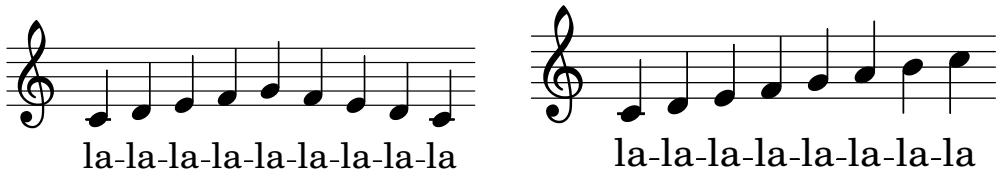


Figure 6.2: *Short musical scales.*

A plot of the spectrogram and the first two formants can be seen in Fig. 6.3 and 6.4. The plots show the up and down scale sung with laryngeal voice (top left), whisper (top right)and EL voice (bottom). In both the whispered and the EL utterance F2 is following the scale up and down. In the male example, also the first formant shows some correlation with the intended pitch. In the laryngeal voice, the second formant shows no movement related to the pitch change. It has previously been claimed, that in a female modal voice the first formant follows the harmonics of the fundamental frequency [ML95a]. We would rather explain this as an effect of the method used to determine formants. For female voices only few harmonics are available, so the LPC estimator locks on the harmonics of the spectrum.

In addition to healthy subjects, alaryngeal speakers were also asked to sing the scale up and down (Fig. 6.2). Most of the patients were not able to perform this task. Only one subject was able to produce a musical scale (see Fig. 6.5). There can be two reasons for this failure. First, not all laryngeal speakers were able to produce a sung EL utterance. One common feature of the successful subjects was their musicality, they were either speech therapists or audio engineers. Second, most of the successful subjects had some time for training, while non of the subjects who where not able to produce a pitch change had the time for training.

Another reason may be that in whispered speech the larynx position might be involved in producing the perceived pitch contour [JCB02]. While movement of the larynx is possible in EL speech of speakers with a larynx, the alaryngeal speaker has no possibility to use the larynx as a means to change the vocal tract configuration. It may, therefore, be harder for an alaryngeal speaker to modify the formants in a

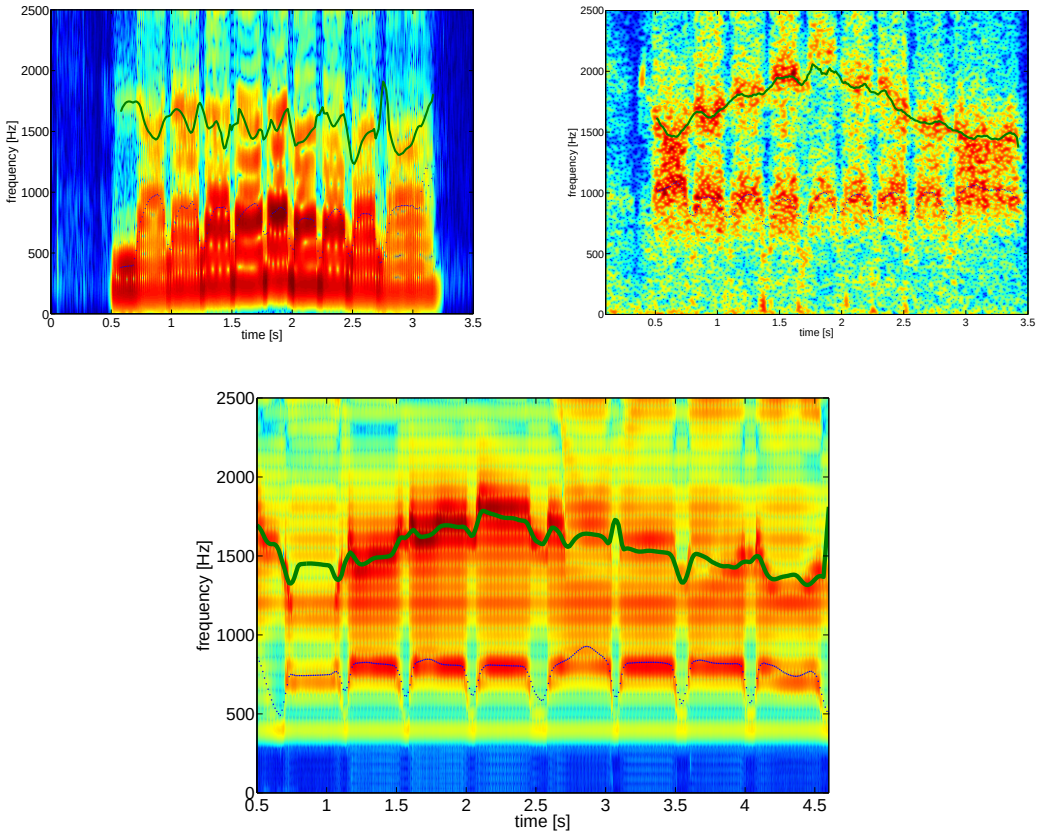


Figure 6.3: *Spectrogram of a musical scale sung by a female person, with superimposed tracks of the formants F1 and F2. Top left: Laryngeal. Top right: Whispered. Bottom: Electro-Larynx.*

way that results in a perceived melody. Against this point can be held that there was one alaryngeal speaker who was able to produce a scale-like utterance (Fig. 6.5).

In Fig. 6.4 one can see that the method to produce the scale is very similar to overtone singing [Kli93]. In opposition to overtone singing, where the formant bandwidth is limited to one harmonic, the bandwidth of the formants are wider.

Emphasis placed on specific words. Another informal experiment was to test whether an EL speaker can place an emphatic accent on specific words in a sentence. The sentence 'Die Katze jagt eine Maus' was recorded three times with an accent placed on either 'Katze', 'jagt', or 'Maus'. In Fig. 6.6 we see a plot of each version of the sentence with a plot of the spectrogram, orthographic transcription and the time-domain signal. In the top figure the word 'Katze' is emphasized and we can see that for the vowel 'a' both F1 and F2 have higher values than in the other two versions of the utterance. The same is true for the word 'jagt' in the middle figure and the word 'Maus' in the lower figure. In addition we can see that the emphasized

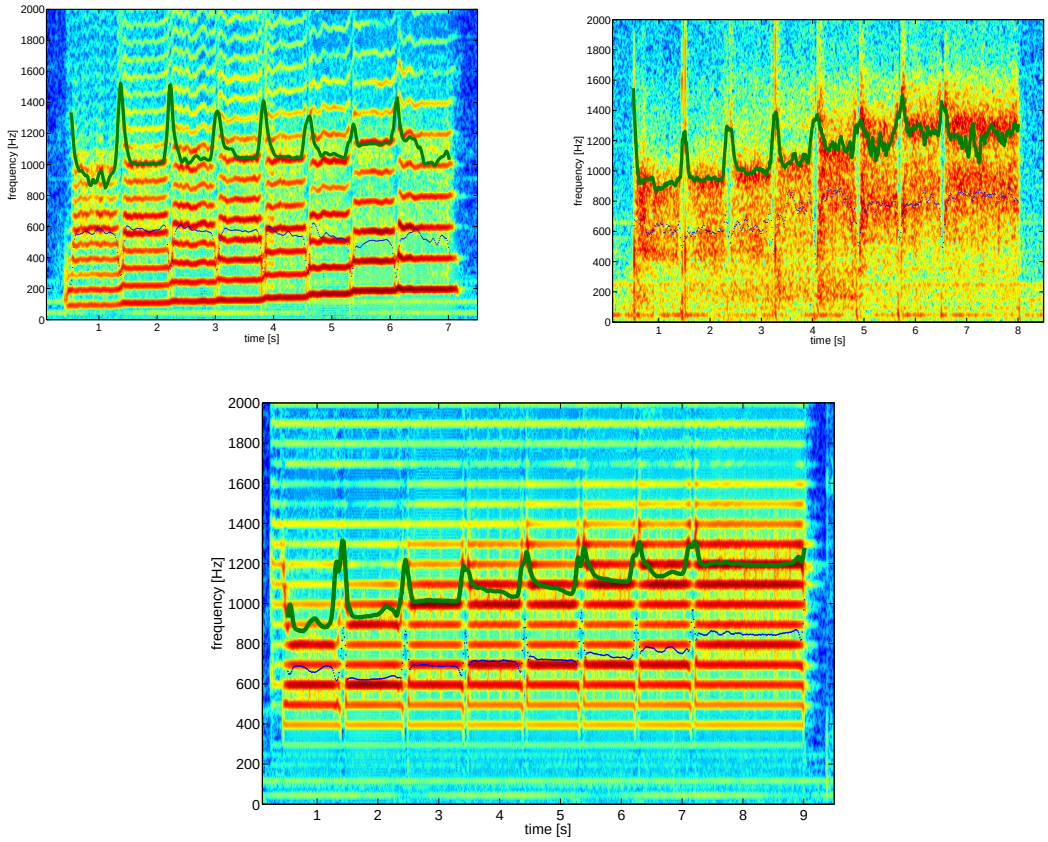


Figure 6.4: *Spectrogram of a musical scale sung by male person with superimposed tracks of the formants F1 and F2. Top left: Laryngeal. Top right: Whispered. Bottom: Electro-Larynx.*

words have higher energy. Most importantly, the location of the accent can not only be seen in the spectrogram, but clearly heard. This suggests that the formants are a cue for alaryngeal prosody. This has not been investigated in the previously mentioned study on alaryngeal pitch [vRdKNQ02].

Question vs. Declaration. The same sentence structure (word order) was used to either produce a question or a declaration. The task to convey the intended prosody (of question vs. declaration) using formants is much more difficult for both laryngeal and alaryngeal EL speakers. This is possibly because the pitch contour for a question is quite complex and even for healthy laryngeal speaker not easy to produce unambiguously. Therefore the difference cannot be seen in the spectrum and the formant tracks as in the previous example.

At least in German and English this is a rather artificial example, because questions and declarations are usually distinguished by using a different word order.

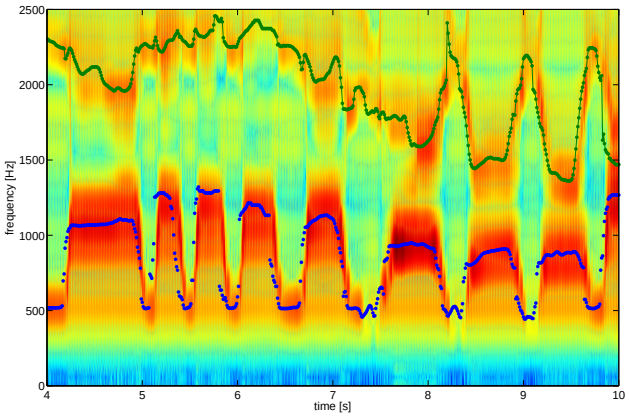


Figure 6.5: *Spectrogram of a musical scale sung by a male alaryngeal speaker using an EL showing F1 and F2 tracks.*

Hypothesis. Based on the above findings, we assume that in alaryngeal speech the formants can be used to convey prosodic information that in modal speech is carried by the F_0 contour. This may be possible because vowels qualities are not only perceived as such when the formants are at a specific frequency, but are defined in a rather broad area of F1/F2 combinations (Fig. 6.7) [PB52]. Prosodic information carried by formants may be limited, however.

Also, we must assume that it is possible to calculate an artificial pitch contour based on formants.

Objection by Source Filter Theory. A possible objection to this hypothesis is that a widely accepted model of speech production, which is the so-called source-filter model, is based on the assumption that source and filter are independent from each other. Therefore, it is not possible to calculate a feature which belongs to the source by analysing the filter. As Fig. 6.3 and 6.4 suggest, this is true for laryngeal utterances. While we see no relation between formants and F_0 in the laryngeal speech utterance, a correlation is observed between the musical scale and second formant F2 in EL speech. As there is no source fundamental frequency, which the human can control, he uses the formants instead to convey at least part of the information that usually is carried by the fundamental frequency contour. So actually, we are not calculating a source feature via the filter, but we take advantage of the fact that a human is able to switch for some prosodic information from the source to the filter.

The following section presents a way to introduce an artificial pitch contour to alaryngeal speech. We focus on EL speech.

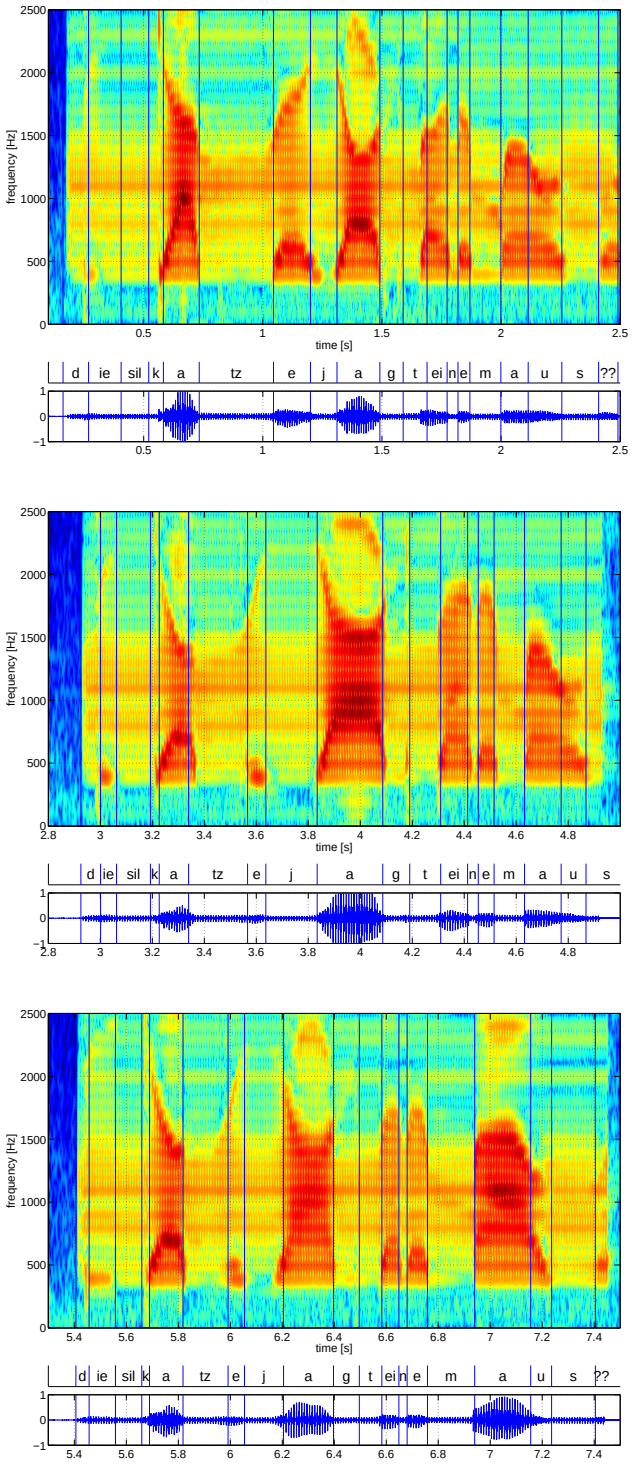


Figure 6.6: *Spectrogram, orthographic transcription and time-domain signal of the sentence 'Die Katze jagt eine Maus' with emphasis on different words spoken by an EL speaker. Top: Katze Middle: jagt. Bottom: Maus.*

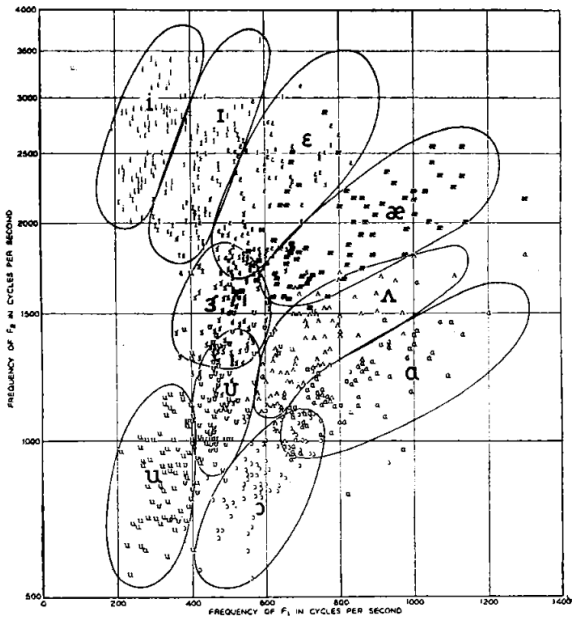


Figure 6.7: *Vowel formant chart of American English (from [PB52]).*

6.3 Pitch Contour from Non-Source Speech Features

The approach rests on the online calculation of an artificial pitch contour from other speech features. This F_0 contour can then be used to either control the EL or post-process the EL speech in real time (Fig. 6.8).

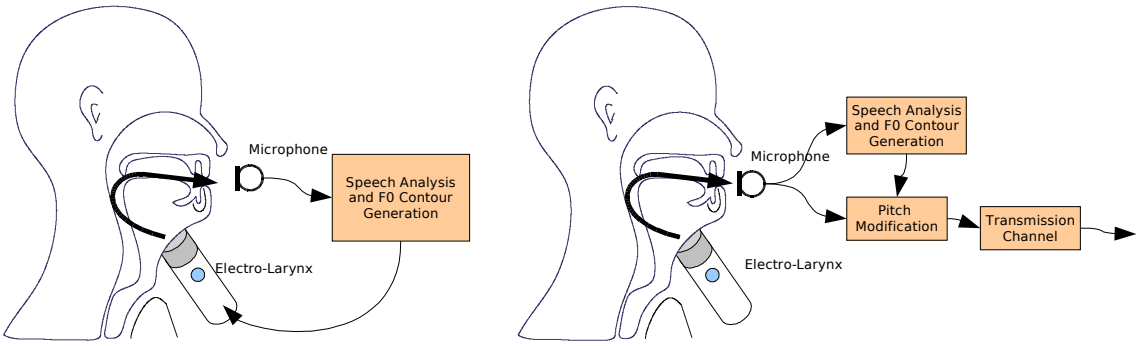


Figure 6.8: *Applications of Pitch Generation. Left: F_0 control of EL. Right: F_0 modification of EL speech for telecommunication.*

Based on the observation of the influence of formants on the perception of prosody, we propose to use the formants as the main information source for the generation of a pitch contour.

6.3.1 Pitch Contour Generation

Given the EL speech production model (see Ch. 5.1.1), we analyze the time-varying impulse response $h_v(\tau, t)$, or equivalently the time-varying frequency response $H_v(\omega, t)$ to determine the speech formants using a parametric spectral estimation method, e.g., Linear prediction (LP) analysis (Fig. 6.9).

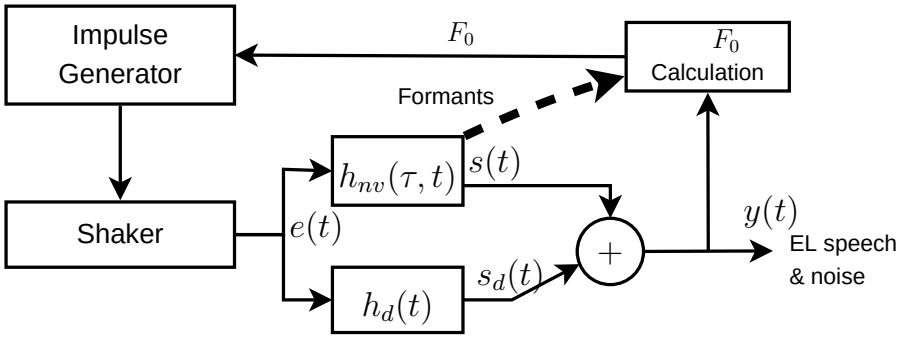


Figure 6.9: *EL speech model. EL signal $e(t)$, time-varying neck-vocal tract impulse response $h_{nv}(\tau, t)$, time-invariant EL direct path impulse response $h_d(t)$, directly radiated sound of EL $s_d(t)$, speech signal $s(t)$. An F_0 contour is calculated by analysing the EL speech signal and determining the formants as shaped by $h_{nv}(\tau, t)$.*

Since we are not able to measure $s(t)$ alone, but $y(t)$, which is the superposition of the EL speech sound $s(t)$ and the Directly radiated electro-larynx (DREL) sound $s_d(t)$, what we estimate is the combination of the impulse responses $h_{nv}(\tau, t)$ and $h_d(t)$. The effect of which is that the formants will be less distinct in the spectrum since the DREL sound provides a steady background noise. We therefore use the multipath separation method, described in the previous chapter, to reduce this noise. (see Fig. 6.10).

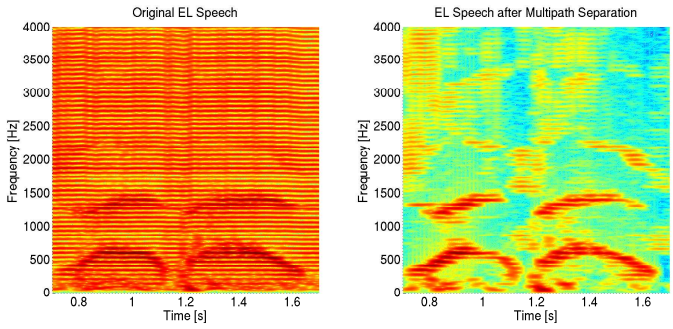


Figure 6.10: *Spectrogram of EL speech. 'zwei drei'. Left: Original. Right: Signal after multipath separation. One can see a significant enhancement of the speech formants.*

Even under ideal conditions, formant tracking is not an easy task. Determining the correct formant tracks is an ongoing research question. As in other areas of speech processing, such as fundamental frequency determination, the many facets of a speech signal make an accurate determination of the formant tracks very difficult. In our case, this is not so much of a problem, since no direct relation exists between the F_0 and formants contours. In fact, the formants are of importance only for speech units that the speaker emphasizes to convey prosodic information.

The question is, given the formant tracks, how to generate a pitch contour. Artificial pitch generation has long been a topic in speech synthesis and research has produced a number of pitch models for human speech production. Unfortunately, for our task, all those models assume to know the complete utterance before generating a contour. A model that is relevant to contour generation is the Fujisaki model [FH84], which has also been extensively studied for German [Mix98]. This model assumes that the pitch contour is a superposition of two components, a word level accent component and an utterance level phrase component (see Fig. 6.11).

While we cannot use this model as is, due to the above mentioned condition of complete knowledge, we can build our pitch contour as a superposition of a short-time word accent and a long-time phrase component. The phrase component models the declination, which has been described previously. For the word accent component, the formant track is scaled to a pitch contour value. The declination $F_p(t)$ is modelled with an exponential scaled by the default frequency F_t, of the EL signal:

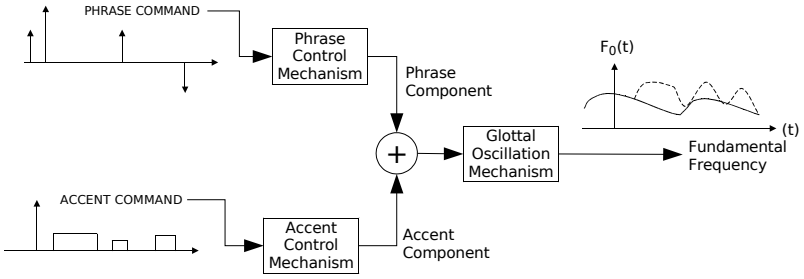


Figure 6.11: *Fujisaki Model. The fundamental frequency is assumed to consist of two components: A phrase component, which changes over a speech phrase and an accent component, which is located on accented words.*

$$\text{EL is} \begin{cases} \text{off:} & F_p(t) = F_t * (1 - e^{-\tau_{rise} t}) \\ \text{on:} & F_p(t) = F_t \; e^{-\tau_{fall} t} \end{cases} \tag{6.1}$$

The time constants, τ_{rise} and τ_{fall} are chosen so that there is a rapid rise and a slow fall of the phrase component. The actual values depend on the switching pattern of the speaker. A starting point would be $\tau_{rise} = 5$ and $\tau_{fall} = 0.1$ After nonlinear and linear smoothing, the second Formant F_2 is rescaled to fit the desired pitch range, e.g., $\pm 30Hz$. The word accent component is modelled as follows.

$$F_a(t) = (F_2(t) - mean(F_2(t)) \cdot \beta + \gamma, \tag{6.2}$$

where β and γ are speaker dependend constants, which are chosen so that the formant frequency is transformed into the default pitch range. The final F_0 contour is calculated by adding the accent and phrase components:

$$F_0(t) = F_a(t) + F_p(t) \tag{6.3}$$

An example of the superposition of phrase component $F_p(t)$ and accent component $F_a(t)$ (offset by mean pitch) resulting in the final pitch $F_0(t)$ can be seen in Fig. 6.12. The phrase component is controlled by the EL activity.

6.3.2 Implementation of Artificial Intonation in EL Voices

To explore the benefits of an artificial pitch contour, we implemented a pitch modification system for telecommunication applications. The method can be embedded in any state-of-the-art signal analysis - resynthesis approach. See Fig. 6.13 for a general framework.

For the actual implementation, several issues have to be dealt with. Those are summarized in Fig. 6.14.

Earlier, we have proposed the multi-path separation approach to remove the directly radiated EL sound from the speech sound prior to pitch modification. The reason is not only the above-mentioned increase in reliability in formant tracking, but also that a variable directly radiated EL sound has been shown to be still more annoying.

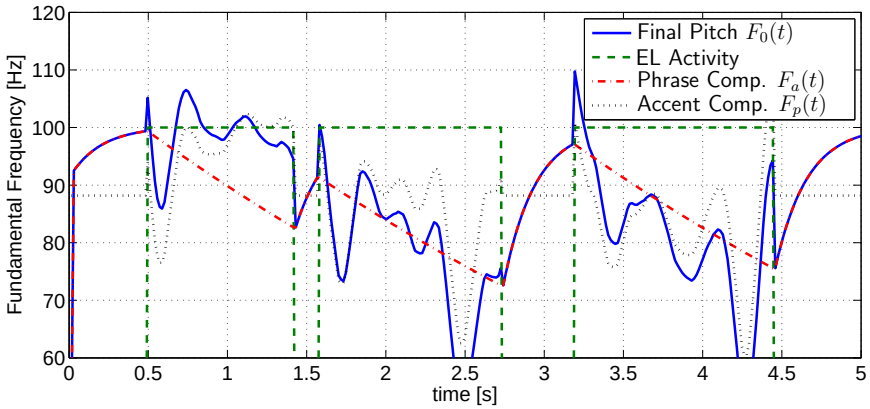


Figure 6.12: *Superposition of phrase component $F_p(t)$ and accent component $F_a(t)$ (offset by mean pitch) gives the final pitch $F_0(t)$. The phrase component is controlled by the EL activity.*

Pitch modification. Concerning pitch modification, we have implemented several possibilities. Inspired by a work to improve esophageal speech [LB06], the voiced pulse model was used as a framework for pitch modification [Hag07]. The results were not satisfactory.

We also tried linear predicition (LP)-PSOLA, in the framework of which a Linear Prediction Coefficient (LPC)-analysis is done and the pitch modification carried out on the residual signal. For unvoiced speech segments, an artificial source signal is generated using white Gaussian noise. This source signal is filtered with the LPC filter and modulated with the energy envelope of the original signal. The latter is used to preserve transient sounds. The artificial unvoiced excitation is used to replace EL speech in unvoiced segments.

For EL speech, PSOLA offers the best result, but has a drawback, which is that it is only usable for EL speech. Esophageal and tracheo-esophageal speech often suffers from highly irregular speech cycles, for which it is not possible to determine the pitch marks reliably.

The alternative is an LPC based system, in which the linear prediction coefficients are analyzed frame-by-frame. The frame is then resynthesized with the new pitch contour using an artificial excitation signal. This approach can be applied to EL speech as well as to Tracheo-esophageal (TE) and ES speech and also aphonic speech, i.e., speech without any voiced excitation.

Hereafter, pitch modification has been carried out using the ordinary Pitch synchronous overlap add (PSOLA) technique.

The PSOLA pitch modification algorithm requires the knowledge of the pitch marks of the EL speech. We have used the *'Praat'* pitch mark algorithm [BW07]. Since the EL signal has a very clear harmonic structure, the choice of the pitch marking algorithm is not critical. To determine formants, we also used the formant tracker provided by *'Praat'*. As both the pitch mark and the formant determina-

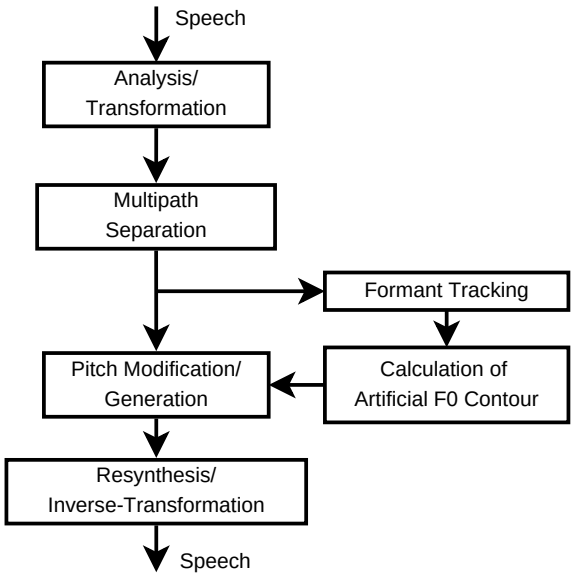


Figure 6.13: *Framework of Pitch Generation from Formants.*

tion algorithm use a Viterbi-based tracking over the whole speech utterance, these algorithms would have to be replaced in real-time systems.

Voicing decision. The artificial pitch is only applied in case of voiced sounds. During pauses or unvoiced sounds, the signal is just copied from the original.

The EL activity is detected by a simple energy detector, because EL speech gives well defined voice activity boundaries (see Fig. 6.15). When the EL is turned on, the dynamic range is only about 15 dB, compared to laryngeal speech for which the range is about 50 dB. For unvoiced sounds, the decision is based on the spectral centroid $C_s[l]$,

$$C_s[l] = \sum_{k=1}^{N/2} f[k] \frac{|X[l,k]|}{\sum_{k=1}^{N/2} |X[l,k]|}, \tag{6.4}$$

where $X[l,k]$ is the discrete Fourier transform of the input signal with frame length N, l is the time index and k the frequency index. $f[k]$ is a vector with the bin frequencies of the Fourier transform. A decision for unvoiced sounds is made when C_s is above a threshold of 2500 Hz. This value has to be adjusted depending on the speaker and the EL device he is using. In addition, the absence of pitch marks for a period longer than the maximal expected period will also be used as a cue for unvoiced sounds.

A voicing decision on the base of present or absent harmonics in the speech signal is not possible, because the typical EL user is not turning off the EL for unvoiced sounds.

Pitch modification should maintain a mean pitch, identical to the constant default pitch. Therefore, the original EL pitch frequency is chosen as the default mean pitch,

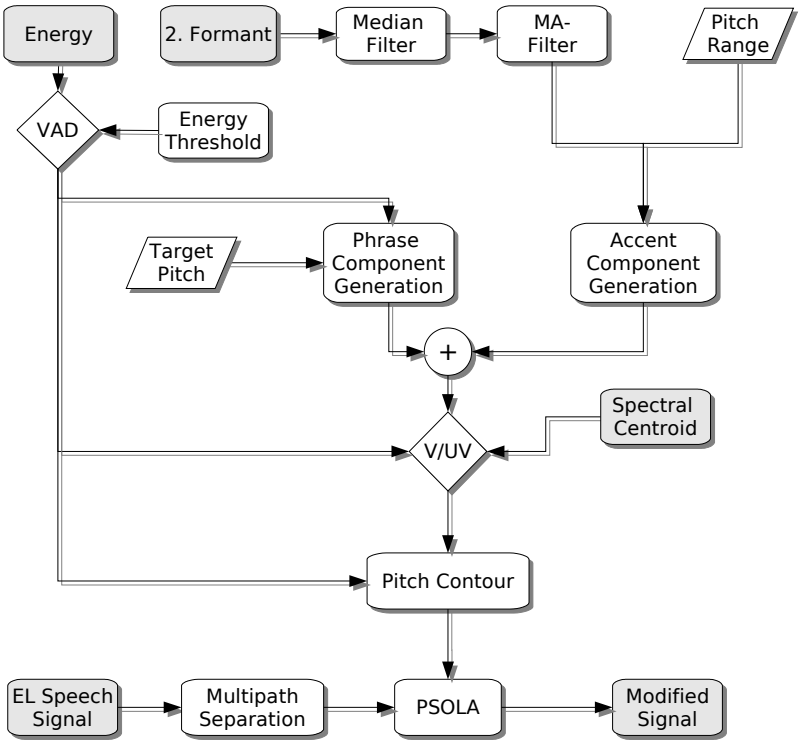


Figure 6.14: *Flowchart of pitch generation and modification.*

to avoid a mismatch between face-to-face and distance communication.

6.3.3 Summary of Operations

- Input EL speech
- Dermine pitch marks
- DREL sound multipath separation
- Word accent component $F_a(t)$:
  - Track formants
  - Smooth of 2nd formant non-linearly and linearly (Median filter, moving average)
  - Remove mean
  - Scale to match target pitch range

$$f_0 = F_0 \cdot \beta$$

- Phrase component $F_p(t)$:
  - Determine EL activity

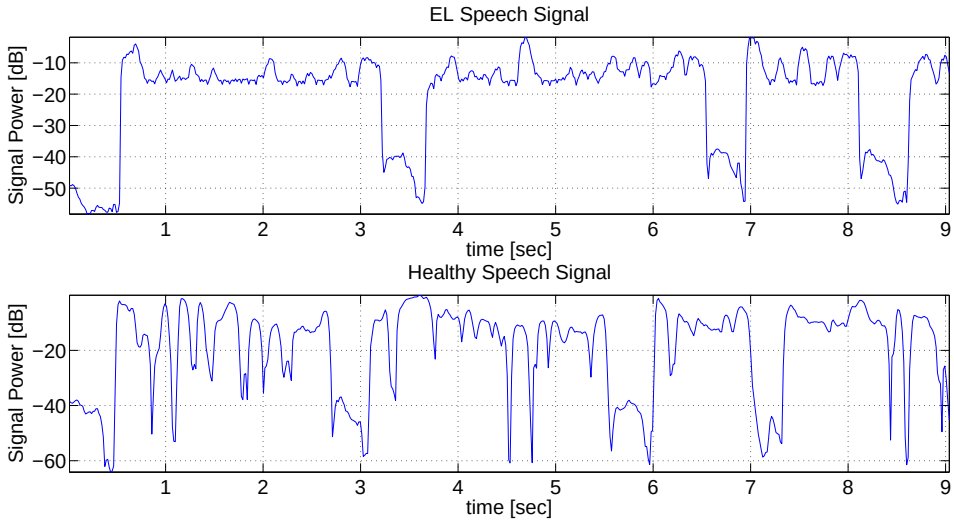


Figure 6.15: *Signal energy. Top: EL speech. Bottom: Healthy speech.*

  - Off → Rise: $x = 1 - e^{-\tau_{rise}t}$
    On → Fall: $x = e^{-\tau_{fall}t}$
  - Scale to target pitch contour (mean, range).

- Add phrase and accent component

$$F_0(t) = F_a(t) + F_p(t)$$

- Generate pitch marks
- Apply PSOLA pitch modification

6.4 Perceptual Evaluation of the Artificial Intonation Contour

In Chapter 3, we have introduced a listening test to evaluate enhanced disordered laryngeal voices. In Chapter 5, we have proposed a listening test to evaluate the multipath separation method. Assessment methods used in telecommunications, such as described in ITU-T standard P.800 [ITU96] and P.85 [ITU94], do not explicitly cover prosodic features of speech.

Even for unprocessed alaryngeal voices, no widely accepted method exists to evaluate the quality of speech, though work is going on to introduce an assessment system for alaryngeal speech [MMVdB+06, MMCB+06]. Existing assessment methods are valid for laryngeal speech only. Therefore, we need to design a test of prosodic quality, which fits the special requirements of EL speech.

We have, therefore, considered other areas of expertise , such as the evaluation of hearing devices. For Cochlear implants (CIs), the success of the devices regarding the transfer of prosodic information has to be tested. A battery of subjective tests for users of a CI has been introduced to evaluate the perception of prosody [MPL+07, MPL+08]. Here, we build on this work and adapt it to the perception of EL speech enhanced by an artificial F_0 contour. Examples from the corpus are either taken from, or inspired by, work testing prosodic perception of CI patients [MPL+08] and the automatic voice translation project VERBMOBIL [NBK+97]. The following categories have been used: intonation minimal pairs, sentence mode, sentence accent and phrasing.

We also include a longer text ('Nordwind und Sonne' - 'north wind and sun') that may enable judging the naturalness of the prosodic features used.

6.4.1 Test Design

We want to determine whether the superimposed F_0 contour helps to increase the accuracy of the recognition of that linguistic information that is conveyed using suprasegmental cues. That is, is it possible to detect sentence mode, or word emphasis? In addition, we want to investigate whether an artificial F_0 contour increases the naturalness of the prosody of a speech utterance. The listening test consists of three parts.

Emphasis Position Test:

The listeners hear a sentence, of which different words have been emphasized. The sentences are produced with a laryngeal voice, whisper, EL voice or an EL voice with superimposed F_0. The listeners hear the same sentence in all four variants.

The subjects are asked, which part of the sentence is emphasized?

E.g.: *Die Katze jagt eine Maus.*

◯ Beginning ◯ Middle ◯ End

Sentence Mode Test:

The listeners hear a sentence, which is either realized as a declaration or a question. The sentences are produced using a laryngeal voice, whisper, EL voice or an EL voice with superimposed F_0. The listeners hear the same sentence in all four variants in randomized order at some time in the listening session. The subjects are asked whether they perceive a question or declaration?

E.g. *Treffen wir uns dann am Hauptplatz*

◯ Question ◯ Declaration

Subjective Quality

In addition, we want to evaluate the listening experience. How is the prosody of the modified speech perceived compared to the original utterance when rated on a Comparison category rating (CCR) scale. The listener is asked: 'How do you like the prosody of the second speech sample compared to the first one?'

Speech Quality:	Score
much better :	3
better :	2
slightly better :	1
about the same:	0
slightly worse :	– 1
worse :	– 2
much worse :	– 3

6.4.2 Recordings

Both laryngectomized (4) and healthy subjects (10) were recorded while using an Electro-Larynx. The recorded material was edited, i.e., the silent periods at the beginning and end of an utterance were removed and recordings that were unfit for further processing were discarded. For more information on the recording conditions, please refer to appendix B.

For the evaluation of emphasis position and sentence mode only the recordings of nine healthy subjects were used, who produced the utterances with whisper, laryngeal and EL voice. We show in chapter 5, page 71, that the results of subjective listening tests for healthy speakers are not significantly different from alaryngeal speakers. For the accent perception evaluation, 108 different utterances were available, i.e., nine people producing four different sentences, placing the emphasis at 3 possible positions. For the sentence mode perception evaluation, 90 utterances were used, i.e, 9 people saying 5 different sentences either as a declaration or question. For a list, please refer to appendix B.2. For the naturalness of prosody test, eleven recordings were available with a fragment of 'Nordwind und Sonne' ('North wind and sun')

Listeners

Most of the listeners had only little or no experience with EL speech. The number of listeners was equal to twelve, three female and nine male. The average age was 28 years. All were native German speakers and had normal hearing. The listeners used studio headphones (AKG K-271) and were asked to adjust the volume to a comfortable level. Every listener had 60 different emphasis position examples

and 40 sentence mode examples to answer. Over all, this gives 720 speech tokens (180 for each production mechanism) for the emphasis position recognition task and 480 speech tokens (120 for each production mechanism) for the sentence mode recognition task which were evaluated.

6.4.3 Results

Emphasis position. The listening test has been processed using a confusion matrix to show how the recognition of the accent position in a senctence depends on the voice. In table 6.2, we see the results for the perception of accent position. In laryngeal speech utterances, almost all of the emphasis positions (97.8 %) were correctly recognized. In whispered speech the recognition rate is still high with 93.9 %. For EL speakers, the rate of correctly identified accent positions goes down to 72.2%. When introducing the artificial F_0 contour to the EL speech, the accuracy of the accent position identification is 69.4 %. While this is lower than for the original EL speech, the difference is not significant. See Fig. 6.16 for a plot of the recognition rate and the corresponding 95 % confidence interval, based on the assumption of a binominal distribution.

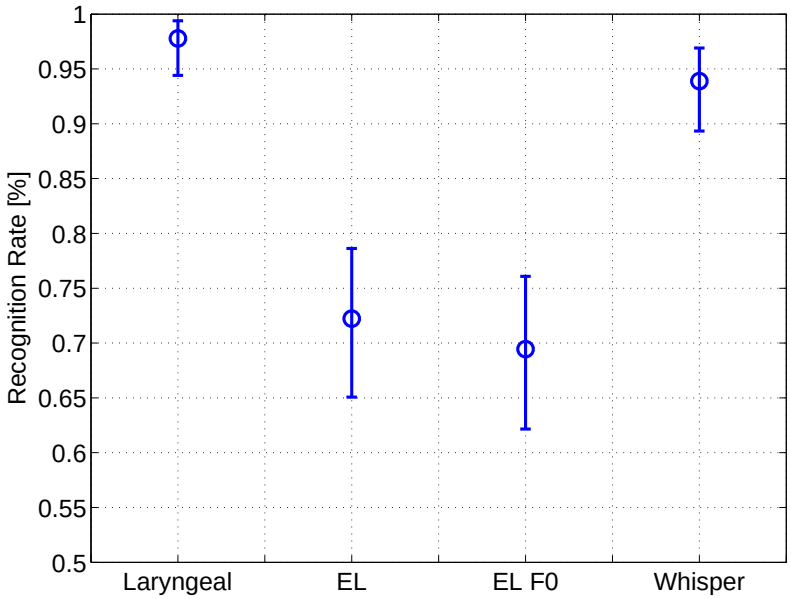


Figure 6.16: *Mean recognition rate of emphasis position and 95 % confidence interval.*

Sentence Mode. In table 6.3, we see the results for the perceived sentence mode presented as a confusion matrix and overall recognition rate. With laryngeal voice, almost all samples (98.3 %) were correctly identified. For whisper, the rate goes down to 77.5 %, when using an EL, the success rate decreases to 65 % which is

Table 6.2: *Confusion matrix and recognition rate for emphasis position perception (in %).*

Laryngeal Speech:

Perceived Accent Position	Actual Accent Positions		
	Beginning	Middle	End
Beginning	32.8	0.0	0.0
Middle	0.6	32.2	0.6
End	0.0	1.1	32.8
Correctly Identified:	97.8		

EL Speech:

Perceived Accent Position	Actual Accent Positions		
	Beginning	Middle	End
Beginning	22.2	6.1	2.8
Middle	2.2	24.4	5.0
End	8.9	2.8	25.6
Correctly Identified:	72.2		

EL Speech with artificial F_0:

Perceived Accent Position	Actual Accent Positions		
	Beginning	Middle	End
Beginning	27.2	8.3	6.1
Middle	1.7	21.1	6.1
End	4.4	3.9	21.1
Correctly Identified:	69.4		

Whisper:

Perceived Accent Position	Actual Accent Positions		
	Beginning	Middle	End
Beginning	32.8	0.6	2.2
Middle	0.0	32.8	2.8
End	0.6	0.0	28.3
Correctly Identified:	93.9		

still above chance level. The recognition rate even further decreases to 49.2 % when applying the artificial F_0 contour. It can be noted that for the EL speakers there is a strong bias toward the recognition of the sentence mode as a declaration, which is even more prominent for the artificial F_0 contour tokens. The recognition rate with the corresponding 95 % confidence interval can be seen in Fig. 6.17.

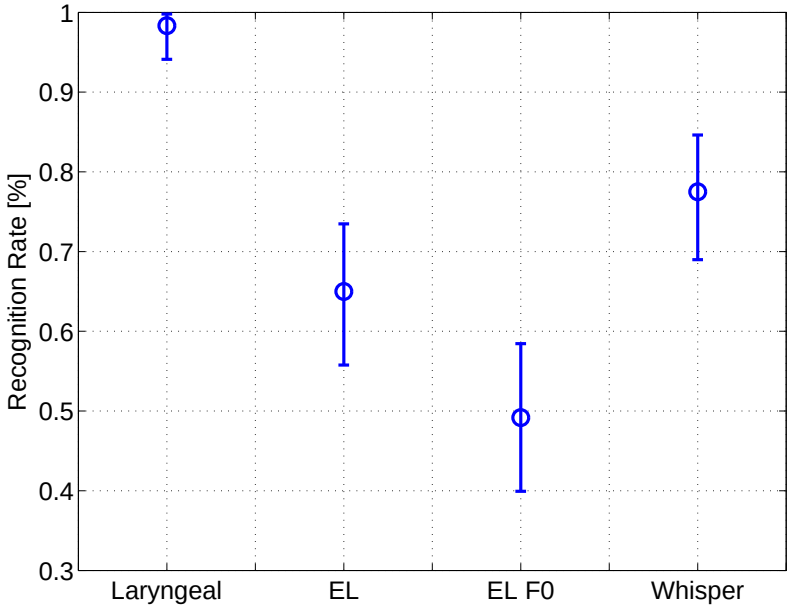


Figure 6.17: *Mean recognition rate of sentence mode and 95 % confidence interval.*

Subjective Quality. The results for the perceived naturalness of the artificial prosody show that the added prosody is strongly perferred. Assuming a normal distribution of the comparison opinion scores, an estimate of the the mean μ, and the standard deviation σ of the Comparison mean opinion score (CMOS) results were calculated. For every mean, a confidence interval CI_{95}, was also determined. Those values superimposed on a histogram are presented in Fig. 6.18. By far most votes were in favor of the artificial prosody. Only few votes either indicated that the artificial prosody was equal to the original one or that the original was superior to the modified version.

6.5 Conclusion

In this chapter, we have presented a method to add a more natural fundamental frequency contour to EL speech. Usually, the EL device is operated at one fixed F_0 value. Some devices offer an alternative frequency value, when pressing a different button. Existing literature suggests that formants contribute to the perceived prosody of whispered speech. Informal experiments have demonstrated that a similar effect may exist for EL speech. Given the relevance of formants for conveying

Table 6.3: *Confusion matrix and recognition rate for sentence mode perception (in %).*

Laryngeal Speech:

Perceived Sentence Mode	Actual Sentence Mode	
	Declaration	Question
Declaration	48.3	0.0
Question	1.7	50.0
Correctly Identified:	98.3	

EL Speech:

Perceived Sentence Mode	Actual Sentence Mode	
	Declaration	Question
Declaration	40.0	25.0
Question	10.0	25.0
Correctly Identified:	65.0	

EL Speech with artificial F_0:

Perceived Sentence Mode	Actual Sentence Mode	
	Declaration	Question
Declaration	37.5	38.3
Question	12.5	11.7
Correctly Identified:	49.2	

Whisper:

Perceived Sentence Mode	Actual Sentence Mode	
	Declaration	Question
Declaration	41.7	14.2
Question	8.3	35.8
Correctly Identified:	77.5	

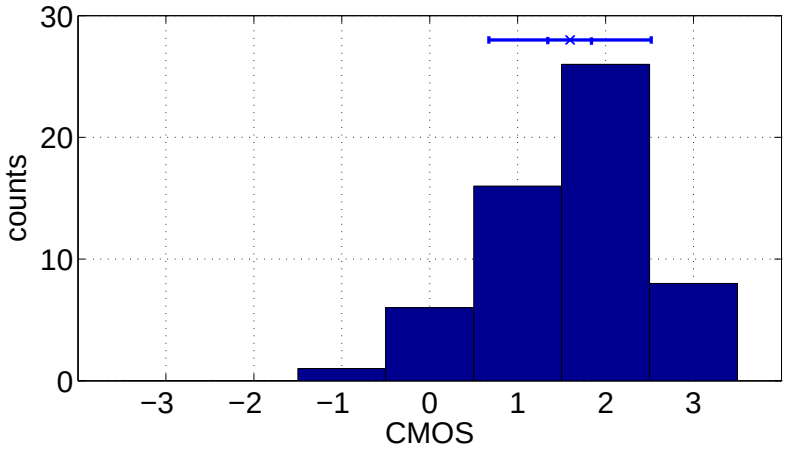


Figure 6.18: *Histogram of Comparison Mean Opinion Score (CMOS) of preference of artificial prosody imposed on a sentence compared to EL speech. On top of the histogram the mean recognition rate, CI_{95} 95 % confidence interval and standard deviation are shown.*

prosody in whispered speech, they are used here as a source signal for an artificial fundamental frequency contour.

We investigated the impact of the artificial F_0 contour on several aspects of the listening experience. First, we examined whether the artificial F_0 improves the the perception of the emphasis position or sentence mode. The results showed that the accuracy of perception of the emphasis position does not significantly change. Whereas the accuracy of the perception of the sentence mode is significantly reduced by the artificial F_0 contour. The artificial F_0 contour only improves the subjective quality of the intonation contour, but this in a very significant way.

This results suggest the conclusion that the before mentioned hypothesis that formants might be used to generate an F_0 contour to enhance prosodic information may be wrong. There may be several reasons. While in some respects, whisper and EL speech may use similar ways to convey prosodic information, there are differences. This can be seen in the significantly better recognition rates for prosodic information in whisper in contrast to EL speech. The ways to convey prosodic information in EL speech are not as simple as only unsing the formants. Duration and loudness are also very important substitute cues for both word emphasis and sentence mode. It may be to optimistic to expect an acoustic feature that is generated by using another acoustic feature as a source signal to increase the perception of linguistic information. The example with alaryngeal singing (Fig. 6.3 – 6.5), which mainly suggested the use of the formants, may be inadequate because this is a very specialized task, that has not a lot to do with speaking. The second example (Fig. 6.6) shows that intensity plays an important role as a cue for the emphasis of a word; the formants may only be a secondary cue. For the sentence mode perception the artificial declination may act against the recognition of questions, since the weak intonation pattern which

should mark the question is overridden by the omni-present declination contour.

Chapter 7

Discussion and Conclusions

To wrap up this thesis, we summarize the topics and discuss the results. We also outline possible future work for alaryngeal voice enhancement.

This thesis deals with several aspects of enhancing disordered voices. After a presentation of the state-of-the-art in disordered speech enhancement (Ch. 1), the first part deals with laryngeal disorders, for which the timbre of the voice is not very far away from modal voices. The second part deals with substitution voices that are much further away from modal human speech. For the enhancement of laryngeal voices (Ch. 3), we have focused on a particular disorder, caused by laryngeal neoplasm. These voices are characterized as very breathy, with a too high fundamental frequency. The enhancement is based on Time-domain pitch-synchronous overlap-and-add (TD-PSOLA) to lower the pitch and period enhancement to reduce breathiness. Listening tests suggest that the perceived speaking effort is reduced. The latter is a term that is defined as the perceived impression of how much effort a given speaker has to make to produce a speech sound. Interestingly, this perception can be modified through signal processing means.

Given the need for accurate pitch mark determination (Ch. 4), which is able to work better with disordered voices, we have investigated an algorithm that takes advantage of methods of dynamical systems analysis. The method is presented so that readers not familiar with dynamical systems analysis are able to follow the key steps and implement the algorithm. The speech signal is embedded into a pseudo state-space and the pitch marks are found at the intersection of the trajectories with a Poincaré plane that is placed perpendicular to the mean flow direction of the trajectories.

This approach has been shown to have advantages when dealing with disordered, and especially diplophonic voices. While a state-of-the-art correlation-based algorithm does not track the alternating cycles of a diplophonic voice, the state-space based approach correctly determines the pitch marks. Another benefit is that this method does not require post-processing, which enables online analysis. The goal of finding an improved pitch mark determination algorithm for TD-PSOLA has not

been reached. The reason is a phase drift of the pitch marks. TD-PSOLA requests pitch marks to be positioned at the energy peaks of voiced speech cycles.

In the second part of the thesis, we move from laryngeal voice disorders to alaryngeal substitution voices, with a special focus on EL speech. We first deal with the problem of the background noise, which is the direct sound radiated from the EL device to the listener (Ch. 5). We present a model of EL speech production and point out the common sound source of the voiced speech signal and the directly radiated noise signal. Previous approaches to solving the problem of direct sound suppression have used standard speech enhancement methods, which assume independent speech and noise signals. Given EL speech production, this assumption cannot hold. Therefore, we view the problem as a separation of different propagation paths. The separation can be achieved by taking into account contrasting temporal properties of the propagation paths. While the speech signal is modulated via the movement of the articulators, the directly-radiated EL signal path is time-invariant. Therefore, the noise is suppressed by high-pass filtering in the modulation frequency domain, which suppresses signals components which are steady or slowly time-varying only.

Subjective listening tests have shown mixed results. Even though the noise is perceived as decreased, this perceived improvement is not reflected in an increased overall acceptance of the modified signal. An analysis of the results shows that a sub-group of listeners disliked the modification, while another appreciated the suppression of the noise sound. For that group the perceived decrease of the background noise has been reflected in a perceived increase of overall quality and a decreased listening effort.

Another major problem of EL speech is the monotonous, robot-like voice, which to a certain extent, is due to the constant fundamental frequency of the device (Ch. 6). Several approaches exist to deal with the problem. The one which has made it into most commercial products is a second button on the EL that allows to use an alternate pitch. Since it is difficult to use it efficiently in a conversation, most patients tend to ignore it. The quest for a more comfortable way to assign a more natural pitch contour to an EL has attracted our attention to whispered speech. Indeed, it is known that through manipulation of the speech formants an impression of a pitch movement can be achieved in whispered speech. Preliminary tests had suggested that this is possible not only in whispered, but also in EL speech. We, therefore, proposed to derive a pitch contour from the formants. Listening tests regarding prosodic function did not show an improvement of the recognition of sentence mode or word emphasis position. However, the artificial prosody was strongly preferred in a CCR Test on a CMOS scale.

A common disadvantage of all methods is that, because of the post processing of the EL speech output, their application is practical only in case of telecommunication and not in face-to-face conversation. In face-to-face conversation, the original and modified signals would be received both and the result would be worse. It is, however, safe to assume that a majority of the speakers are likely to interact both

in a face-to-face as well as a telecommunication setting. This implies that a listener may be confronted with both the unmodified and the modified voice, depending on the setting. This may cause irritation because the enhanced and not enhanced voices of a speaker may sound different. This leads to the constraint that any modification that cannot be applied to every communication setting has to be kept in a range so that the modified voice can still be connected to the EL speaker. This limits, for example, the target F_0 range. The mean F_0 in the experiments has, therefore, been chosen to be equal in the original and the enhanced speech utterances. The same is true for any spectral modification. It should be restricted to differences between the original and enhanced utterances that do not prevent identifying the speaker.

So where should we go from here? In the course of the thesis several issues have been raised which were not solved.

Even after decades of research, the question of pitch mark determination is far from solved. While most state-of-the-art proposals give correct results for a majority of speech utterances, every algorithm is likely to fail in difficult situations. A possible future direction for pitch mark determination could be a hybrid approach of state-space and time-domain methods, where the best of both worlds could be combined.

Concerning the multipath separation, to increase the quality of the EL signal that has been modified in the modulation frequency domain, coherent envelope detection [Sch07] appears promising.

Findings on the use of formants for F0 contour generation leave room for further reseach, only mixed results have been obtained. One possibility would be to include machine learning for F_0 contour generation, based on the automatic learning of the relation between non-source features of an alaryngeal speaker and the F_0 contour of the same sentence spoken by a laryngeal speaker. A further point would be to study the real-time application of F_0 contour generation in the framework of bio-feedback.

While the enhancement of a substitution voice by means of speech signal processing can improve the life of the affected person, it will never be the ultimate answer to the problem. The ultimate goal is to improve the substitution voice so that it is as close as possible to the natural voice, both in terms of sound quality and in terms of control. Work using myoelectrical signals to control an EL device [HHKM99, GHK$^+$04, GHSH07] is a promising approach of research. Other work in that direction is the use of a voice generation device embedded in the tracheo-esophageal shunt [vdTvGL$^+$06]. While the first approach makes heavy use of signal processing, the second is purely biomechanical, where the sound is produced by an oscillating silicone fold. Research in this direction is still at an early stage.

Appendix A

Praat Pitch Marking

This is a short description of the pitch mark determination methods implemented in 'Praat'. The description is from the Praat manual, which is also available online (see [BW07] and is included here for convenience).

As a first step, Praat runs a pitch determination algorithm to get a narrow search range for the pitch marks. It then determines the pitch marks either with an algorithm based on cross correlation or peak picking. The voiced intervals are determined on the base of a voiced/unvoiced decisions in the pitch determination algorithm. For every voiced interval, a number of epochs (or glottal pulses) are found as follows:

Cross-Correlation Algorithm

1. The first point t_1 is the absolute extremum of the amplitude of the sound, between $t_{mid} - \frac{T_0}{2}$ and $t_{mid} + \frac{T0}{2}$, where t_{mid} is the midpoint of the interval, and T_0 is the period at t_{mid}, as can be interpolated from the pitch contour.

2. From this point, we recursively search for points t_i to the left until we reach the left edge of the interval. These points must be located between $t_{i-1} - 1.2 \cdot T_0(t_{i-1})$ and $t_{i-1} - 0.8 \cdot T_0(t_{i-1})$, and the cross-correlation of the amplitude in its environment $[t_i - T_0(t_i)/2; t_i + T_0(t_i)/2]$ with the amplitude of the environment of the existing point t_{i-1} must be maximal (we use parabolic interpolation between samples of the correlation function).

3. The same is done to the right of t_1.

4. Though the voiced/unvoiced decision is initially taken by the Pitch contour, points are removed if their correlation value is less than 0.3; furthermore, one extra point may be added at the edge of the voiced interval if its correlation value is greater than 0.7.

Peak-Picking Algorithm

1. The first point t_1 is the absolute extremum (or the maximum, or the minimum, depending the settings) of the amplitude of the sound, between $t_{mid} - \frac{T_0}{2}$ and $t_{mid} + \frac{T0}{2}$, where t_{mid} is the midpoint of the interval, and T_0 is the period at t_{mid}, as can be interpolated from the pitch contour.

2. From this point, we recursively search for points t_i to the left until we reach the left edge of the interval. These points are the absolute extrema (or the maxima, or the minima) between the times $t_{i-1} - 1.2 \cdot T_0(t_{i-1})$ and $t_{i-1} - 0.8 \cdot T_0(t_{i-1})$.

3. The same is done to the right of t_1.

Appendix B

Voice Recordings

To evaluate algorithms in speech processing one needs speech data. In our case speech data of Electro-Larynx (EL) speakers are necessary.

B.1 Recording Subjects:

The speech database consists of two main parts, alaryngeal and healthy subjects.

Alaryngeal EL speakers. Four alaryngeal subjects were recorded producing 35-40 utterances. From one additional speaker, only a few utterances were usable, because he was not able to follow the recording instructions. Two of the alaryngeal speakers are using the EL as their main means of communication. They are all male and have a mean age of 70 years, which is typical of the largest group of alaryngeal speakers found in clinical practice. The patients used their own EL, namely products from Servox – which has recently been renamed to Servona [Ser] – and Heimomed [Hei].

The recordings were approved by the ethics commission of the Medical University of Graz.

Recording alaryngeal patients presents some specific problems with regard to the recording protocol. Some are discussed here to help follow-up work. For most patients, speaking is hard work so only short recording sessions were possible. A half hour duration for the recordings appeared to be a reasonable upper limit. As mentioned above, most patients are rather old and often need glasses. Two recording sessions were almost unusable because the subjects did not bring their glasses and were hardly able to read the instructions on the monitor. As we discuss later, one aim was to record subtle prosodic differences. Some patients did not understand the instructions even after auditive prompting of the sentence. Some of the EL speakers were very nervous, so the beginning of the recording session was of much lower speech quality than the final part. In that case, the earlier parts of the session were discarded.

Healthy (laryngeal) EL speakers. Since alaryngeal EL speakers are not easily available and not easy to record, ten healthy subjects have been recorded for comparison purposes. They were six female speech & language therapists (SLT) and three male phoniatricians and the author of this thesis. Recording healthy subjects has of course the disadvantage that the subjects are not dependent on speaking with a substitute voice and their vocal tract anatomy differs in that the pharyngeal area of an alaryngeal speaker is changed significantly due to laryngectomy. Nevertheless, by holding their breath, the separation of the trachea from the vocal tract can be simulated even by laryngeal speakers. See Fig. 5.18 for a comparison between results of alaryngeal and healthy subjects. In addition, most of the reference speakers have above-average experience in speaking with an EL, because some are involved in teaching patients how to use an EL. All laryngeal subjects used a Servox EL.

An advantage of healthy speakers is the availability of different speaking modes. To gain insight in the function of intonation in alaryngeal speech and its differences to laryngeal and whispered speech, we recorded the same utterances in three different speaking modes: laryngeal speech, whispered speech and electro-larynx speech.

Recording Setup

We recorded with an omnidirectional head-mounted high-quality condenser microphone (AKG HC577). The head-mounted microphone was chosen to ensure a consistent recording quality, since it guarantees a constant distance between microphone and mouth. The signal was fed into a high quality microphone pre-amplifier and analog/digital converter. The speech data were stored on a laptop. The A/D converter and the computer were connected via a firewire (IEEE 1394) interface. We used an RME Fireface 800, which integrates high-quality microphone pre-amlifiers, A/D conversion and the firewire interface (see Fig. B.1). Recording has been done with a sampling frequency of 48 kHz at a resolution of 24 bits. Microphone, pre-amplifier and A/D converter have an almost flat frequency response from 80Hz-20kHz.

Whenever possible, the recordings were made with the test subject sitting in a sound proof recording booth. The experimenter could observe the test subject through a glass window. When the recording booth was not available, due to a different recording location, the recordings were performed in a quiet office.

The recording software had been designed to record speech corpora. *SpeechRecorder* was developed at the Bavarian archive for speech signals (BAS) of the Institute of Phonetics and Speech Processing at Ludwig-Maximilians-Universität München [DJ04]. Among other features, it offers separate views on different computer displays for the instructor and the speaker. An eXtensible markup language (XML) based recording script can incorporate text and multimedia prompts.

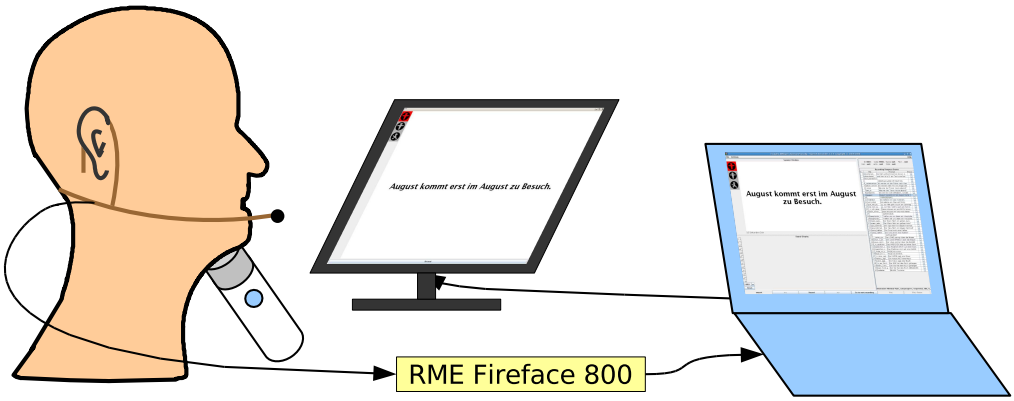


Figure B.1: *Recording Setup. The subject wears a head-mounted microphone. The signal is amplified and A/D converted with an RME Fireface 800 and fed into a notebook via the Firewire interface. Speaker and recording instructor have separate computer displays.*

B.2 Design of the Test Corpus

The test corpus has been designed to meet two goals: Respect the limitations in recording time for the alaryngeal patients and enable prosody-related experiments with the recorded EL speech database

For Cochlear implants (CIs), a battery of tests to evaluate the perception of prosodic information has been proposed [MPL+07, MPL+08]. For the current evaluation, we built on this work and adapted it to the perception of EL speech enhanced with an artificial F_0 contour. The examples from the corpus are either taken from, or inspired by work testing the perception of prosody by CI patients [MPL+08] and the automatic voice translation project VERBMOBIL [NBK+97]. The following categories have been used: Intonation minimal pairs, Sentence mode, Sentence accent and Phrasing.

In addition, we also included a longer text ('Nordwind und Sonne' - 'Northwind and Sun') and an image description task using spontaneous speech, to elicit connected speech utterances.

Recording Protocol All speech material is in German, the native language of the speaker and the listener groups.

Intonation minimal pairs: Placing the accent on different syllables can change the meaning of a word.

- Ich will ihn nicht umfahren sondern umfahren.
- August kommt erst im August zu Besuch.

Figure B.2: *Speech Recorder instructor window.*

- Ich werde mit der Fähre nach Irland übersetzen. vs.
  Ich werde den Text ins Englische übersetzen.
- Wie war der Tenor heute abend? vs.
  Wie war der Tenor heute abend?
- Ich nehme ein Glas Vollmilch. vs.
  Ich nehme ein Glas voll Milch

Sentence mode: We want to test, if the sentence is perceived as a declaration or question.

- Der Mann fährt ein gelbes Auto? vs. Der Mann fährt ein gelbes Auto.
- Der Opa fährt ein blaues Fahrrad? vs. Der Opa fährt ein blaues Fahrrad.
- Die Oma trinkt einen Kaffee? vs. Die Oma trinkt einen Kaffee.
- Treffen wir uns dann am Hauptplatz? vs. Treffen wir uns dann am Hauptplatz!

Word emphasis: For every sentence, three words exist which can be emphasized.

- Der Löwe springt über die Mauer. vs.
  Der Löwe springt über die Mauer. vs.
  Der Löwe springt über die Mauer.

- Das Mädchen sitzt auf einer Bank. vs.
  Das Mädchen sitzt auf einer Bank. vs.
  Das Mädchen sitzt auf einer Bank.

- Der Bär hat den Fisch gefangen. vs.
  Der Bär hat den Fisch gefangen. vs.
  Der Bär hat den Fisch gefangen.

- Die Katze jagt eine Maus. vs.
  Die Katze jagt eine Maus. vs.
  Die Katze jagt eine Maus.

Meaning-changing sentence accent: Sometimes changing the accent gives a different meaning to a sentence.

- Finde ich schon. vs.
  Finde ich schon.

- Dann müssen wir uns noch einen Termin ausmachen. vs.
  Dann müssen wir uns noch einen Termin ausmachen.

Phrasing: Phrasing of a speech utterance is often done by intonation control.

- Ja, zur Not geht's auch am Samstag. vs.
  Ja zur Not. Geht's auch am Samstag?

Musical scale: Is it possible to perceive a change of pitch when trying to sing a short musical scale?

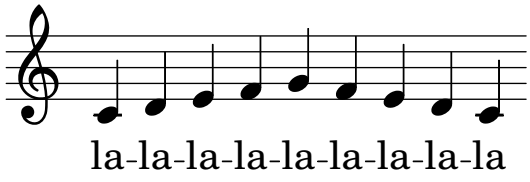


Figure B.3: *Short musical scale.*

Nordwind und Sonne "Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges kam. Sie wurden einig, daß derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzunehmen."

Note: Because for most subjects a long read text was very hard work, for most recordings only the first paragraph of the fable was read.

Image description: To elicit spontaneous speech, an image was presented and the subject was asked to describe the image according to his or her personal liking (Fig. B.4).

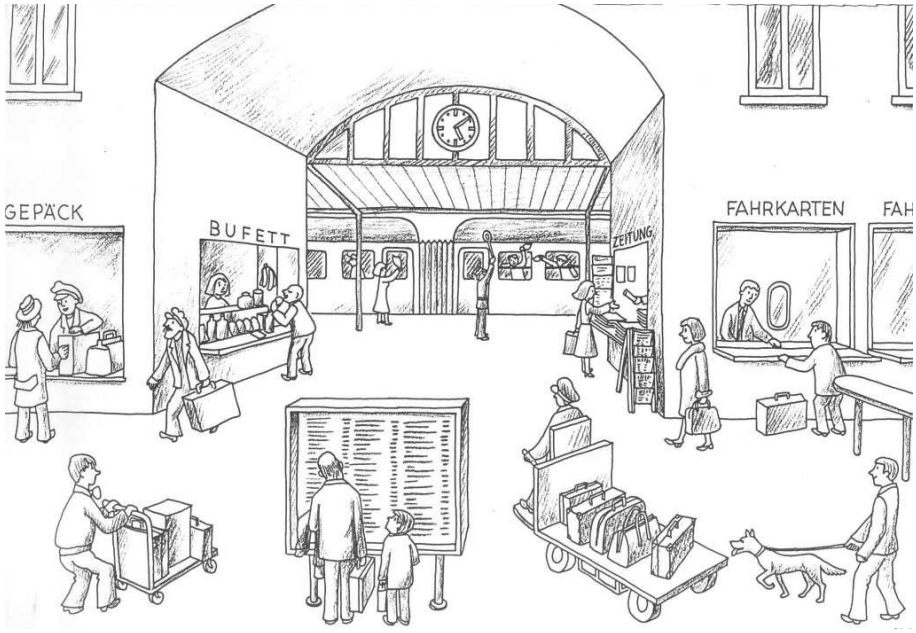


Figure B.4: *Image for free speech prompting.*

Appendix C

Other Work

During my time at SPSC, I have been working on several projects, which due to the wide variety of subjects didn't make it into this thesis. I only give a brief overview here. Details of the projects can be found in the referenced publications.

Water leak detection

The first project was a cooperation with a local company, which wanted to improve the leak detection for water supply pipes. One commonly used method to leak detection employs an acoustical approach. Special microphones, which are placed on the ground, are used to pick the sound emitted by a water pipe. An expericened listener can discern between background noise and the noise produced when water is leaking. Since the sound quality is very bad, and very often disturbed by enviromental noise, like cars, planes, etc. we have employed signal processing methods to enhance the sound the operator has to listen to. One approach was to incorporate the properties of the leakage noise. It can be modelled with stationary gaussian noise, which has a soft peak at an unknown frequency below 2kHz. We used a low-pass filter in the modulation frequency domain to suppress all components which are non-stationary. Further, we used a reference microphone, which picked up the enviromental noise and employed an adaptive filter to suppress stationary enviromental noise [HK02a].

Related Publication:

- Martin Hagmüller and Gernot Kubin. Akustische Punktortung von Wasserleitungsschäden. Technical report, Graz University of Technology - Signal Processing and Speech Communication Laboratory, 2002.

Watermarking for Air Traffic Control

Another project was carried out with Eurocontrol Experimental Centre, Brétigny-sur-Orge, France in cooperation with Frequentis Innovations, Graz, Austria. It

has received the ATC Maastricht Innovation Award at the 14th Annual Air Traffic Control (ATC) Maastricht Conference 2004.

In air traffic control, the voice communication between a controller and all pilots in a delimited airspace are handled on a single VHF channel. To identify their aircraft, pilots have to start all verbal communications with the aircraft call sign. For automatic identification, it is desirable to transmit additional hidden aircraft identification data in time with this voice message over the VHF channel. That means the additional digital data has to be embedded into the analog speech signal. We developed a system for speech watermarking in an air traffic control environment using spread spectrum technology. A digital data string is transmitted every time when a pilot speaks with the air traffic controller. [HHK03, HHKK04, HK05b]. This work has received a quite favorable response and has led to more research.

Related Publications:

- Horst Hering, Martin Hagmüller, and Gernot Kubin. Safety and security increase for air traffic management through unnoticable watermark aircraft identification tag transmitted with the VHF voice communication. In *Proc. 22nd Digital Avionics Systems Conference (DASC)*, Indianapolis, Indiana, October 2003 pp. 4.E.2-1–10.

- Martin Hagmüller, Horst Hering, Andreas Kröpfl, and Gernot Kubin. Speech watermarking for air traffic control. In *Proc. of 12th European Signal Processing Conference*, pages 1653–1656, Vienna, Austria, September 6–10 2004.

- Martin Hagmüller and Gernot Kubin. Speech watermarking for air traffic control. Technical Report EEC Note 2005-05, Eurocontrol Experimental Centre, 2005.

Digital Watermarking for Security Applications

This work was carried out in cooperation with Marcos Faundez-Zanuy from Escola Universitària Politècnica de Mataró, Spain, within the European COST action 277 'Non-linear speech processing'. We proposed a security enhanced speaker identification and verification system based on speech signal watermarking. The system can detect situations in which playback speech, synthetically generated speech, or an impersonator trying to imitate the speech are fooling a biometric system. It is also suitable for forensic experts, who sometimes have to demonstrate in court that a digital recording has not been manipulated or edited. In addition, we demonstrated that this watermark can coexist simultaneously with biometric speaker identification and verification minimizing the mutual effects. [FZHK06, FZHKK06, FZHK07, FZH07]

Related Publications:

- M. Faundez-Zanuy, Martin Hagmüller, Gernot Kubin, and Bastiaan Kleijn. The COST-277 speech database. In *Nonlinear Analyses and Algorithms for Speech Processing*, volume 3817/2005 of *Lecture Notes in Computer Science*, pages 100–107. Springer, 2006.

- Marcos Faundez-Zanuy, Martin Hagmüller, and Gernot Kubin. Speaker verification security improvement by means of speech watermarking. *Speech Communication*, 48(12):1608–1619, December 2006.

- Marcos Faundez-Zanuy, Martin Hagmüller, and Gernot Kubin. Speaker identification security improvement by means of speech watermarking. *Pattern Recognition*, 40(11):3027–3034, November 2007.

- Marcos Faundez-Zanuy and Martin Hagmüller. Sistema de marcas de agua para mejorar la vulnerabilidad de sistemas de reconocimiento de locutores. In *Actas do ill Congreso da sociedade espanola de acustica forense*, pages 167 – 181, Santiago de Compostela, Spain, 2007.

- Marcos Faundez-Zanuy, Jose Juan Lucena-Molina, Martin Hagmüller. Speech watermarking: An approach for the forensic analysis of digital telephonic recordings. *Journal of Forensic Sciences*, accepted for publication [2009].

Recognition of Regional Varieties of German

Research, which has already started with my diploma thesis [Hag01, HK02b], has been continued with the supervision of another diploma thesis [Diz03]. The topic was the recognition of regional varieties of German, i.e., the varieties of German spoken in Germany and in Austria. The features were derived from the prosody of the speech utterance [BDH+04].

Related Publication:

- Micha Baum, Vedran Dizdarevic, Martin Hagmüller, Gernot Kubin, and Franz Pernkopf. Prosody-based recognition of spoken German varieties. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 929–932, Montreal, Canada, May 17–21 2004.

Bibliography

[AJ06] Rym Haj Ali and Sofia Ben Jebara. Esophageal speech enhancement using excitation source synthesis and formant patterns modification. In *Proc. Int. Conf. on Signal-Image Technology & Internet Based Systems (SITIS)*, pages 615–624, Hammamed, Tunesia, December 17–21 2006.

[ANSK88] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. Voice conversion through vector quantization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 655–658, New York City, USA, April 1988.

[APHA96] Takayuki Arai, Misha Pavel, Hynek Hermansky, and Carlos Avendano. Intelligibility of speech with filtered time trajectories of spectral envelopes. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2490–2493, Philadelphia, PA, USA, October 3-6 1996.

[AR77] J.B. Allen and L.R. Rabiner. A unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564, 1977.

[ATNMPM06] G. Aguilar-Torres, M. Nakano-Miyatake, and H. Perez-Meana. Enhancement and restoration of alaryngeal speech signals. In *16th Intern. Conf. on Electronics, Communications and Computers, CONIELECOMP*, pages 31–31, 2006.

[Bag94] Paul Bagshaw. Evaluating pitch determination algorithms, a pitch database. http://www.cstr.ed.ac.uk/research/projects/fda/, 1994.

[BDH+04] Micha Baum, Vedran Dizdarevic, Martin Hagmüller, Gernot Kubin, and Franz Pernkopf. Prosody-based recognition of spoken German varieties. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 929–932, Montreal, Canada, May 17–21 2004.

[Bel07] Irene Velsvik Bele. Dimensionality in voice quality. *Journal of Voice*, 21(3):257–272, May 2007.

[BHD59] HL Barney, FE Haworth, and HK Dunn. An experimental transistorized artificial larynx. *Bell System Technology*, 38:1337–56, 1959.

[BHIB03] Dale H. Brown, Frans J.M. Hilgers, Jonathan C. Irish, and Alfons J.M. Balm. Postlaryngectomy voice rehabilitation: State of the art at the millennium. *World Journal of Surgery*, 27(7):824–831, July 2003.

[Bim03] Frédéric Bimbot, editor. *ISCA Tutorial and Research Workshop on Nonlinear Speech Processing (NOLISP)*, Le Croisic, France, May 20-23 2003.

[BK86] D.S. Broomhead and G.P. King. On the qualitative analysis of experimental dynamical systems. In S. Sarkar, editor, *Nonlinear phenomena and chaos*, pages 113–144. Adam Hilger, Bristol, UK, 1986.

[BLG01] Mary H. Bellandese, Jay W. Lerman, and Harvey R. Gilbert. An acoustic analysis of excellent female esophageal, tracheoesophageal, and laryngeal speakers. *Journal of Speech, Language and Hearing Research*, 44:1315–1320, December 2001.

[BM96] Michael Banbrook and Steven McLaughlin. Dynamical modelling of vowel sounds as a synthesis tool. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1981–1984, Philadelphia, PA, USA, October 1996.

[BQ97] Ning Bi and Yingyong Qi. Application of speech conversion to alaryngeal speech enhancement. *IEEE Transactions on Speech and Audio Processing*, 5:97–105, 1997.

[BW07] Paul Boersma and David Weenink. Praat ver. 4.06, software, downloaded from http://www.praat.org, 2007.

[BWVdHC97] Marc S. De Bodt, Floris L. Wuyts, Paul H. Van de Heyning, and Christophe Croux. Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice*, 11(1):74–80, 1997.

[CEFZM05] Gérad Chollet, Anna Esposito, Marcos Faundez-Zanuy, and Marina Marinaro, editors. *Nonlinear Speech Modeling and Applications - Advanced Lectures and Revised Selected Papers*, volume 3445 of *Lecture Notes in Computer Science*. Springer-Verlag GmbH, 2005.

[CK04] Ann Chang and Michael P. Karnell. Perceived phonatory effort and phonation threshold pressure across a prolonged voice loading task: a study of vocal fatigue. *Journal of Voice*, 18(4):454–466, December 2004.

[CM82] Theo A. C. M. Claasen and Wolfgang F. G. Mecklenbräuker. On stationary linear time-varying systems. *IEEE Transactions on Circuits and Systems*, 29(3):169–184, 1982.

[CM00] Karen MA Chenausky and Joel MacAuslan. Utilization of microprocessors in voice quality improvement: the electrolarynx. *Current Opinion in Otolaryngology & Head & Neck Surgery.*, 8(3):138–142, 2000.

[CSMG97] David Cole, Sridha Sridharan, Miles Moody, and Shlomo Geva. Application of noise reduction techniques for alaryngeal speech enhancement. In *Proc. IEEE TENCON'97*, pages 491–494, Brisbane, Australia, December 1997.

[DFP94] Rob Drullman, Joost M. Festen, and Reinier Plomp. Effect of reducing slow temporal modulations on speech reception. *Journal of the Acoustical Society of America*, 95(5):2670–2680, 1994.

[Diz03] Vedran Dizdarevic. Maschinelles Lernen für die Erkennung von Varianten der gesprochenen deutschen Sprache. Diploma thesis, Graz University of Technology, 2003.

[DJ04] Chr. Draxler and K. Jänsch. Speechrecorder – a universal platform independent multi-channel audio recording software. In *Proc. of the IV. International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 2004.

[EM84] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(6):1109–1121, 1984.

[EM85] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):443–445, 1985.

[EWCM+98] Carol Y. Espy-Wilson, Venkatesh Chari, J. MacAuslan, Caroline Huang, and M. Walsh. Enhancement of electrolaryngeal speech by adaptive filtering. *Journal of Speech, Language and Hearing Research*, 41:1253–1264, 1998.

[Fan97] Gunnar Fant. The voice source in connected speech. *Speech Communication*, 22:125–139, 1997.

[FH84] H. Fujisaki and K. Hirose. Analysis of voice fundamental frequency contours for declarative sentences of japanese. *Journal of the Acoustical Society of Japan*, 5(4):233–241, 1984.

[FZH07] Marcos Faundez-Zanuy and Martin Hagmüller. Sistema de marcas de agua para mejorar la vulnerabilidad de sistemas de reconocimiento de locutores. In *Actas do ill Congreso da sociedade espanola de acustica forense*, pages 167 – 181, Santiago de Compostela, Spain, 2007.

[FZHK06] Marcos Faundez-Zanuy, Martin Hagmüller, and Gernot Kubin. Speaker verification security improvement by means of speech watermarking. *Speech Communication*, 48(12):1608–1619, December 2006.

[FZHK07] Marcos Faundez-Zanuy, Martin Hagmüller, and Gernot Kubin. Speaker identification security improvement by means of speech watermarking. *Pattern Recognition*, 40(11):3027–3034, November 2007.

[FZHKK06] M. Faundez-Zanuy, Martin Hagmüller, Gernot Kubin, and Bastiaan Kleijn. The COST-277 speech database. In *Nonlinear Analyses and Algorithms for Speech Processing*, volume 3817/2005 of *Lecture Notes in Computer Science*, pages 100–107. Springer, 2006.

[FZJE+06] Marcos Faundez-Zanuy, Léonard Janer, Anna Esposito, Antonio Satue-Villar, Josep Roure, and Virginia Espinosa-Duro, editors. *Nonlinear Analyses and Algorithms for Speech Processing: International Conference on Non-Linear Speech Processing, NOLISP 2005*, volume 3817 of *Lecture Notes in Computer Science*. Springer-Verlag GmbH, February 2006.

[GBB97] Erwin Geerts, Antoinette L. Bouhuys, and Gerda M. Bloem. Nonverbal support giving induces nonverbal support seeking in depressed patients. *Journal of Clinical Psychology*, 53(1):35–39, 1997.

[GHK+04] E.A. Goldstein, J.T. Heaton, J.B. Kobler, G.B. Stanley, and R.E. Hillman. Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity. *IEEE Transactions on Biomedical Engineering*, 51(2):325–332, 2004.

[GHSH07] Ehab A. Goldstein, James T. Heaton, Cara E. Stepp, and Robert E. Hillman. Training effects on speech production using a hands-free electromyographically controlled electrolarynx. *Journal of Speech, Language and Hearing Research*, 50(2):335–351, 2007.

[GOG+99] Antoine Giovanni, Maurice Ouaknine, Bruno Guelfucci, Ping Yu, Michel Zanaret, and Jean-Michel Triglia. Nonlinear behavior of vocal fold vibration: The role of coupling between the vocal folds. *Journal of Voice*, 13(4):465–476, 1999.

[GOT99] Antoine Giovanni, Maurica Ouaknine, and Jean-Michel Triglia. Determination of largest Lyapunov exponents of vocal signals: Application to unilateral paralysis. *Journal of Voice*, 13(3):341–354, 1999.

[Gri98] Clifford J. Griffin. Artificial larynx with frequency control. US Patent 5.812.681, September 22 1998.

[Hag01] Martin Hagmüller. Recognition of regional variants of German using prosodic features. Diploma thesis, Graz University of Technology, May 2001.

[Hag07] Martin Hagmüller. Pitch contour from formants for alaryngeal speech. In *International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, pages 205–208, Firenze, Italy, December 13–15 2007.

[HBTS95] H. Herzel, D. Berry, I. Titze, and I. Steinecke. Nonlinear dynamics of the voice: Signal analysis and biomechanical modeling. *CHAOS*, 5(1):30–34, 1995.

[Hei] Heimomed. http://www.heimomed.de.

[HHK03] Horst Hering, Martin Hagmüller, and Gernot Kubin. Safety and security increase for air traffic management through unnoticable watermark aircraft identification tag transmitted with the vhf voice communication. In *Proc. 22nd Digital Avionics Systems Conference (DASC)*, pages 4.E.2–1–10, Indianapolis, Indiana, October 2003.

[HHKK04] Martin Hagmüller, Horst Hering, Andreas Kröpfl, and Gernot Kubin. Speech watermarking for air traffic control. In *Proc. of 12th European Signal Processing Conference*, pages 1653–1656, Vienna, Austria, September 6–10 2004.

[HHKM99] K.M. Houston, R.E. Hillman, J.B. Kobler, and G.S. Meltzner. Development of sound source components for a new electrolarynx speech prosthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 2347–2350, 1999.

[Hir81] M. Hirano. *Clinical Examination of Voice*. Springer, New York, 1981.

[HK02a] Martin Hagmüller and Gernot Kubin. Akustische Punktortung von Wasserleitungsschäden. Technical report, Graz University of Technology - Signal Processing and Speech Communication Laboratory, 2002.

[HK02b] Martin Hagmüller and Gernot Kubin. Erkennung regionaler Varianten des Deutschen unter Verwendung prosodischer Merkmale. In *Fortschritte der Akustik, DAGA02*, Bochum, Germany, March 2002.

[HK03] Martin Hagmüller and Gernot Kubin. Poincaré sections for pitch mark determination in dysphonic speech. In *Proceedings of 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, pages 281–284, Firenze, Italy, December 10 – 12 2003.

[HK05a] Martin Hagmüller and Gernot Kubin. Poincare sections for pitch mark determination. In *Proc. NOLISP*, pages 107–113, Barcelona, Spain, April 19-22 2005.

[HK05b] Martin Hagmüller and Gernot Kubin. Speech watermarking for air traffic control. Technical Report EEC Note 2005-05, Eurocontol Experimental Center, 2005.

[HKM00] Rainer Hegger, Holger Kantz, and Lorenzo Matassini. Denoising human speech signals using chaos-like features. *Physical Review Letters*, 84:3197, 2000.

[HKM01] Rainer Hegger, Holger Kantz, and Lorenzo Matassini. Noise reduction for human speech signals by local projections in embedding spaces. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 48(12):1454–1461, 2001.

[HKS+99] S. Hirano, H. Kojima, K. Shoji, K. Kaneko, I. Tateya, R. Asato, and K. Omori. Vibratory analysis of the neoglottis after surgical intervention of cricopharyngeal myotomy and implantation of tracheal cartilage. *Arch Otolaryngol Head Neck Surg*, 125(12):1335–1340, Dec 1999.

[HM94] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.

[HM99] M. Higashikawa and F. D. Minifie. Acoustical-perceptual correlates of "whisper pitch" in synthetically generated vowels. *Journal of Speech, Language and Hearing Research*, 42(3):583–591, Jun 1999.

[HS83] John N. Holmes and Adrian P. Stephens. Acoustic correlates of intonation in whispered speech. *Journal of the Acoustical Society of America*, 73(S1):S87, May 1983.

[HWA95] H. Hermansky, E.A. Wan, and C. Avendano. Speech enhancement based on temporal processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 405–408, 1995.

[IPJ05] Kevin M. Indrebo, Richard J. Povinelli, and Michael T. Johnson. Subbanded reconstructed phase spaces for speech recognition. *Speech Communication*, 48(7):760–774, 2005.

[ITU94] ITU. A method for subjective performance assessment of the quality of speech voice output devices. ITU-T Recommendation P.85, June 1994. Series P: Telephone Transmission Quality.

[ITU96] ITU. Methods for objective and subjective assessment of quality - Methods for subjective determination of transmission quality. ITU-T Recommendation P.800, August 1996. Series P: Telephone Transmission Quality.

[JCB02] Esther Grabe John Coleman and Bettina Braun. Larynx movements and intonation in whispered speech. Summary of research supported by British Academy Grant SG-36269, 2002.

[JGN97] Hector Javkin, Michael Galler, and Nacy Niedzielski. Enhancement of esophageal speech by injection noise rejection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1207–1210, Munich, Germany, April 1997.

[JLPY03] M.T. Johnson, A.C. Lindgren, R.J. Povinelli, and Xiaolong Yuan. Performance of nonlinear speech enhancement using phase space reconstruction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 920–923, Hong Kong, April 6–10 2003.

[JPP02a] Donald G. Jamieson, Vijay Parsa, and Moneca C. Price. Interactions of speech coders and atypical speech I: Effects on speech intelligibility. *Journal of Speech, Language and Hearing Research*, 45:482–493, June 2002.

[JPP02b] Donald G. Jamieson, Vijay Parsa, and Moneca C. Price. Interactions of speech coders and atypical speech II: Effects on speech quality. *Journal of Speech, Language and Hearing Research*, 45:689–699, August 2002.

[JZM06] Jack J. Jiang, Yu Zhang, and Clancy McGilligan. Chaos in voice, from modeling to measurement. *Journal of Voice*, 20(1):2–17, 2006.

[KAHP99] Noboru Kanedera, Takayuki Arai, Hynek Hermansky, and Misha Pavel. On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication*, 28(1):43–55, May 1999.

[Kie04] Michael Kiefte. Acoustic characteristics of whispered vowels. *Journal of the Acoustical Society of America*, 116(4):2546–2546, 2004.

[KKP+06] R. Kazi, E. Kiverniti, V. Prasad, R. Venkitaraman, C. M. Nutting, P. Clarke, P. Rhys-Evans, and K. J. Harrington. Multidimensional assessment of female tracheoesophageal prosthetic speech. *Clin Otolaryngol*, 31(6):511–517, Dec 2006.

[Kle02] W. Bastiaan Kleijn. Enhancement of coded speech by constrained optimization. In *Proc. of IEEE Workshop on Speech Coding*, Tsukuba, Ibaraki, Japan, October 2002.

[Kli93] F. Klingholz. Overtone singing: Productive mechanisms and acoustic data. *Journal of Voice*, 7(2):118 – 122, 1993. The Voice Foundation's 22nd Annual Symposium.

[KM96] Arun Kumar and S. K. Mullick. Nonlinear dynamical analysis of speech. *Journal of the Acoustical Society of America*, 100(1):615–629, 1996.

[KM05] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. 13(6):1098–1109, Nov. 2005.

[KPK+07] Rehan A. Kazi, Vyas M.N. Prasad, Jeeve Kanagalingam, Christopher M. Nutting, Peter Clarke, Peter Rhys-Evans, and Kevin J. Harrington. Assessment of the formant frequencies in normal and laryngectomized individuals using linear predictive coding. *Journal of Voice*, 21(6):661–668, 2007.

[KS04] Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis.* Cambridge University Press, 2nd edition, 2004.

[Kub95] Gernot Kubin. Nonlinear processing of speech. In W.B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, chapter 16, pages 557–610. Elsevier, 1995.

[Kub97] Gernot Kubin. Poincaré section techniques for speech. In *Proc. of IEEE Workshop on Speech Coding for Telecommunication '97*, pages 7–8, Pocono Manor, PA, September 1997.

[LB03] Jacqueline S. Laures and Kate Bunton. Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions. *Journal of Communication Disorders*, 36(6):449–464, 2003.

[LB06] A. Loscos and J. Bonada. Esophageal voice enhancement by modeling radiated pulses in frequency domain. In *Proceedings of 121st Convention of the Audio Engineering Society*, San Francisco, CA, USA, October 3-6 2006.

[Lin74] H. R. Lindman. *Analysis of variance in complex experimental designs.* W. H. Freeman & Co, San Francisco, 1974.

[LMMR06] M. Little, P. McSharry, I. Moroz, and S. Roberts. Nonlinear, biophysically-informed speech pathology detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1080–1083, 2006.

[LN07] Hanjun Liu and Manwa L. Ng. Electrolarynx in voice rehabilitation. *Auris Nasus Larynx*, 34(3):327–332, September 2007.

[Loh03] J. Lohscheller. *Dynamics of the Laryngectomee Substitute Voice Production.* PhD thesis, Shaker-Verlag, Aachen, Germany, 2003.

[LW99] Jacqueline S. Laures and Gary Weismer. The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language and Hearing Research*, 42(5):1148–1156, 1999.

[LZWW06a] Hanjun Liu, Qin Zhao, Mingxi Wan, and Supin Wang. Application of spectral subtraction method on enhancement of electrolarynx speech. *Journal of the Acoustical Society of America*, 120(1):398–406, 2006.

[LZWW06b] Hanjun Liu, Qin Zhao, Mingxi Wan, and Supin Wang. Enhancement of electrolarynx speech based on auditory masking. *IEEE Transactions on Biomedical Engineering*, 53(5):865–874, May 2006.

[Man99] Iain Mann. *An Investigation of Nonlinear Speech Synthesis and Pitch Modification Techniques.* PhD thesis, University of Edinburgh, 1999.

[Mar01] Rainer Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9(5):504–512, 2001.

[MC02] Robert W. Morris and Mark A. Clements. Reconstruction of speech from whispers. *Medical Engineering & Physics*, 24(7-8):515–520, 2002.

[MDB01] Claudia Manfredi, Massimo D'Aniello, and Piero Bruscaglioni. A simple subspace approach for speech denoising. *Log Phon Vocol*, 26:179–192, 2001.

[MDEWM99] Kun Ma, Pelin Demirel, Carol Espy-Wilson, and Joel MacAuslan. Improvement of electrolaryngeal speech by introducing normal excitation information. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 323–326, Budapest, Hungary, September 1999.

[ME56] Werner Meyer-Eppler. Realization of prosodic features in whispered speech. *Journal of the Acoustical Society of America*, 28(4):760, 1956.

[ME57] Werner Meyer-Eppler. Realization of prosodic features in whispered speech. *Journal of the Acoustical Society of America*, 29(1):104–106, 1957.

[Mel03] Geoffrey Meltzner. *Perceptual and Acoustic Impacts of Aberrant Properties of Electrolaryngeal Speech*. PhD thesis, Massachusetts Institute of Technology, 2003.

[MH99] Kenji Matsui and Noriyo Hara. Enhancement of esophageal speech using formant synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona, March 1999.

[MH05] Geoffrey S. Meltzner and Robert E. Hillman. Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *Journal of Speech, Language and Hearing Research*, 48(4):766–779, August 2005.

[Mix98] Hansjörg Mixdorff. *Intonation Patterns of German - Quantitative Analysis and Synthesis of F0 Countours*. PhD thesis, Technische Universität Berlin, 1998.

[MKH03] Geoffrey S Meltzner, James B Kobler, and Robert E Hillman. Measuring the neck frequency response function of laryngectomy patients: Implications for the design of electrolarynx devices. *Journal of the Acoustical Society of America*, 114(2):1035–1047, Aug 2003.

[ML95a] Dieter Maurer and Theodor Landis. F0-dependence, number alteration, and non-systematic behaviour of the formants in German vowels. *International Journal of Neuroscience*, 83(1):25–44, 1995.

[ML95b] Eric Moulines and Jean Laroche. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16:175–205, 1995.

[MM98] Iain Mann and Steve McLaughlin. A nonlinear algorithm for epoch marking in speech signals using poincare maps. In *Proceedings of the European Signal Processing Conference*, volume 2, pages 701–704, Rhodes Greece, September 1998.

[MM02a] Lorenzo Matassini and Claudia Manfredi. Noise reduction for vocal pathologies. *Medical Engineering & Physics*, 24(7-8):547–552, 2002.

[MM02b] Lorenzo Matassini and Claudia Manfredi. Software correction of vocal disorders. *Computer Methods and Programs in Biomedicine*, 68(2):135–145, 2002.

[MMCB+06] Mieke Moerman, Jean-Pierre Martens, Lise Crevier-Buchman, Else de Haan, Stephanie Grand, Christophe Tessier, Virginie Woisard, and Philippe Dejonckere. The INFVo perceptual rating scale for substitution voicing: Development and reliability. *European Archives of Oto-Rhino-Laryngology*, 263(5):435–439, May 2006.

[MMVdB+06] M. Moerman, J. Martens, M. Van der Borgt, M. Peleman, M. Gillis, and P. Dejonckere. Perceptual evaluation of substitution voices: Development and evaluation of the (I)INFVo rating scale. *European Archives of Oto-Rhino-Laryngology*, 263(2):183–187, February 2006.

[MPL+07] H. Meister, V. Pyschny, M. Landwehr, P. Wagner, M. Walger, and H. von Wedel. Experimente zur Perzeption prosodischer Merkmale mit Kochleaimplantaten. *HNO*, 55(4):264–270, April 2007.

[MPL+08] H. Meister, V. Pyschny, M. Landwehr, P. Wagner, M. Walger, and H. von Wedel. Konzeption und Realisierung einer Prosodie-Testbatterie. *HNO*, 56:340–348, 2008.

[MTM00] Tova Most, Yishai Tobin, and Ravit Cohen Mimran. Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *Journal of Communication Disorders*, 33:165–181, 2000.

[NAW94] T. Nawka, L. C. Anders, and J. Wendler. Die Auditive Beurteilung Heiserer Stimmen nach dem RBH-System. *Sprache - Stimme - Gehör*, 18:130–133, 1994.

[NB93] Robert L. Norton and Robert S. Bernstein. 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, March 1993.

[NBK+97] Elmar Nöth, Anton Batliner, Andreas Kießling, Ralf Kompe, and Heinrich Niemann. Prosodische Information: Begriffsbestimmung und Nutzen für das Sprachverstehen. In *19. DAGM-Symposium Mustererkennung*, pages 37–52, London, UK, 1997. Springer-Verlag.

[NBW+02] E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann. On the use of prosody in automatic dialogue understanding. *Speech Communication*, 36(1-2):45–62, January 2002.

[NP92] Joachim Neppert and Magnús Pétursson. *Elemente einer Akustischen Phonetik*. Buske, Hamburg, 3. edition, 1992.

[NTSS06] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech. In *International Conference of Speech and Language Processing - Interspeech*, pages 1395–1398, Pittsburgh, PA, USA, September 17–21 2006.

[NWWL03] H. J. Niu, M. X. Wan, S. P. Wang, and H. J. Liu. Enhancement of electrolarynx speech using adaptive noise cancelling based on independent component analysis. *Medical and Biological Engineering and Computing*, V41(6):670–678, November 2003.

[PB52] Gordon E. Peterson and H.L Barney. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24:175–184, 1952.

[PBBL02] P.C. Pandey, S.M. Bhandarkar, G.K. Bachher, and P.K. Lehana. Enhancement of alaryngeal speech using spectral subtraction. In *14th International Conference on Digital Signal Processing (DSP), 2002*, volume 2, pages 591–594 vol.2, 2002.

[PPL03] Santosh S. Pratapwar, Prem C. Pandey, and Parveen K. Lehana. 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)*, 2003.

[PY06] A. Del Pozo and S. Young. Continuous tracheosophageal speech repair. In *Proceedings of the European Signal Processing Conference*, Florence, Italy, 2006.

[QW91] Yingyong Qi and Bernd Weinberg. Low-frequency energy deficit in electrolaryngeal speech. *Journal of Speech and Hearing Research*, 34(6):1250–1256, 1991.

[QWBH95] Yingyong Qi, Bernd Weinberg, Ning Bi, and Wolfgang J. Hess. Minimizing the effect of period determination on the computation of amplitude perturbation in voice. *Journal of the Acoustical Society of America*, 97(4):2525–2532, 1995.

[Sch95] Thomas Schreiber. Efficient neighbor searching in nonlinear time series analysis. *International Journal of Bifurcation and Chaos*, 5:349–358, 1995.

[Sch03] Jean Schoentgen. Decomposition of vocal cycle length perturbations into vocal jitter and vocal microtremor, and comparison of their size in normophonic speakers. *Journal of Voice*, 17(2):114–125, June 2003.

[Sch06] Jean Schoentgen. Vocal cues of disordered voices: An overview. *Acta Acustica united with Acustica*, 92(5):667–680(14), September/October 2006.

[Sch07] Steven M. Schimmel. *Theory of Modulation Frequency Analysis and Modulation Filtering, with Applications to Hearing Devices.* PhD thesis, University of Washington, 2007.

[SCM98] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *Speech and Audio Processing, IEEE Transactions on*, 6(2):131–142, 1998.

[Ser] Servona. http://www.servona.de.

[SFB00] V. Stahl, A. Fischer, and R. Bippus. Quantile based noise estimation for spectral subtraction and Wiener filtering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1875–1878 vol.3, 2000.

[SH00] J.A. Schiefer and R. Hagen. Rehabilitation laryngektomierter Karzinompatienten. *Der Onkologe*, 6(1):36–43, January 2000.

[SHN+06] Maria Schuster, Tino Haderlein, Elmar Nöth, Jörg Lohscheller, Ulrich Eysholdt, and Frank Rosanowski. Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. *European Archives of Oto-Rhino-Laryngology*, 263(2):188–193, February 2006.

[SKNBF+06] M. Sliwinska-Kowalska, E. Niebudek-Bogusz, M. Fiszer, T. Los-Spychalska, P. Kotylo, B. Sznurowska-Przygocka, and M. Modrzewska. The prevalence and risk factors for occupational voice disorders in teachers. *Folia Phoniatr Logop*, 58(2):85–101, 2006.

[SLM95] Y. Stylianou, J. Laroche, and E. Moulines. High-quality speech modification based on a harmonic + noise model. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 451–454, Madrid, Spain, September 18-21 1995.

[SS02] Jeffrey P. Searl and Larry H. Small. Gender and masculinity-femininity ratings of tracheoesophageal speech. *Journal of Communication Disorders*, 35(5):407–420, 2002.

[STP03] Jan Svec, Ingo Tietze, and Peter Popolo. Vocal dosimetry: Theoretical and practical issues. In *Proceeding Papers for the Conference Advances in Quantitative Laryngology, Voice and Speech Research (AQL)*, Hamburg, 2003.

[SvH96] Agaath M. C. Sluijter and Vincent J. van Heuven. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100(4):2471–2485, 1996.

[SYC91] Timothy Sauer, James A. Yorke, and Martin Casdagli. Embedology. *J. Stat. Phys.*, 65(3/4):579–616, 1991.

[Tak81] Floris Takens. *Detecting Strange Attractors in Turbulence*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer, New York, 1981.

[Ter02] D.E. Terez. Robust pitch determination using nonlinear state-space embedding. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 345–348, Orlando, Florida, 2002.

[tHCC90] Johan 't Hart, René Collier, and Antonie Cohen. *A perceptual study of intonation – An experimental-phonetic approach to speech melody.* Cambridge Studies in Speech Science and Communication. Cambridge University Press, 1990.

[Tho69] I. B. Thomas. Perceived pitch of whispered vowels. *Journal of the Acoustical Society of America*, 46(2P2):468–470, 1969.

[Tis90] N. Tishby. A dynamical systems approach to speech processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 365–368, Albuquerque, NM, April 1990.

[Tit94] Ingo R. Titze. Workshop on acoustic voice analysis - summary statement. In *Proc. Workshop on Acoustic Voice Analysis*, Denver, Colorado, February 1994.

[UITM94] N. Uemi, T. Ifukube, M. Takahashi, and J. Matsushima. Design of a new electrolarynx having a pitch control function. In *Proceeding of 3rd IEEE Intl. Workshop on Robot and Human Communication, RO-MAN*, pages 198–203, Nagoya, 18-20 Jul 1994.

[vABKvBPH06] Corina J. van As-Brooks, Florien J. Koopmans-van Beinum, Louis C.W. Pols, and Frans J.M. Hilgers. Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. *Journal of Voice*, 20(3):355–368, September 2006.

[vABRKvBH97] Corina J. van As-Brooks, Annemieke M.A. Ravesteijn, Florien J. Koopmans-van Beinum, and Louis C.W. Hilgers, Frans J.M.and Pols. Formant frequencies of dutch vowels in tracheoesophageal speech. *Institute of Phonetic Sciences Proceedings*, 21:143–153, 1997.

[vdTvGL+06] M. van der Torn, C. D. L. van Gogh, I. M. Verdonck-de Leeuw, J. M. Festen, G. J. Verkerke, and H. F. Mahieu. Assessment of alaryngeal speech using a sound-producing voice prosthesis in relation to sex and pharyngoesophageal segment tonicity. *Head & Neck*, 28(5):400–412, 2006.

[VHH98] Peter Vary, Ulrich Heute, and Wolfgang Hess. *Digitale Sprachsignalverarbeitung*. B.G. Teubner Stuttgart, 1998.

[VMT92] H. Valbret, E. Moulines, and J. P. Tubach. Voice transformation using PSOLA technique. *Speech Communication*, 11(2-3):175–187, June 1992.

[vRdKNQ02] M. A. van Rossum, G. de Krom, S. G. Nooteboom, and H. Quené. 'Pitch' accent in alaryngeal speech. *Journal of Speech, Language and Hearing Research*, 45:1106–1118, December 2002.

[WHS80] Bernd Weinberg, Yoshiyuki Horii, and Bonnie E. Smith. Long-time spectral and intensity characteristics of esophageal speech. *Journal of the Acoustical Society of America*, 67(5):1781–1784, 1980.

[Wre98] Birgit Wrentschur. *Die Perzeptuelle Evaluation von Stimmstörungen*. PhD thesis, University of Graz, 1998. Incl. 2 audio CDs with Disordered Voice Samples.

[WSKE96] Jürgen Wendler, Wolfram Seidner, Gerhard Kittel, and Ulrich Eyshold. *Lehrbuch der Phoniatrie und Pädaudiologie*. Georg Thieme Verlag, Stuttgart, 1996.

[Zad50] Lotfi A. Zadeh. Frequency analysis of variable networks. *Proceedings of the IRE*, 38(3):291–299, 1950.