PhD Thesis

Diplophonic Voice

Definitions, models, and detection

conducted at the
Division of Phoniatrics-Logopedics
Department of Otorhinolaryngology
Medical University of Vienna, Austria

and the
Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria

by
Dipl.-Ing. Philipp Aichinger

Examiners:

Univ.-Prof. Dipl.-Ing. Dr.techn. Gernot Kubin, Graz University of Technology, Austria

Dr.habil. Jean Schoentgen (FNRS), Université Libre de Bruxelles, Belgium

Ao.Univ.-Prof. Dr.med.univ. Berit Schneider-Stickler, Medical University of Vienna, Austria

Vienna, December 16, 2014

Abstract

Voice disorders need to be better understood because they may lead to reduced job chances and social isolation. Correct treatment indication and treatment effect measurements are needed to tackle these problems. They must rely on robust outcome measures for clinical intervention studies. Diplophonia is a severe and often misunderstood sign of voice disorders. Depending on its underlying etiology, diplophonic patients typically receive treatment such as logopedic therapy or phonosurgery. In the current clinical practice diplophonia is determined auditively by the medical doctor, which is problematic from the viewpoints of evidence-based medicine and scientific methodology. The aim of this thesis is to work towards objective (i.e., automatic) detection of diplophonia.

A database of 40 euphonic, 40 diplophonic and 40 dysphonic subjects has been acquired. The collected material consists of laryngeal high-speed videos and simultaneous high-quality audio recordings. All material has been annotated for data quality and a non-destructive data pre-selection is applied. Diplophonic vocal fold vibration patterns (i.e., glottal diplophonia) are identified and procedures for automated detection from laryngeal high-speed videos are proposed. Frequency Image Bimodality is based on frequency analysis of pixel intensity time series. It is obtained fully automatically and yields classification accuracies of 78 % for the euphonic negative group and 75 % for the dysphonic negative group. Frequency Plot Bimodality is based on frequency analysis of glottal edge trajectories. It processes spatially segmented videos, which are obtained via manual intervention. Frequency Plot Bimodality obtains slightly higher classification accuracies of 82.9 % for the euphonic negative group and 77.5 % for the dysphonic negative group.

A two-oscillator waveform model for analyzing acoustic and glottal area diplophonic waveforms is proposed and evaluated. The model is used to build a detection algorithm for secondary oscillators in the waveform and to define the physiologically interpretable "Diplophonia Diagram". The Diplophonia Diagram yields a classification accuracy of 87.2 % when distinguishing diplophonia from severely dysphonic voices. In contrast, the performance of conventional hoarseness features is low on this task. Latent class analysis is used to evaluate the used ground truth from a probabilistic point of view. The used expert annotations achieve very high sensitivity (96.5 %) and perfect specificity (100 %). The Diplophonia Diagram is the best available automatic method for detecting diplophonic phonation intervals from speech.

The Diplophonia Diagram is based on model structure optimization, audio waveform modeling and analysis-by-synthesis, which enables a more suitable description of diplophonic signals than conventional hoarseness features. Analysis-by-synthesis and waveform modeling had already been carried out in voice research, but systematic investigation of model structure optimization with respect to perceived voice quality is novel. For diplophonia, the switch between one and two oscillators is crucial. Optimal model structure is a qualitative outcome that may be interpreted physiologically and one may conjecture that model structure optimization is also useful for describing other voice phenomena than diplophonia. The obtained descriptors might be more easily accepted by clinicians than the conventional ones.

Useful definitions of diplophonia focus on the levels of perception, acoustics and glottal vibration. Due to its subjectivity, it is suggested to avoid the sole use of the perceptual definition in clinical voice assessment. The glottal vibration level connects with distal causes, which is of high clinical interest but difficult to assess. The definition at the acoustic level via two-oscillator waveform models is favored and used for in vivo testing. Updating definitions and terminology of voice phenomena with respect to different levels of description is suggested.

Contents

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)

Preface

Research is puzzling. In undergraduate studies, one is told very clearly what puzzles need to be solved, maybe also where to find the right pieces and where to put them. That does not usually happen in doctoral studies.

When I arrived at the Medical University of Vienna, I started to search for puzzle pieces, and I found plenty of them. I started to collect all pieces that looked interesting to me, and I put all those into a little bag. Of course, I found out that most of my pieces did not fit together, maybe because they belonged to different puzzles, or maybe because there were several pieces missing in between them. So I started to collect more and more pieces, and some day I found two pieces that fitted together.

My two pieces were comparable in size and color, so I decided to search for some more that look like them. At some point I had collected one corner piece and a few edge pieces, and I realized that those might be more important than the middle pieces. So I collected more of them, and some day I was able to put ten edge pieces together. I decided to bring my puzzle artifact to a conference, where I met a lot of interesting and inspiring people. I met people who have already built several big and beautiful puzzles, and also people who have just put their first two pieces together. So that was a good experience, and I realized that I am on the right way.

However, I had to learn that not all pieces of my puzzle looked like those that I had. The pieces to look for were comparable in size, but there were also pieces in other colors. That was again a bit more difficult, because I did not know what color they have. Fortunately, some day I found my puzzle box, which was covered under thousands of puzzle pieces. I had not been able to find it in the beginning of my journey, because I was looking for puzzle pieces and not for a box. There were still some puzzle pieces in the box, and most of them belonged to my puzzle. But more importantly, I was able to see my puzzle on the cover of the box. Unfortunately a big part of the box was torn apart, because somebody probably had angrily thrown away the box years ago. Nevertheless, I was able to recognize my puzzle artifacts on the boxes cover, so I could be sure that my puzzle is something that can actually be built.

After some other months of puzzling, I realized that I had to finish some day. I looked at my puzzle and saw that I have put together 34 pieces, which are connected to the puzzle's edges, and that I have several other parts of 3 or 4 pieces, which were separated from the bigger part. When I compared my parts to the cover of the box I was able to approximate where the isolated pieces could be located in relation to the bigger part. All in all I was able to get a quite good impression of how the puzzle should look like.

I spent the next months with taking photos of my puzzle, and I took them from several different perspectives. I took close up photos from the perspective of signal processing experts, from which the puzzle looked very huge. I realized that there were still some very tiny and rare pieces missing, but I decided to leave the puzzle as it is and not to look for pieces that I was not able to find. I also took some photos from the perspective of clinical experts. The puzzle looked rather small from there and maybe a bit to detailed, but for me it felt just right as it was.

Now that I have finished, I hope that people like to watch my pieces and maybe to use them for their own work. I have tried to put together a puzzle that nobody has ever tried to build before. It is far from being finished, but I hope that I have found some cornerstones that I can use in my upcoming work and that may perhaps also inspire some of my colleagues.

Acknowledgements

I would like to thank several institutions and people who made it possible for me to conduct my doctoral studies. First of all I would like to thank my advisors Gernot Kubin, Jean Schoentgen and Berit Schneider-Stickler, for supporting me throughout the creation of my thesis. Gernot guided me to achieve a high level of scientific methodology, Jean provided excellent groundwork, ideas and solutions in modeling, and Berit asked me the most relevant questions from a clinical point of view. Wolfgang Bigenzahn encouraged me to enroll my scientific field at our department and Doris-Maria Denk-Linnert was an anchor during times of difficult decisions. Imme Roesner shared hours of her precious time for data collection and discussions on videokymograms. Matthias Leonhard inspired me with great ideas and Felicitas Feichter and Birgitta Aichstill were always good company. I want to thank Martin Hagmüller and Anna Fuchs for endless discussions on my work, and for providing detailed knowledge about how to conduct interdisciplinary research in the field of sound engineering and medical signal processing. I want to thank Martin Urschler, Fabian Schenk and Christoph Aigner for their work on image segmentation and Paul Berghold for his work on inverse filtering.

I would like to thank the Medical University of Vienna, Department of Otorhinolaryngology, Division of Phoniatrics and Logopedics for affiliating me during the creation of my thesis and the Graz University of Technology, Signal Processing and Speech Communication Lab for advising me. Both institutions provided me with the facilities and resources I needed for conducting the research. I would like to thank the University of Music and Performing Arts, Institute of Electronic Music and Acoustics, where I finished my Master's studies. Alois Sontacchi guided me through the creation of my Master's thesis, which found its way into my PhD-Thesis by means of the masking model. I also want to thank the Richard Wolf GmbH, who provided the high-speed camera and I want to thank all team members of our department for integrating me into their group. I want to thank Michael Döllinger and Denis Dubrovskiy for the Glottis Analysis Tools, Andreas Krenmayr for a very useful tool for handling high-speed videos, Harald Herkner for valuable hints on the methodology of diagnostic accuracy studies and Viktor Hoyos for advices in self-management.

I want to thank the experts that attended the discussion sessions on diplophonia, which are in alphabetical order: Ronald Baken (New York Eye and Ear Infirmary of Mount Sinai), James Daugherty (University of Kansas), Molly Erickson (University of Tennessee), Christian Herbst (Palacký University Olomouc), Michael Johns (Emory Voice Center, Atlanta), Christian Kasess (Austrian Academy of Sciences), Sylvia Moosmüller (Austrian Academy of Sciences), Robert Sataloff (Drexel University College of Medicine, Philadelphia), Nancy Pearl Solomon (Walter Reed Army Medical Center, Washington), Jan Švec (Palacký University Olomouc) and Sten Ternström (KTH Royal Institute of Technology, Stockholm).

I also would like to thank my Master's study colleagues who were excellent company and always helped me out when I needed to stay in Graz overnight during my doctoral studies.

I want to take the opportunity to thank my band mates from No Head on my Shoulders. With some of them I have been playing together for almost 20 years now. It is a pity that I have not been able to spend more time with music making during the last years, and I want to thank them for compensating my lack of time.

My biggest thank however goes to my parents and family, and most importantly to Nathalie. Her never ending patience, kindness, smile and love is the biggest support I can imagine.

1 Introduction

"By three methods we may learn wisdom: First, by reflection, which is noblest; Second, by imitation, which is easiest; and third by experience, which is the bitterest."
– *Confucius*

1.1 Motivation

Verbal communication is one of the most important achievements of human beings. Any communication handicap can lead to reduced job chances, social isolation or loss of quality of live in the long run [1]. Failures in the diagnosis and treatment of patients suffering from voice disorders may result in substantial follow-up costs, and so the care of voice disorders is of great importance for assuring our well-being as a community.

A need for treatment of voice disorders should be recognized in time, treatment should never be administered unnecessarily, and the effectiveness of treatment techniques should be continuously monitored in clinical intervention studies. However, correct indication for treatment and treatment effect measurement are possibly not always accomplished and must be investigated. As a prerequisite for achieving these goals, valid, reliable, objective and accurate methods for the diagnosis of voice disorders are needed.

Diplophonia is a severe and often misunderstood sign of voice disorders. As a subcategory to dysphonia, it is classified into R49 in the ICD-10. Depending on the underlying etiology, diplophonic patients typically receive treatment such as phonosurgery. The absence of diplophonia is being used as an outcome measure in clinical intervention studies [2–13]. It is auditively assessed on connected speech by phoniatricians during clinical anamnesis [14]. Auditory assessment suffers from its subjectivity, and so objectification is needed to agree with principles of evidence-based medicine and scientific methodology.

The terms "hoarseness", "irregularity" and "roughness" are often used as supercategories of diplophonia. Irregularity is defined on waveform properties [15], whereas hoarseness and roughness are perceptually defined [14]. Hoarseness is the overall deviation from normal voice quality, excluding deviations in pitch, loudness and rhythm. It is part of the RBH scale (*Rauigkeit*, *Behauchtheit*, *Heiserkeit*) that is used for perceptual voice assessment in German-speaking clinics. The Grade of hoarseness is part of the GRBAS scale (*Grade*, *Roughness*, *Breathiness*, *Asthenia*, *Strain*) that is used internationally. Roughness is the "audible impression of irregular glottal pulses, abnormal fluctuations in fundamental frequency (F0), and separately perceived acoustic impulses (as in vocal fry), including diplophonia and register breaks".

By summarizing several voice phenomena into the categories irregularity, hoarseness or roughness, the accuracy of voice assessment is reduced. This approach might be legitimate for daily use in clinical practice, but from the signal processing perspective and for moving forward the state-of-the-art more subcategories like diplophonia, fry, creak, glottalization and laryngealization should be used [16, 17]. Those subcategories are fuzzily defined however, which hinders scientific communication. By choosing diplophonia as the research focus, the problem of defining irregular phonation is subdivided into several subproblems, which smooths the way towards its solution [18].

Observed diagnoses of clinically diplophonic patients are unilateral paresis, edema, functional dysphonia, cyst, polyp, sulcus, laryngitis, nodules, scar, benign tumor and bamboo nodes. An unilateral paresis is most common in diplophonic subjects. It may manifest as a malfunctioning of the abduction and adduction required for coordinating respiration/phonation and voiced/unvoiced phonation, and/or asymmetric vocal fold tension. The left and the right vocal fold is normally coupled by airstream modulation and collision, which is reduced if the lateral placement of the vocal folds is abnormal or the vocal fold tension is asymmetric. Decoupled vocal folds often vibrate at different frequencies. The anterior and the posterior part of the vocal folds are coupled by connecting tissue. Anterior-posterior desynchronization in unilateral pareses may mainly be due to asymmetric vocal fold tension. The most common therapies for unilateral pareses are the medialization thyroplasty and vocal fold augmentation.

Other diagnoses are less frequent in diplophonic patients. Tissue abnormalities like edema, cyst, polyp, sulcus, nodules and scar may mainly be caused by mechanical stress to the vocal folds during phonation. Thus, vocal misuse plays a role in the development of these impairments. They can be treated by logopedists, teaching of a proper use of the voice, and by phonosurgery, removing pathologic structures. Functional voice disorders are characterized by abnormal voice sounds in the absence of observable organic impairment. Risk factors are multiple, such as constitutional and hormonal. They can be treated by medication, if they arise from a primary disease, or by logopedic therapy. Psychogenic components are addressed by psychiatric therapy. Bamboo nodes arise from rheumatism and can be treated conservatively or by phonosurgery. Laryngitis may be caused by environmental pollutants or infections and can be treated by medication and avoiding pollutant exposure. Benign tumors are treated by phonosurgery.

In diplophonic vocal fold vibration, two glottal oscillators vibrate at different frequencies. If the eigenfrequencies of the oscillators were known one could focus on the mechanical properties (mass and stiffness), which may enable more precise phonosurgery. However, the observed frequencies do not equal the eigenfrequencies due to coupling. From a physical point of view, one does not exactly know how to affect the oscillations in a desirable way. Because of the lack of sufficient data, the determination of the vocal fold eigenfrequencies is an ill-posed problem and no speculations about eigenfrequencies are made in the present thesis. In lesions like edema, cysts, polyps, nodules, laryngitis and tumors, the most obvious mechanical factor is mass. The oscillator eigenfrequency decreases if mass is increased and stiffness is kept constant. If the mass asymmetry is escalating, the coupling forces may not synchronize the glottal oscillators anymore, and diplophonia may occur. Besides, predominantly in lesions like scars and bamboo nodes, the stiffness parameters may change, which changes the oscillators' eigenfrequency.

Oscillator coupling and combination types allow for further differentiation of the phenomenon. In strong coupling (i.e., entrainment) the frequency ratio of the two oscillators is attracted to a ratio of two small integer numbers (e.g., 1:2, 2:3, 2:2 or 3:3). In such a condition the overlap of the two spectra increases and the fundamental frequencies are likely to merge perceptually, which leads to the perception of one pitch only. The threshold for the transition from two pitches to only one is unknown and certainly not only depends on the frequency ratio. Considering conventional masking models, other factors that play a role are the relative energies and the relative spectral envelope of the oscillators. In weaker coupling, the integer frequency ratios involve larger integers (e.g., 3:4, 7:8, 9:18), and the separate fundamental frequency traces may evolve more or less independently. The spectral overlap of the oscillators is smaller than in strong coupling, which may make the two pitches easier to segregate and detect perceptually. In the theoretical case of "anticoupling", the frequency factor is irrational and the signal is aperiodic, which is sometimes referred to as biphonation. Irrational frequency factors are hard to distinguish from rational ones, because measurement time intervals are finite and the fundamental frequencies of the oscillators are time-variant. To understand the full complexity of the coupling problem, one must take into account that coupling is time-variant and non-linear. Small changes in coupling may change the vibration pattern substantially, which is described in the theory of non-linear dynamics and bifurcations [19]. Oscillator combination can either be additive or modulative. In additive combination, one oscillator waveform is added to another, while in modulative combination, one oscillator is the carrier and the other one is a modulator. In reality, these two types of combination often coexist.

For determining the clinical relevance of diplophonia, its prevalence must be known, but in the lack of a unified definition and acceptable diagnostic test, it is not. Within pathologic speakers, the rate of occurrence may be above 25 %, but may highly depend on the definition or the instrument used for detecting diplophonia [20]. The prevalence of diplophonia in the general population is even more uncertain, because diplophonia may also occur in healthy phonation, which makes interpretation of and conclusions from clinical studies dealing with diplophonia difficult. E.g., if diplophonia occurs in 2 % of the healthy general population and also in 2 % of speakers with voice complaints, detecting diplophonia does not have any clinical relevance, because there is no gain of information in assessing diplophonia. Thus, to determine the prevalence and the clinical relevance of diplophonia, a unified definition and the development of an objective diagnostic test for detecting diplophonia is absolutely indispensable.

In the beginning of research on diplophonic voice, it was investigated what kind of physical phenomena cause double pitch sensation. Today's knowledge about human perception clearly says that double pitch sensation is highly individual, variant, ambiguous and subject to training [21, 22]. The factors influencing the perception of two pitches are manifold, and not necessarily clinically relevant. In a clinical setup, the influencing factors might include room acoustics, background noise, the relative position of the patient to the medical doctor, the medical doctor's general impression of the patient, including visual cues from endoscopic images, and training. From the viewpoint of science and evidence-based medicine, more objective methods for voice assessment are needed.

1.2 Milestones in the literature

The first investigations on double pitch phenomena in voice production were made in 1837 [23]. Müller described the occasional occurrence of two pitches in both vocal fold replica and human excised larynx experiments, when asymmetric vocal fold tensions are applied. 20 years later Merkel described double pitch phenomena during singing [24]. He was occasionally able to produce two simultaneous pitches at the lower end of his falsetto register, and at the lower end of his modal register, most frequently in an interval of an octave. In 1864, Gibb presented several clinical cases with double pitch voice, and introduced the term "diplophonia". Türck [25] introduced the German term "Diphthonie" and Rossbach [26] "Diphthongie" which describe the same phenomenon. It was already clear that the vocal fold tension is not the only parameter that has to be taken into account, as the clinical observations and diagnoses are diverse. Scientists had already realized, that the double pitch phenomenon can arise from two spatially distinct oscillators, which are distributed in the anterior-posterior direction. In 1879, Grützner showed that subharmonic double pitch phenomena can be produced with a siren disc by manipulating its holes in a periodic pattern [27]. He created sound signals that agreed with Titze's definition of diplophonia [28]. In 1927, Dornblüth included the terms "Diplophonie" and "Diphthongie" in his clinical lexicon [29].

The first laryngeal high-speed film of a subharmonic phenomenon in healthy voice was presented by Moore and von Leden [30] in 1958. It is remarkable that the authors were already able to produce spatio-temporal plots at three positions along the vocal folds, which inspired researchers for several decades. The glottis oscillated with alternating period lengths in a 1:2 ratio, which corresponds to a musical interval of an octave. The authors called this phenomenon "glottal fry", and still today it is unclear what relationship the terms "diplophonia" and "glottal fry" may have. Smith [31] performed experiments with vocal fold replicas in the same year, and realized that alternating period amplitudes can arise from superior-inferior frequency asymmetry of the vocal folds, i.e., the inferior part of the vocal folds vibrating at a different frequency than the superior part. Švec documented a case of 3:2 superior-inferior frequency asymmetry and argued that open quotient and flow may distinguish the observed phenomenon from vocal fry [32]. In 1969, Ward [33] was the first to discover that a double pitch phenomenon can arise from left-right frequency asymmetry of the vocal folds. He was able to measure the fundamental frequency ratio (approx. 13:10.25) in a case of voluntarily produced diplophonia by a healthy female subject.

Dejonckere published a milestone article on the analysis of auditive diplophonia in 1983 [20]. He examined 72 diplophonic patients by means of acoustic recordings, electroglottography (EGG), photoglottography and stroboscopy. He showed that diplophonia arises at the level of glottal vibration from oscillators at different frequencies, which results in a beat frequency phenomenon. It manifests at the waveform level in repeating patterns of cycle length and/or cycle peak amplitude.

From that time on, mostly methodical studies with small sample sizes have been published. A milestone in mechanical modeling of diplophonia has been set by Ishizaka in 1976 [34]. He used an asymmetric two-mass model (two-masses for each vocal fold) to obtain diplophonic waveforms. In Cavalli's experiment in 1999, seven trained listeners were able to reliably detect diplophonia in audio material from five diplophonic and five dysphonic subjects [35]. Cavalli was the first to report that auditive diplophonia is not equivalent to the presence of subharmonics visually determined from audio spectrograms. Bergan and Sun showed that perception of pitch and roughness of subharmonic signals depends on modulation type and extent as well as F0 and vowel type, while interrater reliability is low [36, 37]. In 2001, Gerratt contributed a theoretical work on the taxonomy of non-modal phonation and warned that the use of divergent levels of description across studies results in ambiguous definitions and hardly generalizable conclusions [17]. The work has not yet received the attention it demands, and so no unification of definitions or change in publication habits have been achieved. In the same year, spectral analysis of laryngeal high-speed videos was introduced by Granqvist [38]. The method enables visualizing laryngeal structures vibrating at different frequencies in a false color representation. Granqvist analyzed vibrations of the ventricular folds and the arytenoid cartilages, but did not test perceptual cues. Kimura reported two diplophonic subjects with spatially distributed oscillation frequencies, as determined by spectral video analysis [12]. Neubauer introduced empirical eigenfunction analysis for describing oscillation modes and quantitatively measuring spatial irregularity [39]. The procedure was tested on an euphonic example, and on one example of left-right and anterior-posterior frequency asymmetry each. Hanquinet proposed a synthesis model for diplophonia and biphonation in 2005 [40]. An amplitude driving function of a pulsed waveform is sinusoidally modulated. In diplophonia, the modulation frequency is the pulse frequency times a ratio of two small integers. In biphonation, it is an irrational number. Sakakibara proposed a voice source model for analysis-by-synthesis of subharmonic voice in 2011 [41]. It was based on physiological observations in laryngeal high-speed videos of three diplophonic subjects (paresis, cyst, papilloma). Pulsed waveforms were amplitude modulated by sawtooth functions. Their sum was amplitude modulated by sinusoidal functions and drove a vocal tract model. The parameter estimation was manual, and the model was used to conduct perceptual experiments with six listeners, judging roughness and the presence of diplophonia. Sakakibara showed that both diplophonia and roughness are increased for 4:3 and 3:2 frequency ratios as compared to a 2:1 ratio. The amplitude ratio of the pulse waveforms was another factor that affected the perception of diplophonia. The ratio needs to be close to 0 dB, otherwise the weaker waveform is masked by the louder one. This result agrees with principles of current masking models. Alonso recently introduced a synthesis model for subharmonic voice that considers additive oscillator combination [42].

The steps towards the full understanding of diplophonic phonation are becoming smaller, probably because no technique superior to laryngeal high-speed video has been invented yet. It is remarkable that diplophonic phonation has been investigated for nearly 200 years now, partly by means of sophisticated methods, and still it is not really clear what diplophonia is.

1.2.1 Three related PhD-theses

To pinpoint the main research questions and aims of the present thesis, three related PhD theses are analyzed for their content in terms of clinical and technical relevance. Moreover, commonalities and differences with the present thesis are identified. The reviewed theses are Michaelis' "Das Göttinger Heiserkeits-Diagramm - Entwicklung und Prüfung eines akustischen Verfahrens zur objektiven Stimmgütebeurteilung pathologischer Stimmen" [43–45], Mehta's "Impact of human vocal fold vibratory asymmetries on acoustic characteristics of sustained vowel phonation" [46] and Shue's "The voice source in speech production: Data, analysis and models" [47].

Das Göttinger Heiserkeits-Diagramm - Entwicklung und Prüfung eines akustischen Verfahrens zur objektiven Stimmgütebeurteilung pathologischer Stimmen

In 1997, Dirk Michaelis published his dissertation "Das Göttinger Heiserkeits-Diagramm - Entwicklung und Prüfung eines akustischen Verfahrens zur objektiven Stimmgütebeurteilung pathologischer Stimmen" (The Göttingen Hoarseness Diagram - development and evaluation of an acoustic method for objective voice quality assessment of pathologic voice) [43–45]. The thesis aimed at creating an objective method for clinical voice quality visualization and may be grouped into five entities. First, the Glottal-to-Noise Excitation Ratio (GNE) has been introduced. The GNE aims at estimating the level of additive noise components in dysphonic voice. Second, informative representations of conventional hoarseness features together with the GNE have been identified to formulate the Göttingen Hoarseness Diagram. Third, the influence of the vocal tract on jitter and shimmer has been investigated. Fourth, perceptual experiments on the GNE, several perturbation measures, roughness and breathiness have been conducted. Finally, the clinical application of the Hoarseness Diagram has been investigated by means of group effect interpretation, recording protocol shortening and case studies of 48 patients. The following paragraphs give more detailed explanations on Michaelis' investigations.

The GNE correlates Hilbert envelopes obtained from a filterbank and thus reports additive noise in disordered voice, because glottal pulse energy is correlated across frequency bands, whereas noise is not. The GNE is obtained as follows: 1) linear prediction based block wise inverse filtering of the speech signal, 2) filterbank analysis, 3) Hilbert envelope calculation, 4) calculation of across frequency correlation and 5) obtaining the maximum of all correlation coefficients. The GNE has been trained and tested on natural and synthetic audio material. On synthetic material, the Normalized Noise Energy [48] and the Cepstral Harmonic to Noise Ratio [49] are prone to jitter and shimmer, whereas the GNE is not. On natural material, it distinguishes normal voice from whispering. With regard to filterbank bandwidth, GNE1, GNE2 and GNE3 are distinguished (1 kHz, 2 kHz and 3 kHz).

The formulation of the Hoarseness Diagram is based on observations from singular value decomposition and information theory. With regard to singular value decomposition, Michaelis analyzed the data space of 20 hoarseness features, obtained from 1799 analysis intervals of pathological and normal speakers. The features were period length, the waveform matching coefficient, jitter, shimmer and the GNE. The average, the minimum, the maximum and the standard deviation of the period length, the waveform matching coefficient and the GNE were considered. For jitter and shimmer, the perturbation quotients (PQ), as well as the perturbations factors (PF) have been used. With regard to block length K, three versions of the PQ have been considered (PQ3, PQ5 and PQ7). The decomposition resulted in a four-dimensional data space. In a second approach, Michaelis analyzed the data space of four hoarseness features, i.e., the mean waveform matching coefficient (MWC), the shimmer (energy perturbation factor), the jitter (period perturbation factor) and the GNE. This approach resulted in a two-dimensional data space. Data from pathologic speakers were distributed along both axes, whereas normal voice and whispering were located at low and high perturbation values separately.

Michaelis selected the best versions of GNE, jitter and shimmer with respect to information theory. The MWC and numerous versions of jitter and shimmer were considered. The period perturbation quotient (K=3) and the energy perturbation quotient (K=15) added the most information to MWC, as compared to other jitter and shimmer combinations. GNE3 added the most information to the perturbation measures, as compared to other noise features. Thus, the Hoarseness Diagram was obtained by projecting the data space onto two dimensions, namely irregularity (MWC, J3, S15) and noise (GNE3). Rotation and origin shift of the axes enabled convenient visualization.

Michaelis investigated the influence of the vocal tract on perturbation measures by experimenting with EGGs, microphone signals, synthetic glottal signals and synthetic acoustic signal. The correlation of perturbation measures obtained from different signals is fair only. A theory of co-modulation has been proposed, i.e., the amount of shimmer depends on jitter that is modulated in the vocal tract. One can conclude that features obtained at the level of glottal vibration should be used differently than acoustic measures.

The correlation of acoustic measures with perceptual ratings has been tested. Data from 20 trained and untrained raters was used. The correlation of GNE with breathiness turned out to be moderate, whereas the picture is less clear for perturbation measures and roughness. It is unknown whether imperfect correlation is due to failures in measurement or in auditive judgment.

The clinical application of the Hoarseness Diagram has been investigated by means of group effect interpretation, shortening of the recording protocol and case studies. Michaelis showed group effects of vowel type, diagnosis and phonatory mechanism. Different vowel types are located at different positions in the Hoarseness Diagram, which is modeled by multidimensional linear regression and a neuronal network. The different diagnosis were tumor, paresis, benign neoplasm, functional dysphonia and neurological dysphonia. Different phonatory mechanisms of substitution voices have been evaluated. Differentiation between different diagnoses or phonatory mechanisms in substitution voice is not possible.

With regard to removing redundancies, a shortening of the recording protocol has been proposed. Michaelis' recording protocol consists of 28 vowels (seven vowels times four pitch levels). In clinical practice obtaining recordings is expensive, because it is time consuming for patient and examiner. Thus, it has been investigated if the recording protocol can be condensed, without losing relevant information. Linear regression or a neuronal network was used to predict the large recording protocol's average of vowel values from only three vowels. The prediction error was in an acceptable range mostly.

Additionally, case studies of 48 patients have been published. Multiple measurements before, during and after treatment are available, which enables generating hypotheses and designing studies to investigate further clinical interpretations.

Michaelis' thesis and the present thesis have in common that an objective method for clinical voice quality is developed. Signal processing methods are used to develop vowel recording features that correlate with perceptual attributes. Differences are that the present thesis aims at diplophonia, which is a subcategory of irregularity or roughness. Michaelis' irregularity is a conglomerate of three established perturbation features: jitter, shimmer and mean waveform matching coefficient, which depend on correct cycle length detection. However, cycle length detection from irregular voice is often error prone. In laryngeal high-speed videos the voice source can be evaluated more straight forwardly than in the audio signal, hence cycle length detection from laryngeal high-speed videos is investigated in the present thesis. It has been hypothesized that splitting up the problem of measuring irregularity into several smaller subproblems smooths the way towards its solution.

Impact of human vocal fold vibratory asymmetries on acoustic characteristics of sustained vowel phonation

Daryush Mehta published his dissertation "Impact of human vocal fold vibratory asymmetries on acoustic characteristics of sustained vowel phonation" in 2010 [46]. Mehta investigated correlations of features obtained from laryngeal high-speed videos with acoustic features. Three corpora of observed data (normal and pathologic speakers) as well as one corpus of synthesized data have been used. Mehta showed that acoustical perturbations can arise from cycle-to-cycle variability in vocal fold asymmetry, both in natural and simulated data. Steady asymmetry does not influence the tested acoustic measures, including spectral tilt and noise.

Investigated measures of laryngeal high-speed videos were both subjective and objective. Left-right phase asymmetry in digital videokymograms was assessed by three raters on a five-point scale. In objective analysis, one measure for left-right amplitude asymmetry, axis shift, open quotient and period irregularity (i.e., a jitter-like video feature) each and two measures for left-right phase asymmetry were considered. In glottal area waveforms (GAW) the open quotient, the speed quotient, the closing quotient and a plateau quotient were considered.

Acoustic analyses considered perturbation (jitter and shimmer), spectral features and noise. The jitter was the average absolute difference between successive period lengths divided by the average period length in percent. This is the "Jitt" feature in the widespread Multidimensional Voice Program (MDVP). The shimmer was the average absolute difference of peak amplitudes of consecutive periods divided by the average peak amplitude in percent (MDVP's "Shim"). The spectral features were derived from the inverse filtered acoustic signal. The considered features were H1*-H2*, H1*-A1*, H1*-A3* and TL*, where H1* and H2* are the inverse filtered magnitudes of the first and second harmonic, A1* and A3* are the inverse filtered magnitudes of the harmonics next the first and third formant and TL* is a measure for spectral tilt that is obtained from a regression line over the first eight inverse filtered harmonic peaks. The investigated noise feature was the Noise-to-Harmonics Ratio (MDVP's NHR).

Mehta's experiments were carried out on four copora, three built from observed data and one from synthesized data. The observed corpora were one corpus of 52 normal subjects, phonating comfortable and pressed (corpus A), one corpus of 14 pathologic speakers (corpus B) and one corpus of 47 subjects, which were 40 pathologic and seven normal (corpus C). In corpus C, six pathologic speakers brought pre- and post-treatment data, resulting in 53 datasets. The synthesized data was obtained from computational models of the vocal folds and the vocal tract. The vocal fold model was modified from the asymmetric two-mass model of Steinecke and Herzel [19]. The model accounts for nonlinear acoustic coupling, and nonlinear source-filter interaction (level 1: glottal airflow and level 2: tissue dynamics) [50].

In corpus A (52 normal subjects, comfortable and pressed phonation), visual ratings of left-right phase asymmetry were compared to the objective measures with respect to the sagittal position of the kymogram and descriptive statistics of left-right phase asymmetry, left-right amplitude symmetry and axis shift during closure were reported. Mehta showed that his new measure for left-right phase asymmetry correlates stronger to visual ratings than a former version [51]. The correlation was stronger in the midglottis (sagittal position of the kymogram) and weaker toward the endpoints of the main glottal axis. No differences between comfortable and pressed phonation were found.

In corpus B (14 pathologic speakers), objective kymogram measures were compared to normative ranges and their correlation to acoustic features was tested. The objective kymogram measures were left-right phase asymmetry, left-right amplitude asymmetry, axis shift during closure, open quotient and period irregularity. The acoustic features were jitter, shimmer and the NHR. 31 to 62 % of the subjects showed above normal values for phase and amplitude asymmetry as well as for axis shift. The open quotient and the period irregularity fell within normal limits. Correlations were found for standard deviation of left-right phase and amplitude asymmetry. Both measures reflect variability of asymmetry and result in increased acoustic jitter values. The standard deviation of the open quotient also reflects variability of asymmetry but correlates with shimmer.

Model data was compared to observed data (corpora C and D), considering analyses of kymograms, GAWs and acoustical spectra. The measures obtained from kymograms are the left-right phase asymmetry (slight modification to formerly used measure), the left-right amplitude asymmetry and the axis shift. The open quotient, closing quotient and a plateau quotient were obtained from the GAWs. The spectral features were H1*-H2*, H1*-A1*, H1*-A3* and TL*. At first, relationships between the measures in synthesized data were investigated. Over certain ranges, there were linear relationships of the phase asymmetry to the plateau quotient and the axis shift. A monotonic relationship of the closed closing quotient to the phase asymmetry was found. Second, in subject data, only the relationship of phase asymmetry and axis shift could be reproduced, all other observations do not agree with model data.

Divergences between model data and observed data showed the complexity of the apparent physical phenomena and may be explained as follows. Although great effort is put into creating a realistic synthesis model, its parameters could not be estimated appropriately, because they are too many. The number of parameters must be very wisely chosen when observed data should be modeled. Trade offs between model complexity and model fits must be assessed by means of "goodness of fits" measures. The topic of the present thesis is similar to Mehta's, but their scopes do not overlap because:

1. Features used by Mehta rely on cycle detection, which must fail in diplophonic phonation.
2. Vocal fold vibration and waveform patterns are compared to auditive ratings.
3. Frequency asymmetry is considered.

The voice source in speech production: data, analysis and models

Yen-Liang Shue published his dissertation "The voice source in speech production: data, analysis and models" in 2010 [47]. Shue's dissertation is relevant for diplophonia analysis from a methodological point of view. A new GAW model and a new inverse filtering technique were proposed and evaluated. In addition, acoustic features were extracted and correlated to independent variables. The independent variables were voice quality, F0 type, the presence of a glottal gap, gender and prosodic features. All speakers were healthy and the used voice quality types were pressed, normal and breathy. The F0 types were low, normal and high. Evaluations were conducted on corpora of different sizes, which consisted both of sustained phonation and read text.

A new parametric GAW model was trained and evaluated. The GAW cycle model was trained from data of three male and three female healthy speakers with perceptually normal voice. Stable intervals of sustained phonations with different voice quality and F0 types were used for training. The GAWs were manually segmented at cycle borders, and average cycle shapes were obtained. The proposed model was parametric and Liljencrants-Fant (LF) like. It was an integrated version of the first LF equation, which was used for both the opening and the closing. The used parameters were the open quotient, a temporal asymmetry coefficient α, the speed of the opening phase S_{OP} and the speed of the closing phase S_{CP}. The model showed better agreement to observed data than comparable models.

Inverse filtering via codebook search was proposed and evaluated on observed GAWs. Shue jointly estimated parameters of the proposed source model and a conventional six-pole vocal tract model from the acoustic waveform. A larger and a smaller codebook were trained from the proposed parametric voice source model. The smaller one allows for preselecting the coarse shape of the GAW cycle, and the smaller one is used for fine tuning. The vocal tract model was jointly tuned by constrained nonlinear optimization, and its model error was used as optimization criterion for the codebook search. It was shown that the proposed model outperforms the LF model, but that the performance of the procedure crucially relies on the correct estimation of formant frequency and bandwidth.

Acoustic features have been extracted with the custom MATLAB software "VoiceSauce" and evaluated with respect to voice quality and voice source parameters, automatic gender classification and prosody analysis. The acoustic features were F0, energy, cepstral peak prominence, Harmonics-to-Noise Ratio (HNR) and spectral features. In voice quality and voice source analysis, spectral features correlated with the open quotient, breathiness, the asymmetry coefficient α and the speed of closed phase S_{CP}. Energy correlated with loudness and voice intensity, and the cepstral peak prominence and HNR with modality and breathiness. The open quotient was decreased in pressed phonation and increased in breathy phonation. In that study, the open phase was shorter in breathy phonation surprisingly. Automatic gender classification and prosody analysis is less adjacent to the present thesis and thus not reviewed here.

Shue's methods are similar to the presented methods, because he used analysis-by-synthesis based on a voice source and vocal tract model with optimization via minimization of the resynthesis error. Shue's approach, however, differs substantially from the approach pursued in the present thesis because:

1. Average cycle prototypes are not parameterized and used to train a codebook, but extracted directly from the signal under test.
2. The present approaches do not require separating the voice source from the vocal tract, a simpler model structure is therefore sufficient.
3. Two-oscillator waveform models are considered in the present thesis.

A drawback of Shue's thesis may be that the evaluation of the proposed model as compared to other models might be erroneous, because the training data equals the test data. Thus, Shue ran the risk of overfitting because the proposed model received an undue advantage to other models in evaluation.

1.3 The definition of diplophonia

Diplophonia may be defined at three levels of description. The perceptual definition is the most pragmatic and used in clinical practice. Definitions based on the acoustic waveform may be quantitative and enable formulating the Diplophonia Diagram, i.e., an automated procedure for discriminating diplophonic voice from other voice phenomena. Definitions based on vocal fold vibration connect with the mechanical properties of the vocal folds and consequently the distal causes of diplophonia.

At the level of perception, diplophonia is the presence of two pitches [20], which is a rather vague definition. The origin of this definition is in the 19th century and is considered to be traditional. Recent psychoacoustic investigations report complex perceptual effects [21,22]. The presence of two pitches depends firstly on the signal (the stimulus), and secondly on the observer (the subject). The ability to segregate a voice sound into separate pitches is a matter of training, which is often conducted without the aid of accredited sound files and proper play back techniques. In addition, divergent room acoustics and background noise change perceived timbre, which results in high inter- and intrarater variability.

Titze's definition is at the level of waveforms, which can be applied to the level of acoustics. Titze's diplophonia is defined by counting pulses within one metacycle (MC), i.e., a "period-two up-down pattern of an arbitrary cyclic parameter" [52]. Other waveform patterns that rely on pulse counting are "triplophonia", "quadruplophonia" and "multiplophonia". Titze's terms "biphonia", "triphonia" and "multiphonia" denote the presence of two, three or multiple independent sound sources. It is remarkable that, e.g., biphonia can be multiplophonic at the same time. Such a waveform is shown in figure 4.4.

Diplophonic waveforms are described by means of subcycles and metacycles. When two oscillators are summed, the summands' waveform cycles are referred to as subcycles. Meta cycles manifest in temporal fluctuations of the dominant pulse heights, which are pseudo periodic, depending on the frequency ratio of the oscillators and its evolvement. The meta cycle length can fluctuate, and very likely does in natural signals, because the oscillators' frequencies and their ratio vary in time. The variation depends on the oscillators' coupling strength, laryngeal muscle tension and subglottal pressure.

On the glottal vibration level, diplophonia is defined as the presence of two distinct oscillations at different frequencies. In the search for a robust and clinically relevant outcome measure, the assessment of vocal fold vibration is central. Medical doctors are interested in the question how abnormal voice is produced and its underlying etiology, because a better understanding of voice production would facilitate the development of more effective treatment techniques. The presence of diplophonia is assessed by detecting left-right and anterior-posterior frequency asymmetries in laryngeal high speed videos. In the case of superior-inferior frequency asymmetry, the observation of the secondary frequency is impeded because it is partly or entirely hidden from view. A hidden oscillator affects the glottal area in a purely modulative way, i.e., its additive contribution to the waveform is zero.

All definitions of diplophonia are problematic in practical use. The perceptual definition via the presence of two pitches suffers from its subjectivity. The waveform definition suffers from the unsolved problems of cycle detection and sound source independence decision. The definition at the glottal vibration level suffers from the constraints of the observation methods, i.e., laryngeal high-speed videos offer two-dimensional projections of the opaque three-dimensional vocal fold tissue. To obtain insight in the diplophonia phenomenon, all three levels of description are addressed in this thesis.

1.4 Methodical prerequisites

Testing aims at detecting a certain target condition in a sample. In a diagnostic context, "the term test refers to any method for obtaining additional information on a patient's health status. It includes information from history and physical examination, laboratory tests, imaging tests, function tests and histopathology. The condition of interest or target condition can refer to a particular disease or to any other identifiable condition that may prompt clinical actions, such as further diagnostic testing, or the initiation, modification or termination of treatment" [53]. Test outcomes indicate clinical actions to avoid clinical consequences such as death or suffering, ideally via predefined rules. As an ultimate scientific goal, it must be proven if an individual patient benefits from the administration of a diagnostic test, and what the benefit is quantitatively, because unnecessary testing should be avoided.

A good diagnostic test has to respect several quality criteria. The main criteria in test theory are validity, reliability and objectivity. A valid test measures what it is supposed to measure, and influencing factors that are probably biasing the results (i.e., covariates) are either controlled to be constant between groups, or are adjusted for. E.g., when measuring the intelligence quotient, a subject under test has to solve several logical tasks. Such a test can only measure how subjects behave in this particular test situation, with all covariates as they are. Covariates can be: affinity for solving IQ tasks, practice, vigilance, age, education, patience and cultural background. Generalization (i.e., deductive conclusions) of test results to other life situations might be problematic or even invalid. Very wisely, intelligence is defined as the magnitude that is measured by the IQ test, which solves parts of the validity problem. However, economical validity may still be small because of the divergence between laboratory setups and real world setups. In the clinical assessment of voice disorders, the same is true. Conclusions from single descriptors to general voice health conditions may lack validity.

Regarding reliability, measurement divergences between raters (intra and inter), test procedures, test repetitions and test devices need to be assessed [54]. Both systematic and random measurement errors may occur. Systematic errors (e.g., a rater giving systematically higher rates than his/her colleague) can be adjusted for if their magnitudes are known. Random errors need to be kept as small as possible right from the beginning of the experiment, because they cannot be removed from the results afterwards.

Objective test results only depend on the test object (i.e., the patient under test) and not on the test subject (e.g., the measurement device/procedure, or the reader of the test). In an objective diagnostic test, any effect that does not directly stem from the patient is avoided. One may see here, that the described concepts of validity, reliability, objectivity and systematic/random measurement errors are dependent on each other.

Diagnostic table

In diagnostic accuracy studies, the performance of a diagnostic test is evaluated with respect to a known true target condition. The target condition and the test outcome have two possible values, namely present or absent and positive or negative. In a diagnostic study, positive subjects must be available for which the target condition is present, and negative subjects for which the target condition is absent. The negative group should consist of subjects that typically need to be distinguished from positive subjects.

The performance of a test is evaluated by setting up a cross table with four cells. The columns of the table denote the target condition, whereas the rows denote the test outcome. The entries of the table are absolute counts or rates of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). In a true positive, the target condition is present and the test result is positive, which is desirable. The second desirable situation is a negative test result, given the condition is absent. Undesirable events are positive tests if the condition is absent (false positives) and negative tests of the condition is present (false negatives). Table 1.1 illustrates an example of a diagnostic table.

		Target condition	
		Present	Absent
Test	Positive	TP	FP
	Negative	FN	TN

Table 1.1: Example of a diagnostic table. True positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

Several summary measures are calculated from diagnostic tables. The measures express statistical properties of the observed data distribution and are used to draw conclusions for clinical decision making. The measures are the sensitivity, the specificity, the accuracy, the prevalence (aka pretest probability), the positive predictive value (aka posttest probability), the positive likelihood ratio, the negative predictive value and the negative likelihood ratio. All measures are derived from absolute counts of TP, TN, FN and FP. The measures range from 0 to 1 or from 0 % to 100 %.

Sensitivity (SE)

The sensitivity is the probability for a positive test outcome, given the target condition is present.

$$SE = \frac{TP}{TP + FN} \tag{1.1}$$

Specificity (SP)

The specificity is the probability for a negative test result, given the target condition is absent.

$$SP = \frac{TN}{TN + FP} \tag{1.2}$$

Accuracy (ACC)

The accuracy is the probability for a correct test result.

$$ACC = \frac{TN + TP}{TP + FP + FN + TN} \tag{1.3}$$

Prevalence (PR) aka pretest probability (PRP)

The prevalence equals the proportion of positive subjects or samples in a cohort. The pretest probability is the probability for a present target condition, independently of the test outcome. In other words, the pretest probability is the probability of randomly picking a positive subject or sample from a cohort. The values of the prevalence and the pretest probability are equal, but the term prevalence is used for proportions, whereas the term pretest probability expresses probability. A proportion is a property of observed data, whereas probability is a property of a random process.

$$PR = \frac{FN + TP}{TP + FP + FN + TN} \tag{1.4}$$

Positive predictive value (PPV) aka posttest probability (POP)

The positive predictive value aka posttest probability is the probability for a present target condition, given a positive test outcome. High positive predictive values reflect good tests, but must always be seen in relation to the prevalence, which is better achieved by the positive likelihood ratio.

$$PPV = \frac{TP}{TP + FP} \tag{1.5}$$

Positive likelihood ratio (PLR)

Probabilities can also be expressed in odds, which are calculated from equation 1.6. An exemplary probability of 10 % results in odds of $\frac{1}{9}$, which means that on average 1 positive event happens while 9 negative events. The positive likelihood ratio is actually an odds ratio and given by equation 1.7.

$$\text{Odds(P)} = \frac{\text{P}}{1 - \text{P}} \tag{1.6}$$

$$PLR = \frac{\text{Odds}(PPV)}{\text{Odds}(PR)} = \frac{SE}{1 - SP} \tag{1.7}$$

Negative predictive value (NPV)

The negative predictive value is the probability for an absent target condition, given a negative test outcome.

$$NPV = \frac{TN}{TN + FN} \tag{1.8}$$

Negative likelihood ratio (NLR)

The negative likelihood ratio expresses how the odds for the absence of a target condition change from their pretest value, if a single subject or sample under test is tested negative.

$$NLR = \frac{\text{Odds}(NPV)}{\text{Odds}(1 - PR)} = \frac{SP}{1 - SE} \tag{1.9}$$

Confidence intervals

The above measures are estimates of the true proportions. Confidence intervals express the certainty of the estimates. Binomial distribution is assumed and the intervals are calculated by an iterative algorithm [55].

Receiver operating characteristic curves and optimal threshold selection

The receiver operating characteristic (ROC) curve is a way for depicting the test performance of a cut-off threshold classifier with respect to its threshold. The curve is useful for interpreting the overlap of two data distributions used for classification. It is applied for determining optimal cut-off thresholds.

The ROC curve shows the sensitivity on its y-axis and 1 - specificity on its x-axis and relates to different values of the threshold. Figure 1.1 shows an example of an ROC curve. The plot also shows the line of pure guessing (dashed green) and selected threshold values (red crosses). The boxes on the bottom right show the area under the curve (AUC), the optimal threshold and the sensitivity and specificity at the optimal threshold. As a rule of thumb, an AUC of 0.7 or more reflects a meaningful test paradigm. A 10-fold bootstrap validation is used throughout the thesis for calculating the confidence intervals of the AUC [56].

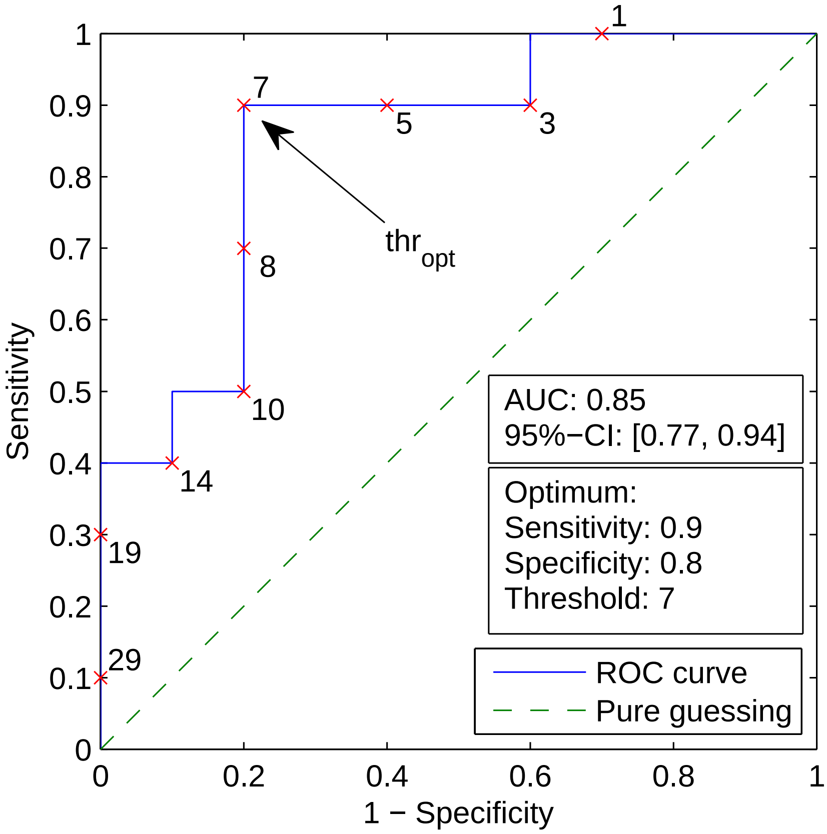


Figure 1.1: Example of an receiver operating characteristic (ROC) curve.

The optimal threshold is found by minimizing the geometrical distance of the ROC curve to the virtual optimal point, i.e., the upper left corner in the coordinate system (SE = 1, SP = 1). The distance D equals $\sqrt{(1-SE)^2+(1-SP)^2}$ and is a function of the cut-off threshold. The optimal threshold is found at the point of the ROC for which D is minimal. Unless stated otherwise, the optimal thresholds in this thesis are derived by minimizing D. Other methods for optimizing the threshold are, e.g., maximal $F1$ measure from pattern recognition ($F1 = 2 \cdot \frac{PPV \cdot SE}{PPV+SE}$), equal error rate ($SE = SP$), maximal PLR or moving a straight line from the upper left corner toward the curve. The last option is explained and used in chapter 3.

Interpretation

All presented performance measures reflect a good diagnostic test if they are high. The majority of the measures are probabilities or proportions, which can go from 0 % to 100 %. The PLR and the NLR are measures for likelihood ratio and can take any positive number.

The sensitivity, the specificity and the accuracy should be (far) above 50 %. If they are below 50 %, one should check if an inverse test has been administered. An inverse test would test for the absence of the condition instead for the presence. The prevalence aka pretest probability is a number that purely reports a property of the cohort and not a property of the test. It cannot be influenced by test design, but needs to be taken into account for adjusting thresholds and interpreting results. The proportion of diplophonic samples in the used database is most likely not equal to the prevalence in general population, which is discussed in section 2.4 and chapter 6. The positive predictive value aka posttest probability highly depends on the prevalence and thus should be interpreted in relation to it. The positive likelihood ratio relates the pretest probability and the posttest probability by dividing their odds and is thus a very valuable feature, because it expresses how the odds for the target condition change if a single subject or sample under test is tested positive. A PLR of 1 reflects pure guessing, and a PLR below 1 reflects an "inverse" test. As a rule of thumb, a PLR of 2 or more is considered to reflect a meaningful test. The advantage of the PLR in comparison to other measures is that it evaluates how much additional information about the presence of a target condition is obtained by administering a test, while taking prevalence into account. It is difficult to achieve high posttest probability for very rare pathologies, but the gain of information obtained by administering the test is reflected by the PLR.

1.5 Aims and Hypotheses

The aim of this thesis is to work towards an automated method for the clinical detection of diplophonia. A diagnostic study of sensitive and specific discrimination of diplophonia from other voice phenomena is conducted.

It is hypothesized that:

1. typical vocal fold vibration patterns of diplophonic phonation can be identified,
2. the patterns can be automatically detected from laryngeal high-speed videos,
3. diplophonia can be automatically detected from audio waveforms,
4. the detectors outperform related approaches,
5. the test rationales are physiologically interpretable, and
6. a solid ground truth for the presence of diplophonia can be defined.

1.6 Overview

In chapter 2, a database that consists of laryngeal high-speed videos with simultaneous high-quality audio recordings is presented. The database enables investigating vocal fold vibration patterns, glottal area waveforms and audio recordings with regard to perceptual cues. In chapter 3, automatic detectors that are based on recent publications [57, 58] are introduced. The detectors are designed for discriminating diplophonic from non-diplophonic vocal fold vibration patterns. The detection paradigms are discussed with respect to a geometric model of diplophonic vocal fold vibration, which improves the understanding of spatially complicated vibration patterns. Recent publications on audio waveforms [59,60] are presented in chapter 4. The state-of-the art of disordered voice analysis and synthesis is moved forward by modeling diplophonic waveforms accurately. The pursued strategies increase the knowledge of physical correlates of perceptual cues of disordered voice. In chapter 5, the waveform model is used to design the Diplophonia Diagram, which relates diplophonic and non-diplophonic audio waveforms to physiologically interpretable scales [60]. Finally, the ground truth is checked from a probabilistic viewpoint by combining information from three independent descriptors via latent class analysis.

List of publications

The thesis is based on material that has partially been presented in the listed publications. The author's contributions to the publications 1-3 and 5-6 were the basic idea, theoretical analysis, experimental design, data collection, software implementation, interpretation of the results and paper writing. The author's contributions to the publications 4 and 7 were the basic idea, experimental design, data collection and the interpretation of the results. The publications are listed in chronological order.

1. "Describing the transparency of mixdowns: the Masked-to-Unmasked-Ratio," in 130th Audio Engineering Society Convention, 2011. [61].
2. "Double pitch marks in diplophonic voice," in IEEE International Conference on Acoustics, Speech, and Signal Processing, 2013, pp. 7437-7441. [59].
3. "Spectral analysis of laryngeal high-speed videos: case studies on diplophonic and euphonic Phonation," in Proceedings of the 8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, 2013, pp. 81-84. [57].
4. "Automatic glottis segmentation from laryngeal high-speed videos using 3D active contours," in Medical Image Understanding and Analysis, 2014, pp. 111-116. [62].
5. "Comparison of an audio-based and a video-based approach for detecting diplophonia," Biomedical Signal Processing and Control, in Press. [58].
6. "Towards objective voice assessment: the Diplophonia Diagram," Journal of Voice, accepted. [60].
7. "Automatic high-speed video glottis segmentation using salient regions and 3D geodesic active contours," Annals of the British Machine Vision Association, submitted. [63].

2 A database of laryngeal high-speed videos with simultaneous high-quality audio recordings

2.1 Motivation

A prerequisite for the development of a method for detecting diplophonia is the investigation of phonatory mechanisms, typically by means of laryngeal high-speed videos. Several publications based on fairly large databases of laryngeal high-speed videos are available [64–76], but unfortunately no database is open to the scientific community, probably because publishable databases are expensive to obtain and administrate. In many scientific disciplines it is common to publish databases, but not so in research on disordered voice. Examples of published databases in speech research are [77–79], but also medical databases with patient related data have been opened for retrieval by the scientific community [80–83]. Publishing databases increases the comparability of research results and makes scientific communication more effective. Publishing databases of laryngeal high-speed videos would professionalize voice research and increase the quality of scientific studies.

Analyzing laryngeal high-speed videos seems to be a fruitful approach to investigating diplophonic voice, because the amount of information available in the video exceeds that in the audio signal. Because diplophonia is assessed perceptually, the additional acquisition of high-quality audio recordings is mandatory. Quality criteria that should be striven for are high sampling rates (at least 44.1 kHz), anechoic room acoustics, low levels of background noise, high quantization resolution (at least 16 bits), high quality microphones (e.g., professional condenser microphones with symmetrical wiring), high quality pre-amplifiers and recorders. This chapter documents the design, creation and administration of a database of laryngeal high-speed videos with synchronous high-quality audio recordings. The database is highly valuable for answering research questions addressed in this thesis.

2.2 Design

2.2.1 Data collection

Between September 2012 and August 2014, 80 dysphonic subjects have been recruited among the outpatients of the Medical University of Vienna, Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics. 40 of these subjects were diplophonic and 40 were non-diplophonic[1]. All those subjects have a Grade of Hoarseness greater than 0, as determined by phoniatricians during medical examination. The presence of diplophonia is a dichotomic and subject-global attribute of voice, and has been determined by perceptual screening of the dysphonic subjects. Additionally 40 subjects that constitute the euphonic group have been recruited via public announcement in Vienna, Austria. The data collection has been approved by the institutional review board of the Medical University of Vienna (1810/2012 and 1700/2013). The study design has been chosen to enable the development of a sensitive and specific method for detecting diplophonia both in dysphonic subjects and in the general population.

Table 2.1 shows the diagnoses with respect to the clinical groups. One observes a high number of pareses in the diplophonic group and a high number of dysfunctional dysphonia in the dysphonic group. The risk for pareses is significantly increased in the diplophonic group (odds ratio 9.15, CI [1.91, 43.90], calculated by [84]), but the presence of diplophonia should not be used to derive the diagnosis of a patient because the confusion rates are high.

	Euphonic	**Diplophonic**	**Dysphonic**
Unknown	40	1	1
Laryngitis (acute/chronic)	0	2	8
Sulcus	0	2	2
Nodules	0	1	3
Polyp	0	3	2
Edema	0	6	5
Cyst	0	4	2
Scar	0	1	2
Paresis	0	13	2
Dysfunction	0	5	12
Benign tumor	0	1	0
Bamboo nodes	0	1	0
Neurological	0	0	1

Table 2.1: Number of diagnoses per clinical group.

A laryngeal high-speed camera and a portable audio recorder were used for data collection. The camera was a HRES Endocam 5562 (Richard Wolf GmbH) and the audio recorder was a TASCAM DR-100. The frame rate of the laryngeal high-speed camera was set to 4 kHz. The CMOS camera sensor had three color channels, i.e., red, green and blue. Its spatial resolution was 256x256 (interpolated from red: 64x128, green: 128x128, blue 64x128). The light intensity values were quantized with a resolution of 8 bits.

Two microphones have been used for recording audio signals. A headworn microphone AKG

[1] For convenience, in the remainder of the thesis "diplophonic dysphonic" is referred to as "diplophonic" and "non-diplophonic dysphonic" is referred to as "dysphonic".

HC 577 L was used to record the voice of the subjects. The loudest noise source in the room was the cooling fan of the camera's light source. A lavalier microphone AKG CK 77 WR-L was put next to the fan to record its noise. Both microphones were used with the original cap (no presence boost) and with windscreens AKG W77 MP. The microphones were connected to phantom power adapters AKG MPA V L (linear response setting) and the portable audio recorder (headworn: left channel, lavalier: right channel). The sampling rate was 48 kHz and the quantization resolution was 24 bits. The subjects' sessions mostly consisted of more than one video and were entirely microphone recorded. The audio recordings were carried out by the author. The sound quality was continuously monitored during the recordings with AKG K 271 MK II headphones. The audio files are saved in the uncompressed PCM/WAV file format.

Figure 2.1a shows the footprint of the recording room, which is a standard room for ENT examinations. The figure shows the subject and the medical doctor who were sitting on chairs, the sound engineer, the high-speed camera, the lavalier microphone, the head-worn microphone and the portable recorder. The other items in the room were not used in the experiment, but are depicted for documentation. The background noise was 48.6 dB(A) and 55.5 dB(C), measured with a PCE-322A sound level meter (IEC 61672-1 class 2), with temporal integration set to "slow". Figure 2.1b is taken from [64] and shows the midsagittal view of an endoscopic examination. The tip of the tongue was lightly held by the medical doctor when he inserted the endoscope into the mouth of the subject way back to the pharynx. The larynx was illuminated and filmed by the endoscopic camera and the medical doctor previewed the camera pictures on a computer screen. He adjusted the position of the camera for optimal sight of the vocal folds. Once an optimal position was achieved the medical doctor instructed the subject to phonate an /i/, which positioned the epiglottis so as to allow direct sight of the vocal folds[2].

2.048 seconds of video material were continuously stored in a first-in-first-out loop. When the medical doctor decided to store video data permanently, he pushed the trigger on the handle of the camera. The loop was then stopped and the video was watched in slow motion and pre-evaluated for video quality. The experimenters either decided to copy the video to a long-term storage or to dismiss the current video and to take a new one, depending on the achieved video quality. The stored video data have been included in the database. Tables 2.2 and 2.3 show the numbers of stored videos per clinical group and per subject.

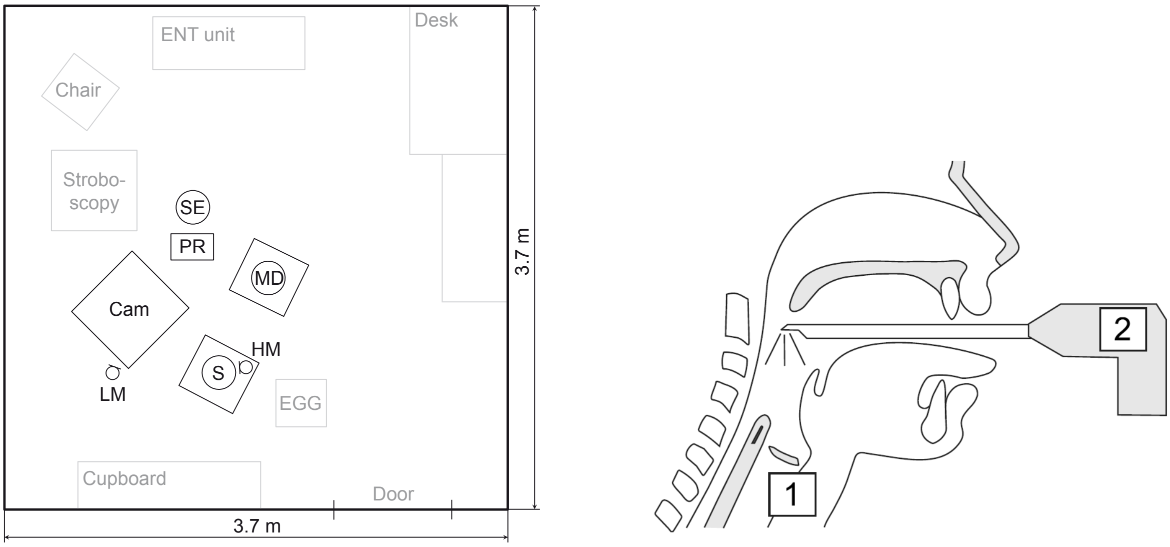
	Nr. of videos	**Proportion (%)**
Euphonic	132	35.2
Diplophonic	123	32.8
Dysphonic	120	32

Table 2.2: Number of videos per clinical group.

	Min	**Max**	**Average**	**Std**
Euphonic	0	5	3.3	1.181
Diplophonic	0	6	3.075	1.3085
Dysphonic	1	6	3	1.281

Table 2.3: Number of videos per subject and clinical group. Minimum, maximum, average, standard deviation.

[2] The produced sound is schwa-like, due to the endoscope and the lowered position of the tongue.



(a) Footprint of the recording room. Ear, nose and throat (ENT) unit, sound engineer (SE), portable audio recorder (PR) (TASCAM DR-100), laryngeal high-speed camera (Cam) (HRES Endocam 5562, Richard Wolf GmbH), medical doctor (MD), headworn microphone (HM) (AKG HC 577L), subject (S), lavalier microphone (LM) (AKG CK 77 WR-L), electroglottograph (EGG).

(b) Midsagittal view of the subject and the endoscope, taken from [64]. 1: vocal folds, 2: endoscope.

Figure 2.1: Footprint of the recording room and midsagittal view of an endoscopic examination.

2.2.2 Data annotations

The collected data have been annotated for voice quality, video to audio synchronization, video quality and audio quality.

Voice quality

To learn voice quality annotation, three scientific all-day seminars on the detection of diplophonia were conducted. Several experts on voice and signal processing attended the meetings, which were in alphabetical order: Wolfgang Bigenzahn (Medical University of Vienna), Martin Hagmüller (Graz University of Technology), Christian Herbst (Palacký University Olomouc), Christian Kasess (Austrian Academy of Sciences), Gernot Kubin (Graz University of Technology), Sylvia Moosmüller (Austrian Academy of Sciences), Berit Schneider-Stickler (Medical University of Vienna), Jean Schoentgen (Université Libre de Bruxelles) and Jan Švec (Palacký University Olomouc). Additionally, within the New Investigator Research Forum of The Voice Foundation, an expert discussion (25 minutes) was conducted with several experts on voice assessment, which were in alphabetical order: Ronald Baken (New York Eye and Ear Infirmary of Mount Sinai), James Daugherty (University of Kansas), Molly Erickson (University of Tennessee), Michael Johns (Emory Voice Center, Atlanta), Robert Sataloff (Drexel University College of Medicine, Philadelphia), Nancy Pearl Solomon (Walter Reed Army Medical Center, Washington) and Sten Ternström (KTH Royal Institute of Technology, Stockholm). There have been intensive discussions on the definition of diplophonia. The results from these discussions were faithfully aggregated by the author and used to learn annotating voice quality with respect to the presence of diplophonia.

The author has annotated all phonated audio recordings with available video for voice quality. The voice quality groups are euphonic, diplophonic and dysphonic. Diplophonia has been defined as the simultaneous presence of two different pitches or a distinct impression of beating. Simple pitch breaks have not been considered to be diplophonic. When the perceptual determination of the presence of diplophonia was doubtful, the audio waveforms and spectrograms were visually inspected. The spectral criterion was the presence of two separate harmonic series or the presence of metacycles in the waveform. The audio signal only was available to the annotator.

Audio to video synchronization

The audio files have been synchronized to the video by visually matching their waveforms to the waveforms of the audio that was recorded with the camera's inbuilt microphone. The camera was equipped with a Sennheiser KE 4-211-1 microphone. It was fixedly mounted on the endoscope, approximately 16 cm from the tip. It could not be used to carry out perceptual or acoustic analysis, because the microphone was pre-amplified with a harsh automatic gain control (integrated circuit, SSM2165) and filtered with a 2 kHz low pass filter. It is synchronized to the video signal, which enabled synchronizing the high-quality audio to the video, up to the difference of the microphone positions. This difference is approximately 10 cm or 0.3 ms at $c = 343 \frac{\mathrm{m}}{\mathrm{s}}$ or $\frac{1}{6}$ of a cycle length at a high phonation frequency of 600 Hz.

Praat [85] was used to annotate the audio files for synchronization points. Synchronization points have been set at corresponding local maxima of the high-quality audio and the camera's inbuilt audio. Figure 2.2 shows examples of audio waveforms together with their synchronization markers. In the shown example the local maximum in the KE 4-211-1 audio is approximately at 0.603 s with respect to the video file beginning, and the corresponding maximum in the HC577 L audio is at 447.484 s with respect to the audio file beginning.

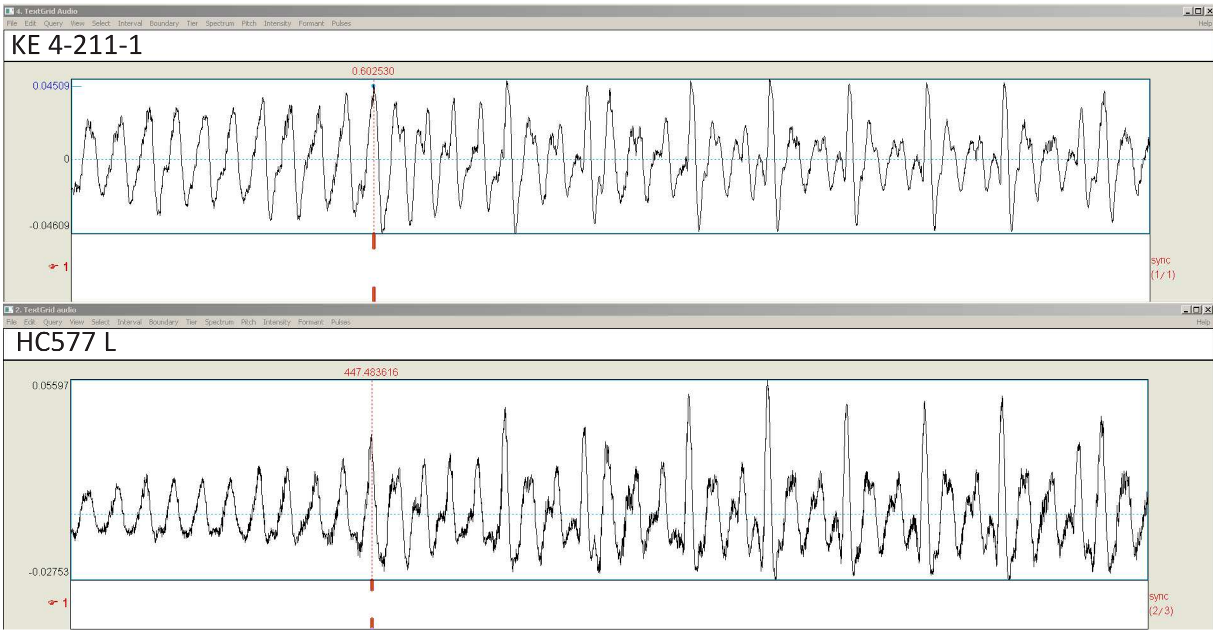

4. TextGrid Audio
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses
Help
KE 4-211-1
0.602530
0.04509
0
-0.04609
1
sync
(1/1)
2. TextGrid audio
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses
Help
HC577 L
447.483616
0.05597
-0.02753
1
sync
(2/3)

Figure 2.2: Synchronization of the camera's inbuilt audio (KE 4-211-1) and the high-quality audio (HC577 L). The synchronization markers denote corresponding local maxima.

Video quality

A problem of laryngeal high-speed videos is the inconsistent video quality. A pre-selection based on video quality has been performed to allow correct analyzes. The video quality has been evaluated with respect to the visibility of relevant structures and the presence of artifacts. The used criteria, the classes and the labels are shown in table 2.4. The videos have been annotated with the Anvil tool [86], which allows for placing time interval tiers considering time-variant video quality.

The visibility of the vocal fold edges and vocal fold vibration is important for facilitating spectral video analysis (cf. chapter 3) and extracting the glottal area waveform (cf. chapter 4). To ensure that no visualizable distinct glottal oscillations are hidden, a sufficiently large proportion of the glottal gap must be visible. Parts of the vocal fold edges can be hidden behind the epiglottis, the epiglottis' tubercle, the false vocal folds, the aryepiglottic folds or the arytenoid structures. The criteria used for judging the visibility of the glottal gap were the visibility of the anterior commissure and the visibility of the processus vocales. The processus vocales act as nodes in the vocal fold oscillation modes, and thus are adequate for orientation in the anatomy when oscillation patterns are observed and interpreted.

Criteria	Classes (labels)	Acronyms
Visibility		
Vocal fold edges	Visible (1) Not visible (0)	A1
Vocal fold vibration	Visible (1) Not visible (0)	A2
Anterior commissure	Visible (1) Not visible (0)	A3
Processus vocales	Both visible (2) One visible (1) None visible (0)	A4
Artifacts		
Blurring	Severe (2) Mild (1) Absent (0)	B1
Reduced contrast	Severe (2) Mild (1) Absent (0)	B2
Mucus on the vocal folds	Connecting the vocal folds (3) Blurring the vocal fold edges (2) Visible (1) Not visible (0)	B3
Extraglottal mucus yarn artifact	Blurring the vocal fold edges (2) Present (1) Absent (0)	B4
False vocal fold artifact	Hiding the vocal fold edges (1) Absent (0)	B5
Aryepiglottic folds artifact	Hiding the vocal fold edges (1) Absent (0)	B6

Table 2.4: Overview of the video quality annotation criteria.

The presence of blurring and/or reduced contrast, false vocal fold artifacts, mucus that blurs the vocal fold edges and a remote mucus yarn may impede spectral video analysis or the extraction of the glottal area waveform. Blurring occurred if the camera was not focused on the glottal gap, or if there were fluid artifacts on the endoscope. A false vocal fold artifact occurred when the vocal fold edges are partly hidden behind the false vocal folds. Reduced contrast occurred for several reasons. It was not always possible to illuminate the larynx sufficiently due to inter-individual anatomical differences, for instance, the epiglottis was not fully lifted and in subjects with lowered larynx the light reaching the vocal folds was weaker. Blurring and reduced contrast have been labeled as absent, mild or severe.

The selection of usable video intervals has been based on a logical expression. The logical expressions ensure that only video intervals with sufficient visibility are analyzed and that disadvantageous combinations of video artifacts are excluded.

Logical expressions, acronyms from table 2.4, video frame is usable if:

(A2 = 1) AND
((A3 = 1) OR (A4 = 1) OR (A4 = 2)) AND
NOT(B1 = 2) AND
NOT(B2 = 2) AND
NOT((B1 = 1) AND (B2 = 1)) AND
NOT(B3 = 2) AND
NOT(B3 = 3) AND
NOT(B4 = 1) AND
NOT(B4 = 2) AND
NOT(B5 = 1).

Audio quality

Audio quality has been evaluated with respect to the presence of artifacts and synchronizability. Parts of some recordings contain the medical doctor's voice giving instructions to the subject and only the remaining parts can be used for audio analyzes. Other artifacts are background noise or stem from microphone contact. Artifacts were successively minimized as they had been recognized in the course of the study, but were not fully avoidable.

Figure 2.3 shows an example of the audio quality annotation using Praat [85] with one point tier and five interval tiers. The "sync" tier has been used for synchronizing the audio and the video. The "Video" tier denotes the temporal borders of the high-speed video. The "Phonation" tier denotes the temporal borders of the phonation(s) and is bounded to the video borders. The "Voice quality" tier labels the voice quality, which has been assessed for all audio material within phonation borders. The "Examiner" tier denotes time intervals in which the examiner's voice is audible, and the "Artifact" tier labels other audio recording artifacts.

The video in the example is the fourth of the recording session, thus the sync point and the Video interval are labeled "4". Within the video borders, two intervals of phonation are observed, with a phonation break in between. The phonation intervals are labeled "1", denoting phonation within video borders. For all phonation within video borders, the voice quality is labeled in the "Voice quality" tier (1: euphonic, 2: diplophonic and 3: dysphonic). The examiner's voice is labeled in the "Examiner" tier ("1"), and background noise is labeled in the "Artifact" tier ("1").

Logical expressions, acronyms from table 2.5, audio sample is usable if:

EXIST(C1) AND
NOT(C2 = 0) AND
(D1 = 1) AND
((D2 = 1) OR (D2 = 2) OR (D2 = 3)) AND
NOT(E1 = 1) AND
NOT(E2 = 1).

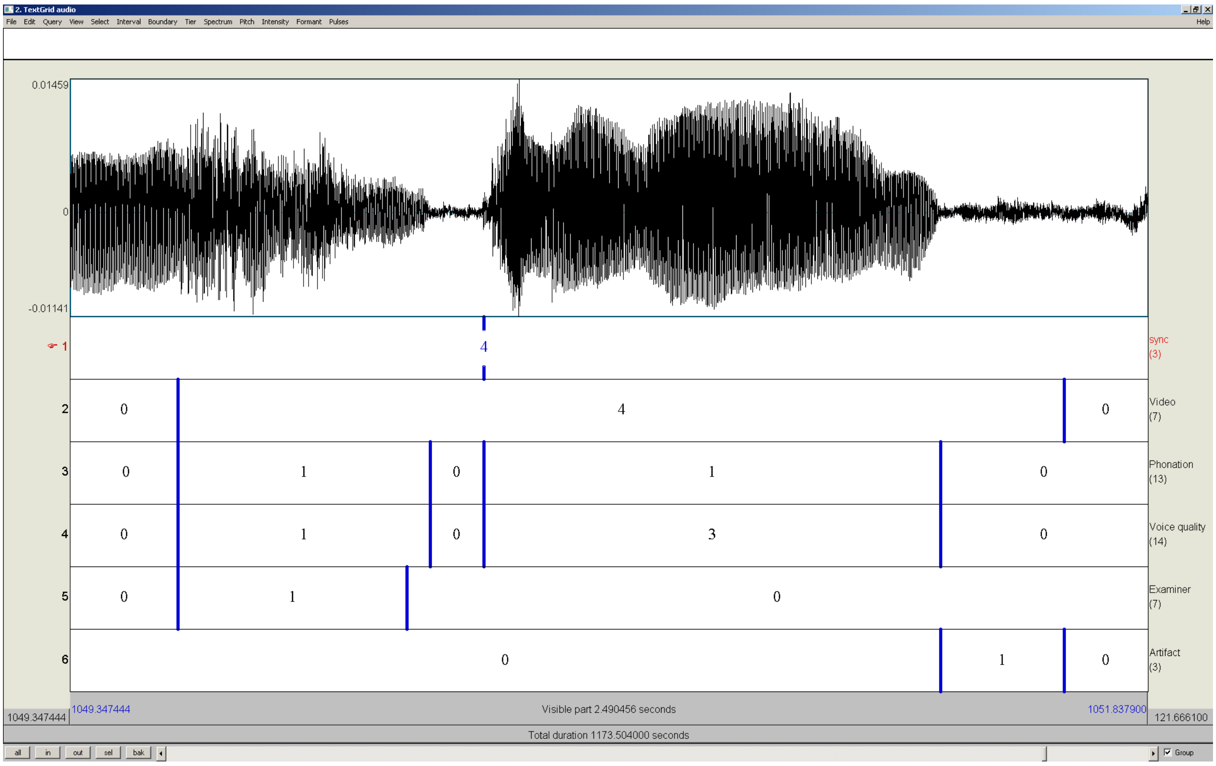


Figure 2.3: An example of annotated audio material.

Criteria	Classes (labels)	Acronyms
Video relation		
sync Video	(Video ID) Absent (0) Present (Video ID)	C1 C2
Auditive assessment		
Phonation	Absent (0) Present (1)	D1
Group	Absent (0) Euphonic (1) Diplophonic (2) Dysphonic (3)	D2
Audio artifacts		
Examiner Artifact	Absent (0) Present (1) Absent (0) Present (1)	E1 E2

Table 2.5: Overview of the audio annotation criteria.

2.2.3 Corpora

The previous sections explain how the video and audio material has been annotated with respect to their usability for further analyzes. To allow for joint analyzes of audio and video material, unusable video and audio have been excluded and only uninterrupted intervals of different voice qualities are used. Figure 2.4 shows the analysis interval selection on the base of joint data quality pre-selection and voice quality annotation. Subplot 1 shows the borders of the video file, subplot 2 vocal fold visibility, subplot 3 shows the visibility of vocal fold vibration, subplot 4 shows the presence of video artifacts. Subplots 5 to 7 refer to audio annotation, namely the presence of phonation, the voice quality and the presence of audio artifacts. Finally, subplot 8 shows the borders of the resulting analysis intervals. All intervals shorter than 0.125 s are discarded.

Several corpora were built between September 2012 and August 2014. Table 2.6 gives an overview of the corpora, which are shown row wise. The columns specify the characteristics, which are: the number of subjects, videos and intervals, the range of the intervals lengths, the signal type and the date. The seventh column shows the corpus ID. The last column denotes the chapters of this thesis, in which the corpora are referred to. The corpora grew incrementally as the available clinical data grew in time. In the corpora 8, 98 and 99 no video quality based pre-selection has been performed, because only audio material has been used for analysis. In early corpora (ID 88, 89, 82, 92), intervals for analysis have been selected without the aid of video quality annotation. In later corpora (ID 46, 99), the video annotations have been used to make the selection more explicit. The corpora 88 and 89 are used for training the model structures introduced in chapters 3 and 4. The remaining corpora are used for testing and evaluation.

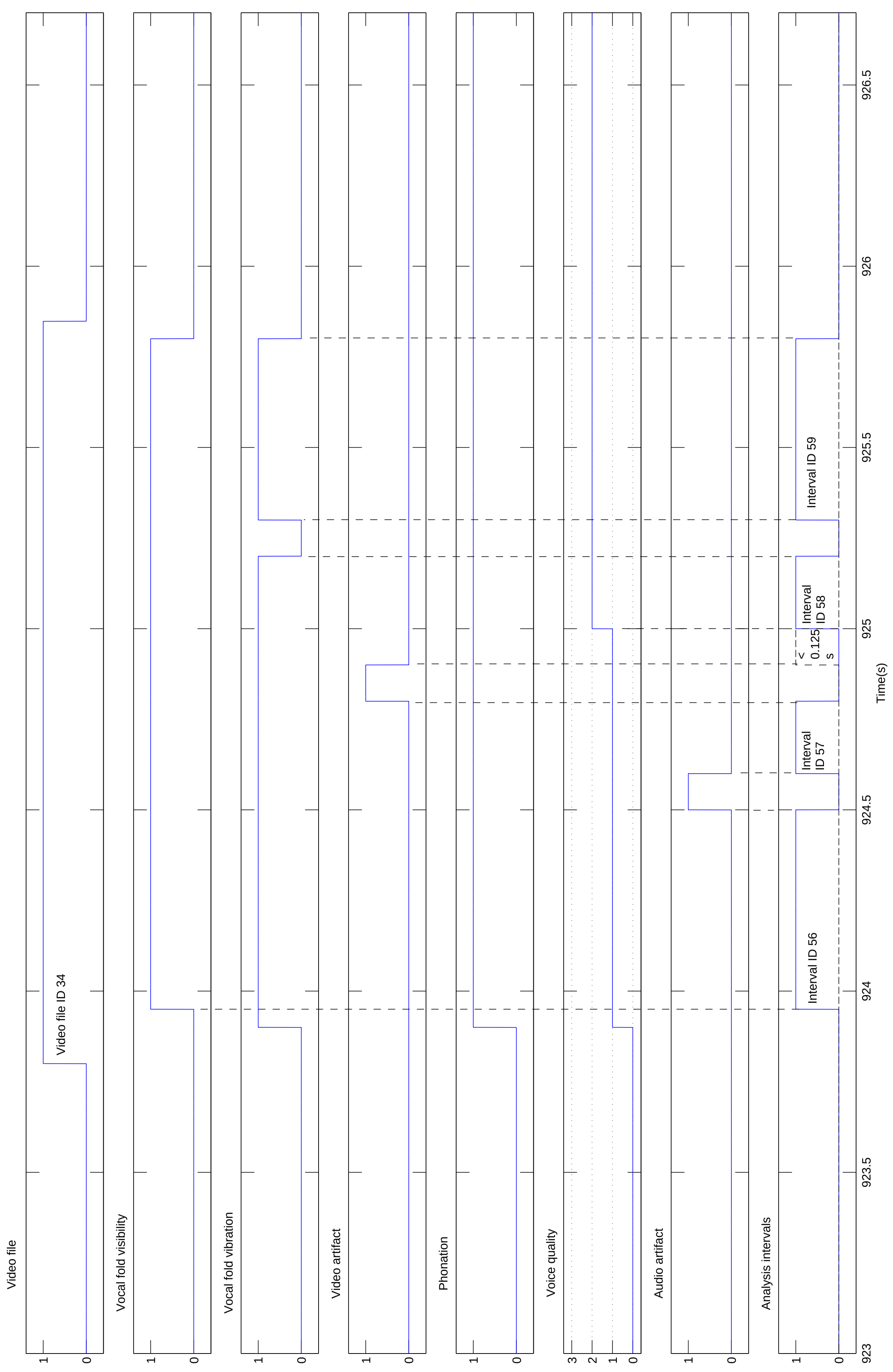


Figure 2.4: Example for the analysis interval selection from joint data quality pre-selection and voice quality annotation.

Subjects (E./Di./Dy.)	Videos (E./Di./Dy.)	Intervals (E./Di./Dy.)	Interval lengths	Signal type	Date	ID	Chapter
0/1/0	0/1/0	0/1/0	300 ms	Glottal area waveform (GAW)	20.09.2012	88	4
1/1/0	1/1/0	1/1/0	126.5 ms	GAW	13.03.2013	89	4
8/11/1	8/16/1	10/10/10	125 ms	Spatially cropped video, GAW	11.07.2013	82	2
8/8/0	9/10/0	10/10/0	100 - 125 ms	Spatially cropped video	24.09.2013	92	3
15/19/10	17/27/10	21/20/20	132.8 - 2047.8 ms	Spatio-temporal plot (STP), raw video	02.05.2014	46	3
30/28/22	75/66/44	96/55/127	125.4 - 2048 ms	Audio	12.06.2014	8	4, 5
9/23/13	12/49/23	0/55/70	125.4 - 2048 ms	Audio	29.09.2014	98	4, 5
29/23/27	54/39/45	65/28/84	128.7 - 2047.4 ms	Raw video, audio	29.09.2014	99	4

Table 2.6: Overview of the corpora. E. ... euphonic, Di. ... diplophonic, Dy. ... dysphonic.

2.2.4 Data structure and analysis framework

A data structure and analysis framework that had to meet several requirements has been established. The requirements were the anonymization of subject related data, the availability of a professional backup system, the application of a non-destructive system for data selection, the integration into clinical practice and the management of data and code. Figure 2.5 gives an overview of the data structure and analysis framework. The structure is organized into input data (yellow), executable code (green), output data (red) and procedures (blue).

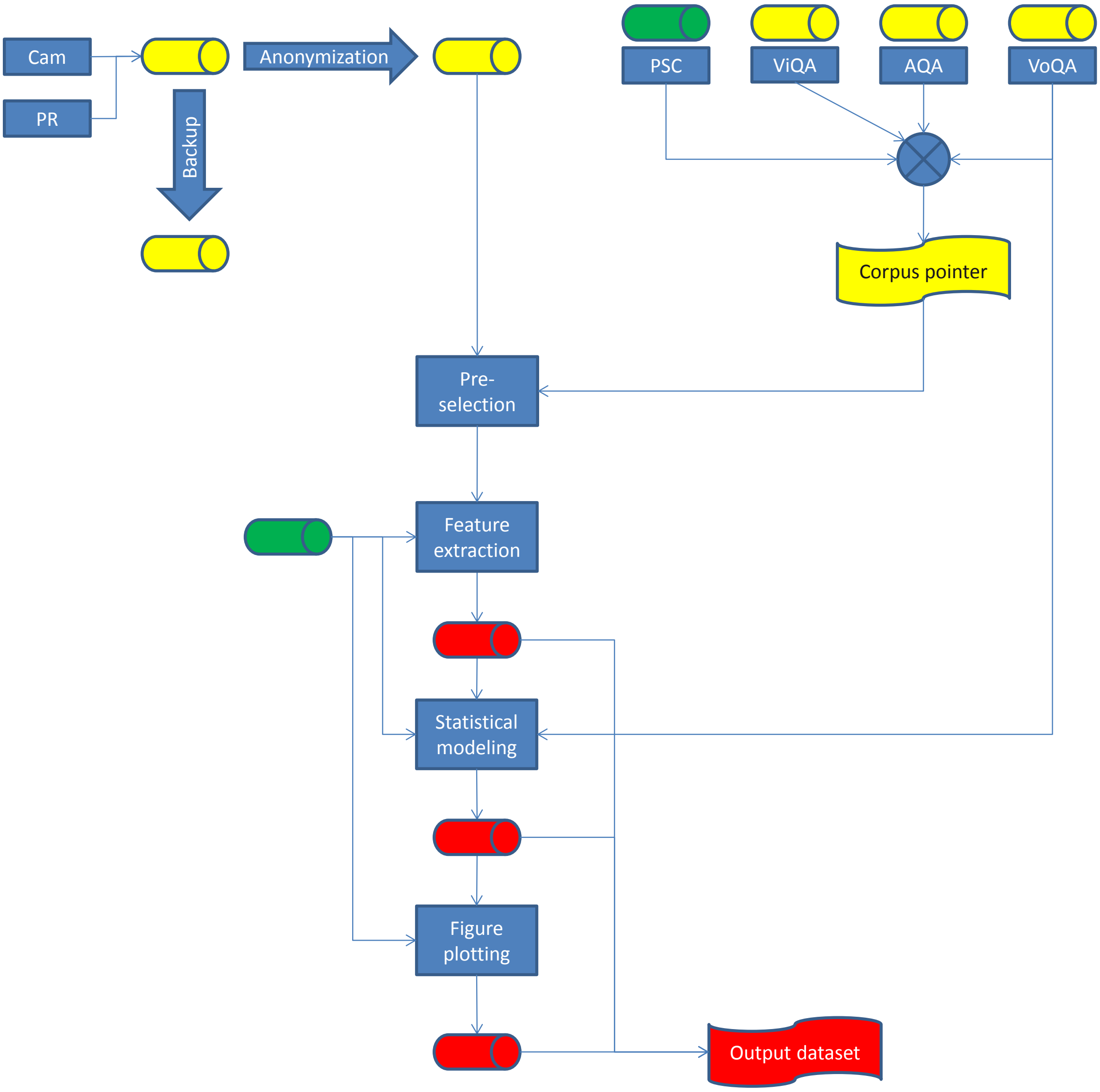


Figure 2.5: Overview of the data structure and analysis framework. Input data volumes are in yellow, executable code volumes are in green, the output data volumes are in red and procedures are in blue. Laryngeal high-speed camera (Cam), portable audio recorder (PR), pre-selection combinations (PSC), video quality annotations (ViQA), audio quality annotations (AQA), voice quality annotations (VoQA).

The laryngeal high-speed camera fed a person database, who's real names are visible to the clinician. This feature has been protected for clinical reasons. The audio files from the recorder have been integrated into the camera's database. The database has been mirrored on an institutional backup server of the Medical University of Vienna. For scientific purposes the database had to be anonymized to protect the subjects' identities. The anonymized database has been mirrored onto a secondary harddisk, from which all scientific queries are performed. An anonymization list relates the primary keys of the camera's database to the primary keys of the mirrored database.

As shown in section 2.2.2, the video and audio files have been annotated for video quality, audio quality and voice quality. Selection criteria that were used for creating corpora from these annotations are given as logical expressions (i.e., the selection combinations). The selection criteria and their combination are subject to change, unused data are therefore not deleted from the database (i.e., non-destructive pre-selection).

MATLAB was used as analysis environment because it provides capabilities for indexing and processing data arrays, feature extraction, statistical modeling and generating vector graphics. MATLAB queried the database of the laryngeal high-speed camera and carried out data analyses in three steps. In step one (feature extraction) the audio/video data were analyzed. In step two (statistical modeling) the data distributions were statistically modeled with respect to their voice quality. In step three (figure generation) the results were visualized. To ensure reproducibility, the experimenter took snapshots of all input data pointers and executables. All snapshots were copied into a tagging folder of a Subversion repository. Additionally, the experimenter maintained a logbook of all performed experiments.

2.3 Evaluation

The evaluation of the pursued strategies was carried out in two steps, namely a pilot study on a smaller database and a confirmatory study on a larger database. In the pilot study, the clinical group allocation, the sample size and hypotheses, the effectiveness of the data collection, the data quality and the database structure were tested.

Clinical group allocation

The clinical group allocation has been tested with respect to the appropriateness for answering the research questions. Meaningful classification results were produced from the pilot study data, and so the group allocation was kept for the confirmatory study. Some discussions have taken place about clinically defining subgroups of dysphonia, which would have complicated and prolonged the data collection.

Sample size and hypotheses

Two versions of the sample size planning were designed. Version 1 was based on strict criteria for demanding high sensitivity and specificity with subsequent estimation of the sample size. Version 2 was based on proposing a minimally achievable sample size and estimating the expected statistical outcome.

Working hypothesis 1 was that the spectral modality M of the high-speed video (cf. chapter 3) differs between diplophonic and euphonic phonation. Table 2.7 shows the classification results, with $M \neq 1$ indicating a positive test, and $M = 1$ indicating a negative test. Working hypothesis 2 was that the difference of synthesis qualities SQ_1 and SQ_2 (cf. chapter 5) distinguishes diplophonic from dysphonic phonation. Table 2.8 shows the classification results for an optimal cut-off threshold classifier.

		Voice quality	
		Diplophonic	Euphonic
M	$\neq 1$	9	2
	$= 1$	1	8

(a) Diagnostic table.

	Estimate	CI
SE (Eq. 1.1)	90 %	[55.5 %, 99.7 %]
SP (Eq. 1.2)	80 %	[44.4 %, 97.5 %]
ACC (Eq. 1.3)	85 %	[62.1 %, 96.8 %]
PR (Eq. 1.4)	50 %	[27.2 %, 72.8 %]
PPV (Eq. 1.5)	81.8 %	[48.2 %, 97.7 %]
PLR (Eq. 1.7)	4.5	[1.28, 15.81]
NPV (Eq. 1.8)	88.9 %	[51.7 %, 99.7 %]
NLR (Eq. 1.9)	8	[1.21, 52.69]

(b) Performance measures.

Table 2.7: Results of the pilot study: Diplophonic versus euphonic phonation, with spectral video modality serving as predictor. 95 % confidence intervals (CI).

		Voice quality	
		Diplophonic	Dysphonic
$\mathrm{SQ}_1 - \mathrm{SQ}_2$	$> 0.23\ dB$	8	3
	$\leq 0.23\ dB$	2	7

(a) Diagnostic table.

	Estimate	CI
SE	80 %	[44.4 %, 97.5 %]
SP	70 %	[34.8 %, 93.3 %]
ACC	75 %	[50.9 %, 91.3 %]
PR	50 %	[27.2 %, 72.8 %]
PPV	72.7 %	[39 %, 94 %]
PLR	2.67	[0.99, 7.22]
NPV	77.8 %	[40 %, 97.2 %]
NLR	3.5	[0.95, 12.9]

(b) Performance measures.

Table 2.8: Results of the pilot study: Diplophonic versus dysphonic phonation, with synthesis quality difference serving as predictor. 95 % confidence intervals (CI).

In version 1 of the sample size planning, the procedure ”Sample size for one proportion” was used [87, 88]. The estimation was chosen to be ”exact” and the hypothesis test was designed to be one-sided. The α was set to 0.05 and the β was set to 0.8, which are commonly used parameters. The α expresses the desired level of significance, i.e., the probability for falsely detecting a significant effect. The β expresses the power, i.e., the probability of correctly detecting a significant effect.

For working hypothesis 1 the desired lower limit of the confidence interval of the sensitivity estimator was chosen to be 80 % and that of the specificity estimator to be 70 %, which are the estimates minus 10 %. Those demands resulted in a sample size of 99 diplophonic and 119 euphonic phonation intervals. For working hypothesis 2 the desired lower limit of the confidence interval of the sensitivity estimate was chosen to be 70 % and that of the specificity estimate to be 60 %. Those demands resulted in a sample size of 119 diplophonic and 142 dysphonic intervals. Taking the maximum of the estimated sample sizes results in a study design of 119 diplophonic intervals, 142 dysphonic intervals and 119 euphonic intervals. It would not have been feasible to record such a large number of intervals within the given time frame, man power and financial resources.

In version 2 of the sample size planning, a sample size of 30 intervals per group was assumed (three times the pilot study sample size) and the prospects of the statistical outcomes were estimated. The PLR has been chosen as the statistical outcome criterium. For example, a PLR of 2 would mean that the odds for the presence of diplophonia in a randomly selected interval would double if it is tested positive.

Test 1 (spectral modality) of the pilot study resulted in a PLR of 4.50, with a 95% confidence interval of [1.28, 15.81], and test 2 (synthesis quality) resulted in a PLR of 2.67, with a 95% confidence interval of [0.98, 7.22]. The PLR estimate and its confidence interval were calculated with the online software MedCalc [88]. Both estimates for the PLR are above 2, which would indicate a reasonable test performance. However, the lower confidence interval limits are 1.28 (close to guessing), and 0.98 (pure guessing).

To estimate how the confidence interval limits change when more data are added, it is assumed that future observations will behave as past observations, i.e., that the PLR estimates equal the true population PLRs. The hypothesized data distribution is obtained by multiplying all counts of tables 2.7 and 2.8 by three, which gives tables 2.9 and 2.10. The confidence intervals rose to 2.18 and 1.5. A PLR of 2.18 would be acceptable but a PLR of 1.5 would still be suboptimal. The lower confidence interval however rose from 0.98 to 1.5. Hence, the proposed sample size was kept in the confirmatory study and the overall sample size (pilot study + confirmation study) resulted in 40 intervals per group. At the time of planning the confirmation study, the number of intervals that can be obtained per subjects was unknown, because the database pre-selection had not been implemented yet. A conservative estimate is one interval per subject, which results in the overall sample size of 40 subjects per group. In corpus 99, which was based on the described selection procedure from audio and video quality, 177 intervals from 120 subjects (on average 1.475 intervals per subject) were obtained. Therefore assuming one interval per subject was appropriate.

		Voice quality	
		Diplophonic	Euphonic
M	$\neq 1$	27	6
	$= 1$	3	24

(a) Diagnostic table.

	Estimate	CI
SE	90 %	[73.5 %, 97.9 %]
SP	80 %	[61.4 %, 92.3 %]
ACC	85 %	[73.4 %, 92.9 %]
PR	50 %	[36.8 %, 63.2 %]
PPV	81.8 %	[64.5 %, 93 %]
PLR	4.5	[2.18, 9.3]
NPV	88.9 %	[70.8 %, 97.6 %]
NLR	8	[2.7, 23.8]

(b) Performance measures.

Table 2.9: Hypothesized distribution for sample size planning of confirmatory study: Diplophonic versus euphonic phonation, with spectral video modality serving as predictor. 95 % confidence intervals (CI).

		Voice quality	
		Diplophonic	Dysphonic
$\mathrm{SQ}_1 - \mathrm{SQ}_2$	$> 0.23\ dB$	24	9
	$\leq 0.23\ dB$	6	21

(a) Diagnostic table.

	Estimate	CI
SE	80 %	[61.4 %, 92.3 %]
SP	70 %	[50.6 %, 85.3 %]
ACC	75 %	[62.1 %, 85.3 %]
PR	50 %	[36.8 %, 63.2 %]
PPV	72.7 %	[54.5 %, 86.7 %]
PLR	2.67	[1.5, 4.74]
NPV	77.8 %	[57.7 %, 91.4 %]
NLR	3.5	[1.65, 7.43]

(b) Performance measures.

Table 2.10: Hypothesized distribution for sample size planning of confirmatory study: Diplophonic versus dysphonic phonation, with synthesis quality difference serving as predictor. 95 % confidence intervals (CI).

Effectiveness of data collection

The effectiveness of the data collection has been evaluated with respect to the appropriateness of the recorded material for testing the hypotheses. In the pilot study an additional audio corpus and a logopedic anamnesis including a voice range profile had been recorded. The audio corpus contained recordings of sustained vowels at different loudness levels and pitches, standard text readings, continuous conversation elicited via written prompts, counting at different loudness levels and sung notes at different loudness levels and pitches. The logopedic anamnesis consisted of auditive perceptual rating (RBH), measurement of the s/z ratio, administering questionnaires, measuring a voice range profile, the dysphonia severity index (DSI) [89], the Göttingen Hoarseness Diagram and querying some general information on the subject. Testing the relation of the presence of diplophonia with such clinical data would be necessary to gain full insight into the clinical relevance of diplophonia. Nevertheless, only laryngeal high-speed videos with synchronous audio recordings were recorded in the confirmation study, because data collection and data handling are very time consuming and should only be done if hypotheses with good prospects for testing can be formulated.

Data quality

Although the image quality of laryngeal high-speed videos and the audio quality of microphone recordings are crucial factors for analyses, analysis intervals are most commonly selected subjectively in voice research. This strategy was kept in the pilot study and evaluated on its scientific appropriateness. In the confirmation study video quality and audio quality annotation criteria were formulated, which increased the explicitness of the selection.

Database structure

With incrementally growing databases one runs the risk of over- or underestimating the appropriate amount of data administration. The amount of administration must be assessed continuously when a new database structure and analysis framework is built. However, it is also important not to adapt the administration procedures to often, but to continuously keep track of desired features and occasionally implement them.

In the beginning of the thesis, only a small amount of recordings was available. Small databases can easily be handled manually, but conclusions on small corpora are criticizable in terms of their significance and validity. As the number of observation grows, one should have a database that is easy to feed and query and which is non-redundant as well.

If changes in the analyzes are necessary, previous experiments should be reproducible. It may be sufficient to make a snapshot of all the data and archive them as soon a study is finished. If the data size is of the order of Tera bytes, which is the case for laryngeal high-speed videos, this costs a lot of harddisk space, which may not be available. Thus, snapshots of data pointers only were made instead of creating redundant copies of the files. All small files were version controlled and tagged on a Subversion repository.

2.4 Discussion and conclusion

The design and the evaluation of a database of laryngeal high-speed videos with simultaneous high-quality audio recordings have been described. This section summarizes the original contributions, discusses limitations and gives an outlook.

The study design aimed at a group of patients that can be recruited from clinical practice by perceptual screening and that is reproducibly detectable by using signal processing methods in a physiologically interpretable way. Added values are that the data collection and analysis have been performed at the same center and that the database and the analysis framework have been integrated into clinical practice. Medical and technical know-how have been combined, which required interdisciplinary communication.

Laryngeal high-speed videos with simultaneous high-quality audio recordings have been obtained and purposefully pre-selected. In contrast to what has often been practiced in the field of disordered voice research [90], the analysis intervals were not selected subjectively and/or destructively. Instead, video and audio quality criteria that allow for reproducible and non-destructive pre-selection of analysis have been proposed. The quality of the collected audio material exceeds that of most other work in the field of disordered voice research. The audio material was recorded with a high-quality microphone and high quality recorder, a high sampling rate of 48 kHz and a high bit resolution of 24 bits. In contrast to clinical group allocation in terms of subject global voice quality, audio quality annotation aims at assessing voice quality interval wise.

However, there are also some limitations that should be taken into account when the data are analyzed and conclusions are drawn. First, statistical interpretation maybe difficult. The distributions in the study design are not the same as the distributions in the general population. Because diplophonia is a rare phenomenon in the general population, the classification thresholds may have to be adjusted for population prevalence. The database was built from sustained phonations during rigid endoscopic high-speed laryngoscopy, which imposes unnatural phonatory conditions. Moreover, the recordings are 2.048 seconds short and so the ecological validity of the collected data can be questioned. Several stages of pre-selection needed to be passed through, the number of intervals differs between subjects and the interval lengths differ, which may harm the assumption of independence in statistical testing.

Second, the recording and playback conditions were not perfect. Experimental conditions in a hospital are similar to field conditions rather than laboratory conditions. Often it was not possible to carry out video recordings repeatedly because rigid telescopic high-speed video laryngoscopy is demanding for the subject, which impacts on achieved video quality. Hospitals are not equipped with professional recording studios, and so noise may disturb the recordings. No reference listening room was available for the audio judgments. The annotations were therefore performed with headphones, which eliminates the room acoustics of the listening room that is unrelated to the subject's condition. The timbre at the position of the microphone is different from the timbre anywhere else in the room, and so the perceptual impression during the recording may differ from the impression during playback.

Future work should focus on addressing the limitations of the database. To increase representativeness and ecological validity, studies that account for the prevalence of diplophonia in spoken language of the general population are needed. Perceptual effects in the clinical environment need further investigation with calibrated recording and playback setups that account for different room acoustics. It is also important to investigate what data quality is sufficient for what kind of analysis. Moreover, data handling that allows for publishing the database should be established.

A state-of-the-art database of laryngeal high-speed videos with synchronous high-quality audio recordings has been created successfully. The database has proven to be highly valuable for answering the central research questions of the thesis.

3

Spatial analysis and models of vocal fold vibration

3.1 Motivation and overview

This chapter describes how spatial analysis and modeling is used to identify diplophonic vocal fold vibration patterns. The presented approaches are motivated by the clinical wish to understand the etiologies and distal causes of diplophonic phonation. In other disciplines one may want to know what information is conveyed by phonation auditorily, but in a clinical context one may prefer to focus on glottal conditions. However, clinicians are still being trained to use perceptual methods and the origin of the diplophonic timbre may remain unknown in the absence of other tools. In contrast to auditive detection, analyzing laryngeal high-speed videos aims at detecting spatially separate oscillations with different fundamental frequencies, which may in parallel with the auditive approach help in clinical decision making.

Known approaches to spectral video analysis with respect to time are adapted and used for deriving scalar features that describe diplophonia in recorded data [12, 38, 41, 57–59, 72, 91, 92]. The features that are introduced are the Frequency Image Bimodality (FIB) and the Frequency Plot Bimodality (FPB). The FIB is an automated approach that is applied to raw video data. The intensity time-series of each pixel is normalized and transformed to the frequency domain. The FIB is obtained from a Frequency Image that shows the spatial distribution of spectrally dominant frequencies across the pixel coordinates in false colors. The FIB is the minimal spectral magnitude of the diplophonic frequency.

The FPB is obtained from spatially segmented videos (i.e., spatio-temporal plots). The trajectories of the vocal fold edges are transformed to the frequency domain. The FPB is obtained from a Frequency Plot that shows the spatial distribution of spectrally dominant frequencies along the vocal folds' edges. The FPB is the spatial proportion of the diplophonic frequency along the vocal fold edges in dB.

Four experiments on the automatic detection of diplophonia from laryngeal high-speed videos are presented. Two experiments deal with the FIB that is tested on two databases. In the smaller database only diplophonic and euphonic intervals are used. In the larger database all three types of phonation (euphonic, diplophonic, dysphonic) are used and the impact of the analysis block length and the region of interest (ROI) cropping are investigated. In the experiments on FPB only the large database is used, once with block wise (short-time) results and once with analysis fragment averages.

Frequency asymmetry, of which three types exist, is referred to as ”glottal diplophonia”. Types that are fairly understood are anterior-posterior and left-right frequency asymmetry. In anterior-posterior frequency asymmetry the anterior and the posterior part of the vocal folds vibrate at different frequencies, whereas in left-right frequency asymmetry the left and the right vocal fold vibrate at different frequencies. In both types the oscillators are visible in the laryngeal high-speed video. Less well understood is the superior-inferior type, because the inferior oscillation is likely to be occasionally or permanently hidden under superior tissue. Superior-inferior asymmetry is discussed by geometric modeling of asymmetric vocal fold vibration.

3.2 Frequency analysis of pixel intensity time series

3.2.1 Video spectrum model description

The proposed procedure [57] is based on groundwork by Granqvist and Lindestad [38]. The algorithm aims at determining the presence of secondary oscillation frequencies (i.e., diplophonic oscillators) from laryngeal high-speed videos, and uses Frequency Images of vocal fold vibration. The Frequency Images show the spatial distribution of the dominant frequencies of the pixel intensity time-series, which are thresholded by their spectral magnitude. The scalar feature FIB is the minimal spectral magnitude threshold, for which only one oscillation frequency is observed in the video. The algorithm for determining the FIB is structured as follows.

1. Region of Interest (ROI) cropping
2. Normalization of the pixel intensity time series
3. Time-windowing of the pixel intensity time series
4. Discrete Fourier transform (DFT)
5. Identification of the maximal spectral magnitudes
6. High-pass filtering
7. Spectral magnitude thresholding
8. Creation of the Frequency Images
9. Extraction of the Frequency Image Bimodality (FIB)
10. Time-averaging of the FIB (only confirmatory corpus)

Let $I_{(x,y,n)}$ denote the three-dimensional video data array, where x is the lateral position, y is the sagittal position, n is the discrete time index and N is the block length. The video data are manually cropped for the region of interest (ROI), which is a rectangular image section that contains the glottal gap only. The intensity time series are normalized, which reduces the influence of luminosity. Each intensity time series is divided by its time average. The subtraction of 1 yields zero-mean signals. Figure 3.1 illustrates the extraction and normalization of intensity time series.

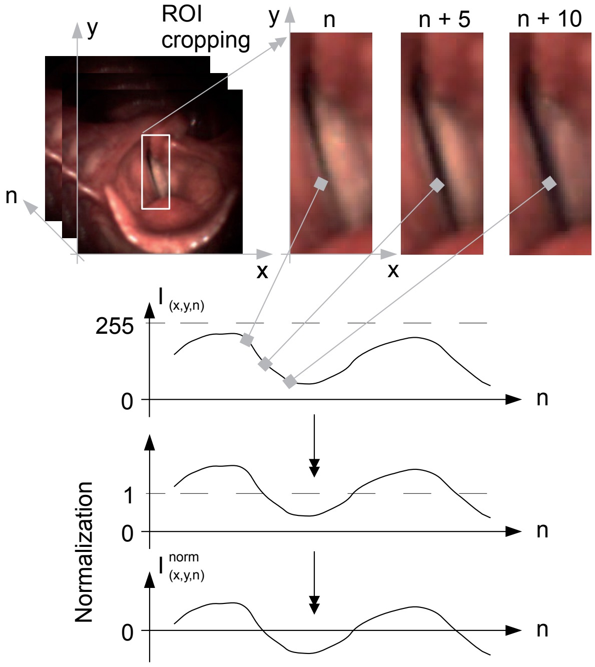


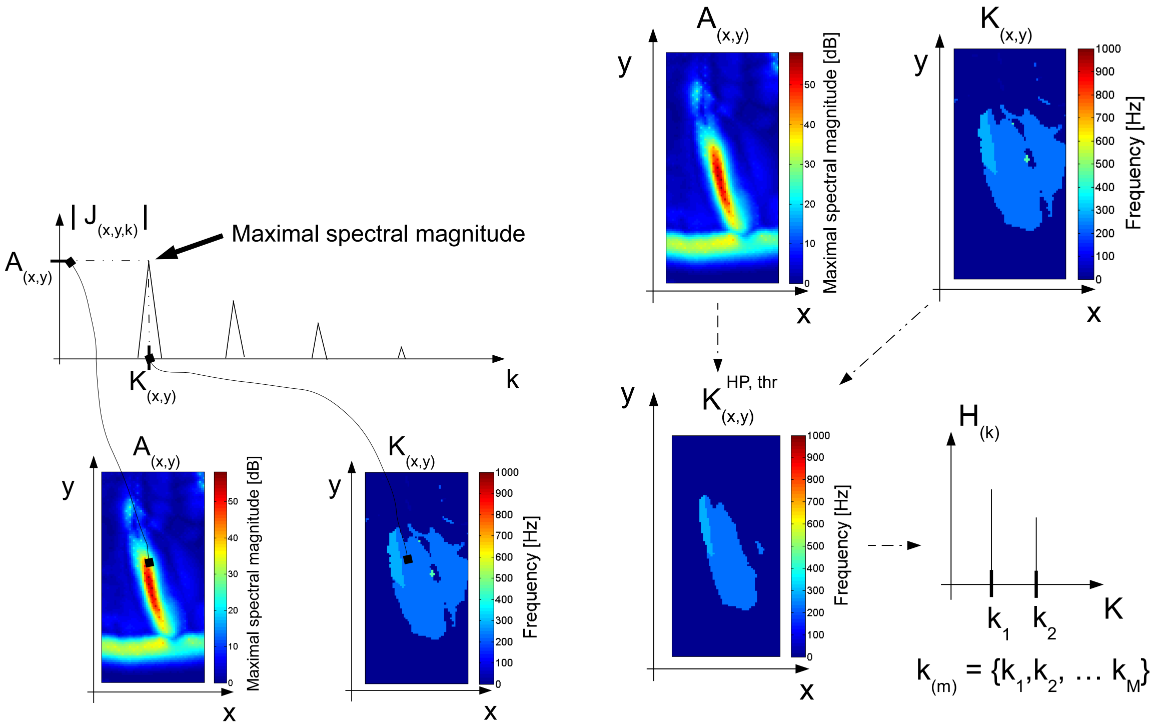
Figure 3.1: Extraction of a pixel intensity time series $I_{(x,y,n)}$ and its normalization.

$$I^{norm}_{(x,y,n)} = \frac{I_{(x,y,n)}}{\frac{1}{N} \cdot \sum_{n=1}^{N} I_{(x,y,n)}} - 1 \tag{3.1}$$

The normalized intensity time series $I^{norm}_{(x,y,n)}$ are time-windowed (Kaiser window, $\beta = 0.5$). The pixel spectra $J_{(x,y,k)}$ are obtained by discrete Fourier transform of the windowed time series, where k denotes discrete frequency. The dominant frequencies (i.e., the frequency at the maximal spectral magnitude) are determined for each spectrum. The spatial distribution of the dominant frequency is the Frequency Image $K_{(x,y)}$.

$$K_{(x,y)} = \underset{k}{\operatorname{argmax}} \left| J_{(x,y,k)} \right| \tag{3.2}$$

The spatial distribution of the dominant frequencies is high-pass filtered at 70 Hz and small spectral magnitudes are discarded, because low frequencies and small spectral magnitudes are considered to be irrelevant for detecting diplophonia. A magnitude threshold increases or decreases the number of detected frequencies in the Frequency Image $K_{(x,y)}$. Figures 3.2a and 3.2b and equations 3.3 and 3.4 illustrate the spectral magnitude thresholding. The frequencies are summarized in the frequency histogram H_k, with M distinct oscillation frequencies.



(a) Extraction of the spatial distributions of the maximal spectral magnitude $A(x, y)$ and the dominant oscillation frequency $K_{(x,y)}$ (equations 3.2 and 3.3).

(b) Creating the high-passed and thresholded spatial distribution of the dominant oscillation frequency $K^{HP,thr}_{(x,y)}$ and the histogram of distinct frequencies H_k. (equation 3.4).

Figure 3.2: Obtaining Frequency Images from laryngeal high-speed videos.

$$A(x,y) = \max_{k} \left|J_{(}x,y,k)\right| \tag{3.3}$$

$$K^{HP,thr}_{(x,y)} = \begin{cases} K_{(x,y)}, & \text{if } K_{(x,y)} \geq 70 \text{ Hz} \cap A(x,y) \geq thr \\ \text{NaN}, & \text{if } K_{(x,y)} < 70 \text{ Hz} \cup A(x,y) < thr \end{cases} \tag{3.4}$$

3.2.2 Video spectrum model parameter estimation

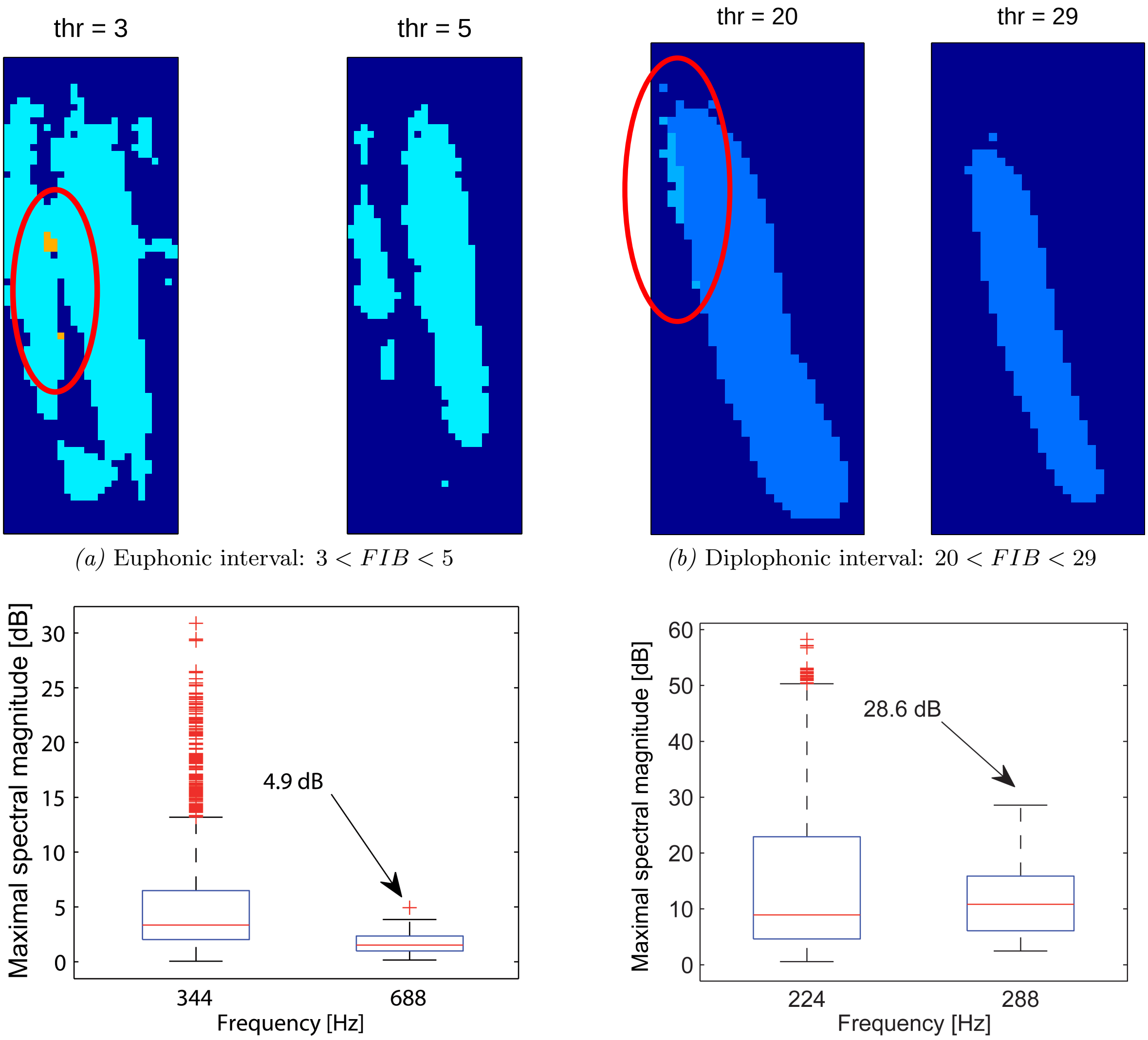
The Frequency Image Bimodality (FIB) is a threshold-based spectral magnitude feature that estimates the presence of secondary oscillators in vocal fold vibration. The FIB is obtained via Frequency Images. A spectral magnitude threshold *thr* enables discarding irrelevant oscillations from Frequency Images. The choice of this threshold depends on objective criteria of relevancy, which are derived hereafter.

The number M of spatially distinct oscillation frequencies is a function of the spectral magnitude threshold. The FIB is defined as the minimal threshold for which the number of detected oscillator frequencies is 1.

$$FIB = min\left\{thr \in \mathbb{R}^{+} | M_{(thr)} = 1\right\} \tag{3.5}$$

Figures 3.3b and 3.3a show the high-passed and thresholded spatial distributions of the dominant oscillation frequencies of a diplophonic interval and an euphonic interval. Each interval is analyzed using two different thresholds[3]. In the diplophonic interval two distinct frequencies exist at a threshold of 20, but only one at a threshold of 29. In the euphonic interval there are two distinct frequencies at a threshold of 3, but only one at a threshold of 5. The FIB is somewhere in between the depicted thresholds and needs to be estimated. Exactly one number fulfills the requirement of equation 3.5 in an observed video interval. Figures 3.3d and 3.3c show the boxplots of the discussed cases, with amplitude on the y-axis and frequency on the x-axis. The FIB equals the maximal amplitude at the secondary frequency (288 and 688 Hz). If a threshold greater than the FIB is used for creating a Frequency Image, only primary frequencies (224 and 344 Hz) are visible.

[3] The values for the thresholds have been chosen for the purpose of visualization and do not have any further meaning.



(a) Euphonic interval: $3 < FIB < 5$

(b) Diplophonic interval: $20 < FIB < 29$

(c) Euphonic interval: FIB = 4.9 dB

(d) Diplophonic interval: FIB = 28.6 dB

Figure 3.3: Determination of the Frequency Image Bimodality from Frequency Images and maximal spectral magnitude boxplots.

3.2.3 Evaluation

The FIB is tested and evaluated with regard to its ability to detect videos of diplophonic intervals in two corpora, namely the test corpus (ID 92) [57] and the confirmatory corpus (ID 46).

Test corpus (ID 92)

In the test corpus ten diplophonic and ten euphonic video intervals of length 400-500 frames (i.e., 100-125 ms at 4000 frames per second) have been auditively selected from the database. Figure 3.4 shows the boxplots and the ROC curve when the FIB is used for detecting diplophonia. The boxplots reveal a group effect and a t-test confirms its significance. The ROC curve in figure 3.4b shows that the optimal cut-off threshold for separating diplophonic from euphonic fragments is 7. Table 3.1a shows the classification performance in absolute numbers of video fragments and table 3.1b shows the performance measures of the classifier and their confidence intervals. The performance measures are promising but suffer from large confidence intervals, which reflect the small sample size.

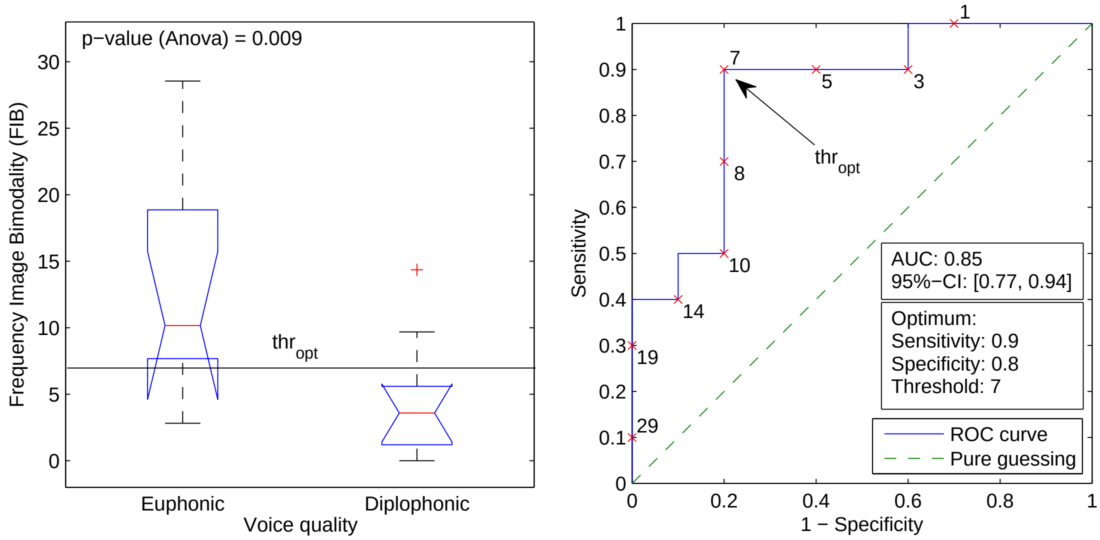
		Voice quality	
		Diplophonic	Euphonic
FIB	> 7	9	2
	< 7	1	8

(a) Diagnostic table.

	Estimate	CI
SE	90 %	[55.5 %, 99.7 %]
SP	80 %	[44.4 %, 97.5 %]
ACC	85 %	[62.1 %, 96.8 %]
PR	50 %	[27.1 %, 72.8 %]
PPV	81.8 %	[48.2 %, 97.7 %]
PLR	4.5	[1.28, 15.81]
NPV	88.9 %	[51.7 %, 99.7 %]
NLR	8	[1.21, 52.69]

(b) Performance measures.

Table 3.1: Diagnostic test evaluation, diplophonic versus euphonic phonation. The Frequency Image Bimodality serves as predictor.



(a) Boxplots, optimal threshold thr_{opt} and t-test.

(b) ROC curve and optimal threshold thr_{opt}.

Figure 3.4: Frequency Image Bimodality versus voice quality (diplophonic vs. euphonic).

Three out of 20 videos are incorrectly classified. Figure 3.5 shows the Frequency Images of these, with the spectral magnitude threshold set equal to the optimal cut-off threshold (7). The Frequency Images for the false positive videos show secondary oscillations at integer multiples of the primary frequency. Visual inspection of the video fragments revealed that these artifacts stem from multiple vocal folds reflections within one cycle. A traveling mucosal wave may reflect the light beam of the light source more often than once, which is hard to interpret because no three-dimensional data of the vocal fold surface are available.

Another source of false decisions is camera movement. In the Frequency Image of the false negative video fragment, only a small amount of pixels show oscillations that are above threshold. Visual inspection of the video revealed that camera movement leads to left/right and forward/backward movement of the glottal gap in the video. The spectral energy is spatially smeared and the maximal spectral magnitudes fall below threshold.

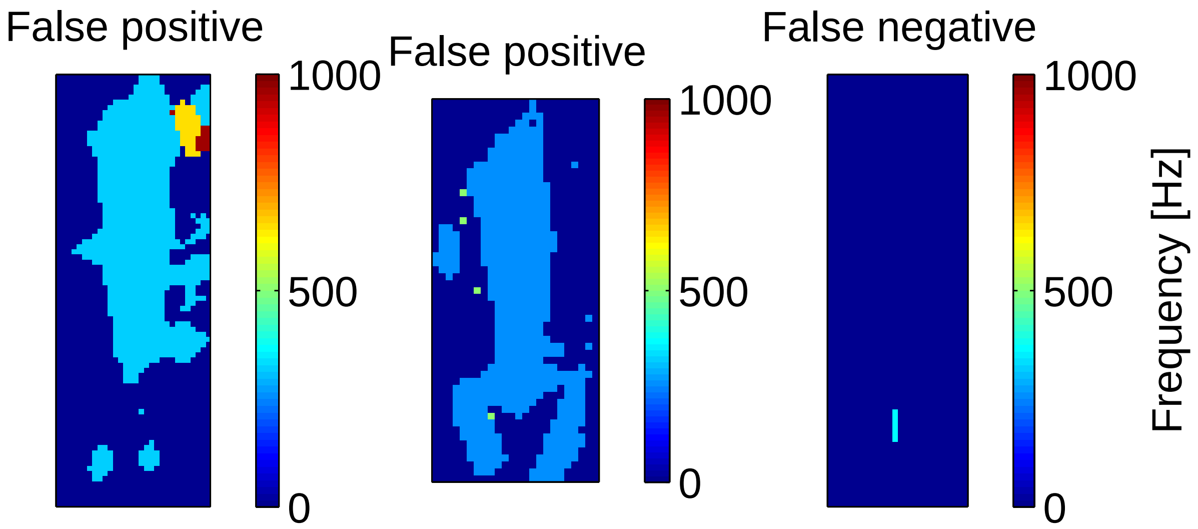


Figure 3.5: Frequency Images of false decisions from Frequency Image Bimodality classification. The spectral magnitude threshold is set equal to the optimal cut-off threshold.

Confirmatory corpus (ID 46)

To evaluate the validity of the FIB it is tested on a larger corpus. Corpus 46 is well suited, because it is pre-selected for video quality. In contrast to corpus 92, the videos are not spatially cropped prior to analysis. To suppress artifacts from camera movement, the analyzed intervals are shorter (128 frames or 64 ms). The FIB is time-averaged over video intervals.

The same tendencies as in the test corpus are observed, which suggest that the FIB is a valid method for detecting diplophonia. Figure 3.6 shows the boxplots of the FIB versus voice quality and the p-value of the Kruskal-Wallis test. The FIB is significantly increased in the diplophonic group compared to the other two groups (euphonic and dysphonic), which confirms the expected relationship between FIB and diplophonic voice quality. Figure 3.7 shows the ROC curves for the FIB classifier. Both negative groups (euphonic or dysphonic) are depicted. The classification performances of both classifiers are comparable, which is reflected by similar sensitivities (90 % and 80 %), specificities (66.7 % and 70 %) and AUCs (0.76 and 0.73). The ROC curves show a very steep decrease of sensitivity in a threshold range from 2.6 to 3.3, which reflects the large number of diplophonic videos in that range. The algorithm for searching the optimal threshold obtains low specificity, because increasing the threshold would have disastrous consequences for the sensitivity.

Tables 3.2 and 3.3 show the classification outcome and the performance measures. The confidence intervals are smaller than in the test corpus, but far from narrow, which suggests that larger corpora are necessary for statistical validation.

The camera movement problem can successfully be addressed by choosing shorter analysis windows ($N = 128$ frames instead of 500 frames). It appears that window length is crucial for analysis, which must be adjusted for if FIB values are to be compared. The optimal thresholds are lower in the confirmatory set (2.3 and 2.6, compared to 7), because shorter time-windows are used for analysis. FIB is a threshold of spectral magnitude, which decreases if the block length is decreased. If the shorter window length is adjusted for, optimal thresholds of $2.3 \cdot \frac{128}{500} = 9$ and $2.6 \cdot \frac{128}{500} = 10.2$ are obtained, which are closer to 7.

The results from the confirmatory corpus suggest that spatial cropping is not crucial for classification accuracy. Data pre-selection was carried out, which may be a prerequisite for accurate analysis of uncropped videos.

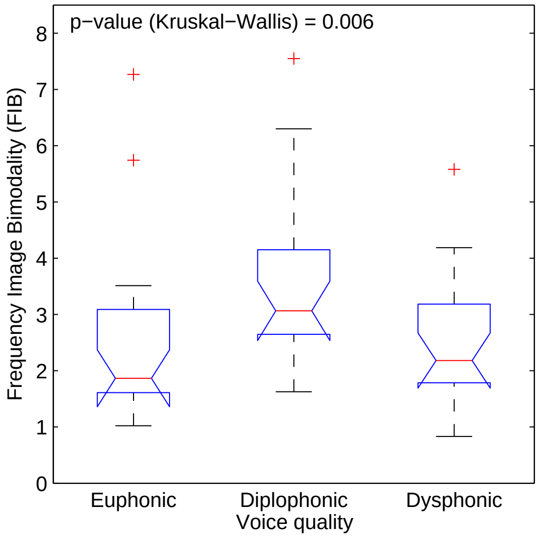
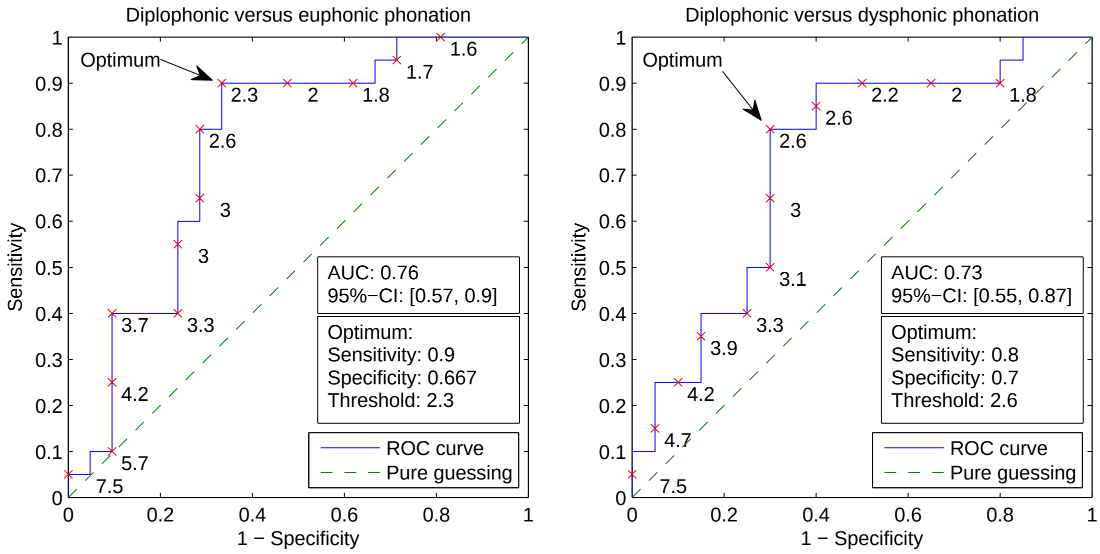


Figure 3.6: Boxplots of Frequency Image Bimodality versus voice quality, testing on corpus 46.



(a) Diplophonic versus euphonic phonation. *(b)* Diplophonic versus dysphonic phonation.

Figure 3.7: ROC curves for predicting voice quality from corpus 46, predictor: Frequency Image Bimodality.

		Voice quality	
		Diplophonic	Euphonic
FIB	> 2.3	18	7
	< 2.3	2	14

(a) Classification outcome.

	Estimate	CI
SE	90 %	[68.3 %, 98.8 %]
SP	66.7 %	[43 %, 85.4 %]
ACC	78 %	[62.4 %, 89.4 %]
PR	48.8 %	[32.9 %, 64.9 %]
PPV	72 %	[50.6 %, 87.9 %]
PLR	2.7	[1.45, 5.04]
NPV	87.5 %	[61.6 %, 98.4 %]
NLR	6.67	[1.73, 25.7]

(b) Performance measures.

Table 3.2: Classification outcome, diplophonic versus euphonic phonation, with Frequency Image Bimodality as predictor. Confirmatory corpus (ID 46).

		Voice quality	
		Diplophonic	Dysphonic
FIB	> 2.6	14	4
	< 2.6	6	16

(a) Classification outcome.

	Estimate	CI
SE	70 %	[45.7 %, 88.1 %]
SP	80 %	[56.3 %, 94.3 %]
ACC	75 %	[58.8 %, 87.3 %]
PR	50 %	[33.8 %, 66.4 %]
PPV	77.8 %	[52.3 %, 93.6 %]
PLR	3.5	[1.39, 8.8]
NPV	72.7 %	[49.8 %, 89.3 %]
NLR	2.67	[1.32, 5.39]

(b) Performance measures.

Table 3.3: Classification outcome, diplophonic versus dysphonic phonation, with Frequency Image Bimodality as predictor. Confirmatory corpus (ID 46).

3.3 Frequency analysis of glottal edge trajectories

In the previous section two-dimensional Frequency Images are derived from three-dimensional video data. In this section one-dimensional Frequency Plots are derived from two-dimensional video data (i.e., from glottal edge trajectories). It is hypothesized that analyzing trajectories is less prone to mucosal wave artifacts than analyzing pixel intensity time-series because the mucosal wave is not present in the trajectories.

The algorithm is structured as follows:

1. Extraction of the glottal edge trajectories.
2. Mean removal from the trajectories.
3. Short-time Fourier transform.
4. Plotting of the dominant frequencies along the vocal fold edges.
5. Determination of the Frequency Plot Bimodality (FPB).

The trajectories have been extracted from the videos with the program "Glottis Analysis Tools" via spatial glottis segmentation (figure 3.8) [93, 94]. The dashed yellow line is the main glottal axis, the red and the blue curves are the glottal edges. The time-variant separation of the glottal edge pixels from the main axis are the glottal edge trajectories (red and blue horizontal lines). They have been obtained for 256 positions along the main glottal axis via spatial interpolation. More details on the extraction of the trajectories can be found in [94].

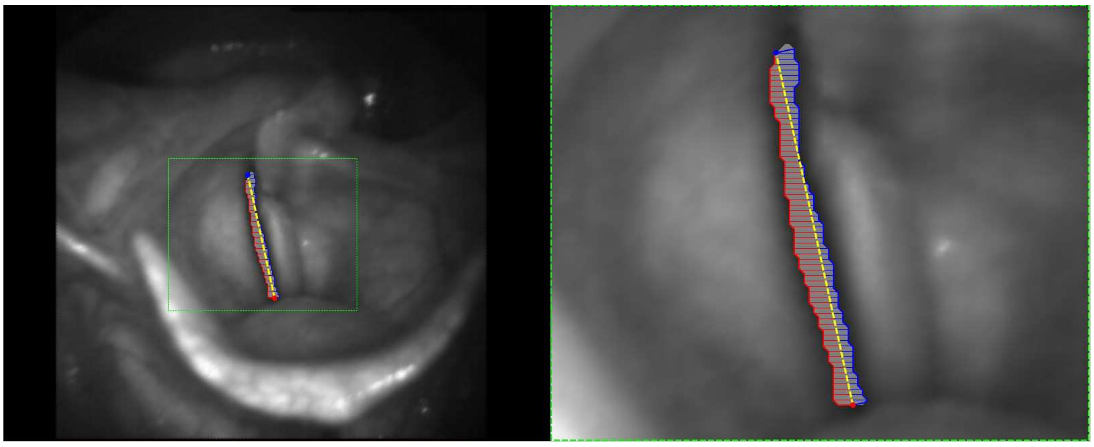


Figure 3.8: Extraction of the glottal edge trajectories. The dashed yellow line is the main glottal axis, the red and the blue curves are the vocal fold edges. The time-variant separation of the glottal edge pixels from the main axis are the glottal edge trajectories (red and blue horizontal lines). They have been obtained for 256 positions along the main glottal axis via spatial interpolation.

3.3.1 Video spectrum model description

Let $d_{y'}^{(x')}(n)$ denote the trajectories, where y' is the position along the main glottal axis (posterior: 1, anterior: 256), x' labels the positions of the left and right vocal fold edges and n is the video frame index. The trajectories are blocked with respect to time which yields $d_{y'}^{(x')}(n', i)$, where i is the block index and n' the block-relative video frame index. The block length N equals 240 video frames (i.e., 60 ms at a video frame rate of 4000 frames per second) and the overlap is 50 %. Figure 3.9 shows an example of a spatio-temporal plot (STP) of diplophonic vocal fold vibration, with the left and the right vocal fold vibrating at different frequencies.

$\delta_{y'}^{(x')}(n', i)$ denotes the blocked and zero-mean trajectories, where the mean is computed over n', separately for each position x' and y' and each block i .

$$\delta_{y'}^{(x')}(n', i) = d_{y'}^{(x')}(n', i) - \overline{d_{y'}^{(x')}}(n', i) \tag{3.6}$$

$\Delta_{y'}^{(x')}(k, i)$ are the short-time Fourier transforms of the trajectories, where k is the discrete frequency and $w_{n'}$ is a 240 frames time-domain Hann window (equation 3.7, figure 3.10). Both vocal fold frequencies are visible in the majority of the spectra, which is due to the coupling. The dominant frequency $\kappa_{y'}^{x'}(i)$ is the frequency at the maximal spectral magnitude and obtained as a function of positions x' and y' (figure 3.11a).

$$\Delta_{y'}^{(x')}(k, i) = \underset{n' \to k}{\mathrm{DFT}} \left\{ \delta_{y'}^{(x')}(n', i) \cdot w_{n'} \right\} \tag{3.7}$$

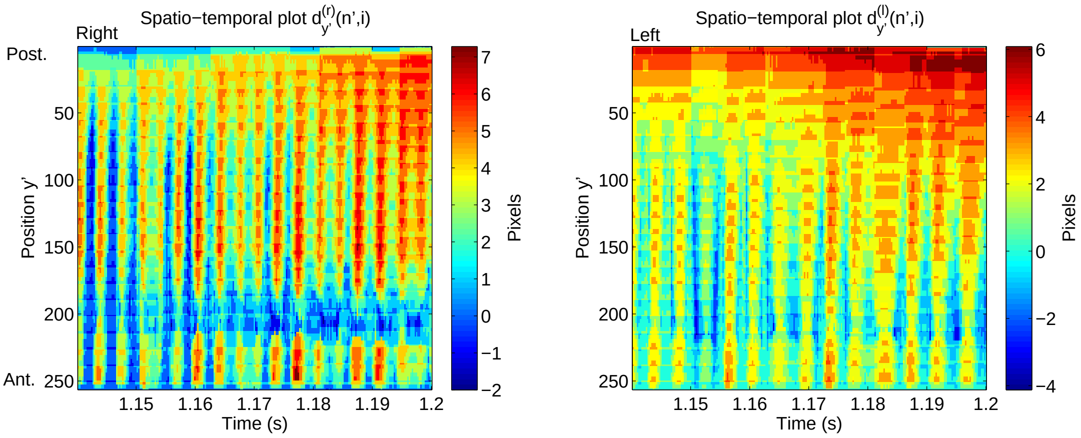


Figure 3.9: Example of a spatio-temporal plot of diplophonic phonation, with color-coded distances of the glottal edges from the main glottal axis. y' denotes the position along the main glottal axis.

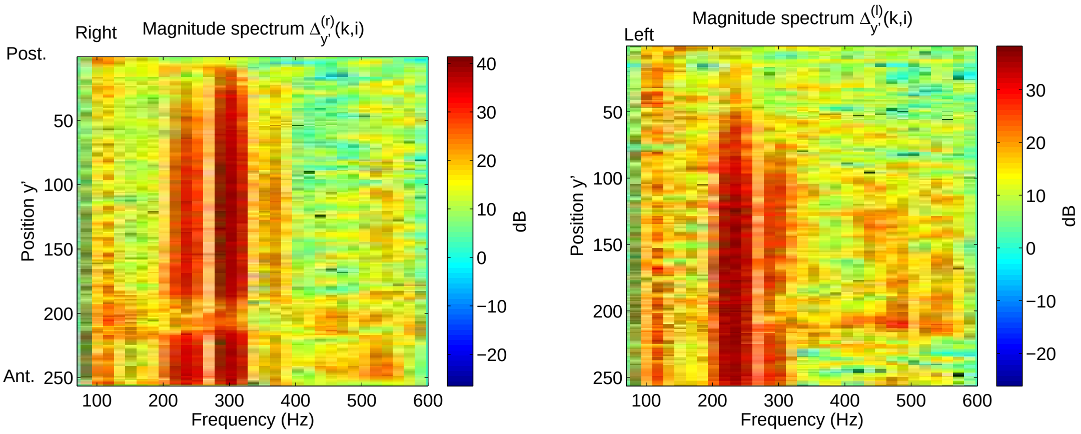
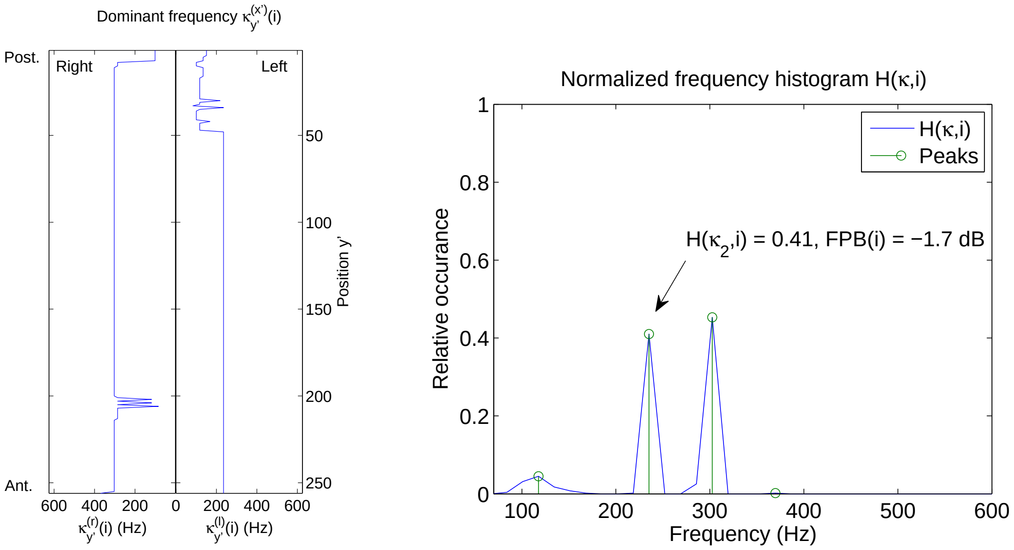


Figure 3.10: Short-time Fourier transforms $\Delta_{y'}^{(x')}(k,i)$ of the mean free glottal edge trajectories. y' denotes the position along the main glottal axis.



(a) *(b)*

Figure 3.11: Frequency Plot and normalized frequency histogram. y' denotes the position along the main glottal axis.

.

3.3.2 Video spectrum model parameter estimation

The FPB (Frequency Plot Bimodality) quantifies the spatial distribution of dominant frequencies along the vocal fold edges. It is obtained from the Frequency Plot (figure 3.11a). The rationale of the FPB is published in [58] and described hereafter.

The Frequency Plot is summarized in the normalized frequency histogram $H(\kappa, i)$ (figure 3.11b). The second highest peak $H(\kappa_2, i)$ is obtained by peak picking and κ_2 is its diplophonic frequency. The peaks are represented by circle headed stems in the normalized frequency histogram. The FPB is obtained by relating $H(\kappa_2, i)$ to 0.5 on a dB scale (equation 3.8).

$$FPB(i) = 20 \cdot \log\left(\frac{H(\kappa_2, i)}{0.5}\right) \text{dB} \tag{3.8}$$

The choice of 0.5 is justified by the desired range of FPB. For perfect bimodality, i.e., two equally high peaks in the normalized frequency histogram, 0 dB are desired. Any deviation from perfect bimodality results in a reduction of FPB towards negative values. One possibility of a deviation is a decrease of the second highest peak, while the highest peak is increased. Another possibility is a decrease of the second highest peak while other peaks are increased (tertiary, quaternary etc.), which indicates other oscillation patterns than diplophonia. Considering 512 trajectories (256 positions y' for each vocal fold) the smallest possible FPB is −48.2 dB, because $H(\kappa_2, i) \geq \frac{1}{512} \hat{=} -48.2$ dB, and thus the FPB is set to −50 dB if there is only one peak in $H(\kappa, i)$.

3.3.3 Evaluation

The ability of FPB to detect diplophonic vocal fold vibration from a corpus of euphonic, diplophonic and dysphonic phonation video intervals is tested. Results from two experiments are presented, the first one on block wise analysis [58], and the second one on interval averages. A sample of 21 euphonic, 20 diplophonic, 20 dysphonic video intervals has been drawn from the database. The intervals are between 0.1328 and 2.048 seconds long (corpus ID 46).

Block wise classification

The analysis has been carried out for blocks of 240 frames with 50 % overlap. The results are obtained for 753 euphonic, 443 diplophonic and 616 dysphonic blocks.

Figure 3.12 shows the boxplots of FPB versus voice quality. A group effect is observed, i.e., diplophonic video blocks have higher FPB values than euphonic and dysphonic blocks ($p < 0.001$). The overlap between the euphonic and the diplophonic group is smaller than the overlap between the diplophonic and the dysphonic group.

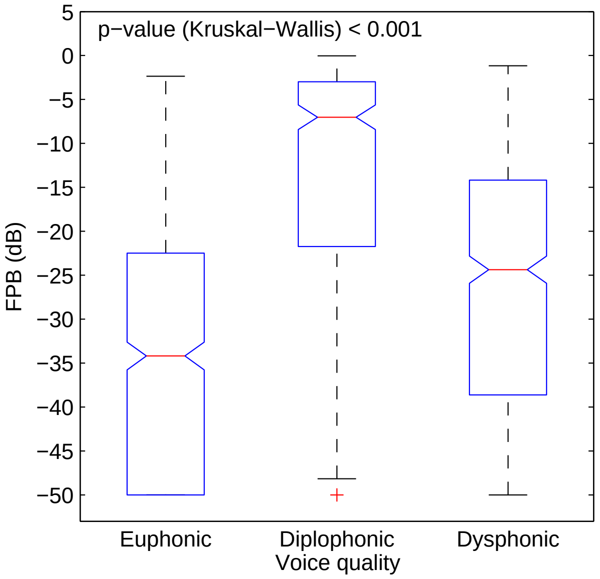
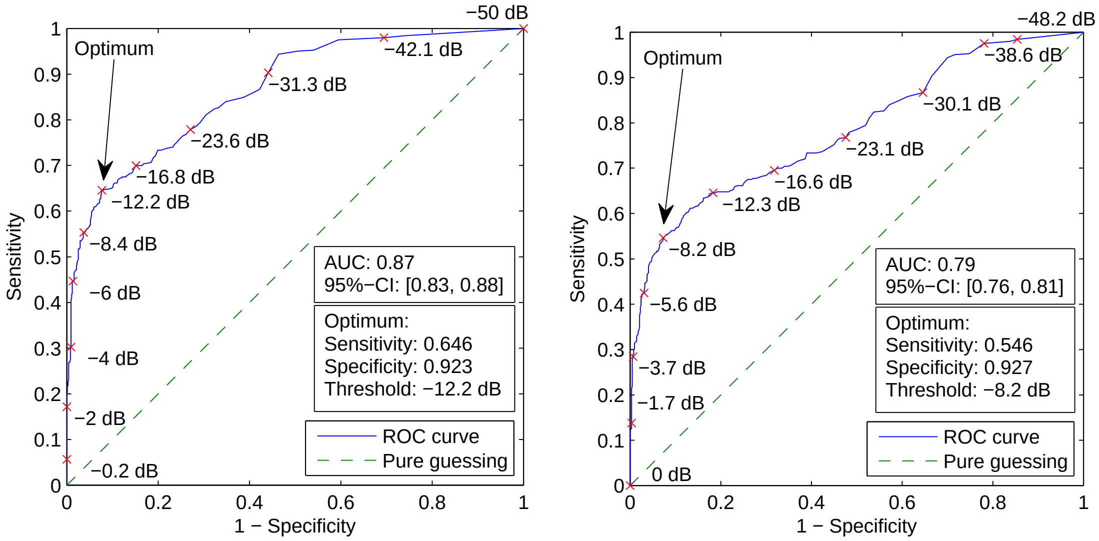


Figure 3.12: Boxplots of the block wise Frequency Plot Bimodality versus voice quality.

Block wise classification from the proposed FPB is not perfect, and the distinction between diplophonia and dysphonia is more challenging than the distinction between diplophonia and euphonia. The ROC curve for the euphonic negative group shows a better performance than the ROC curve for the dysphonic negative group (figure 3.13). The AUCs are at 0.87 and 0.79 and differ quantitatively, which reflects the overlap of the data distributions. The optimal thresholds are -12.2 dB (diplophonic versus euphonic) and -8.2 dB (diplophonic versus dysphonic). The specificities are comparably high at 92.3 and 92.7 % and the sensitivities are lower and differ quantitatively (64.6 and 54.6 %). Tables 3.4 and 3.5 summarize test performances.

The optimal thresholds in this section have been obtained by moving a straight line with slope S from the upper left corner of the ROC (i.e., the ultimate optimum) to the right and to the bottom, until it intersects the curve. The slope S is given by $S = \frac{N}{p}$, where N is the number of negative observations and P the number of positive observations. The optimal threshold is at the point of intersection [56]. This approach to threshold determination takes into account that if the number of negative observation exceeds the number of positive ones, the specificity estimates are more reliable than sensitivity estimates.



(a) Diplophonic vs. euphonic.

(b) Diplophonic vs. dysphonic.

Figure 3.13: ROC curves of the block wise Frequency Plot Bimodality predicting voice quality.

		Voice quality	
		Diplophonic	Euphonic
FPB(i)	≥ -12.2 dB	286	58
	< -12.2 dB	157	695

(a) Classification outcome.

	Estimate	CI
SE	64.6 %	[59.9 %, 69 %]
SP	92.3 %	[90.2 %, 94.2 %]
ACC	82 %	[79.9 %, 84.2 %]
PR	37 %	[34.3 %, 39.8 %]
PPV	83.1 %	[77.8 %, 86.9 %]
PLR	8.39	[6.49, 10.8]
NPV	81.6 %	[78.8 %, 84.1 %]
NLR	2.61	[2.3, 2.96]

(b) Performance measures.

Table 3.4: Classification performance at the optimal cut-off threshold, block wise Frequency Plot Bimodality (FPB), euphonic negative group. 95 % confidence intervals (CI).

Classification from interval averages

This section shows results of voice quality classification from video fragment averages of the FPB (equation 3.8) as compared to block wise results from the previous section. Figure 3.14 shows the boxplots for the averaged FPB versus voice quality. The variances of the averaged FPB are smaller than those of the block wise FPB because extreme values average out. The confidence intervals (notches) are larger due to the smaller sample size. Figure 3.13 shows the ROC curves for the euphonic negative group and the dysphonic negative group. In contrast to the ROC curves obtained from block wise results, the curves do not differ significantly for the euphonic and dysphonic negative groups. The optimal thresholds are lower than those for block wise classification. Tables 3.6 and 3.7 summarize test performances.

		Voice quality	
		Diplophonic	Dysphonic
FPB(i)	≥ -8.2 dB	242	45
	< -8.2 dB	201	571

(a) Classification outcome.

	Estimate	CI
SE	54.6 %	[49.9 %, 59.3 %]
SP	92.7 %	[90.3 %, 94.6 %]
ACC	76.8 %	[74.1 %, 79.3 %]
PR	41.8 %	[38.8 %, 44.9 %]
PPV	84.3 %	[79.6 %, 88.3 %]
PLR	7.48	[5.57, 10]
NPV	74 %	[70.7 %, 77 %]
NLR	2.04	[1.84, 2.27]

(b) Performance measures.

Table 3.5: Classification performance at the optimal cut-off threshold, block wise Frequency Plot Bimodality (FPB), dysphonic negative group. 95 % confidence intervals (CI).

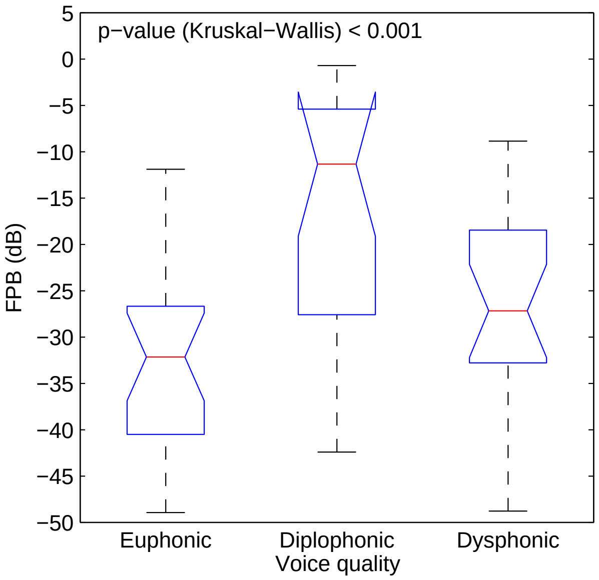


Figure 3.14: Boxplots of interval averages of the Frequency Plot Bimodality versus voice quality.

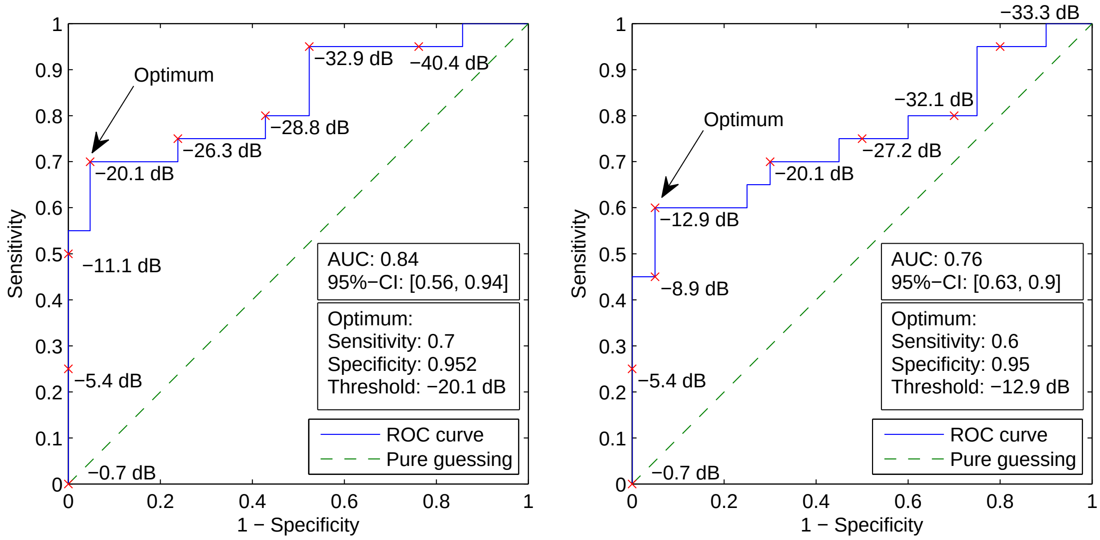
		Voice quality	
		Diplophonic	Euphonic
FPB(i)	≥ -20.1 dB	14	1
	< -20.1 dB	6	20

(a) Classification outcome.

	Estimate	CI
SE	70 %	[45.7 %, 88 %]
SP	95.2 %	[76.1 %, 99.2 %]
ACC	82.9 %	[67.9 %, 92.8 %]
PR	48.8 %	[32.9 %, 64.9 %]
PPV	93.3 %	[68.1 %, 99.8 %]
PLR	14.6	[2.12, 100]
NPV	76.4 %	[56.4 %, 91 %]
NLR	3.17	[1.61, 6.24]

(b) Performance measures.

Table 3.6: Classification performance at the optimal cut-off threshold, interval average Frequency Plot Bimodality, euphonic negative group. 95 % confidence intervals (CI).



(a) Diplophonic vs. euphonic.

(b) Diplophonic vs. dysphonic.

Figure 3.15: ROC curves of the interval average Frequency Plot Bimodality predicting voice quality.

		Voice quality	
		Diplophonic	Dysphonic
FPB(i)	≥ −12.9 dB	12	1
	< −12.9 dB	8	19

(a) Classification outcome.

	Estimator	CI
SE	60 %	[36.1 %, 80.8 %]
SP	95 %	[75.1 %, 99.2 %]
ACC	77.5 %	[61.6 %, 89.2 %]
PR	50 %	[33.8 %, 66.2 %]
PPV	92.9 %	[64 %, 99.8 %]
PLR	12	[1.72, 83.8]
NPV	70.4 %	[49.8 %, 86.2 %]
NLR	2.38	[1.38, 4.1]

(b) Performance measures.

Table 3.7: Classification performance at the optimal cut-off threshold, interval average Frequency Plot Bimodality, dysphonic negative group. 95 % confidence intervals (CI).

3.4 Discussion

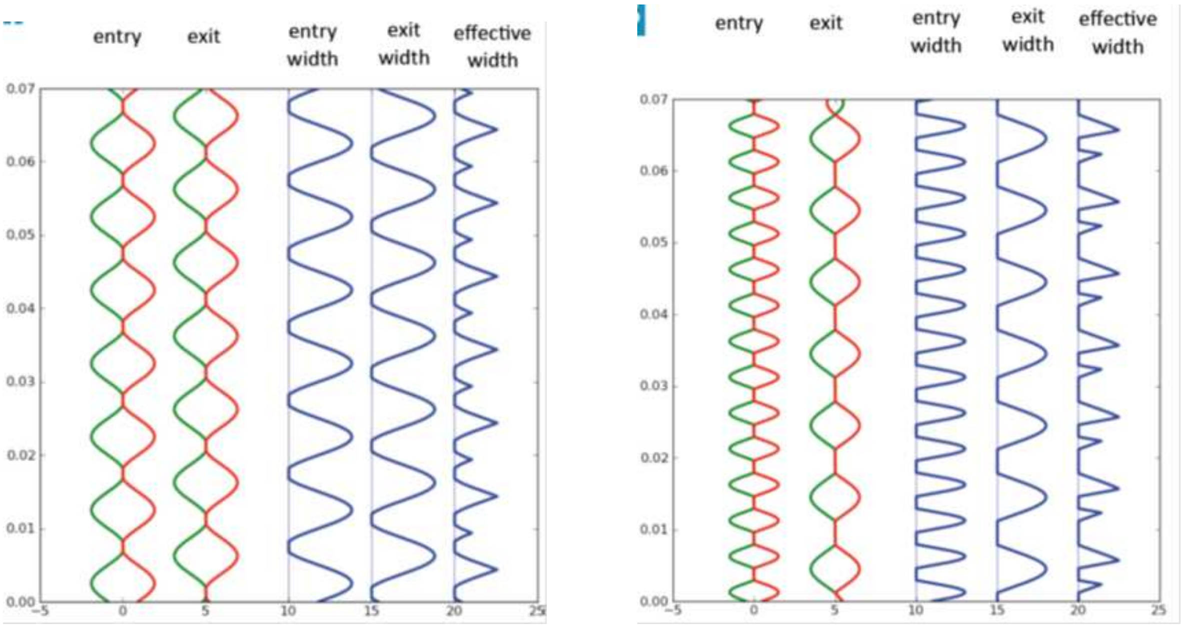
Spectral representations of glottal edge trajectories in the presence of frequency asymmetry in vocal fold vibration are not fully understood, because two problems limit the interpretability of the analyzes. First, laryngeal high-speed videos are three-dimensional projections of four-dimensional phenomena. Second, the vocal fold tissue is opaque and only visible glottal edges are observable. Hereafter, a geometrical model of vocal fold vibration is used to simulate several types of asymmetric vocal fold vibration and some clinically observed kymograms are shown.

3.4.1 Titze's phase-delayed overlapping sinusoids model

A geometric phase-delayed overlapping sinusoids (PDOS) model is used to gain insight into the complexity of asymmetric vocal fold vibration. Titze's geometric PDOS model has been generalized for frequency asymmetry [95]. The effective glottal width, the glottal area and pseudo-kymographic patterns are simulated.

Superior-inferior asymmetry

Figure 3.16 shows that unequal glottal cycles can arise both from superior-inferior phase asymmetry and superior-inferior frequency asymmetry. Neither the FIB nor the FPB are able to distinguish these phenomena, but only frequency asymmetry (figure 3.16b) should test positive on glottal diplophonia.



(a) Superior-inferior phase asymmetry - No diplophonia - 1:1 double pulsing

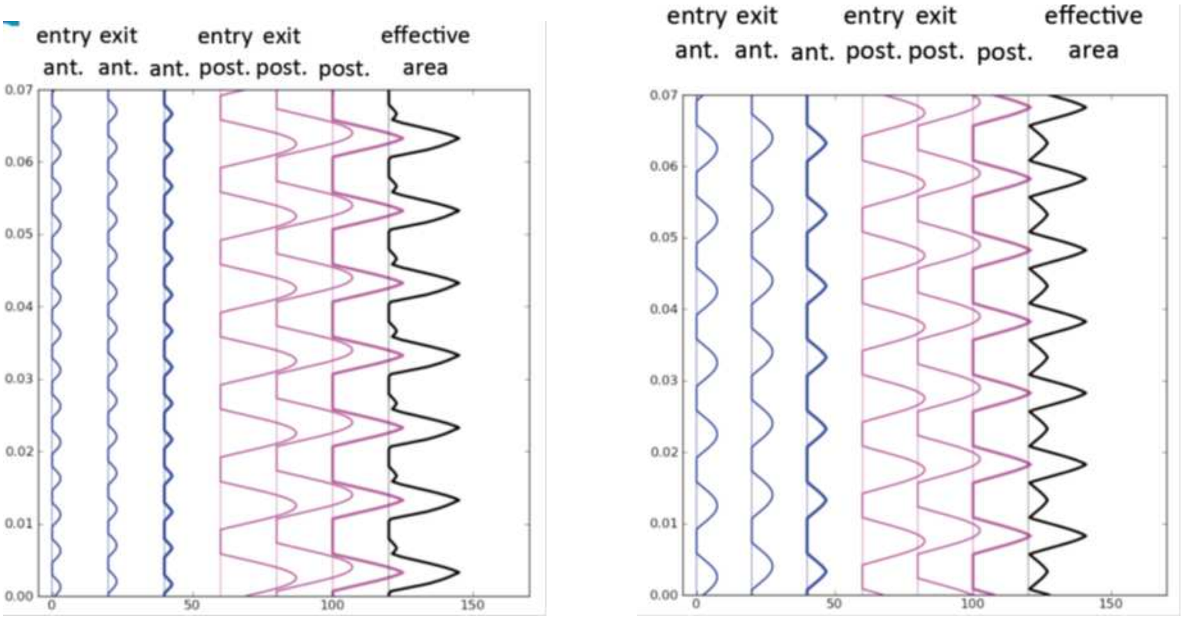
(b) Superior-inferior 1:2 frequency asymmetry - Diplophonia

Figure 3.16: Simulation of superior-inferior asymmetry with a phase-delayed overlapping sinusoids (PDOS) model, modified from [96].

Spectral representations of subharmonic patterns similar to those shown in figure 3.16b were investigated by Malyska et al. [97]. In metacyclic vibration, the partials' magnitudes of the spectrum are multiplied by a continuous envelope function $M(\omega)$, where $M(\omega)$ is the Fourier transform of one metacycle. The dominant spectral frequency depends on the fundamentals' positions relative to the spectral envelope $M(\omega)$, which is not fully understood. Malyska investigated spectral representations of scaled and time-displaced versions of unit pulses. Figures 6 and 8 to 10 in [97] show that subharmonic frequencies (white headed stems) are weak compared to fundamentals (black headed stems). Thus it is likely that the correct fundamental is found by the global spectral maximum selection in the proposed algorithms if unit pulses are replaced by more realistic pulses.

Anterior-posterior asymmetry

As opposed to double pulsing due to superior-inferior asymmetry, patterns of double pulsing can also arise from anterior-posterior asymmetry (figure 3.17), where one distinguishes phase asymmetry and frequency asymmetry. FIB and FPB succeed in detecting glottal diplophonia (figure 3.17b), because the separate frequencies are spatially distributed along the anterior-posterior glottal axis.



(a) Anterior-posterior 1:2 frequency asymmetry

(b) Anterior-posterior phase asymmetry

Figure 3.17: Simulation of anterior-posterior asymmetry with a phase-delayed overlapping sinusoids (PDOS) model, modified from [96].

Left-right asymmetry

In left-right 1:2 frequency asymmetry, the edge of the right vocal fold (red) vibrates at two times the left edge frequency (figure 3.18). The right trajectory experiences clipping due to the collision, which introduces subharmonic frequencies in its spectrum. The FIB and FPB are expected to detect the true fundamental frequencies by selecting the maximal spectral magnitude.

More examples

Figure 3.19 shows left-right 2:3, 4:5 and 1:1.43 frequency ratios. Collision clipping causes subharmonic patterns that are more complicated with regard to former cases. It is likely that the meta fundamental is the dominant frequency in the spectrum, but the spectral representations of these signals are not fully understood.

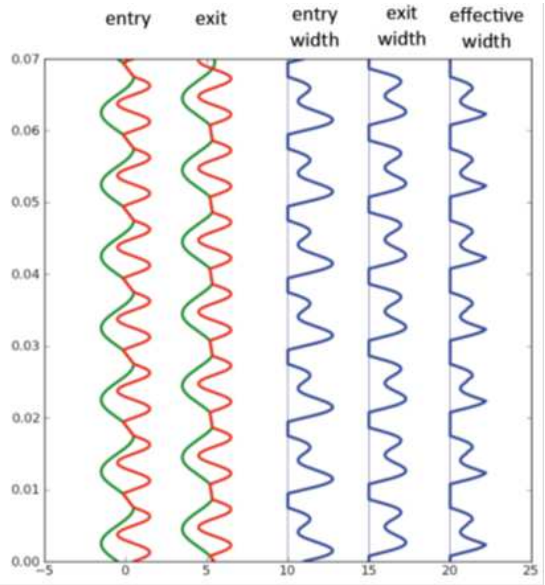
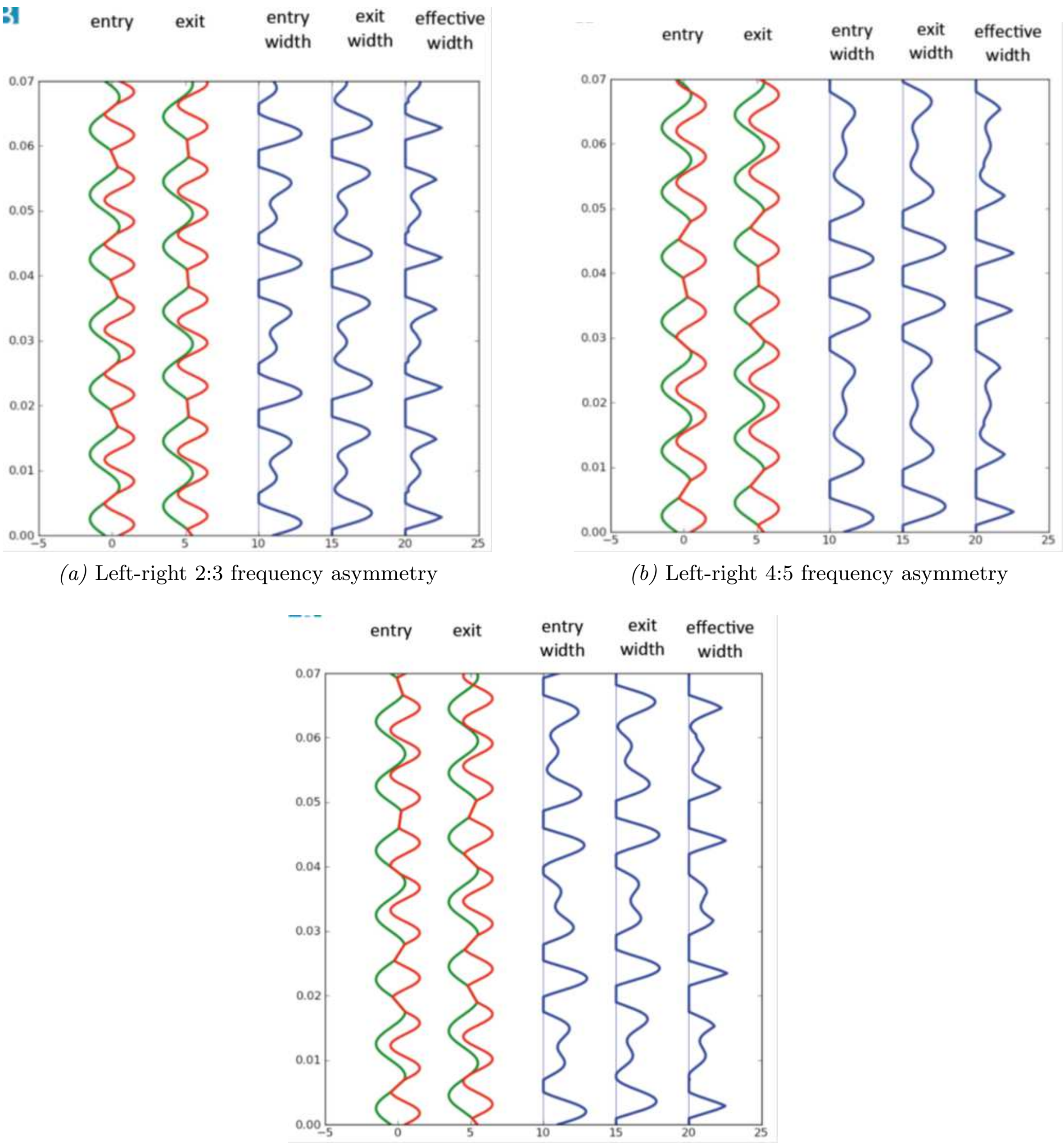


Figure 3.18: Simulation of 1:2 left-right asymmetry with a phase-delayed overlapping sinusoids (PDOS) model, modified from [96].



(a) Left-right 2:3 frequency asymmetry

(b) Left-right 4:5 frequency asymmetry

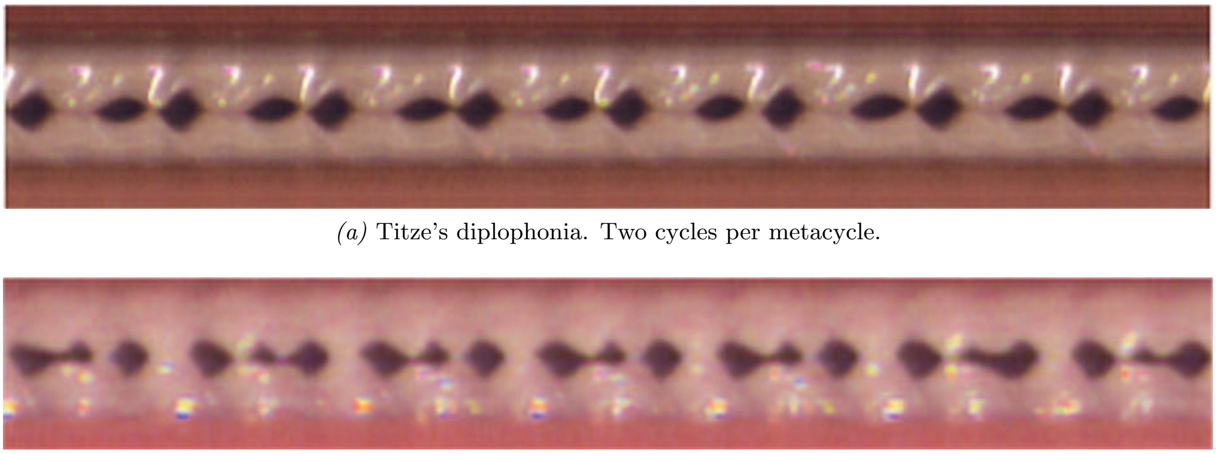
(c) Left-right 1:1.43 frequency asymmetry

Figure 3.19: Simulation of left-right frequency asymmetry with a phase-delayed overlapping sinusoids (PDOS) model, modified from [96].

3.4.2 Digital kymograms

This section shows some kymograms of diplophonic phonation to demonstrate the complexity of the phenomenon. Left-right and anterior-posterior frequency asymmetry can be directly observed from laryngeal high-speed videos of diplophonic vocal fold vibration. The observation of superior-inferior asymmetry is more difficult, because inferior edges are likely to be permanently or occasionally hidden by the superior tissue.

Figure 3.19 shows kymograms of subharmonic metacycle patterns. The technique of kymography and digital multiplane kymography is explained in [98, 99]. In that figure no spatial asymmetry is visible in the kymograms. With regard to Schoentgen's geometrical model it is not clear whether the metacyclic patterns arise from superior-inferior frequency asymmetry or left-right n:n frequency asymmetry. Superior-inferior phase shift is less likely because the inferior edges cannot be seen in the kymogram and the cycle lengths are unequal.



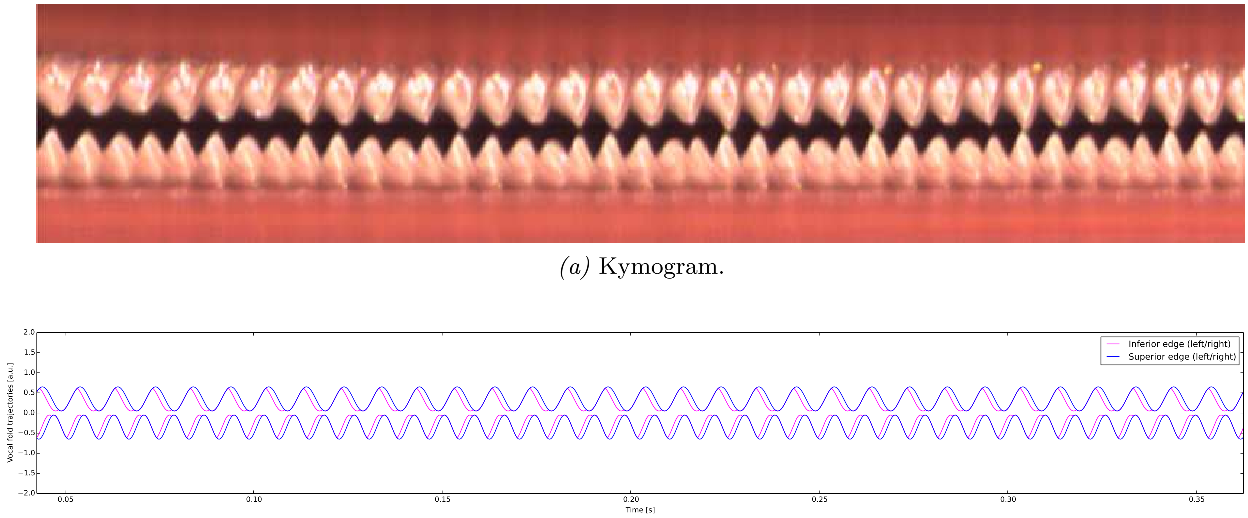
(a) Titze's diplophonia. Two cycles per metacycle.

(b) Titze's triplophonia. Three cycles per metacycle.

Figure 3.20: Metacycles observed in kymograms. Brightness and contrast were corrected.

Figure 3.21 shows a kymogram and a qualitative model fit of left-right 5:4 frequency asymmetry. One sees that the superior and inferior edges are phase shifted, which is correctly simulated in the model. However, one also sees that the model might benefit from adding more cross sections in the superior-inferior direction and from modulation capabilities (amplitude, frequency, pulse shape). Pixelizing the model would avoid unobservable double peaks medially. Both the FIB and the FPB would succeed in detecting diplophonia in this video.

Figure 3.22 shows a kymogram of anterior-posterior 1:2 frequency asymmetric phonation. This observed case is similar to the modeled case in figure 3.17b but no closure is observed and anterior-posterior phase asymmetry exists. The spatial amplitude of the secondary oscillator is very small, which may lead to failure to detect by the FIB and the FPB.



(a) Kymogram.

(b) Simulation with a phase-delayed overlapping sinusoids (PDOS) model.

Figure 3.21: Left-right frequency asymmetry (4:5).

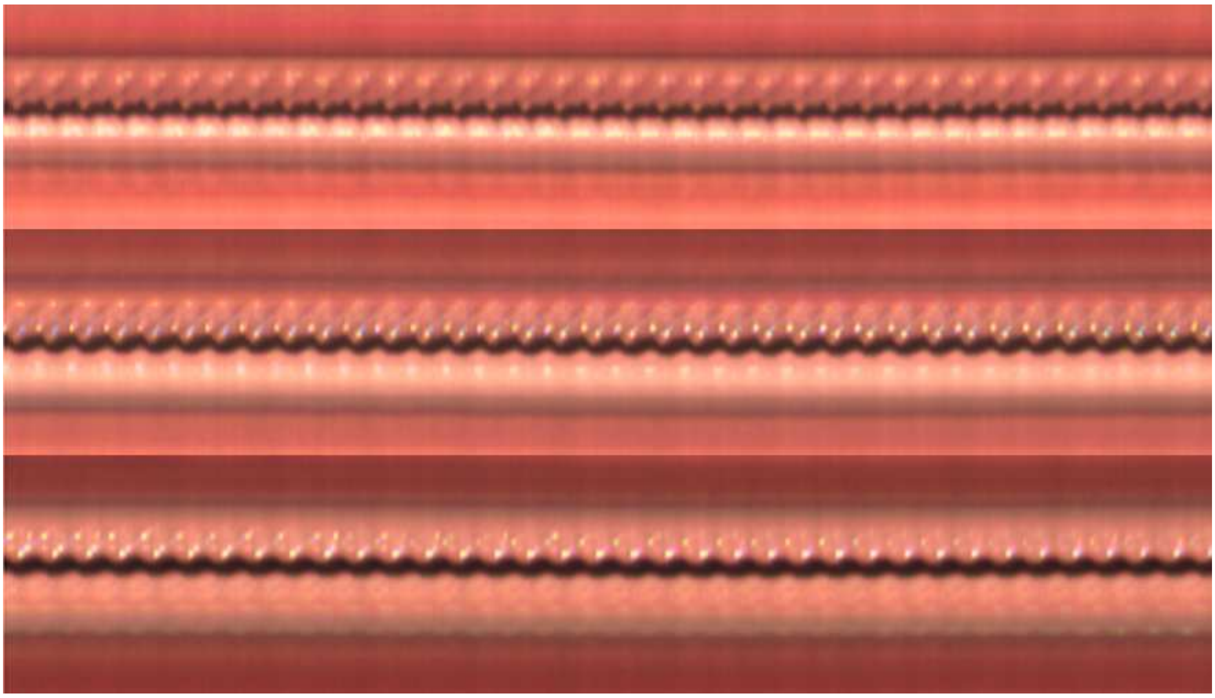


Figure 3.22: Anterior-posterior frequency asymmetry (2:1) + phase asymmetry. Anterior: top, posterior: bottom.

3.5 Conclusion

Investigating spatial aspects of vocal fold vibration connects with the etiologies of possible disorders. Semi-automatic and automatic procedures for detecting the presence of secondary oscillations in the vocal folds have been proposed, evaluated and discussed with respect to a geometric vocal fold model and measurement artifacts [58, 59]. It has been shown that the proposed procedures are likely to be valid for left-right and anterior-posterior frequency asymmetry, and in the absence of measurement artifacts, likely to succeed in detecting these kinds of asymmetries. The procedures are not able to distinguish superior-inferior frequency asymmetry from superior-inferior phase asymmetry. The discussion on superior-inferior asymmetry offers the potential to resolve the definition problem in chapter 1: alternating patterns in phonation pulses can be explained by superior-inferior asymmetry. Both frequency and phase asymmetry are possible, but only frequency asymmetry is glottal diplophonia.

The literature suggests that diplophonia discovered in the audio domain is similar to diplophonia discovered in the in the video domain [12], but the presented data show that this is not always true. The voice quality annotation has been carried out solely on the base of the audio material and several reasons exist for divergences between levels of description. First, turbulent or extraglottal sound sources possibly exist. Turbulent noise does not have a harmonic structure, but can cause a pitch sensation [100]. "Glottal whistle" may also arise from glottal turbulence, but produces periodic sound pressure variations by interacting with vocal tract resonances [101]. The perceived periodic sound pressure variations are similar to labial whistling and cannot be seen in the video. Other examples of extraglottal sound sources are the false vocal folds and the aryepiglottal folds. The amplitude based FIB does not detect false vocal fold or aryepiglottic folds vibration, because their contrast with the surrounding tissue is lower than the contrast of the glottal gap. The FPB is not able to detect extraglottal vibrations either, because they are not represented in STPs. Extraglottal sound sources that are outside of the camera's viewing angle are not detectable at all. In addition, auditive diplophonia arising from small amplitude vibrations of the vocal folds may not be detectable from laryngeal high-speed videos due to the coarse spatial camera resolution. The FIB fails in detecting such vibrations because the pixel intensity fluctuations are subliminal and the FPB fails if the oscillation spectral intensity falls below the spatial quantization noise floor.

Analysing three-dimensional video data (FIB) has one advantage and one disadvantage. The advantage is that raw video data can be analyzed automatically. No spatial segmentation of the glottal gap is needed, and the ROI cropping may not be necessary because the FIB classification has been tested successfully on uncropped video data. However, the procedure suffers from low specificity, which is mainly due to mucosal wave artifacts.

Analysing two-dimensional video data (FPB) has one advantage and two disadvantages. The advantage is that the FPB is not prone to the mucosal wave problem. One disadvantage is that semi-automatic spatial segmentation of the glottal gap needs to be done as a preliminary step to FPB analysis. Another disadvantages is that the classifier achieves rather low sensitivity.

4 Two-oscillator waveform models for analyzing diplophonia

4.1 Motivation

The previous chapter introduces strategies for modeling and analyzing the spatial patterns observed in laryngeal high-speed videos. That approach is cause-oriented, because it enables understanding the underlying mechanisms of phonation. In this chapter, procedures for investigating the glottal area and the audio waveform are proposed and evaluated. These procedures are output-oriented, because they enable understanding the acoustic output of disordered phonation and its perception. An analysis-by-synthesis paradigm for automated detection of diplophonia from observed waveforms is pursued.

Fundamental frequency is an essential input to synthesis procedures and cycle marks[4] express fundamental frequency by identifying cycle-synchronous events in the waveform. Figure 4.1 shows an audio waveform with speech cycles and cycle marks that were determined via two different approaches [102]. The first is the one of Praat [85] and the second is based on Poincaré sections. Praat misses every second pulse in the diplophonic interval of alternating amplitudes and cycle lengths (2.46 - 2.54 s). The estimated fundamental frequency drops from approximately 90 Hz to 45 Hz (i.e., the approximate frequency of the metacycle) at 2.46 s and returns back at 2.58 s. In contrast, the Poincaré approach detects all pulses. Neither of the approaches is correct or wrong because no ground truth for diplophonic phonation is available. The correct value of the fundamental frequency and the positions of the cycle marks depend on the definition of fundamental frequency, which is imposed by the extraction algorithms in the shown example. The extraction of double cycle marks for diplophonic voice [59] accounts for the ambiguity of the definition of the fundamental frequency and describes additively combined primary and secondary oscillators. The approach is presented in section 4.2.

[4] The term ”pitch marks” is often used synonymically to ”cycle marks”, but should be avoided because pitch is a perceptual attribute.

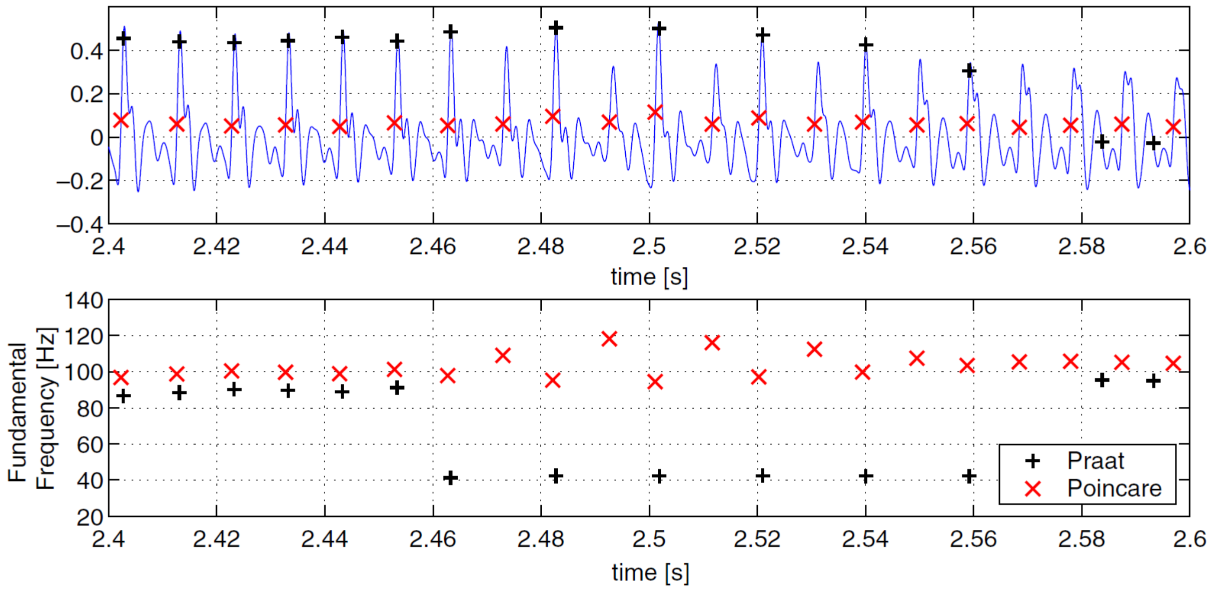


Figure 4.1: A diplophonic speech waveform together with its fundamental frequencies and cycle marks, determined with Praat [85] and from Poincaré sections, taken from [102].

A synthesizer for disordered voice sounds has been published by Schoentgen and co-workers [40, 103]. In an early version, diplophonic waveforms are simulated by periodic modulation of amplitude and/or frequency. The modulator is sinusoidal and its frequency is the signals' fundamental divided by a small integer ratio. Figure 4.2 is taken from [40] and shows an example of a simulated phonatory excitation signal with simultaneous amplitude and frequency modulation. In this example the modulator's frequency is at half the signals' fundamental frequency and results in period-two alternating amplitudes and period lengths. Such a pattern (Titze's diplophonia) can be explained by 2:2 entrainment.

The presence of modulation oscillators with its fundamental at the signal's fundamental divided by a small integer ratio is a special case of diplophonia. Secondary oscillators with nearly irrational frequency ratios are observed in clinical diplophonia more frequently. It is believed that the superior-inferior coupling of oscillators is usually stronger than the coupling in other kinds of asymmetry, which leads to small rational frequency ratios and strictly consonant intervals (1:2, 2:3). Consonant intervals are auditorily less salient than dissonant intervals and thus less likely to be detected perceptually. Therefore consonant diplophonia is much less frequent in the used database than dissonant diplophonia.

Figure 4.3 shows an observed glottal area waveform acquired from clinically diplophonic phonation (corpus ID 88). The waveform is qualitatively evaluated with regard to its time-domain waveform properties and modeled with two different approaches. The first one is Hanquinet's modulative approach [40] and the other one is an additive approach [59]. The model parameters are adjusted manually to obtain waveform models that visually fit the observed waveform. The temporal fine structure of the observed waveform can only be modeled with the additive approach.

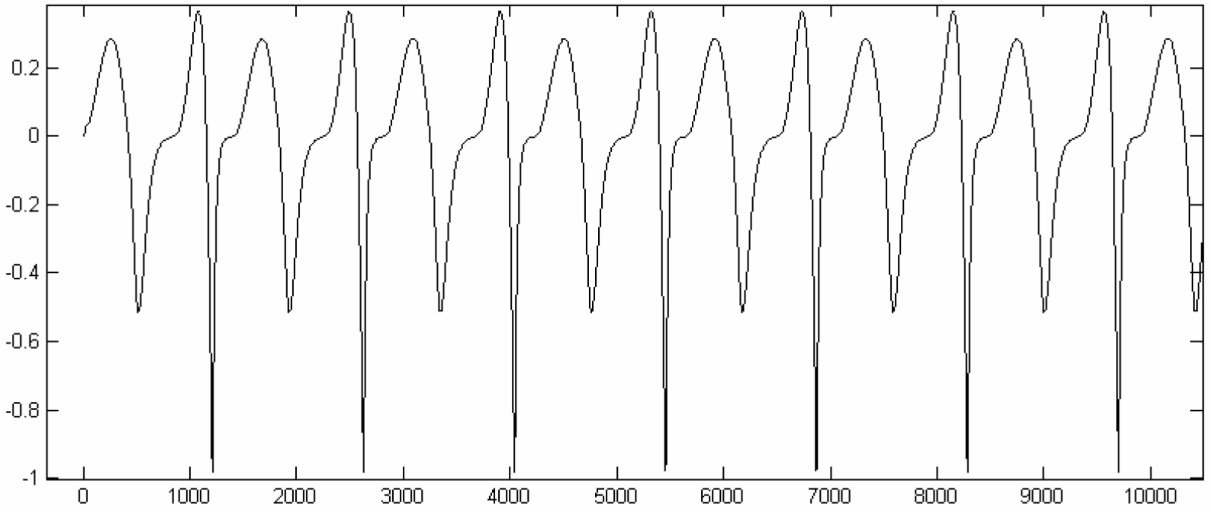


Figure 4.2: Simulation of a diplophonic phonatory excitation signal with sinusoidal amplitude and frequency modulation, taken from [40].

The time-domain properties that can be observed are:

1. The presence of major and minor pulses
2. Fluctuations of the major pulse height
3. Fluctuations of the residual glottal gap
4. Occasional excursions in major pulses

The glottal area waveform shows large fluctuations of the major pulse height. Height dips are marked with downward facing black arrows and its maxima are marked with tilted red arrows. The periodic pattern of pulse heights enables splitting the waveform into metacycles. The metacycle length increases from four major pulses per metacycle in the beginning to nine major pulses per metacycle in the end. The metacycle length is constant in the last three metacycles. A similar fluctuation is observed in the waveform's minima. The residual glottal gap fluctuates at the same frequency as the major pulse height. The upwards facing arrows mark the maximal residual glottal gap in each metacycle. The maximal residual glottal gaps are located $1\frac{1}{2}$ and $2\frac{1}{2}$ cycles after each major pulse height dip.

The upper subplot of figure 4.4 shows the stationary interval of the signal (1.833 - 1.981 s). It contains three metacycles of nine major pulses each, denoted by downward facing black and red arrows. Minor pulses in between the major pulses are marked by upwards red arrows. The occasional excurvations of major pulses are marked by upward black arrows. The waveform model later shows that all major pulses are excurvated in theory, but only those that are visible are marked by arrows here.

The middle subplot shows the model waveform, generated with Hanquinet's modulative approach. Hanquinet's amplitude driving function is sinusoidally modulated with a modulation frequency of $\frac{1}{9}$ times the signals' fundamental. The cycle shape template is a Liljencrants-Fant (LF) model cycle shape [104]. The synthesizer is able to produce a fluctuation of the major pulse height and a fluctuation in the residual glottal gap. In the model waveform the maximal residual glottal gap is always situated $\frac{1}{2}$ cycle after the major pulse height minimum. In the observed waveform the maximal residual glottal gap is situated one or two cycles later. Other differences from the observed waveform are the absence of minor pulses in between the major ones and the absence of the major pulses' excurvations.

The additive waveform model is shown in the lower subplot of figure 4.4. Two independent waveforms with a frequency ratio of 17:9 are additively combined. The dominant oscillator is generated with an LF cycle shape, and the second one with a sinusoidal cycle shape. The synthesis paradigm (polynomial modeling) is explained in section 4.3. One sees that the proposed model is able to copy the fine structure of the observed waveform. All described waveform properties can be modeled, i.e., the height fluctuations of major pulses, the fluctuations of the residual glottal gap, the occasional presence of minor pulses and the occasional presence of pulse excurvations. The noise in the observed waveform is not considered in the model. Other divergences are that the position of the first maximal residual glottal gap is one cycle late and that the exact shapes of excurvations differ from the observed patterns. These divergence stem from the a priori assumption with regard to the cycle shapes, which will dropped in section 4.2. These qualitative observations suggest that oscillators combine additively in the observed diplophonic waveform.

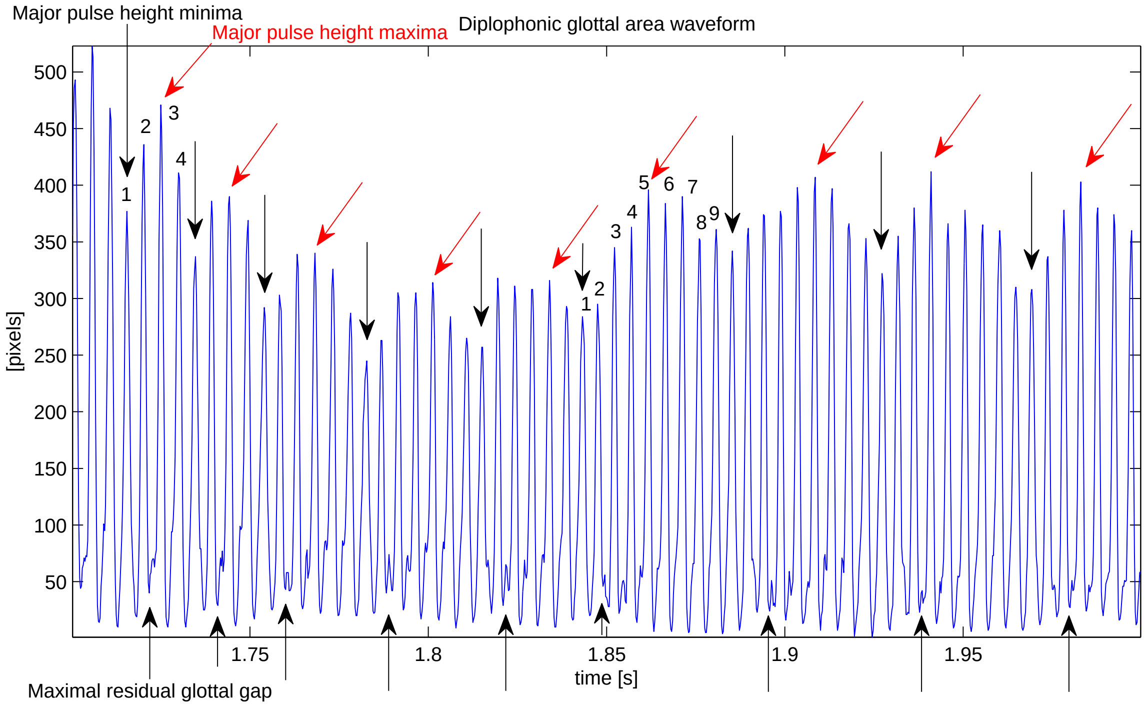


Figure 4.3: Glottal area waveform, acquired from a clinically diplophonic phonation. The arrow annotations indicate the phases of metacycles. Downwards black arrows mark major pulse height dips, upwards black arrows mark maxima of the glottal gap residue and red arrows mark major pulse height maxima.

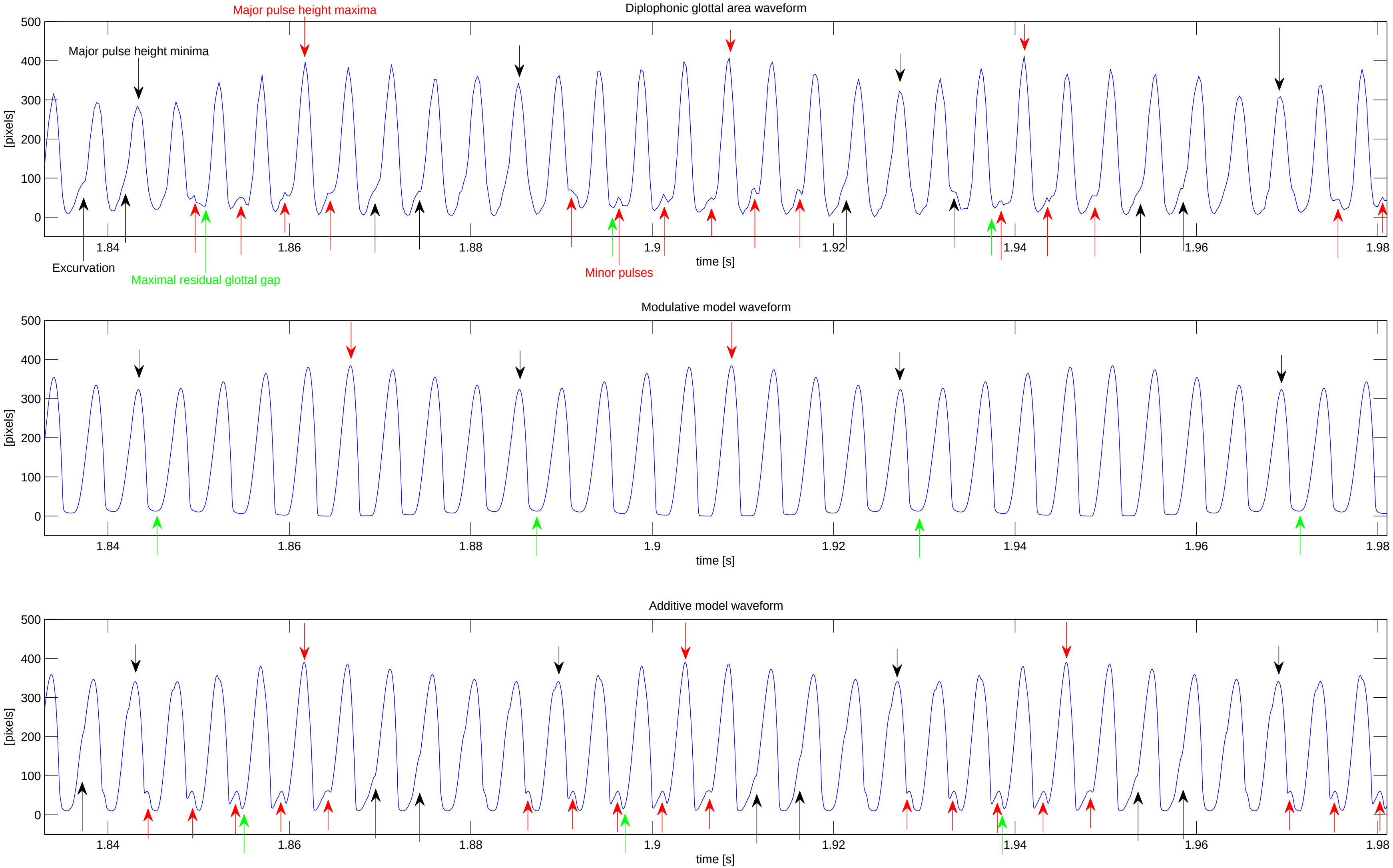


Figure 4.4: Observed glottal area waveform and its waveform models. A modulative and an additive waveform model are fitted to the observed waveform by manual parameter adjustment. Downwards black arrows mark major pulse height dips, downwards red arrows mark major pulse height maxima, green arrows mark maxima of the residual glottal gap, upwards red arrows mark minor pulses and upwards black arrows mark excurvations of the major pulses.

4.2 Unit-pulse FIR filtering for modeling glottal area waveforms

The previous section suggests that an additive waveform model is needed for simulating diplophonic phonation. Manual parameter adjustment enabled producing waveforms that imitate the detailed time-domain properties of a diplophonic glottal area waveform. A model structure that enables automatic estimation of the model parameters is a prerequisite for automated detection. The model is introduced and evaluated on euphonic and diplophonic glottal area waveform intervals (corpus ID 89).

4.2.1 Model description

Figure 4.5 illustrates the unit-pulse FIR filtering model, which is a cycle mark based waveform model [59]. M independent unit-pulse oscillators with period lengths N_m are taken into account. The fundamental frequencies k_m are restricted to the range of [60, 700] Hz and the period lengths in samples are rounded to even numbers. The phases of the unit-pulse oscillators are controlled by the parameters Δl_m. The unit-pulse trains u_m are convolved with filter coefficients r_m, which are the length-N_m pulse shapes of the oscillator waveforms d_m. The maxima of r_m are centered, and so the positions of the unit-pulses u_m coincide with the positions of maximal amplitudes of d_m. The individual oscillator waveforms d_m are added together to give the glottal area waveform d. The noisy glottal area waveform d' is observed in the laryngeal high-speed video. The zero-mean white Gaussian noise η is the spatial quantization noise of glottal area waveform extraction and is uncorrelated with the oscillators. All model parameters are constant within one synthesis block.

Equations 4.1 to 4.8 explain the model, where n is the time index, n' is the relative time index, N is the length of the synthesis block, m is the oscillator index, f_s is the sampling frequency, μ is the pulse index and l_m are the filter coefficient indices.

Frequencies:

$$k_m \in \mathbb{R} \mid 70\,\text{Hz} \leq k_m \leq 600\,\text{Hz}, \quad \text{where } m = 1, 2, ... M \tag{4.1}$$

Pulse lengths:

$$N_m = 2 \cdot \left\lfloor \frac{f_s}{2 \cdot k_m} \right\rceil \tag{4.2}$$

Unit-pulse:

$$\delta(n') = \begin{cases} 1, & ... \; n' = 0 \\ 0, & ... \; n' \neq 0 \end{cases}, \quad n' \in \mathbb{Z} \tag{4.3}$$

Unit-pulse series:

$$u_m(n) = \sum_m \delta(n - \mu \cdot N_m - \Delta l_m), \quad \mu \in \mathbb{Z}, \quad n = 0, 1, 2, ... N - 1 \tag{4.4}$$

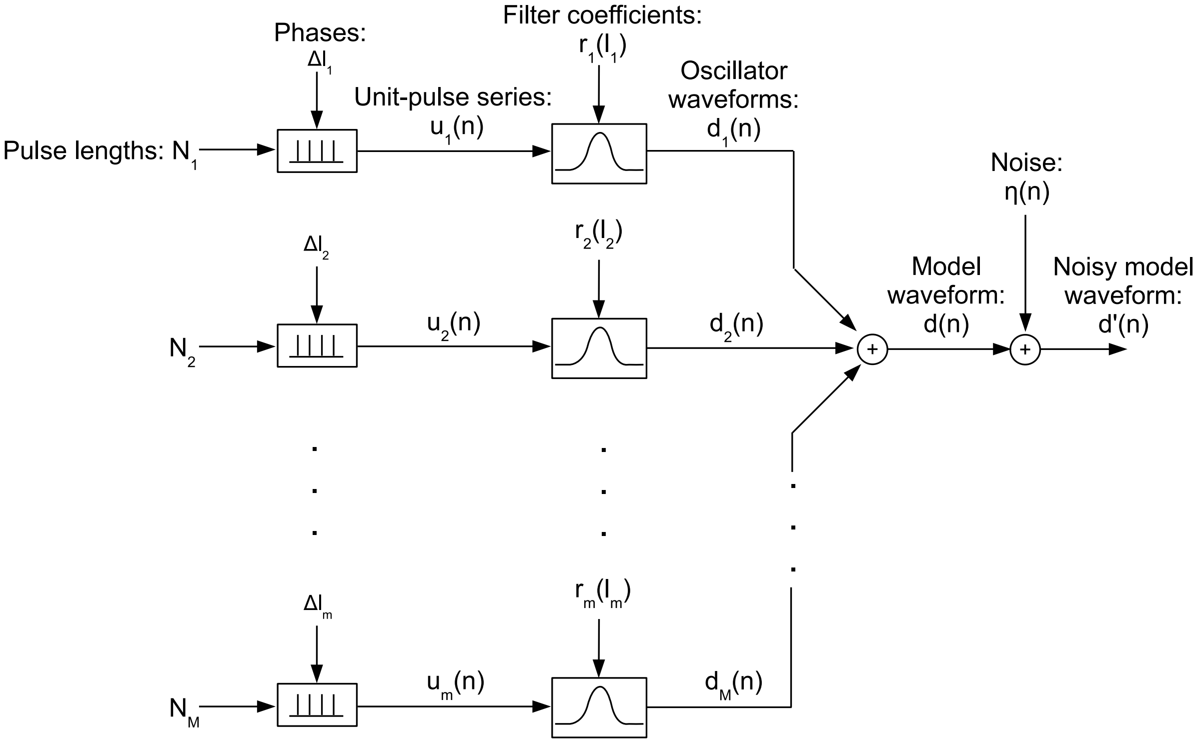


Figure 4.5: Block diagram of the model structure.

Filter coefficients:

$$r_m(l_m) \in \mathbb{R}, \quad \text{where } l_m = -\frac{N_m}{2}+1, -\frac{N_m}{2}+2, ... -1, 0, +1, ... \frac{N_m}{2}-2, \frac{N_m}{2}-1$$
$$\text{and } \underset{l_m}{\operatorname{argmax}} \{r_m(l_m)\} = 0 \tag{4.5}$$

Oscillator waveforms:

$$d_m(n) = \sum_{l_m} u_m(n) \cdot r_m(n - l_m) \tag{4.6}$$

Model waveform:

$$d(n) = \sum_{m=1}^{M} d_m(n) \tag{4.7}$$

Noisy model waveform:

$$d'(n) = d(n) + \eta(n) \tag{4.8}$$

4.2.2 Parameter estimation and resynthesis

The noisy glottal area waveform d' is obtained from the laryngeal high-speed video and resampled at 50 kHz. The parameter estimation and resynthesis aims at estimating the model parameters Δl_m and r_m, while M, k_m and N_m are assumed to be known from spectral video analysis (section 3.2).

The parameter estimation and resynthesis is structured as follows:

1. Extract the glottal area waveform
2. Resample at 50 kHz
3. Generate the phase-uncompensated unit-pulse trains
4. Extract the phase-uncompensated FIR-filter coefficients
5. Determine the phase shift
6. Generate the phase-compensated unit-pulse trains
7. Extract the phase-compensated FIR-filter coefficients
8. Go to step 5 and repeat if the phase shift is not 0.
9. Create the individual oscillator waveforms
10. Add the individual oscillator waveforms together
11. Subtract the waveform model from the observed waveform

Figure 4.6 shows the block diagram of the parameter estimation, where the circumflexes denote estimators. The position of the unit-pulses with respect to the observed waveform is used for visually evaluating the procedure in section 4.2.3, therefore the unit-pulses must be phase-compensated. Unit-pulse oscillators are driven with period lengths $\hat{N}_m$ and generate the phase-uncompensated unit-pulse trains $\hat{u}_m$ (equation 4.10). The phase-uncompensated pulse shapes $\hat{r}'_m$ are obtained by cross correlating $\hat{u}_m$ with the noise corrupted desired signal d' (equation 4.11). The phase shift $\hat{\Delta}\hat{l}_m$ is the lag at which $\hat{r}'_m$ is maximal. The phase-compensated unit-pulse trains $\hat{u}_m$ are obtained by shifting $\hat{u}'_m$ by $\hat{\Delta}\hat{l}_m$. The phase compensation is iteratively repeated until the maxima of $\hat{r}_m$ are at $\hat{l}_m = 0$. The individual oscillators $\hat{d}_m$ are obtained by convolving the unit-pulse trains $\hat{u}_m$ with the FIR-filter coefficients $\hat{r}_m$. All individual oscillator waveforms are added together to give the glottal area waveform model $\hat{d}$. The model is subtracted from its observed counterpart, yielding the time domain error signal e.

$$\hat{N}_m = 2 \cdot \left\lfloor \frac{f_s}{2 \cdot \hat{k}_m} \right\rceil, \quad m = 1, 2, ... \hat{M} \tag{4.9}$$

$$\hat{u}'_m(n) = \sum_m \delta(n - \mu \cdot \hat{N}_m), \quad \mu \in \mathbb{Z}, \quad n = 0, 1, 2, ... N - 1 \tag{4.10}$$

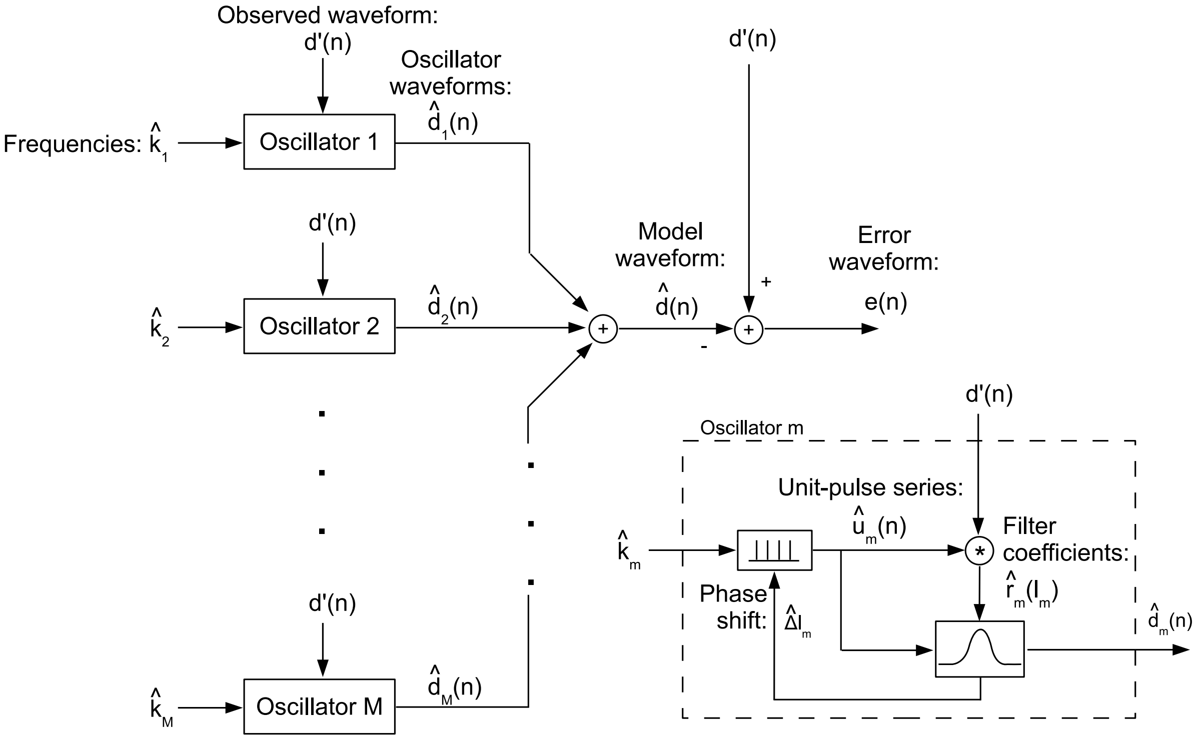


Figure 4.6: Block diagram of the parameter estimation.

$$\hat{r}'_m(\hat{l}_m) = \frac{1}{\sum_n \hat{u}'_m(n)} \cdot \sum_n \hat{u}'_m(n) \cdot d'(n+\hat{l}_m), \quad \text{where}$$

$$\hat{l}_m = 1-\frac{\hat{N}_m}{2}, 2-\frac{\hat{N}_m}{2}, \ldots -1, 0, +1, \ldots \frac{\hat{N}_m}{2}-2, \frac{\hat{N}_m}{2}-1 \tag{4.11}$$

$$\hat{u}_m(n) = \sum_m \delta(n-\mu\cdot\hat{N}_m-\hat{\Delta}\hat{l}_m), \quad \mu\in\mathbb{Z}, \quad n=0,1,2,\ldots N-1 \tag{4.12}$$

$$\hat{\Delta}\hat{l}_m = \underset{\hat{l}_m}{\operatorname{argmax}}\left\{\hat{r}'_m(\hat{l}_m)\right\} \tag{4.13}$$

$$\hat{r}_m(\hat{l}_m) = \frac{1}{\sum_n \hat{u}_m(n)} \cdot \sum_n \hat{u}_m(n)\cdot d'(n+\hat{l}_m) \tag{4.14}$$

$$\hat{d}_m(n) = \sum_{\hat{l}_m} \hat{u}_m(n)\cdot\hat{r}_m(n-\hat{l}_m) \tag{4.15}$$

$$\hat{d}(n) = \sum_{m=1}^{\hat{M}} \hat{d}_m(n) \tag{4.16}$$

$$e(n) = d(n) - \hat{d}(n) \tag{4.17}$$

4.2.3 Evaluation and validation

The unit-pulse FIR-filtering model is illustrated on euphonic and diplophonic intervals, 506 frames or 126.5 ms long each (corpus ID 89). The audio signals stem from the camera's inbuilt microphone. The model is evaluated by both visual and quantitative comparison of the model waveform with its observed counterpart and is validated by visual comparison of the estimated cycle marks with the empiric glottal area and audio waveforms.

Figure 4.7 shows the results of the spectral video analysis. The fundamental frequency of the euphonic interval is 221.8 Hz and the fundamentals of the diplophonic interval are 213.9 Hz and 404 Hz. Figure 4.8 shows the estimated pulse shapes of the euphonic and diplophonic phonation intervals. The pulse shape estimated from the euphonic interval looks typical. A short constant/closed phase at the pulse edges alternates with a evolving open phase in the center of the pulse. Pulse shape r_1 of the diplophonic waveform has a longer constant/closed phase at its edges. Pulse shape r_2 looks like a sinusoid, it has no constant/closed phase and it is shorter in time and smaller in amplitude.

Visual inspection of the diplophonic video reveals anterior-posterior asymmetry. The anterior part of the left vocal fold is separated from its posterior companion by a polyp. The posterior part of the vocal folds vibrates and closes normally (oscillator 1), which is reflected in r_1. The posterior part of the left vocal fold (oscillator 2) vibrates faster than the anterior part, which shortens r_2 compared to r_1. No prolonged posterior closed phase exists, because the polyp prohibits the collision of the vocal folds anteriorly. A full closure is achieved seldom and shortly only, in times when amplitude minima of oscillator 1 coincide with amplitude minima of oscillator 2. Oscillator 2 vibrates sinusoidally, which is reflected in r_2.

Figures 4.9 and 4.10 show the waveform modeling results. The upper subplots show the observed waveform and its models. The middle subplots show the normalized unit-pulse trains and the filter outputs. To clarify the spatial origins of the oscillators, the colors of the lines and stems in the middle subplots are identical to the colors of the corresponding regions of the spectral video analysis (figure 4.7). The lower subplots show the time-domain modeling error e. One oscillator models the euphonic waveform (figure 4.9) and two oscillators model the diplophonic waveform (figure 4.10).

The upper subplot of figure 4.9 illustrates the high accuracy of the euphonic model, because it is similar to the observed signal. The waveform model is quantitatively evaluated via the relative root mean square error (relRMSE, [dB]), which relates the energy level of the error waveform to the energy level of the observed waveform. It would become $-\infty$ if the error was zero, which is practically never achieved because of noise. If the model waveform would be zero, the error would equal the empiric waveform and relRMSE would equal 0 dB. It is desirable to have a small modeling error, which is reflected in negative numbers. The relRMSE is approximately

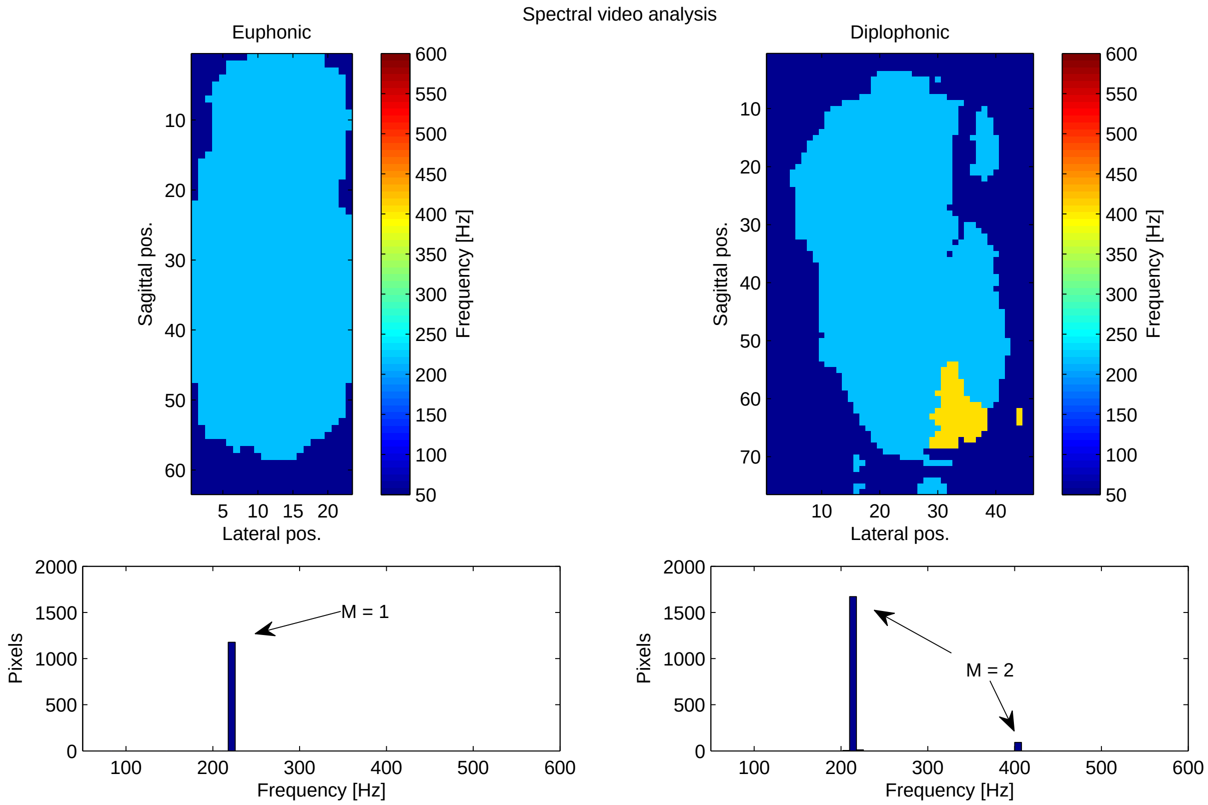
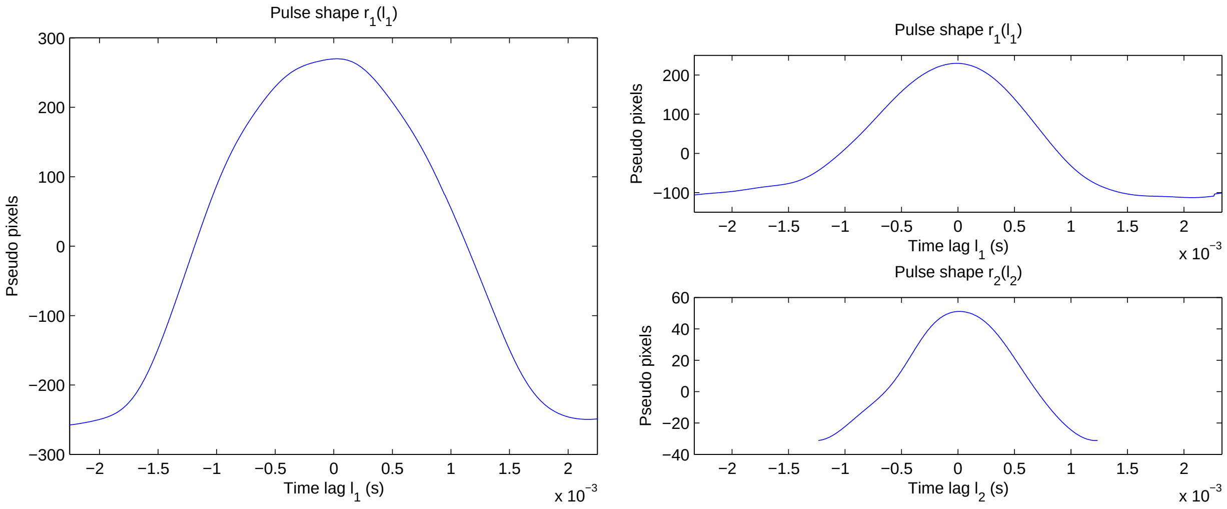


Figure 4.7: Estimation of the fundamental frequencies from spectral video analysis.



(a) Euphonic.

(b) Diplophonic.

Figure 4.8: Measured pulse shapes from one euphonic and one diplophonic phonation interval.

-12.3 dB in the shown example.

$$\text{relRMSE} = 20 \cdot \log \left(\frac{\sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} e(n)^2}}{\sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} d'(n)^2}} \right) \tag{4.18}$$

The modeling error of the euphonic waveform is larger at the onset and offset of the depicted interval than at its center. At onset the pulses of the model are early, i.e., they are shifted to the left with respect to the observed pulses. In the middle the model's pulses coincide with the empiric pulses, resulting in a small modeling error. Near offset the model's pulses are late, i.e., the are shifted to the right with respect to the observed pulses. These divergences stem either from fluctuations of fundamental frequency within the analysis block or from a coarse frequency resolution in the fundamental frequency estimation. Divergences arising from the first problem are addressed by using shorter analysis blocks and divergences arising from the latter problem are addressed by introducing zero-padding in spectral analysis. Both strategies are described in section 4.3.

The upper subplot of figure 4.10 shows that the model of the diplophonic waveform is accurate. The model is capable of imitating all properties of the observed signal that are discussed in section 4.1. Although the waveform model is more complex for diplophonia than for euphonia, the modeling error relRMSE is -10.40 dB exceeding that of the euphonia model. If only one oscillator would have been used in the diplophonic waveform model, the error would have been larger (-8.81 dB). The numbers indicate that the accuracy of the diplophonic waveform model increases when a second oscillator is added, whereas one oscillator is sufficient for modeling euphonic phonation.

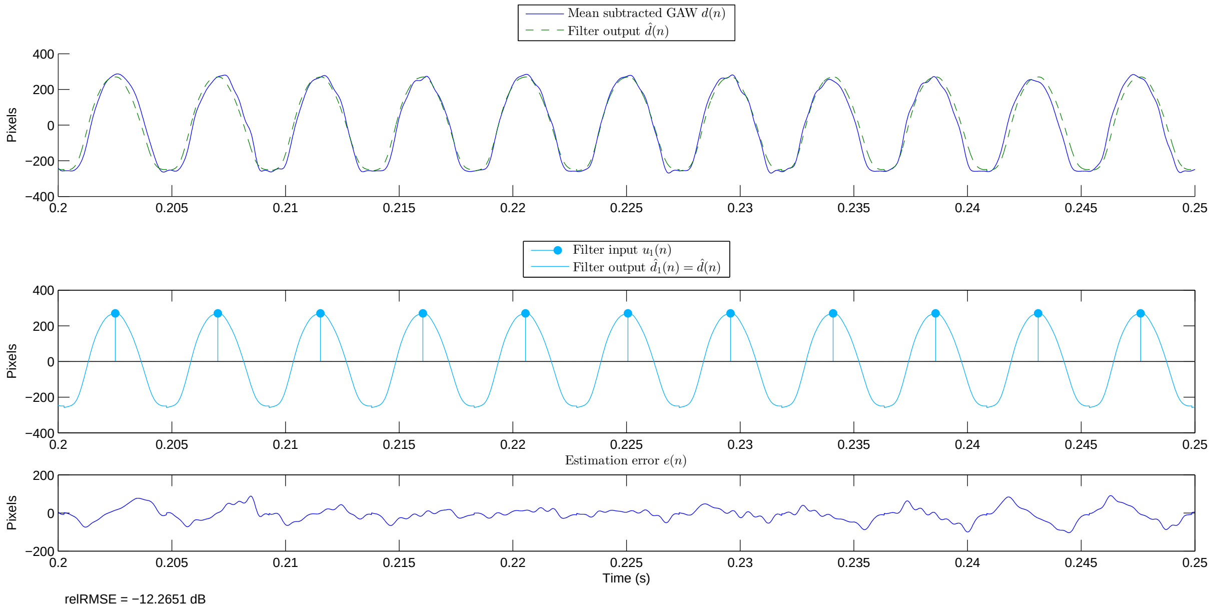


Figure 4.9: Waveform modeling summary of the euphonic interval (zoomed in).

Evaluation via visual comparison of cycle marks and waveforms

The model is evaluated by comparing the phase-compensated unit-pulse trains $\hat{u}_m$ to the observed glottal area and audio waveforms. The temporal positions of the unit-pulses denote times of maximal oscillator amplitude and enable explaining the height fluctuations of major pulses, the occasional presence of minor pulses in-between the major ones and the occasional excurvation of the major pulses. The term "double pitch marks" is introduced for the phase-compensated unit-pulse trains of two-oscillator waveform models [59].

Figure 4.11 shows the glottal area and audio waveforms, together with their estimated cycle

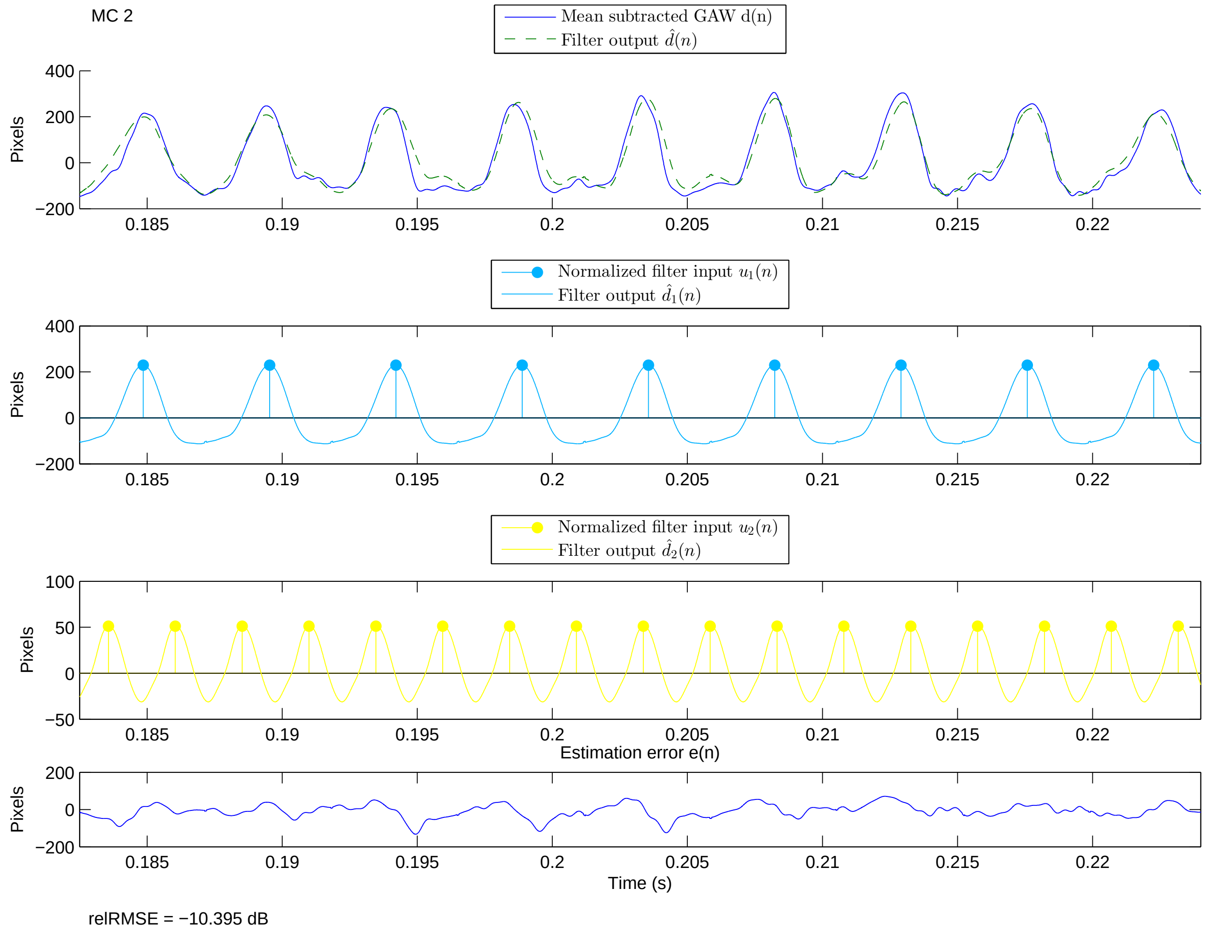


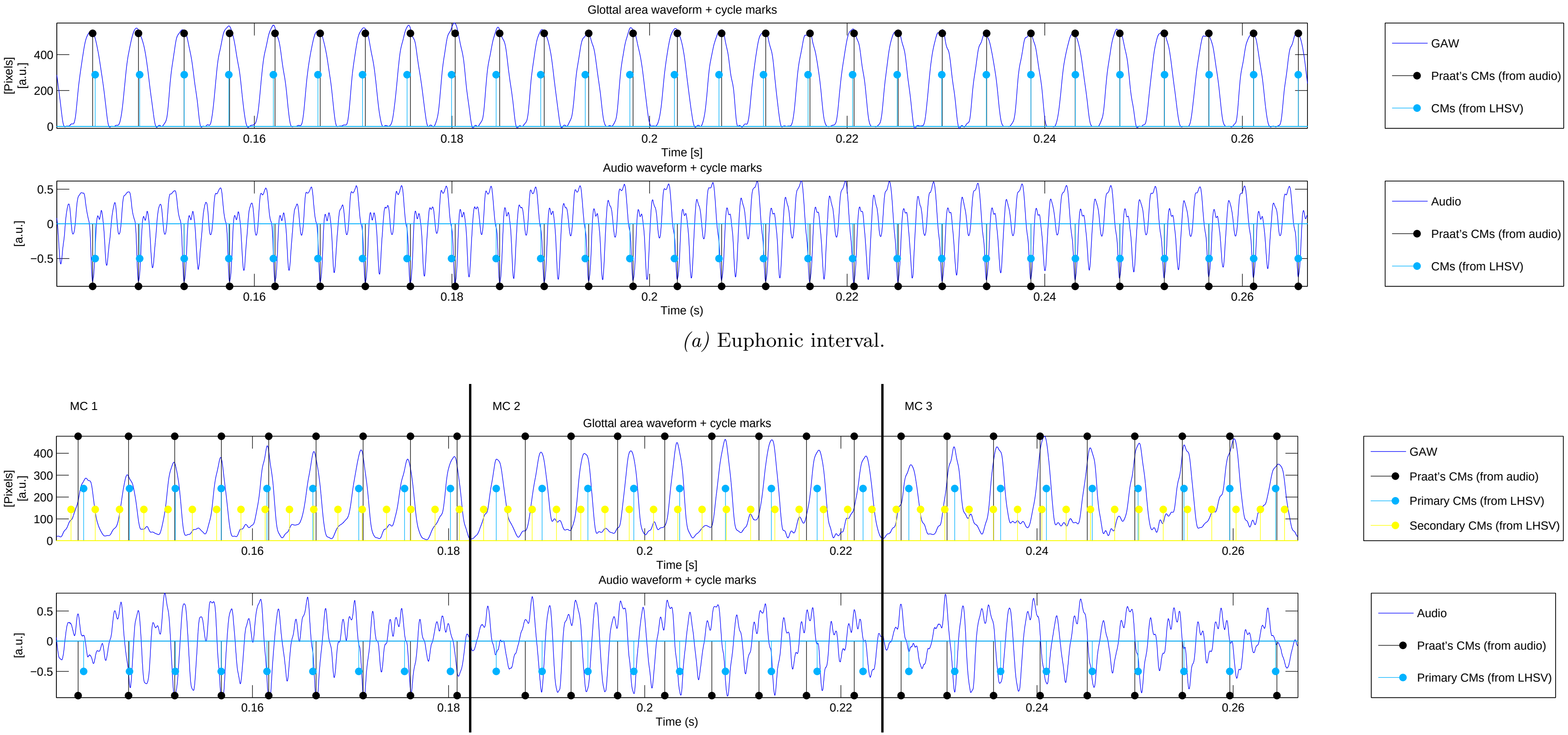
Figure 4.10: Waveform modeling summary of meta cycle 2 of the diplophonic interval.

marks. Praat's [105] cycle marks are extracted from the audio signal for comparison. Figure 4.11a shows that Praat's and the proposed cycle marks are valid for euphonic voice because all cycle marks approximately coincide with positive peaks in the glottal area waveform and with cycle synchronous negative peaks in the audio signal.

Figure 4.11b shows the diplophonic glottal area and audio waveform together with its cycle marks. Praat loses synchronization to the waveforms in the mid-metacycle. In contrast, the primary cycle marks retrieved by unit-pulse FIR-filtering succeed in recognizing the positions of the waveform's major pulses. In the glottal area waveform the stems denote time instances of the oscillators' maximal positive amplitudes.

At the first major pulses within each metacycle, the oscillators are exactly out of phase. Destructive interference leads to minimal height of the major pulses. In the middle of each metacycle, the oscillators are in phase and the height of the major pulses is increased by constructive interference. Yellow stems coincide with minor pulses in the waveform. Yellow stems that are close to blue stems cause excurvation because the secondary oscillator pulses disturb the flanks of the primary oscillator pulses. Yellow stems that are very close to blue stems result in a pulse height increase because the secondary oscillator pulses are absorbed by the primary oscillator pulses.

The metacycles in the audio waveform are reported by a fluctuation of the signal peak amplitudes, similar to that observed in the glottal area waveform. The primary cycle marks retrieved with unit-pulse FIR-filtering coincide with cycle synchronous negative peaks in the audio signal. The negative peaks at the metacycle boundaries are blurred by the vocal tract resonances and thus cannot be seen. Praat's cycle marks agree with the primary cycle marks in the first and last metacycles, but fail to agree in the mid-metacycle. The secondary cycle marks are observed in the audio signal only via the overall amplitude decrease at the metacycle boundaries. This suggests that analyzing laryngeal high-speed videos gains qualitative insight into the temporal fine structure of vocal fold vibration in the presence of two or more distinct oscillator frequencies, as opposed to analyzing audio recordings.



(a) Euphonic interval.

(b) Diplophonic interval.

Figure 4.11: Glottal area and audio waveforms together with their estimated cycle marks (CMs). The colors of the cycle marks are taken from their corresponding vocal fold region in spectral video analysis (figure 4.7). Metacycles (MCs).

4.3 Polynomial modeling of audio waveforms

In the previous section unit-pulse FIR filtering for building two-oscillator waveform models of diplophonic phonation is introduced. It has been assumed that the model parameters are constant within each analysis block. A violation of this assumption can lead to discontinuous cycle boundaries in resynthesized oscillators, which are broadband disturbances and undesired in synthesis for auditive evaluation. To relax this assumption and limit the bandwidth of the candidate oscillators, the FIR filters are replaced by polynomial oscillator models. The polynomial model has been proposed by Schoentgen and co-workers [40], and a simplified version is used here. It is further assumed in the last section that the correct number of oscillators and their fundamental frequencies are known a priori. This assumption is relaxed by introducing a heuristic oscillator selector for the joint estimation of the number of sources and the waveform model parameters.

The model is evaluated with respect to voice quality group effects of quantitative waveform modeling errors. A database of 55 diplophonic and 223 non-diplophonic audio waveform intervals (corpus ID 8) and a critical corpus of 55 diplophonic and 70 dysphonic intervals (corpus ID 98) are used. Different versions of diplophonic model waveforms are visually compared to their observed counterpart. Their individual oscillators and the error waveform are auditively evaluated and analyzed by means of a mutual masked loudness model for multiple sources [61]. Moreover, fundamental frequencies extracted with three different methods are evaluated on a corpus of 65 euphonic, 28 diplophonic and 84 dysphonic intervals (corpus ID 99).

4.3.1 Model description

A waveform model for diplophonic and euphonic audio waveforms is proposed (equations 4.19). d_m are the oscillator waveforms, m is the oscillator index, n is the discrete time index, a and b the real and imaginary parts of the oscillators' Fourier coefficients, P is the number of partials, p the partial index, ω_m is the angular frequency and k_m is the fundamental frequency, which is in the range of $[70, 600]$ Hz. f_s is the sampling frequency, d is the waveform model combining M oscillators (1 for euphonia and 2 for diplophonia) and η is random noise, which consists of aspiration noise, background noise and measurement noise. The spectral magnitude at the fundamental is greater or equal -15 dB with regard to the overall maximum of the waveform's spectrum. All model parameters are constant within one synthesis block, the maximal partial frequency is below the Nyquist frequency, the oscillators are independent and η is uncorrelated with the oscillators.

$$\begin{aligned}
d_m(n) =& \frac{1}{2} \cdot a_{m,0} + \sum_{p=1}^{P} \left[a_{m,p} \cdot \cos\left(\omega_m \cdot p \cdot n\right) + b_{m,p} \cdot \sin\left(\omega_m \cdot p \cdot n\right) \right], \quad \text{where} \\
& \omega_m = \frac{2\,\pi\,k_m}{f_s}, \quad 2 \cdot \pi \cdot 70\,\text{Hz} \leq \omega_m \leq 2 \cdot \pi \cdot 600\,\text{Hz}\ \forall\, m, \quad \text{and} \\
& 20 \cdot \log\Big(\sqrt{a_{m,1}^2 + b_{m,1}^2}\Big) \geq \\
& \max_p \left\{ 20 \cdot \log\Big(\sqrt{a_{m,p}^2 + b_{m,p}^2}\Big),\, 20 \cdot \log(a_{m,0}) \right\} - 15\,\text{dB} \quad \forall\, m, \quad p = 2, 3, \ldots P
\end{aligned} \tag{4.19}$$

$$d(n) = \sum_{m=1}^{M} d_m(n) \tag{4.20}$$

$$d'(n) = d(n) + \eta(n) \tag{4.21}$$

4.3.2 Parameter estimation and resynthesis

The parameter estimation and resynthesis aims at estimating the model parameters M, ω, a and b from the observed noisy audio waveform d'.

The parameter estimation and resynthesis is structured as follows:

1. Resample the audio waveform at 50 kHz
2. Spectral peak pick fundamental frequency candidates
3. Generate unit-pulse trains
4. Extract the pulse shapes via cross correlation
5. Approximate the pulse shapes with finite Fourier series
6. Transform polynomial coefficients
7. Resynthesize individual oscillators
8. Build all possible one- and two-oscillator models
9. Select the optimal one- and two-oscillator models

Figure 4.13 shows an example of a normalized magnitude spectrum of a diplophonic audio signal and its estimated fundamental frequency candidates. The fundamental frequency candidates $\hat{k}_\gamma$ are obtained by peak picking in the waveform's magnitude spectrum (figure 4.13). To save computation time during oscillator selection, the maximal number of oscillator candidates is limited to 12. If the number of candidates obtained via peak picking exceeds 12, the 12 with the lowest frequencies are retained. The block length N is 60 ms with 50 % overlap, the window is rectangular and the signal is zero padded to obtain a frequency step of 0.1 Hz (i.e., $5 \cdot 10^5$-point DFT). The fundamental frequency candidates are restricted to the range [70, 600 Hz] and only pulses that exceed -15 dB with regard to the global maximum are taken into account.

The pulse shapes r_γ of the candidate oscillators $\hat{d}_\gamma$ are estimated by cross correlating unit-pulse trains u_γ with the observed audio waveform d' (cf. section 4.2.2). Equations 4.22 to 4.25 explain the building of the candidate pulse shapes, where n' is the relative time index, μ is the pulse index, δ is a unit-pulse, N_γ is the period length of the γ^{th} candidate oscillator and l_γ is the cross correlation time lag. The length of r_γ is kept equal to N_γ.

$$\delta(n') = \begin{cases} 1, & \dots\ n' = 0 \\ 0, & \dots\ n' \neq 0 \end{cases}, \quad n' \in \mathbb{Z} \tag{4.22}$$

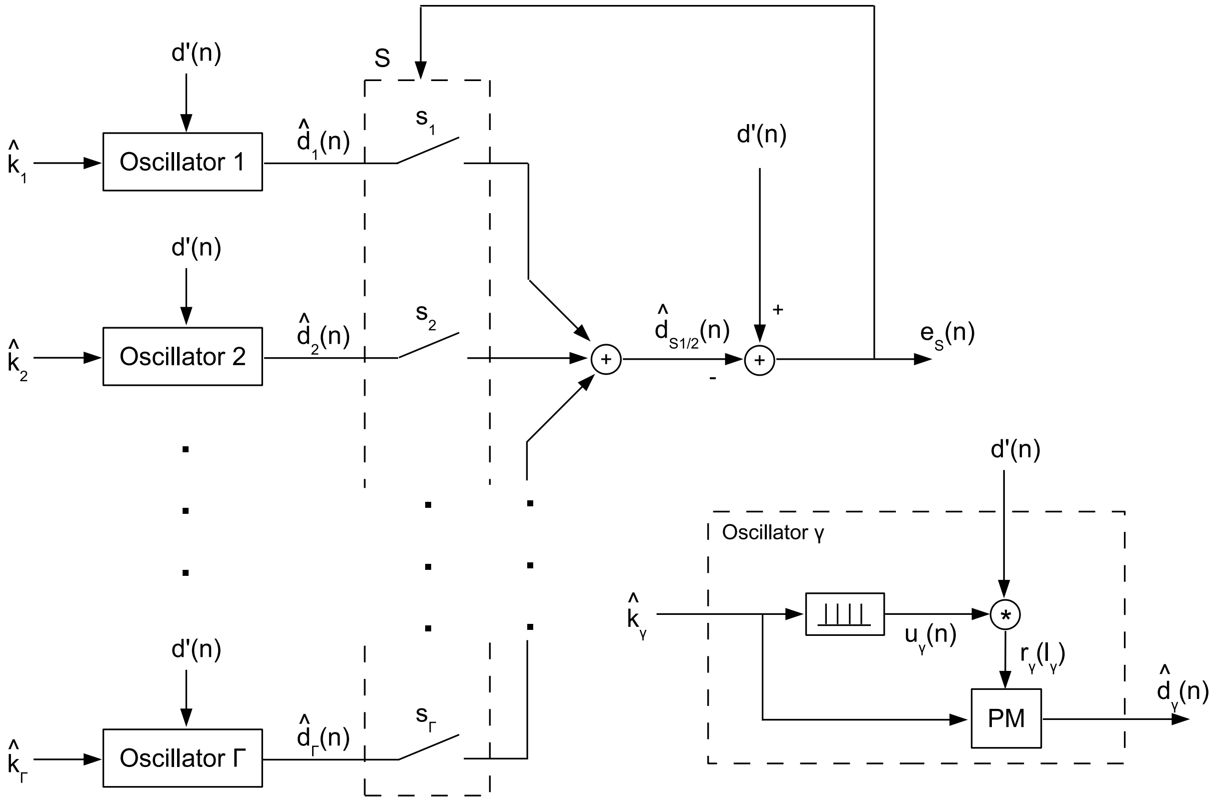


Figure 4.12: Block diagram of the parameter estimation.

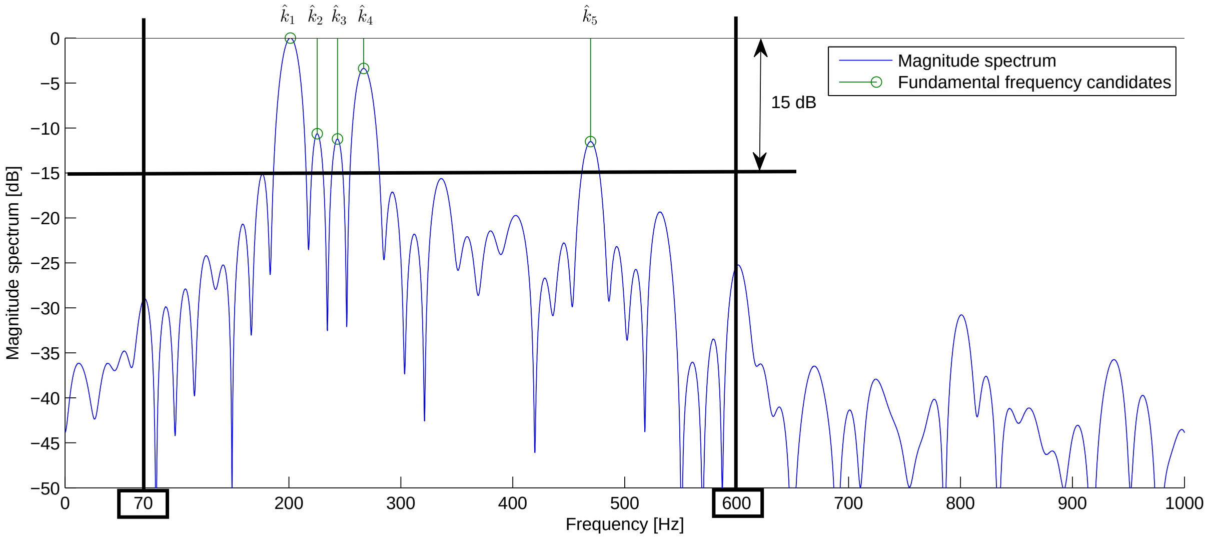


Figure 4.13: An example of a normalized magnitude spectrum of a diplophonic audio signal block together with its estimated fundamental frequency candidates.

$$N_\gamma = 2 \cdot \left\lfloor \frac{f_s}{2 \cdot \hat{k}_\gamma} \right\rceil, \quad \gamma = 1, 2, \dots \Gamma \tag{4.23}$$

$$u_\gamma(n) = \sum_\gamma \delta(n - \mu \cdot N_\gamma), \quad \mu \in \mathbb{Z}, \quad n = 0, 1, 2, ... \, N - 1 \tag{4.24}$$

$$r_\gamma(l_\gamma) = \frac{1}{\sum_n u_\gamma(n)} \cdot \sum_n u_\gamma(n) \cdot d'(n + l_\gamma)$$
$$\text{where } l_\gamma = 1 - \frac{N_\gamma}{2}, 2 - \frac{N_\gamma}{2}, ... - 1, 0, +1, ... \frac{N_\gamma}{2} - 2, \frac{N_\gamma}{2} - 1 \tag{4.25}$$

The pulse shapes r_γ are approximated by a finite Fourier series (equation 4.27) [106]. $r_\gamma(l_\gamma)$ is a real valued signal and thus has a symmetric spectrum. The number of unique Fourier series coefficients $a_{\gamma,p}$ and $b_{\gamma,p}$ equals the period length N_m and the length of $r_\gamma(l_\gamma)$. The first and last terms in equation 4.27 represent the mean of $r_\gamma(l_\gamma)$, where $\hat{\omega}_\gamma$ is the angular frequency. Half of its energy is located at zero frequency and the other half at the Nyquist frequency. The bracketed term represents the Fourier oscillators.

$$\hat{\omega}_\gamma = 2 \cdot \pi \cdot \hat{k}_\gamma \tag{4.26}$$

$$\begin{aligned} r_\gamma(l_\gamma) = \frac{1}{2}\hat{a}_{\gamma,0} + \sum_{p=1}^{\frac{N_\gamma}{2}-1} \Big[\hat{a}_{\gamma,p} \cdot \cos(\hat{\omega}_\gamma \cdot p \cdot l_m) + \hat{b}_{\gamma,p} \cdot \sin(\hat{\omega}_\gamma \cdot p \cdot l_m)\Big] + \\ \frac{1}{2}\hat{a}_{\gamma,\frac{N_\gamma}{2}} \cdot \cos(\hat{\omega}_\gamma \cdot p \cdot l_m) \end{aligned} \tag{4.27}$$

The Fourier coefficients $\hat{a}_{\gamma,p}$ and $\hat{b}_{\gamma,p}$ are obtained by the DFT of $r_\gamma(l_\gamma)$ (equations 4.28 and 4.29) [106]. The Fourier coefficients are truncated to $P = 10$ partials, which allows for a reasonable model quality, while limiting the bandwidth and saving computation time.

$$\hat{a}_{\gamma,p} = \frac{2}{N_\gamma} \cdot \sum_{l_\gamma = 1 - \frac{N_\gamma}{2}}^{N_\gamma - 1} r_\gamma(l_\gamma) \cdot \cos(\hat{\omega}_\gamma \cdot p \cdot l_m), \quad \text{where } p = 0, 1, \, ... \, P \tag{4.28}$$

$$\hat{b}_{\gamma,p} = \frac{2}{N_\gamma} \cdot \sum_{l_\gamma = 1 - \frac{N_\gamma}{2}}^{N_\gamma - 1} r_\gamma(l_\gamma) \cdot \sin(\hat{\omega}_\gamma \cdot p \cdot l_m), \quad \text{where } p = 1, 2, \, ... \, P \tag{4.29}$$

The polynomial coefficients $a'_{\gamma,p}$ and $b'_{\gamma,p}$ of the candidate pulse shape are obtained by equations 4.30 and 4.31, where × is the multiplication of a matrix and a vector, and $\odot$ is the element wise multiplication of two vectors. The matrices M_e and M_o are derived from the Pascal triangle, where M_e is $P{+}1{\times}P{+}1$ and M_o is $P{\times}P$. The candidate oscillators $\hat{d}_\gamma$ are obtained by resynthesis equation 4.34. The derivation and the prove of the described transformation is found in [107].

$$a'_{\gamma,p} = \frac{1}{2} \cdot M_e^{-1} \times (a_{\gamma,p} \odot v_p), \quad \text{where } p = 0, 1, \ldots P, \text{ and } v_p = 2^p \tag{4.30}$$

$$b'_{\gamma,p} = M_o^{-1} \times \left(b_{\gamma,p} \odot v'_p\right), \quad \text{where } p = 1, 2, \ldots P, \text{ and } v'_p = 2^{(1-p)} \tag{4.31}$$

$$M_e = \begin{bmatrix} 1 & 0 & 2 & 0 & 6 & 0 & 20 & 0 & 70 & \ldots \\ 0 & 1 & 0 & 3 & 0 & 10 & 0 & 35 & 0 & \ldots \\ 0 & 0 & 1 & 0 & 4 & 0 & 15 & 0 & 56 & \ldots \\ 0 & 0 & 0 & 1 & 0 & 5 & 0 & 21 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 1 & 0 & 6 & 0 & 28 & \ldots \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 7 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 8 & \ldots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{4.32}$$

$$M_o = \begin{bmatrix} 1 & 0 & 1 & 0 & 2 & 0 & 5 & 0 & 14 & \ldots \\ 0 & 1 & 0 & 2 & 0 & 5 & 0 & 14 & 0 & \ldots \\ 0 & 0 & 1 & 0 & 3 & 0 & 9 & 0 & 28 & \ldots \\ 0 & 0 & 0 & 1 & 0 & 4 & 0 & 14 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 1 & 0 & 5 & 0 & 25 & \ldots \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 6 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 7 & \ldots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \ldots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{4.33}$$

$$\hat{d}_\gamma(n) = \sum_{p=0}^{P} a'_{\gamma,p} \cdot \cos^p(\hat{\omega}_\gamma \cdot n) + \sin(\hat{\omega}_\gamma \cdot n) \cdot \sum_{p=1}^{P} b'_{\gamma,p} \cdot \cos^p(\hat{\omega}_\gamma \cdot n) \tag{4.34}$$

The heuristic oscillator selector searches for the optimal one-oscillator and two-oscillator waveform model. Waveform models $\hat{d}_S$ and their errors e_S are obtained for all possible one- and two-oscillator additive waveform combinations S. The minimal error one- and two-oscillator models are selected and their errors are given in equations 4.36 and 4.37, where S_1 and S_2 are the minimal error oscillator combinations.

$$e_S(n) = d'(n) - \hat{d}_S(n) \tag{4.35}$$

$$\mathrm{relRMSE}(S_1) = 20 \cdot \log \left(\frac{\sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} e_{S_1}(n)^2}}{\sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} d'(n)^2}} \right) \tag{4.36}$$

$$\mathrm{relRMSE}(S_2) = 20 \cdot \log \left(\frac{\sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} e_{S_2}(n)^2}}{\sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} d'(n)^2}} \right) \tag{4.37}$$

4.3.3 Evaluation

The model is evaluated with regard to waveform similarity and error measures. Additionally, auditory observations and results from loudness modeling and fundamental frequency extraction are described.

Waveform similarity and error measures

Figure 4.14 shows an example of a diplophonic audio waveform snippet. The upper subplot shows the recorded audio signal d', the middle subplot shows the optimal one-oscillator model waveform $\hat{d}_{S_1}$ and the lower subplot shows the optimal two-oscillator model waveform $\hat{d}_{S_2}$. The model waveforms are evaluated qualitatively by visual comparison of the modeled waveforms with the recorded waveform. One sees that the waveform of the two-oscillator model is a better copy of the recorded waveform than the one-oscillator model.

The model waveforms are evaluated quantitatively by obtaining the intervals' average relRMSE. The error of the optimal one-oscillator model waveform relRMSE(S_1) is -4 dB and the error of the optimal two-oscillators model waveform relRMSE(S_2) is -8.8 dB, which confirms the visual evaluation.

These results suggest that it is possible to detect the presence of diplophonia from a recorded audio signal, which is evaluated on a database of 278 phonation intervals (55 diplophonic, 223 non-diplophonic). Figure 4.15 shows the boxplots of the one-oscillator model error relRMSE(S_1) and the two-oscillator model error relRMSE(S_2), with regard to voice quality. The numbers are interval averages, i.e., the average block results of separate voice qualities. relRMSE(S_1) is higher in the diplophonic group than in the other groups, which suggests that one oscillator is insufficient for modeling diplophonic audio waveforms. Although the model error relRMSE(S_2) is comparable in all groups, it is exploited for building a classifier in chapter 5, because the model waveforms of the diplophonic group benefit from adding a second oscillator on average, while model errors of the non-diplophonic intervals tend to increase when a second oscillator is added.

Auditory observations

The extracted oscillators and the model error waveform are evaluated auditorily. Figure 4.16 shows the oscillator waveforms of an example of a diplophonic phonation and its model error waveform. The sound files `d1.wav`, `d2.wav`, `e.wav`, `d1+d2.wav`, `d1+e.wav`, `d2+e.wav`, `d1+d2+e.wav` can be downloaded at [108]. The blocked waveforms have been extracted from the

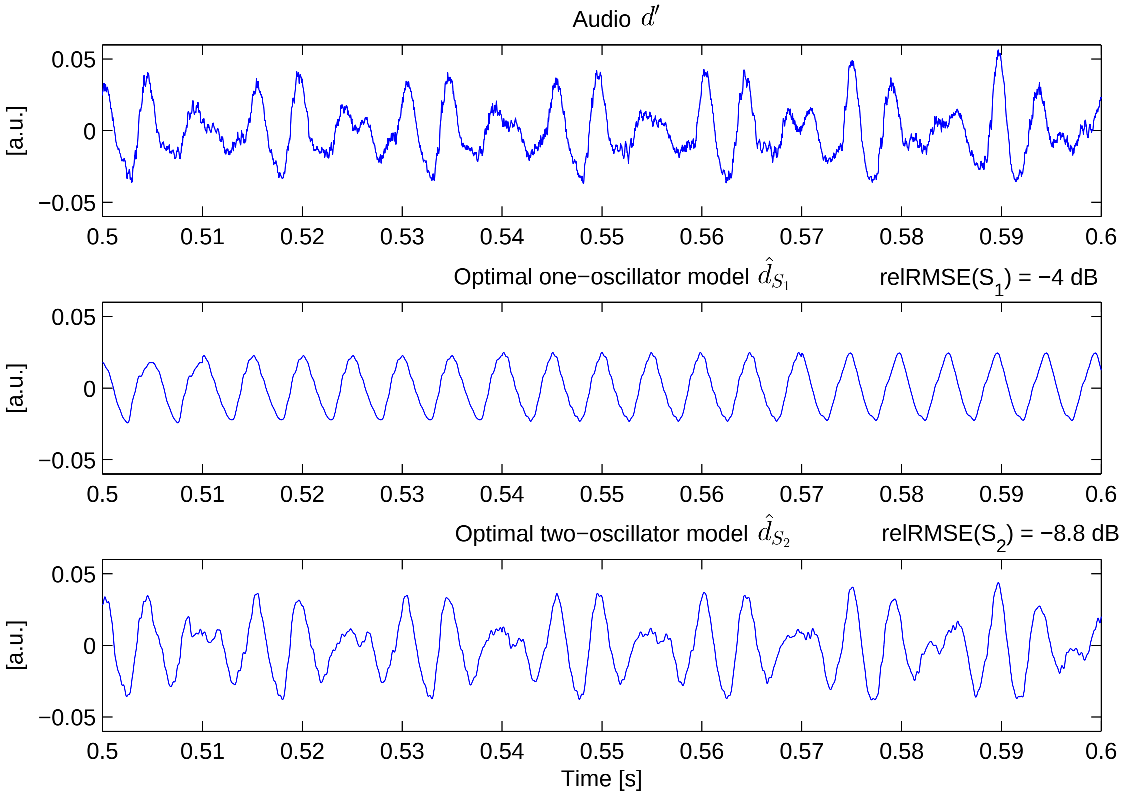


Figure 4.14: Diplophonic audio waveform and its best one-oscillator and two-oscillator model waveforms.

recorded signals as described, and resynthesized with overlap-and-add (Hann window, overlap of 50 %). $\hat{d}_1$ is the oscillator waveform corresponding to the lower pitch of the diplophonic example and $\hat{d}_2$ corresponds to the higher pitch. The model error (lower subplot) looks noisy.

`d1.wav` is the waveform of the low pitch oscillator and `d2.wav` is the waveform of the high pitch oscillator, i.e., the diplophonic components. One can auditively identify the separate pitches of the oscillators from these waveforms. `d1+d2.wav` is the sum of the harmonic oscillators and enables evaluating the perceived musical interval simultaneously. It is time-variant but approximately a major third. The waveform `d1+e.wav` can be seen as an enhanced version of the diplophonic audio recording, because the perceived degree of diplophonia is decreased. As compared to the recorded waveform, the high pitch auditory stream is louder in `d2+e.wav`, because it is not masked by the low pitch stream. In the euphonic interval, the diplophonic component is zero.

`e.wav` is the model error waveform that sounds noisy, hoarse and harsh. Narrowband energy in the waveform model error exists, which arises from oscillator coupling, i.e., the combination tones of the oscillators. The model error consists of aspiration noise, plus the combination tones which are remainders of the coupling/co-modulation of the oscillators, plus artifacts from modulation noise. In the model error waveform, one hears a pitch that is not audible in the recorded waveform, because it is masked.

Figure 4.17 shows the spectrograms of the extracted oscillations and the error waveform. The oscillations have harmonic structure and the error is noise plus some narrowband components. The narrowband energy in the error signal is a linear combination of the oscillators' partial frequencies and explained by oscillator coupling.

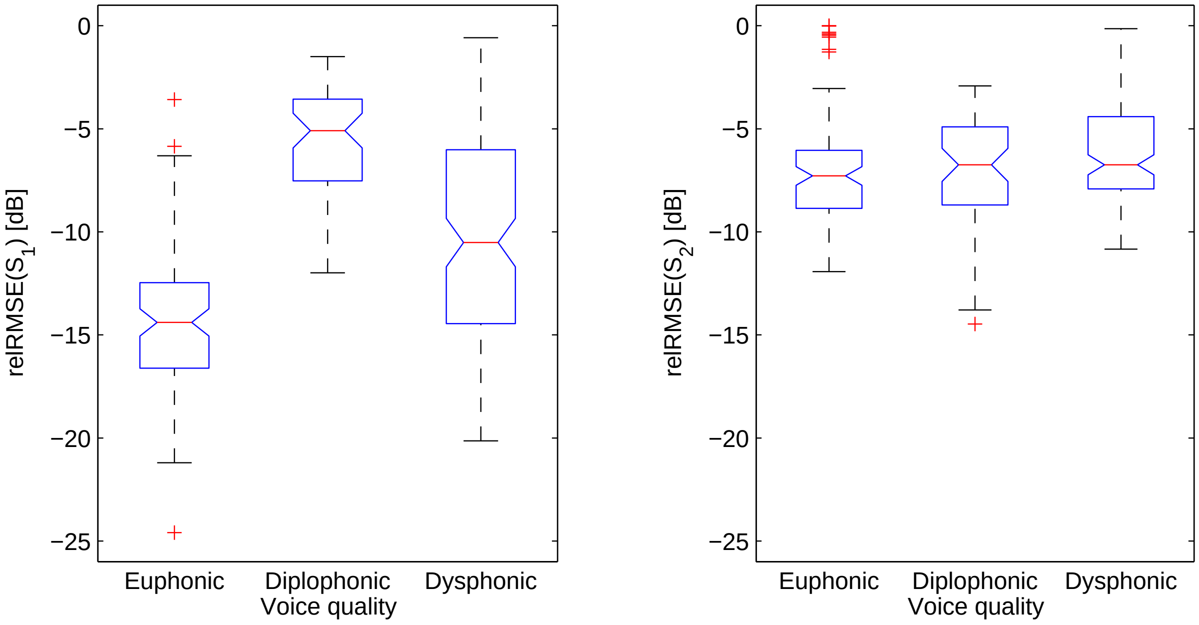


Figure 4.15: Boxplots of the waveform model errors versus voice quality.

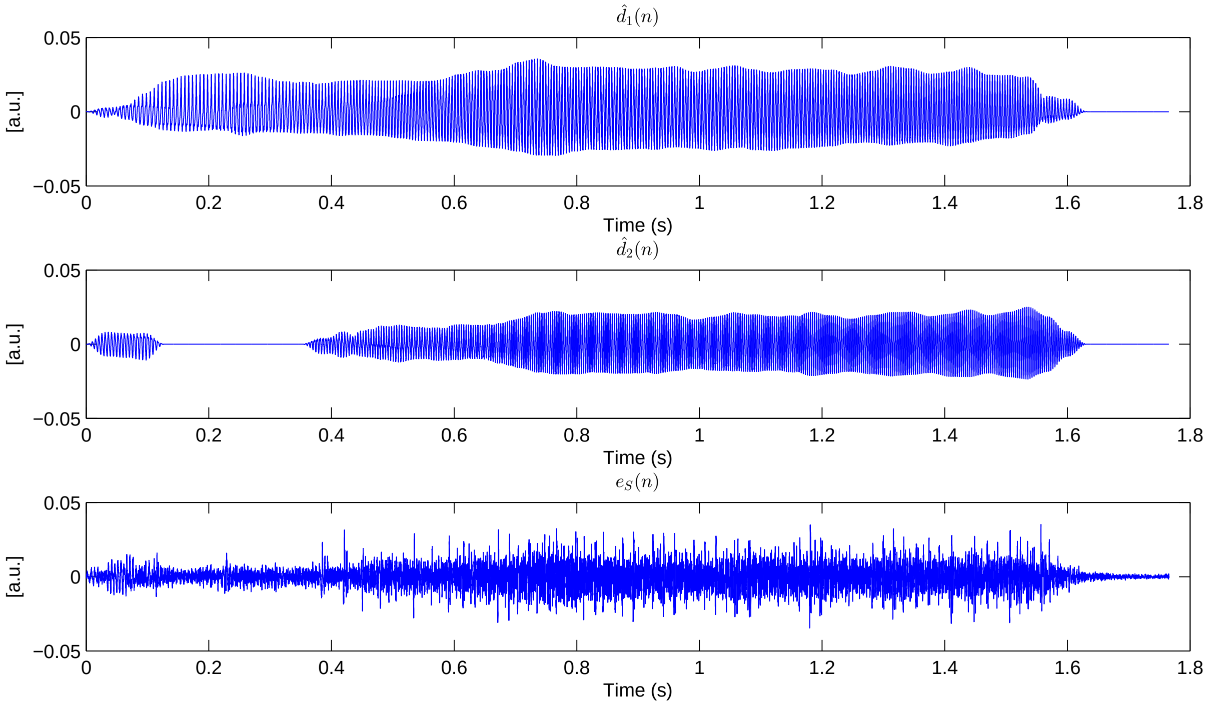


Figure 4.16: Extracted oscillator waveforms of an example of a diplophonic phonation and its model error waveform. The upper subplot shows the first oscillator waveform, the middle subplot shows the second oscillator waveform and the lower subplot shows the error waveform.

Loudness modeling

The predicted loudness curves of the individual waveforms are obtained from a multi-source model of mutual masked loudness [61]. Figure 4.18 shows the masked short-term loudness. A

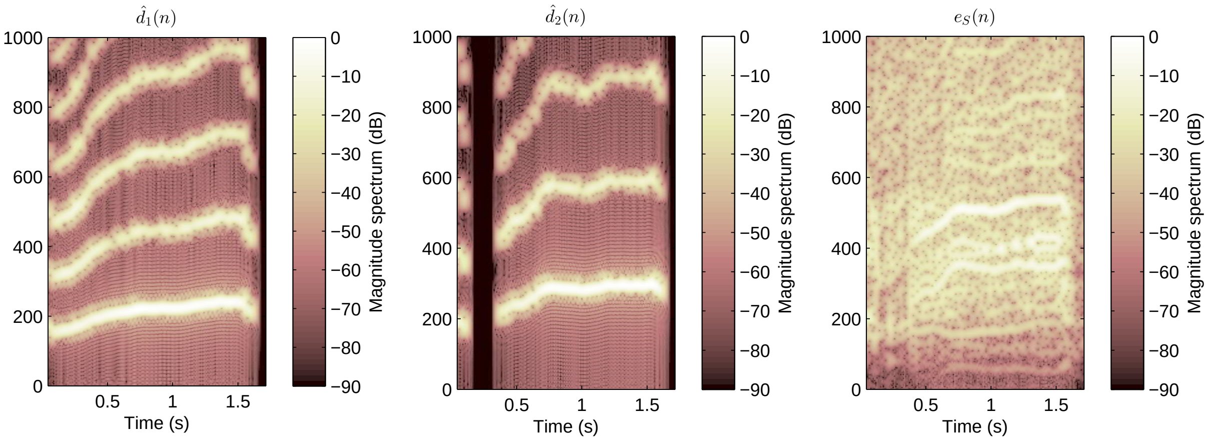


Figure 4.17: The oscillators' and the model error waveforms' spectrograms.

masking model yields the masked loudness of each of the three signals [109, 110]. The average sound pressure level of the observed waveform was defined to be 70 dB(A) in binaural headphone playback. Three temporal intervals are distinguished: At onset the error waveform is the loudest (red line), the secondary oscillator (green line) is less loud at 0.15 sone and the primary oscillator is suppressed. In the quasi modal interval the primary oscillator is the loudest of the three signals. The loudness of the error waveform is approximately 0.8 sone, and the second oscillator is absent. In the interval from 0.4 to 1.55 s the error waveform is dominant and both oscillators are audible. The loudness of the primary oscillator is at approximately 0.8 sone and the loudness of the secondary oscillator at approximately 0.4 sone. At phonation offset, loudness decreases and the second oscillator shortly exceeds the loudness of the primary oscillator. It is illustrated that sound sources of diplophonic phonation, which can be segregated by humans, can also be segregated automatically and fed into a perceptual model that estimates the single source contributions to the overall loudness.

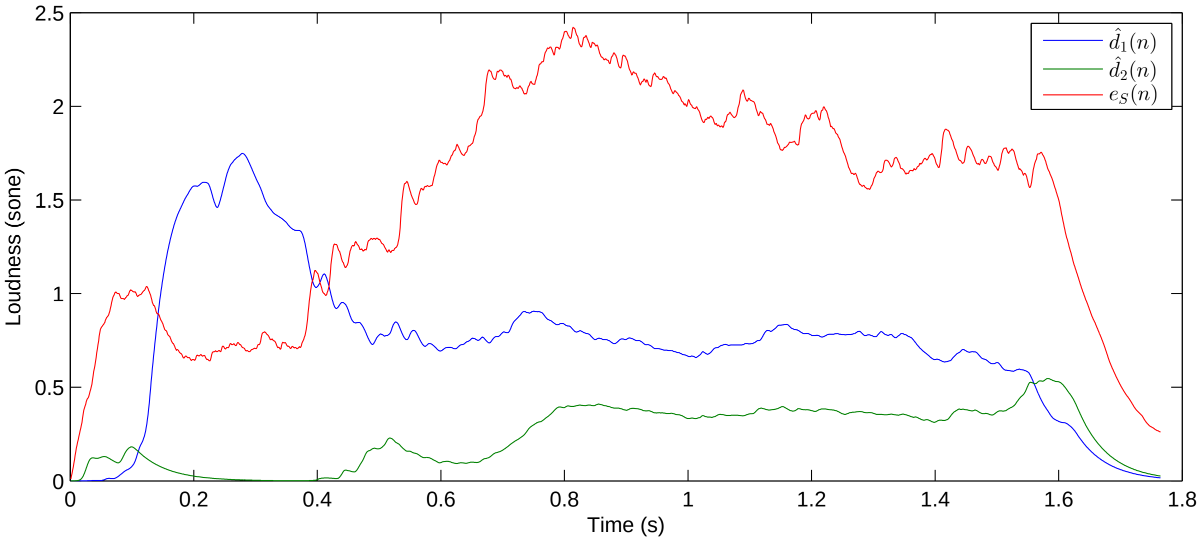


Figure 4.18: Estimated short-term masked loudness [61, 109, 110] of the separated oscillators and the model waveform. Distinct auditory streams are separated from a diplophonic phonation example and single source contributions to the overall loudness is estimated. The audio files can be downloaded from the internet [108].

Fundamental frequency extraction

The performance of fundamental frequency extraction is investigated in an experiment with 65 euphonic, 28 diplophonic and 84 dysphonic intervals (corpus ID 99). Fundamental frequency extraction via polynomial waveform modeling is compared to a state-of-the-art fundamental frequency extractor for noisy mixtures of two speakers [111].

The fundamental frequency extracted by spectral video analysis serves as ground truth (chapter 3). The block length is 128 video frames and the amplitude filtering threshold is set to 2.45, which is the average of the optimal thresholds determined in chapter 3 (2.3 and 2.6). The lowest frequency is 70 Hz, and if more than two estimates for the fundamental frequency are obtained, the lowest two are chosen.

Figure 4.19 shows the spectrogram of an example of diplophonic phonation together with the fundamental frequencies extracted by polynomial waveform modeling and spectral video analysis. In the beginning of the interval the phonation is euphonic. There is clear harmonic structure, but also a high level of noise. The fundamental frequency is approximately 170 Hz and the partials are at integer multiples of the fundamental frequency. Both spectral video analysis and polynomial modeling are able to extract the fundamental frequency. The diplophonic interval appears after a bifurcation at approximately 0.35 s. From there on the spectral characteristics become more complicated, because two fundamental frequencies coexist with coupling products. The polynomial modeling finds both fundamentals whereas the spectral video analysis fails from 0.35 to 0.6 s.

Figure 4.20 shows the audio spectrogram together with the fundamental frequency extracted by Wu's method [111] and spectral video analysis. Wu's method fails for this example of diplophonic voice. It either tends to obtain some frequency in between the true fundamentals or the frequency of the meta cycle. The fundamental frequencies are extracted by autocorrelation, which may be suboptimal for diplophonic phonation. In chapter 5, another approach that is based on autocorrelation suffers from a similar problem.

E_{01}, E_{02}, E_{10}, E_{12}, E_{20}, E_{21} are block error rates that express the extraction of the wrong number of fundamental frequencies. E.g., if there is one fundamental in the analysis block and the algorithm extracts two, E_{12} counts an error. E_{Gross} is the error rate of gross fundamental frequency errors, i.e., the rate of blocks in which the number of fundamental frequencies is correct, but the relative frequency error of either estimate exceeds 20 %. E_{Total} is the sum of all above mentioned rates. If no total error is observed, E_{Fine} is calculated, which is the relative frequency error in percent. If there are two fundamental frequencies, it is the sum of the relative frequency errors. Details on the calculation of the error measures can be found in [111].

Table 4.1 summarizes the error rates with respect to the method of extraction. Some of the error rates are high, i.e., E_{12}, E_{21}, and E_{Gross} for polynomial waveform modeling and E_{21} and E_{Gross} for Wu's method, which results in high total error rates for both methods. E_{Fine} is tested with a two-sided Wilcoxon rank-sum test, because the number of observations differs between the compared groups. All other error rates are tested with an eight-fold Bonferroni corrected two-sided paired Wilcoxon signed rank test. The rate of total errors is significantly higher for Wu's approach (47.18 and 57.53 %).

The polynomial waveform modeling prefers two fundamental frequencies over one, which is reflected by E_{12} (12.12 %). The high rates of E_{21} (22.22 %) may stem from mucosal wave artifacts in spectral video analysis. Wu's method has an even higher E_{21} (38.11 %), because only one

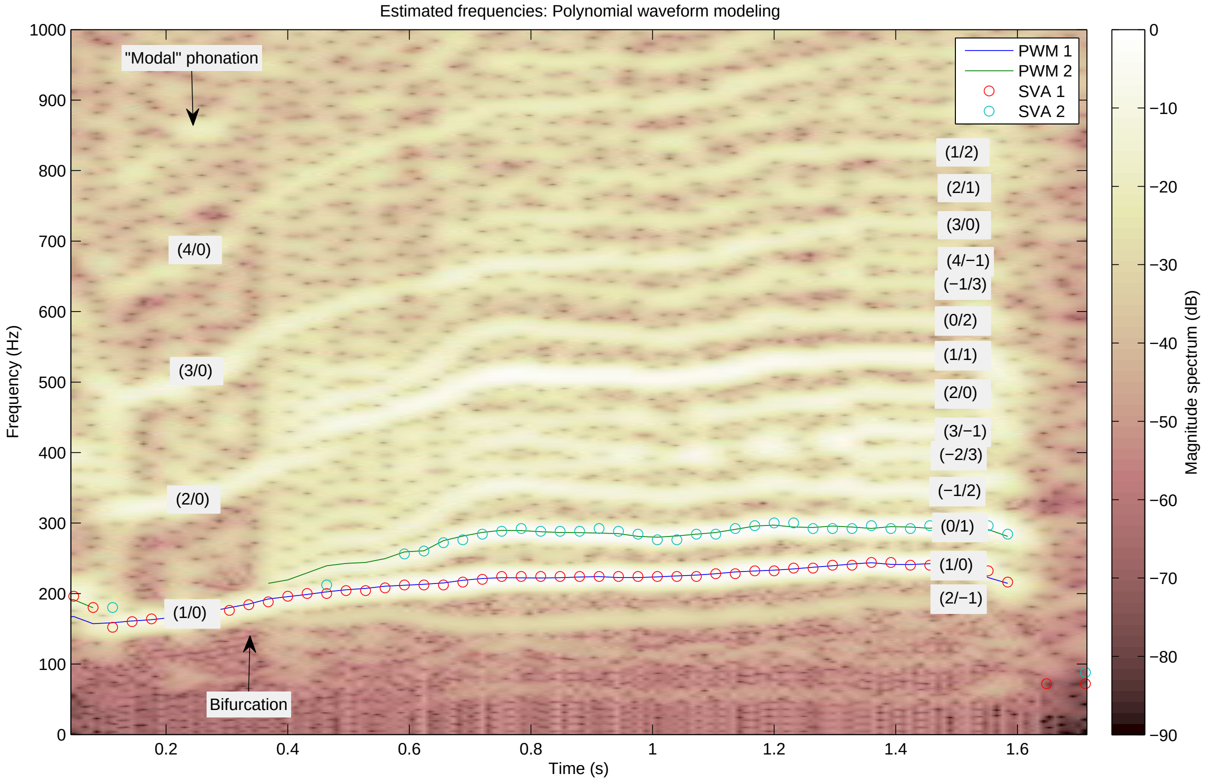


Figure 4.19: Audio spectrogram and fundamental frequency estimates for an example of diplophonic phonation. The used fundamental frequency extractors are polynomial waveform modeling (PWM) and spectral video analysis (SVA). The frequencies of the sinusoidal components of the spectrogram are interpreted by means of linear frequency combinations in the format (a/b), with the sinusoidal frequency $k = a \cdot k_1 + b \cdot k_2$ (right side of the figure).

fundamental frequency is often extracted for diplophonic voice, which confirms the observation made in figure 4.20.

Table 4.2 shows E_{Total} with regard to voice quality. Wu's method is wrong (95.74 %) for diplophonic intervals, but polynomial waveform modeling achieves at least 46.28 %. The errors are tested with a three-fold Bonferroni corrected two-sided paired Wilcoxon signed rank test.

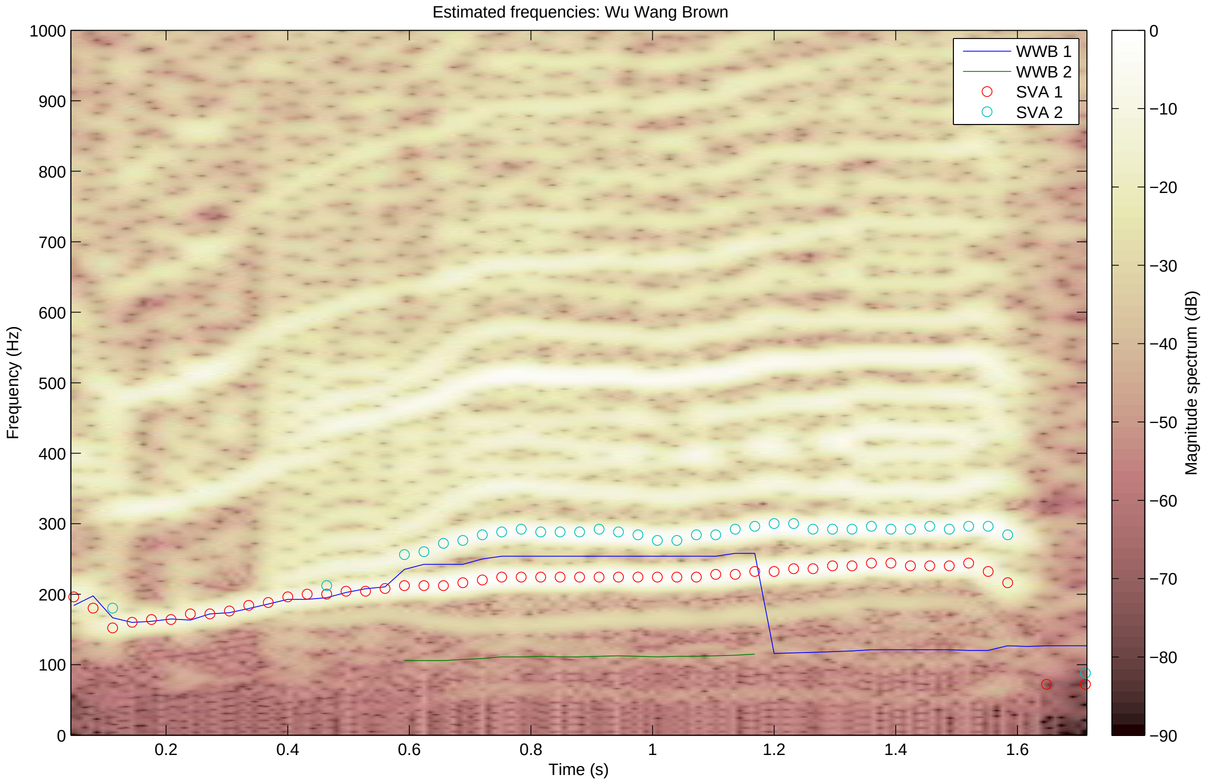


Figure 4.20: Audio spectrogram and fundamental frequency estimates for an example of diplophonic phonation. The used fundamental frequency extractors are Wu's method (WWB) and spectral video analysis (SVA).

	PWM (%)	WWB (%)	p-value
E_{01}	0.12	0.87	$= 0.013$
E_{02}	0.68	0.05	< 0.001
E_{10}	1.29	3.63	< 0.001
E_{12}	12.12	1.23	< 0.001
E_{20}	1.5	3.34	< 0.001
E_{21}	22.22	38.11	< 0.001
E_{Gross}	9.25	10.29	n.s.
E_{Total}	47.18	57.53	$= 0.009$
E_{Fine}	2.06	1.25	n.s.

Table 4.1: Median error rates of multiple fundamental frequency extraction. Comparison of polynomial waveform modeling (PWM) with Wu's method (WWB) [111]. Wilcoxon signed rank test, paired, two-sided, eight-fold Bonferroni corrected. E_{Fine} is tested with a two-sided Wilcoxon rank sum test.

(Medians)	PWM (%)	WWB (%)	p-value
E_{Total}, euphonic intervals	37.5	39.37	n.s.
E_{Total}, diplophonic intervals	46.28	95.74	< 0.001
E_{Total}, dysphonic intervals	54.97	58.83	n.s.

Table 4.2: Total error rates of fundamental frequency extraction with respect to voice quality.

4.4 Discussion and conclusion

Two waveform models that yield physiologically interpretable results for diplophonic phonation have been proposed, pilot-tested and evaluated on observed data. Unit-pulse FIR filtering was used for analyzing one euphonic and one diplophonic glottal area waveform. A ground truth for double cycle marks and fundamental frequencies in diplophonic voice has been established and the fine structure fidelity of the diplophonic waveform model was demonstrated [59]. The parameter extraction can handle uncorrelated additive noise, which occurs in glottal area waveforms and in audio waveforms of disordered phonation. The waveform parameterization exploits periodicity and yields a parsimonious signal representation. The results also illuminate limitations of cycle-based analyses (e.g., jitter or shimmer), which rely on a one-oscillator assumption.

Polynomial waveform modeling has been proposed, which improves unit-pulse FIR filtering in two ways. First, the number of oscillators and the audio waveform model are jointly estimated. In unit-pulse FIR-filtering, it was assumed that the number of oscillators and their fundamental frequencies are known from spectral video analysis. Under this assumption, the added diagnostic value of audio waveform modeling over spectral video analysis is small, because the presence of diplophonia can be detected by counting the number of oscillators in spectral video analysis alone. Second, cycle discontinuities that occur in unit-pulse FIR filtering when the model parameters evolve block internally have been taken into account by band limitation during resynthesis.

Three large data experiments on polynomial waveform modeling were carried out. First, group effects of model error measures have been illustrated for 55 diplophonic and 223 non-diplophonic audio waveforms. The model error measures are used as computational markers of voice quality in chapter 5. Second, one diplophonic example has been analyzed in more detail, namely by auditive evaluation of the different waveform model components and by modeling the mutual masked short-term loudness of these components. Sound sources in that example have been segregated automatically and fed into a perceptual model that estimates the single source contributions to the overall loudness. Third, a corpus of 28 diplophonic and 149 non-diplophonic phonation intervals has been used for evaluating fundamental frequency extraction. A state-of-the-art algorithm fails for diplophonic phonation, because its algorithm parameters are optimal for tracking a mixture of two independent speakers [111], while polynomial waveform modeling achieves smaller error rates.

Limitations of the waveform models arise from the assumptions of the oscillators' additivity, uncorrelatedness and the presence of spectral energy at their fundamentals. In reality, additivity and co-modulation of oscillators coexist and the correlation of the oscillators is not zero. Thus the waveform model solutions are not optimal under all circumstances. The assumption that significant spectral energy is present at the oscillator fundamental frequencies could be criticized from a perceptual point of view, because the perceived pitch can correspond to a frequency that has no significant spectral energy [112].

Suggestions for future work are possible. Automatic fitting of waveform models to observed data is not limited to diplophonic voice but may also be attempted on creak and non-diplophonic roughness. The combination of additive and modulative oscillators may outperform the additive-only approach to model diplophonic waveforms in some cases. The video and audio data need ground truth annotation of fundamental frequency, which requests expert knowledge of waveform patterns in disordered phonation and constitutes a workload of many man-hours. Algorithmically separated signal components might be tested for ear training of logopedists and phoniatricians. The separation approach might also be tested for auditory stream segregation in dysphonic voices [21]. Methods inspired by joint multiple fundamental frequency estimation and source separation should in the future be used to overcome limitations arising from the assumptions that the proposed parameter estimation procedures rely on. However, it must be kept in mind that the a priori knowledge available on voices identified as diplophonic is limited and qualitative only. A promising direction of joint multiple fundamental frequency estimation and source separation with only qualitative knowledge about the sources is taken in [113].

5 Diagnostic tests and their interpretation

In the previous chapters, novel features from laryngeal high-speed videos and audio recordings are obtained. In this chapter, selected available features are evaluated on their ability to detect diplophonia in audio signals. In section 5.1, formerly published features are used to train binary threshold classifiers and evaluated with regard to their underlying assumptions. In section 5.2 the features obtained from waveform modeling are visualized in a physiological interpretable way, via the Diplophonia Diagram. The Diplophonia Diagram is used for automated detection of diplophonia by means of a binomial logistic regression model. Finally, the problem that the voice quality annotation may be an imperfect ground truth is addressed by constructing a reference standard via latent class analysis.

5.1 Conventional features

To discuss the status-quo of objective clinical assessment of diplophonic voice, five conventional features are tested with regard to their ability to detect diplophonia. They are jitter [114], shimmer [115, 116], the Harmonics-to-Noise Ratio (HNR) [85], the Göttingen features for irregularity and noise [44] and the Degree of subharmonics (DSH) [117]. Binary threshold classifiers are trained for each feature and evaluated with regard to their validity. The corpus consists of 278 phonation intervals, from which 96 are euphonic, 55 are diplophonic and 127 are dysphonic (corpus ID 8).

5.1.1 Jitter and shimmer

Analysis

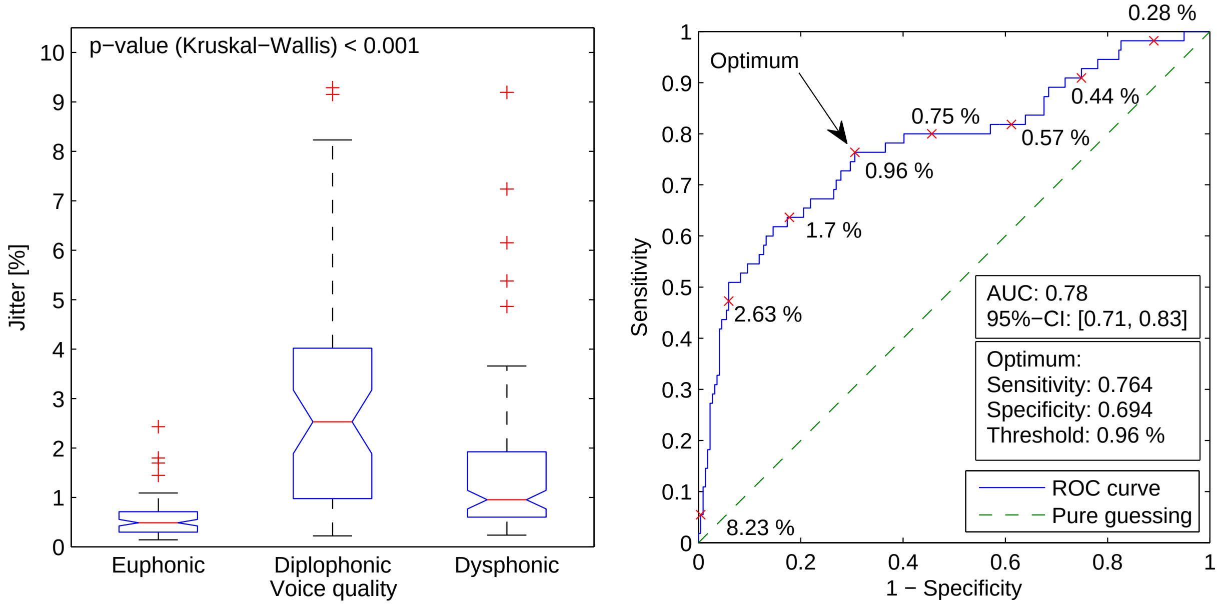
Jitter is a measure of the temporal fluctuations of cycle length and has been introduced by Liebermann in 1961 [114]. Jitter is given in percent and its simplest version is used. The average absolute difference of adjacent cycle lengths divided by the average cycle length. Shimmer is a measure of temporal fluctuations of cycle peak amplitudes. The term has been introduced as a synthesis parameter by Wendahl in 1966 [115]. Shimmer is a cover term for all kinds of cycle peak amplitude fluctuations, of which many types exist. The used version of the shimmer is 20 times the average absolute base-10 logarithm of the ratio of adjacent cycle peak amplitudes. It is given in dB.

The open source software Praat [85] is used to determine jitter and shimmer of audio waveforms. At first, the cycle length is determined via the subroutine ”To Pitch (cc)”. All the parameters are set to their standard values. Cycle-length candidates are obtained via forward cross-correlation. A post-processing algorithm obtains the cheapest temporal path through the candidates and delivers the fundamental frequency contour. The subroutine ”To PointProcess (cc)” is used to estimate the underlying cycle marks. The jitter is obtained using the subroutine ”Get jitter (local)” and the shimmer is obtained using the subroutine ”Get shimmer (local_dB)” with all its parameters set to their standard values.

In the popular Multi-Dimensional Voice Program (MDVP) of KayPENTAX, the corresponding features for jitter and shimmer are called ”Jitt” and ”ShdB”. It must be noted, that MDVP may obtain different results than Praat, because the extraction of cycle length and cycle peak amplitudes is different.

Evaluation and discussion

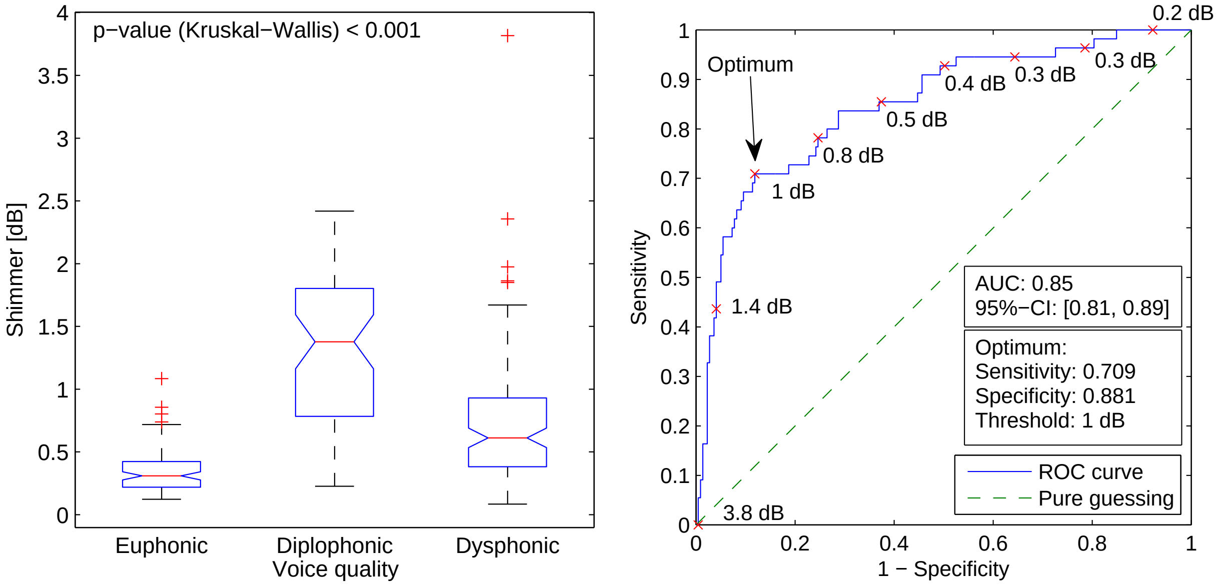
Figure 5.1a shows the boxplots for jitter with respect to voice quality. The average of the jitter as well as its within group variation is larger in the diplophonic group than in the other two groups. The ROC curve (figure 5.1b) shows that the optimal threshold for separating diplophonic from non-diplophonic voices is at 0.96 %. A sensitivity of 76.4 % and a specificity of 69.4 % are achieved. The boxplots of shimmer look comparable to jitter (figure 5.2). The optimal threshold is 1 dB, which yields a lower sensitivity (70.9 %) and a higher specificity (88.1 %) than jitter.



(a) Boxplots.

(b) ROC curve, diplophonic versus non-diplophonic phonation.

Figure 5.1: Evaluation of jitter versus voice quality.



(a) Boxplots.

(b) ROC curve, diplophonic versus non-diplophonic phonation.

Figure 5.2: Evaluation of shimmer versus voice quality.

At first sight, these results suggest that jitter and shimmer are useful for distinguishing diplophonic from non-diplophonic phonation intervals. However, one may question their validity for diplophonic and dysphonic voices. Both features are based on the assumption that the waveform is produced by one oscillator only, which does not hold for diplophonic voices and subgroups of dysphonic voices, i.e., voices with high levels of noise or strong modulative components. Figure 5.3 shows an example of a diplophonic audio waveform together with its spectrogram and Praat cycle marks. The waveform is composed of metacycles and the cycle mark determination fails. Cycle mark positions are quasi random in adjacent metacycles, and therefore jitter and shimmer values are random. Jitter is estimated to be high (7.677 %), owing to the fluctuations of the cycle mark distances. Shimmer is also estimated to be high (2.091 dB), owing to the cycle-to-cycle fluctuations of peak amplitudes.

Figure 5.4 shows another example of a diplophonic audio waveform together with its spectrogram and cycle marks. Although there are severe divergences from strict periodicity in the micro structure of the waveform, the negative peaks are quasi-periodic, and so are the estimated cycle marks. Two oscillations are present in the waveform, and the algorithm detects a subharmonic fundamental. Jitter and shimmer are low (0.433 % and 0.412 dB) owing to the quasi-periodicity of the waveform's macro structure.

Figure 5.5 shows an example of a dysphonic audio waveform together with its spectrogram and cycle marks. There are even more severe divergences from perfect periodicity than those shown in figure 5.4. Noise, but no distinct second fundamental frequency can be seen in the spectrogram. The macro structure of the waveform is pseudo-periodic however, thus jitter is low (0.707 %) and shimmer is moderate (1.167 dB).

The presence of metacycles and noise may mislead the fundamental frequency measurement and thus the measurement of jitter and shimmer. One is sometimes tempted to expect high jitter and shimmer if divergences from normal voice quality are auditorily perceived, but the

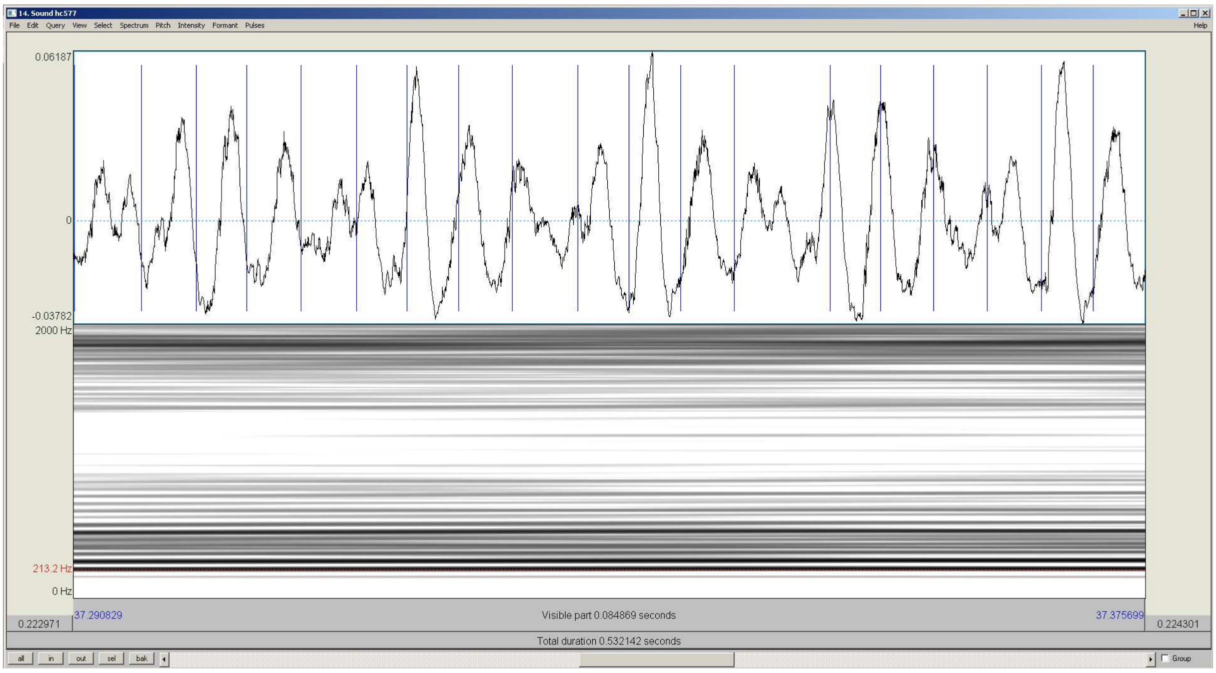


Figure 5.3: Audio waveform, cycle marks and spectrogram of a diplophonic interval where the jitter and shimmer measurement fails. The jitter is estimated to be 7.677 % and the shimmer is estimated to be 2.091 dB.

relation of the perturbation measures with perceptual cues is fuzzy. The example in figure 5.3 sounds noisy and rough, and the interval of the two pitches is at a slightly detuned major third. The example shown in figure 5.4 sounds less noisy, but the pitch interval is much more dissonant (slightly detuned tritone). The example in figure 5.5 sounds rough and noisy, but not diplophonic. Due to the lack of validity of cycle-based perturbation measures for two-oscillator waveforms, the presence of diplophonia should not be tracked by means of jitter or shimmer.

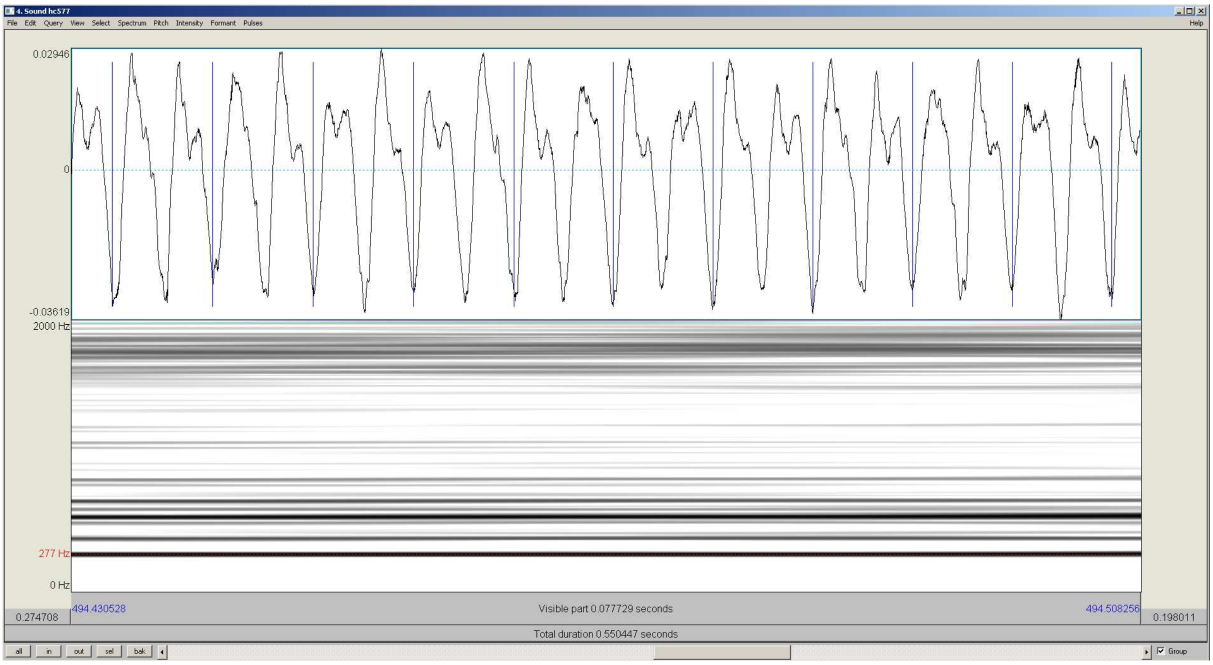

4. Sound hc577
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses
Help
0.02946
0
-0.03619
2000 Hz
277 Hz
0 Hz
0.274708
494.430528
Visible part 0.077729 seconds
494.508256
0.198011
Total duration 0.550447 seconds
all
in
out
sel
bak
Group

Figure 5.4: Audio waveform, cycle marks and spectrogram of a diplophonic interval where the jitter and shimmer measurement fails. Two oscillators produce the waveform, and the algorithm detects the subharmonic fundamental. The jitter is estimated to be 0.433 % and the shimmer is estimated to be 0.412 dB.

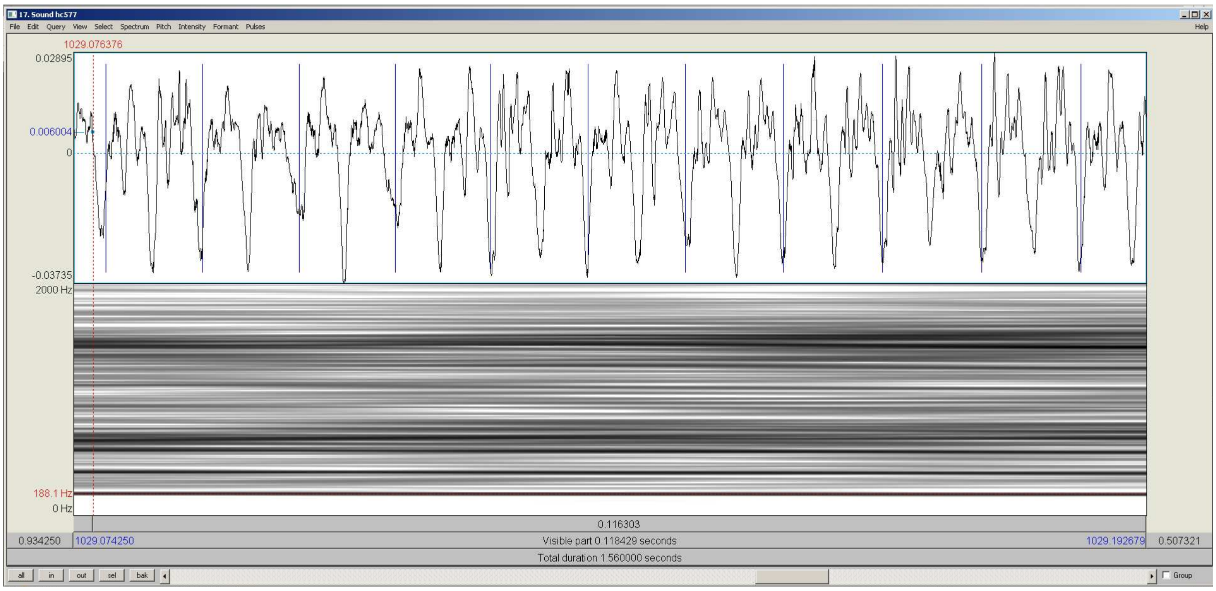


Figure 5.5: Audio waveform, cycle marks and spectrogram of a dysphonic interval where the jitter and shimmer measurement fails, because the algorithm detects a subharmonic fundamental. The jitter is estimated to be 0.707 % and the shimmer is estimated to be 1.167 dB.

5.1.2 Göttingen irregularity feature

Analysis

The Göttingen irregularity feature has been introduced in 1998 by Michaelis et al. [44]. It is part of the "Göttingen Hoarseness Diagram", a clinically evaluated open source algorithm for voice assessment. The irregularity feature is a linear combination of the average waveform matching coefficient, the PPQ jitter (period perturbation quotient, $K = 3$) and the EPQ shimmer (energy perturbation quotient, $K = 15$) [44]. The fundamental frequency is measured by the waveform matching algorithm [118, 119]. Equation 5.1 [44] calculates the irregularity measure, where MWC is the mean waveform matching coefficient, j_3 is the jitter (in percent) and s_{15} is the shimmer (in percent). The MWC is the correlation of all corresponding pairs of adjacent cycles' samples, averaged over the whole analysis interval. The perturbation quotient is given in equation 5.2 [44], where N is the number of cycles within the analyzed interval, K is the block length (3 for jitter and 15 for shimmer), and $\nu(\mu)$ is the cycle feature of the μ^{th} cycle (cycle length or energy). The energy is the sum of the squared μ^{th} cycle's samples.

Smaller values of MWC or higher values of j_3 or s_{15} lead to higher values of irregularity (equation 5.1). Thus, the Göttingen irregularity is in principle similar to jitter and amplitude shimmer, i.e., invalid when estimated from two-oscillator waveforms.

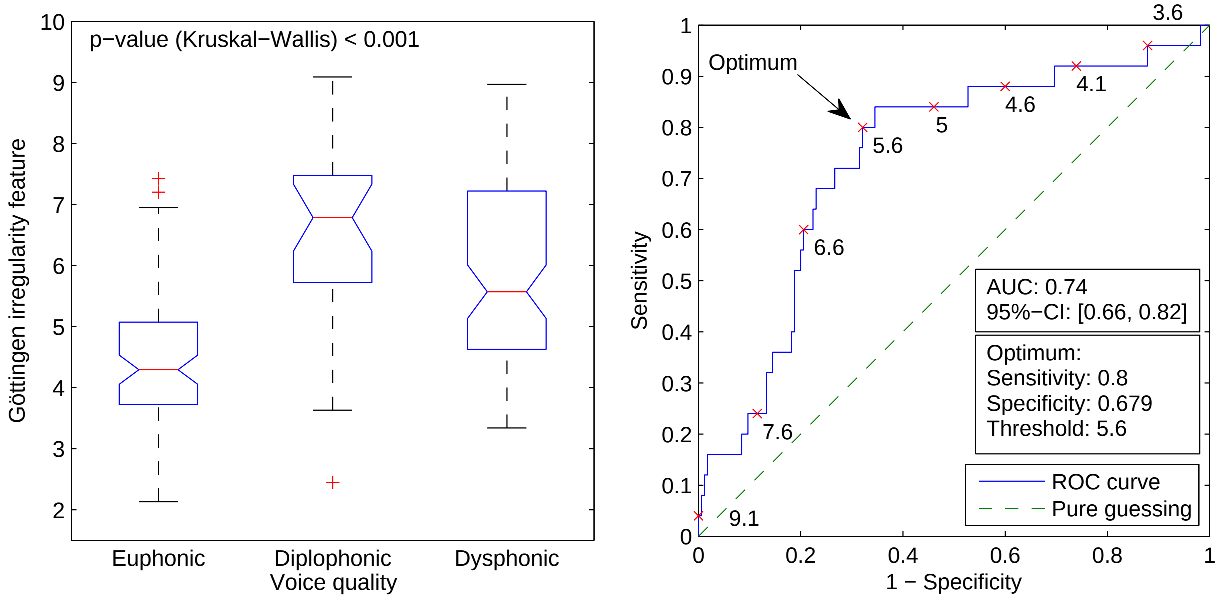
$$\text{Irregularity feature} = 5 + \frac{1}{\sqrt{3}} \cdot \left(\frac{\log(1 - \mathrm{MWC}) + 1.614}{0.574} + \frac{\log(j_3) + 0.374}{0.645} + \frac{\log(s_{15}) - 0.757}{0.368} \right) \tag{5.1}$$

$$\mathrm{PQ} = \frac{100\,\%}{N - K} \cdot \sum_{\mu=\frac{K-1}{2}}^{N-\frac{K-1}{2}-1} \left| \frac{\nu(\mu) - \frac{1}{K} \cdot \sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} \nu(\mu + k)}{\frac{1}{K} \cdot \sum_{k=-\frac{K-1}{2}}^{\frac{K-1}{2}} \nu(\mu + k)} \right| \tag{5.2}$$

Evaluation and discussion

Figure 5.6 shows the boxplots and the ROC curve for the irregularity feature versus voice quality. The group effects are similar to the group effects for jitter and shimmer. At a cut-off threshold of 5.6, a sensitivity of 80 % and a specificity of 67.9 % are achieved. This performance seems to be promising, but given the reservations with regard to jitter and shimmer, it is suggested that the irregularity is not meaningful for two-oscillator waveforms.

The Göttingen irregularity is obtained for intervals that are longer than 520 ms only. 190 out of the 278 intervals satisfy this criterion. 78 of the 190 intervals are euphonic, 25 are diplophonic and 87 are dysphonic. The size of the confidence intervals of the observed average estimates reflects the smaller sample size and the variance of the distributions.



(a) Boxplots.

(b) ROC curve, diplophonic versus non-diplophonic phonation.

Figure 5.6: Evaluation of the Göttingen irregularity versus voice quality.

5.1.3 Göttingen noise feature

Analysis paradigm

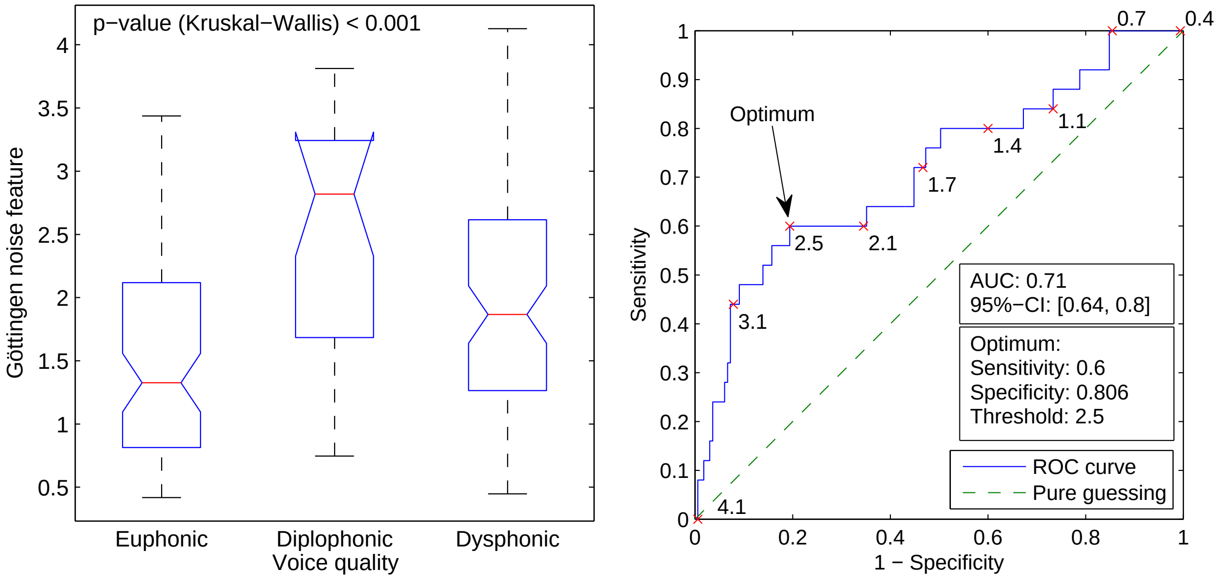
In the "Göttingen Hoarseness Diagram", the Gottal-to-Noise Excitation Ratio (GNE) is a term in the noise feature expression (equation 5.3). The GNE reports additive noise in disordered voice [43]. It was introduced by Michaelis in 1997, and measures correlation of Hilbert envelopes obtained from a filterbank. The rationale is that glottal pulse energy is correlated, whereas the noise energy is uncorrelated. The GNE is obtained as follows: 1) Block wise inverse filtering of the speech signal, based on linear prediction, 2) filterbank analysis, 3) Hilbert envelope calculation, 4) calculation of the correlation across frequency and 5) obtaining the maximum of all correlation coefficients. GNE3 is obtained using a filterbank bandwidth of 3 kHz.

$$\text{Noise feature} = 1.5 + \frac{0.695 - \text{GNE3}}{0.242} \tag{5.3}$$

Evaluation and discussion

Figure 5.7 shows the boxplots of the noise feature versus voice quality. In accordance with the Göttingen irregularity feature, only 190 intervals are used. The dysphonic phonation intervals were expected to have higher noise than euphonic intervals, which is indeed confirmed. The noise feature values are higher in the diplophonic group than in the other two groups. Possible explanations are given as follows.

Firstly, diplophonic voice may on average be more noisy than non-diplophonic voice. Secondly, the increase may also be explained by some selection bias that may have happened during data collection. The length of the phonation intervals was limited to 2.048 seconds, and so the examiners strictly selected diplophonic phonation intervals from clinically diplophonic patients. In the clinically dysphonic group, the selection has been less strict to obtain a broader variety of voice qualities. Therefore the noise in the diplophonic intervals may be higher than in the dysphonic intervals, which may confound the statistics. For methodical reasons the increased noise level in diplophonic voice should not be used as an indicator for the presence of diplophonia, because noise is a different physical phenomenon than diplophonia.



(a) Boxplots.

(b) ROC curve, diplophonic versus non-diplophonic phonation.

Figure 5.7: Evaluation of the Göttingen noise feature versus voice quality.

5.1.4 Harmonics-to-Noise Ratio

Analysis

Another noise feature is the Harmonics-to-Noise Ratio (HNR). It promises to estimate the signal noise energy in relation to its harmonic energy. Praat's subroutine "To Harmonicity (cc)" is used to obtain the time-dependent HNR, with analysis parameters set to their standard values. The subroutine "Get mean" is used to calculate the time-average over the phonation interval.

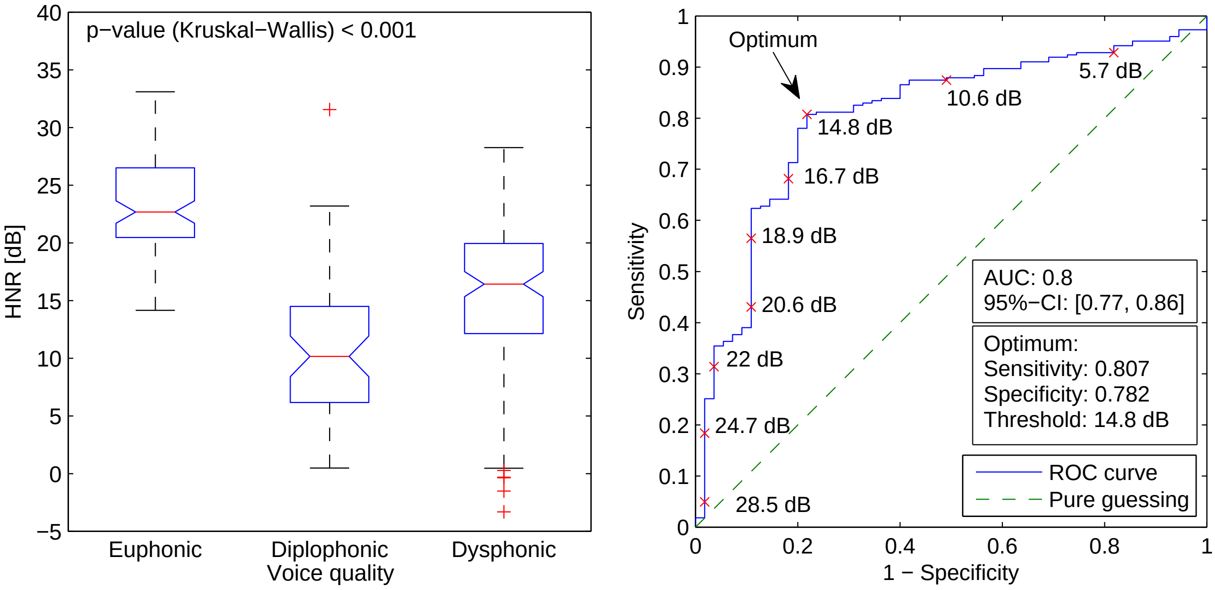
Praat's algorithm estimates the HNR in the lag domain [105]. The HNR is given in equation 5.4, where r' denotes the normalized autocorrelation function of the signal and $\tau_{\max}$ is the lag of its local maximum. A decrease of peak height in the autocorrelation function is interpreted as an effect of noise.

$$\mathrm{HNR}\,[dB] = 10 \cdot \log\left(\frac{r'(\tau_{\max})}{1 - r'(\tau_{\max})}\right) \tag{5.4}$$

Evaluation and discussion

Figure 5.8a shows the boxplots of the HNR with regard to voice quality. The euphonic group has the highest values and the diplophonic group the lowest. The overlap between groups is smaller as compared to the Göttingen noise, which is reflected by increased performance. At the optimal threshold of 14.8 dB the sensitivity is 80.7 % and the specificity is 78.2 %.

In addition to the arguments for the group effects that were given with respect to the Göttingen noise (higher noise levels in diplophonic voice and selection bias), a third factor may increase the group effect of HNR. Boersma's signal model assumptions do not hold for diplophonic waveforms. The local maximum of the autocorrelation function decreases when two oscillator waveforms are added. This effect may be wrongly interpreted by the user as noise. The Göttingen noise has been explicitly related to additive noise, whereas Praat's algorithm has been tested on sinusoidal signals + noise disregarding secondary oscillators and modulation noise. Therefore the Göttingen noise is more trustworthy than Praat's HNR when diplophonic voice is analyzed.



(a) Boxplots.

(b) ROC curve, diplophonic versus non-diplophonic phonation.

Figure 5.8: Evaluation of the Harmonics-to-Noise Ratio (HNR) versus voice quality.

5.1.5 Degree of subharmonics

Analysis

The Degree of subharmonics (DSH) is a feature for detecting ambiguous fundamental frequency estimates. It has been introduced by Deliyski in 1993 [117] and integrated in the popular Multi-Dimensional Voice Program (MDVP) of KayPENTAX. The DSH has been implemented in MATLAB [58]. The procedure estimates the fundamental frequency twice. Each estimator uses different thresholds for frequency detection. The signal is considered to be subharmonic if the fundamental frequency detectors obtain different results. Figure 5.9 shows the block diagram of the DSH extraction.

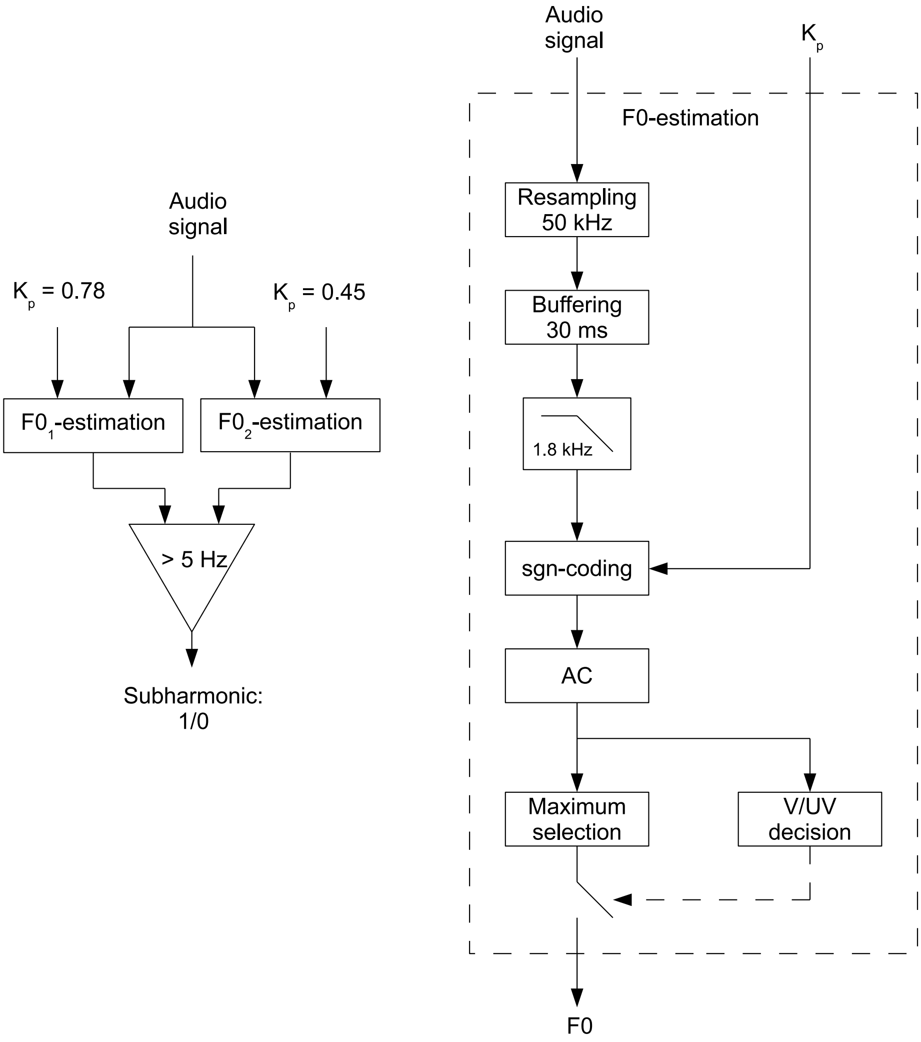
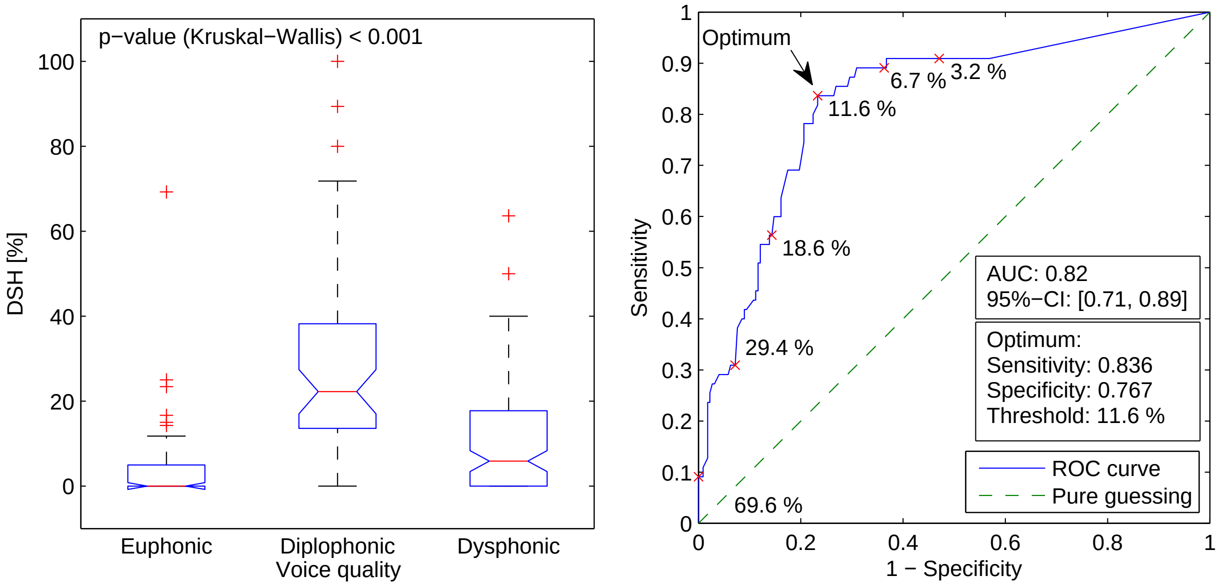


Figure 5.9: Block diagrams of the Degree of subharmonics (DSH) extraction. F0: Fundamental frequency, AC: Autocorrelation, V/UV: voiced/unvoiced, sgn-coding: see text.

In each of the fundamental frequency estimators the signal is resampled to 50 kHz, buffered in 30 ms blocks and lowpass filtered with a cut-off frequency of 1.8 kHz. The signal is sgn-coded using the coding threshold K_p. The sgn-coded signal is 0 if the signal's absolute value is below the threshold and it is -1 or +1 according to the sign of the signal if its absolute value exceeds the threshold. The fundamental frequency of the sgn-coded signal is determined using the autocorrelation, once with $K_p = 0.78$, and once with $K_p = 0.45$. A signal is considered to be voiced if the height of the normalized lagged autocorrelation peak exceeds 0.27. A signal block is considered to be subharmonic if the variation of K_p changes the fundamental frequency estimate or the voicing condition. The DSH is the rate of subharmonic analysis blocks in percent.

Evaluation and discussion

Figure 5.10a shows the boxplots for the DSH with regard to voice quality, revealing large group effects. Figure 5.10b shows the ROC curve. The sensitivity at the optimal threshold of 11.6 % is good (83.6 %) and the specificity is fair (76.7 %).



(a) Boxplots.

(b) ROC curve, diplophonic versus non-diplophonic phonation.

Figure 5.10: Evaluation of the Degree of subharmonics (DSH) versus voice quality.

Figure 5.11 shows results for an analysis block of diplophonic phonation. The upper subplot shows the audio signal, the sgn-coding thresholds and the measure of a subharmonic time interval (13.3 ms). Subplots 2 and 3 show the two sgn-coded signals. Subplots 3 and 4 show their autocorrelation functions together with their cycle length estimates. In stage one ($K_p = 0.78$, subplots 2 and 4) a sub-periodicity is dominant and an artificial cycle length is obtained (marginally unvoiced). In stage two ($K_p = 0.45$, subplots 3 and 5) a shorter cycle length is obtained, which suggests the presence of subharmonics.

To investigate the validity of the fundamental frequency estimates obtained by DSH, the audio spectrogram and the fundamental frequency estimates are reported (figure 5.12). The fundamental frequency estimates are incorrect and most often agree with an artificial low estimate (at times 0.3 s - 0.6s, 0.85 s) or with a frequency in-between the true fundamentals (rest of the time). The DSH is 21.4 % in this interval.

The DSH aims at determining the ambiguity of fundamental frequency measurements rather than measuring the correct values of the fundamental frequencies. However, the DSH is considered to inform on the presence of diplophonia, because of two reasons. Compared to other conventional features, the DSH is the only one that considers two fundamental frequency estimates within one analysis block. Second, the DSH detects cyclic peak height fluctuations. Thus, it is used in section 5.3 when a reference standard for the presence of diplophonia is proposed.

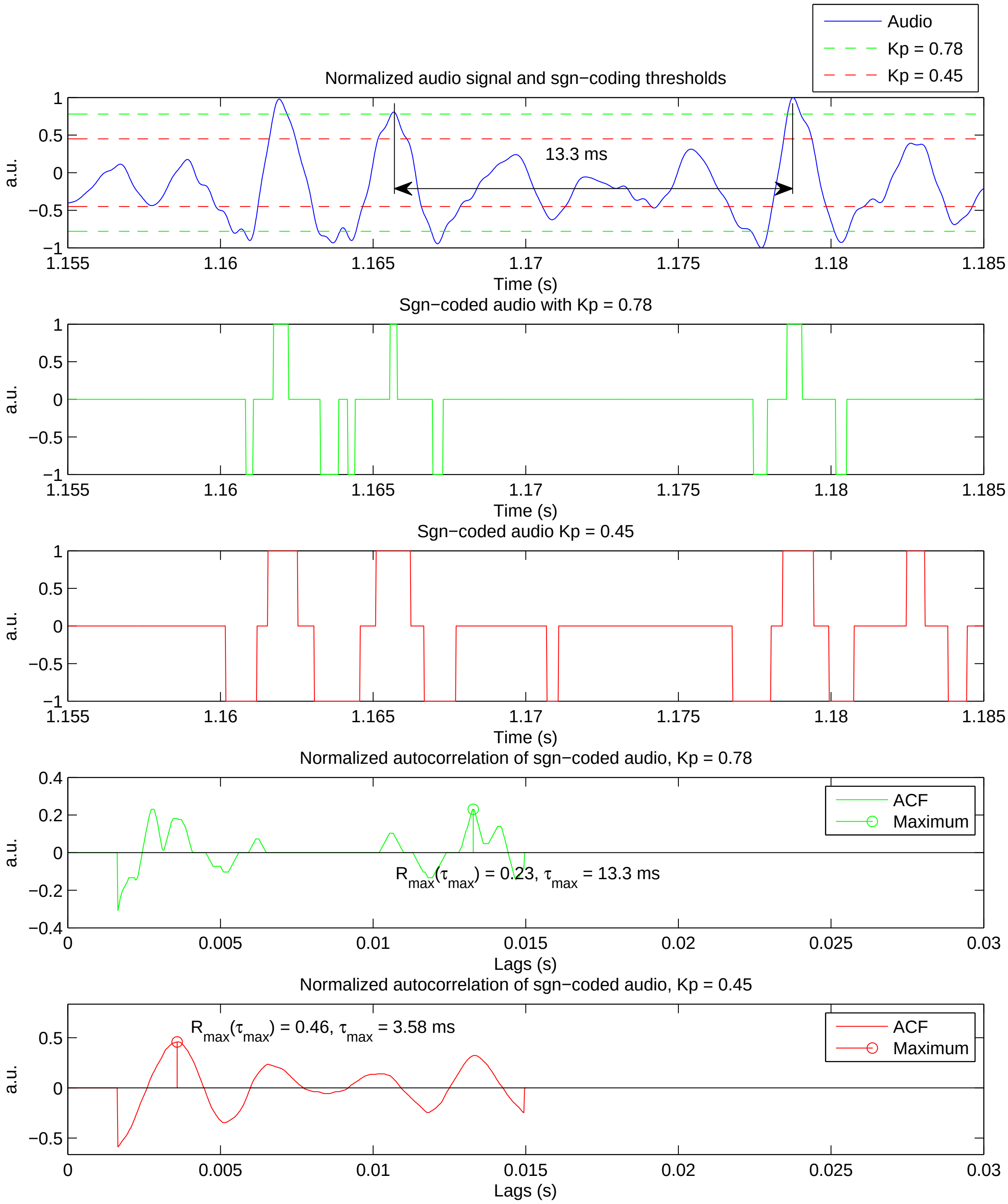


Figure 5.11: Results for the Degree of subharmonics (DSH) analysis of one 30 ms block of diplophonic phonation. Subplots 2 and 3 show sgn-coded signals. Subplots 3 and 4 show their autocorrelation functions together with their cycle length estimates. In sgn-coding version one (K_p = 0.78, subplots 2 and 4) a sub-periodicity is dominant and an artificial cycle length is obtained. In sgn-coding version two (K_p = 0.45, subplots 3 and 5) a shorter cycle length is obtained, which suggests the presence of subharmonics.

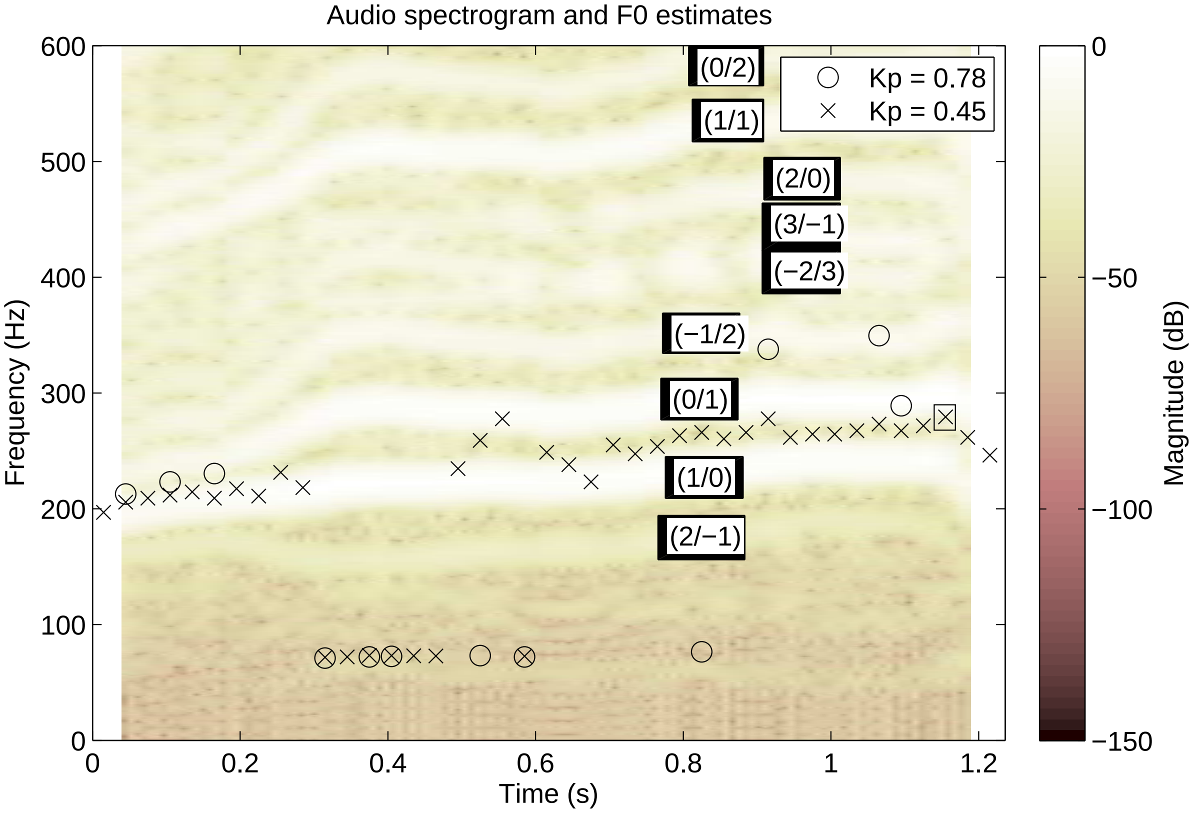


Figure 5.12: Diplophonic phonation: audio spectrogram and fundamental frequency estimates (from Degree of subharmonics measurement). The rectangle at 1.17 s marks the analysis block of figure 5.11.

5.1.6 Discussion and conclusion

Several available features have been evaluated with respect to their ability to detect diplophonia. The tested features were jitter, shimmer, the Harmonics-to-Noise Ratio (HNR), the Göttingen features and the DSH. Jitter, shimmer and the HNR have been calculated with Praat [85], the Göttingen features with the available release of the Göttingen Hoarseness Diagram [15] and the DSH with a custom MATLAB implementation [58].

The majority of the features enabled building cut-off threshold classifiers that yield promising performance measures. However, the validity of the features may be questioned, because they rely on the assumption that exactly one oscillator produces the observed waveform. If this assumption is violated, the feature values are incorrect and they fail to report valid results, risking misinterpretations.

The DSH is the only feature that considers two fundamental frequency estimates within one analysis block. Although their estimates are wrong, the procedure detects cyclic peak height fluctuations and thus delivers valid information on the presence of diplophonia. The DSH is used in section 5.3 when constructing a reference standard.

5.2 The Diplophonia Diagram

Except for the DSH, all features investigated in section 5.1 assume that the signal under test has either one or no detectable fundamental frequency, which is not true for diplophonic voice. The present section explains how taking into account two additive oscillators in a signal model enables automated detection of diplophonia. The model structure discussed in section 4.3 is used to obtain a Diplophonia Diagram [60].

5.2.1 Binomial logistic regression

Diplophonia is detected by means of binomial logistic regression [120], which is trained to predict the dichotomic class labels from continuous predictors. The class labels are the voice quality annotations described in section 2.2.2, and the euphonic and the dysphonic group are pooled into one non-diplophonic group. The predictors are the modeling error measures of the best one-oscillator waveform model and the best two-oscillator waveform model (section 4.3.2). 278 phonation intervals are used for testing and evaluation. 96 of the intervals are euphonic, 55 are diplophonic and 127 are dysphonic (corpus ID 8).

The errors of the optimal one-oscillator and two-oscillator waveform models relRMSE(S_1) and relRMSE(S_2) are negative numbers (section 4.3.3). The synthesis quality of the best one-oscillator (SQ_1) and the best two-oscillator models (SQ_2) are obtained by inverting the sign.

$$\mathrm{SQ}_1 = -\mathrm{relRMSE}(S_1), \quad \mathrm{SQ}_2 = -\mathrm{relRMSE}(S_2) \tag{5.5}$$

The regression coefficients B_1, B_2 and B_3 are obtained by inserting the dichotomic class labels and the corresponding synthesis qualities SQ_1 or SQ_2 into the MATLAB subroutine mnrfit. The non-diplophonic group is the reference group. The log odds for the interval under test to be diplophonic is expressed by equation 5.6 [121]. B_1 is the intercept coefficient that expresses the log odds for diplophonia if both predictors SQ_1 and SQ_2 are zero. B_2 and B_3 express how much the log odds of diplophonia change if SQ_1 or SQ_2 increase by 1 dB. Equation 5.7 gives the probability of an interval to be diplophonic.

$$\ln\left(\frac{P(d)}{1-P(d)}\right) = B_1 + B_2 \cdot \mathrm{SQ}_1 + B_3 \cdot \mathrm{SQ}_2 \tag{5.6}$$

$$P(d) = 1 - \frac{1}{1 + \mathrm{e}^{(B_1 + B_2 \cdot \mathrm{SQ}_1 + B_3 \cdot \mathrm{SQ}_2)}} \tag{5.7}$$

A 10-fold cross validation is applied. The corpus is split up randomly into 10 approximately equally large subsets. 10 regression models are trained on 9 subsets and tested on the remaining one. The final model coefficients and the classification performances are obtained by averaging over the 10 models. Table 5.1 shows coefficients B_1, B_2 and B_3 together with their confidence intervals. The confidence intervals are +/- 1.96 times the average model coefficient standard errors that are returned by mnrfit. They exclude the value 0, therefore all coefficients have a significant effect on the predicted probability of diplophonia. To obtain the probability of the presence of diplophonia $P(d)$ from the predictors SQ_1 and SQ_2, the coefficients are inserted into equation 5.7. B_1 is negative, and so the predicted probability is small for $\mathrm{SQ}_1 = 0$ and

$SQ_2 = 0$. It becomes even smaller when SQ_1 is increased, because B_2 is negative. In contrast, B_3 is positive and an increase in SQ_2 increases the probability. The predicted probability with respect to the predictors SQ_1 and SQ_2 is color coded in figure 5.13.

	Coefficient	CI
B_1	-2.719	[-4.217, -1.221]
B_2	-1.205	[-1.589, -0.822]
B_3	1.612	[1.049, 2.175]

Table 5.1: The estimated model coefficients of the binomial logistic regression and their 95 % confidence intervals (CI). The confidence intervals are the coefficients +/- 1.96 times the average standard errors that are returned by the MATLAB function mnrfit.

	Estimate	CI
SE	76.4 %	[63.0 %, 86.8 %]
SP	98.7 %	[96.1 %, 99.7 %]
ACC	94.2 %	[90.8 %, 96.7 %]
PR	19.8 %	[15.3 %, 25.0 %]
PPV	93.3 %	[81.7 %, 98.6 %]
PLR	58.8	[18.6, 186.2]
NPV	94.4 %	[90.6 %, 97 %]
NLR	4.2	[2.6, 6.7]

Table 5.2: Performance measures of the logistic regression model. 95 % confidence intervals (CI).

5.2.2 Evaluation and discussion

The logistic regression model is evaluated on the corpus that has been used for coefficient estimation and 10-fold cross validation. Figure 5.13 shows the scatter plot of SQ_1 and SQ_2, together with the color coded probability for the presence of diplophonia $P(d)$. All information relevant for detecting diplophonia can be found in the ranges [0 15 dB], therefore the axes of the Diplophonia Diagram are bounded. The markers denote the positions of the phonation intervals in the feature space, where each marker is located at interval averages. x markers denote euphonic voice, o markers denote diplophonic voice, and + markers denote dysphonic voice. The dotted line is the line of equal synthesis quality, i.e., models for intervals lying on this line neither do improve nor degrade if a second oscillator is considered. Models of diplophonic intervals often profit from adding the seconds oscillator to the waveform model. Therefore their markers are situated above the line of equal synthesis quality. Waveform models of non-diplophonic intervals mostly degrade if a second oscillator is added and so their markers are situated below the line of equal synthesis quality.

The decision border is obtained by selecting the most probable class for each location across the data space (P(d) = 0.5). It is slightly tilted with respect to the line of equal synthesis quality. The tilt either stems from random variations that are not accounted for by the corpus or from a systematic effect. The systematic difference of the intercept B_1 from zero is an argument for a systematic effect on the line tilt. A dysphonic cluster of noisy intervals with very low SQ_1 and SQ_2 exists, because the signal model is not valid for those. These intervals lie by chance above or below the line of equal synthesis quality, which flattens the decision border. In addition, no non-diplophonic intervals with large SQ_1 and SQ_2 exist, which further flattens the decision border. The sparseness of non-diplophonic intervals with large SQ_1 and SQ_2 is due to the fact that non-diplophonic voices that obtain a high SQ_1 also obtain a low SQ_2, because non-diplophonic signals cannot be appropriately modeled by two independent oscillators in the proposed procedure.

Figure 5.14 shows the boxplots for the predicted probabilities of diplophonic and non-diplophonic phonation intervals. Non-diplophonic intervals have low probabilities and diplophonic intervals have high probabilities, which confirms the validity of the regression model.

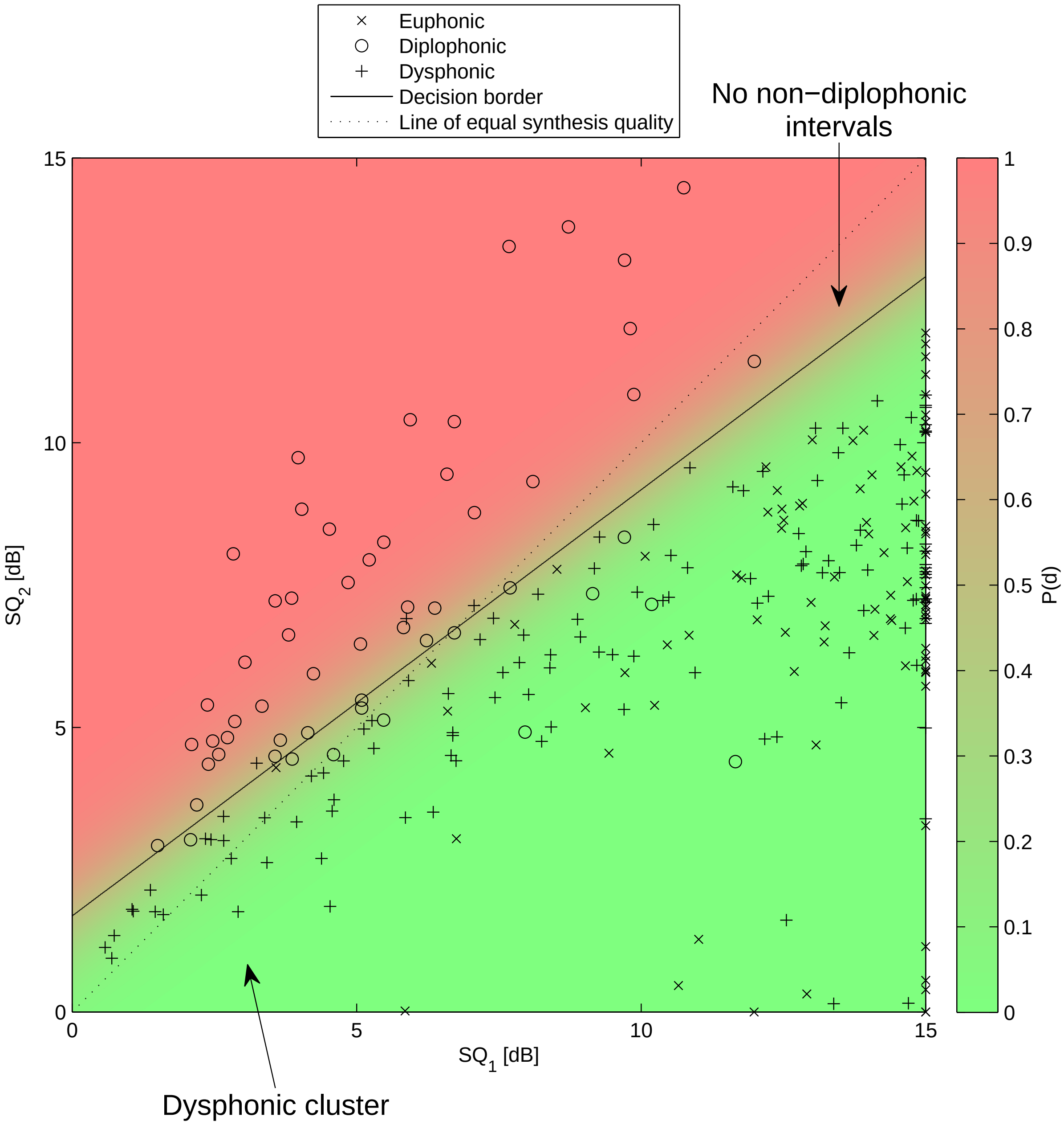


Figure 5.13: The Diplophonia Diagram. P(d): predicted probability of diplophonia.

Table 5.3 shows the diagnostic table of the Diplophonia Diagram classifier. One observes 42

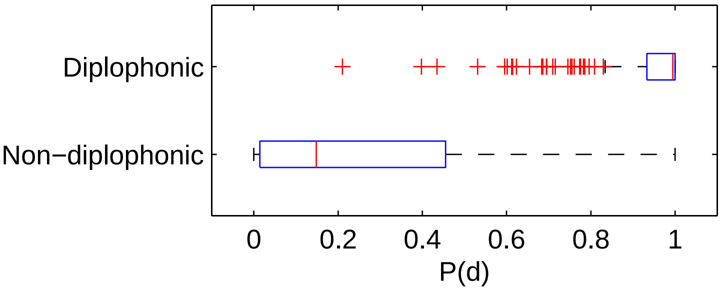


Figure 5.14: Predicted probability of diplophonia versus observed presence of diplophonia.

true positives, 220 true negatives, 3 false positives and 13 false negatives. Table 5.2 summarizes the performance measures, which have been introduced in chapter 1. The following observations must be taken into account should the test be used to trigger clinical actions.

		Diplophonia	
		Present	Absent
P(d)	> 0.5	42	3
	< 0.5	13	220

Table 5.3: Diagnostic table of the Diplophonia Diagram.

The sensitivity is 76.4 %, which means that approximately 3 out of 4 diplophonic intervals are detected by the test. The specificity is 98.7 %, which means that only 13 out of 1000 non-diplophonic phonation intervals test positive. The relation between the sensitivity and specificity must be interpreted with respect to the prevalence. If the target condition is rather rare, as is the case for diplophonia, false positive tests must be avoided. Otherwise it would be more likely that a positively tested interval is a false positive than a true positive and the accuracy of the test would decrease. The positive predictive value is 93.3 %, which means that from 1000 positively tested intervals, 933 are truly diplophonic. If the target condition would be rather common, 93.3 % would not be satisfying. For example, if the pretest probability would have been 90 %, there would not be much benefit from administering the test. The positive likelihood ratio takes into account this circumstance and reports how the odds change when the test is administered. The positive likelihood ratio of 58.8 means that the odds for the presence of diplophonia increase 58.8-fold if an interval receives a positive test result. The pretest odds are approximately 1:4, i.e., 1 out of 5 intervals is diplophonic. The posttest odds are much higher, namely approximately 14:1, i.e., from 15 intervals that receive a positive test result 14 are truly diplophonic. The negative predictive value and the negative likelihood ratios play similar roles when the absence of diplophonia should be detected. These numbers are given here for completeness sake and are relevant for treatment effect measurement, which is beyond the scope of this thesis.

5.2.3 Evaluation on a critical corpus

The Diplophonia Diagram is now validated on a restricted corpus. All available diplophonic voice intervals are used as the positive group and only dysphonic intervals with $SQ_1 < 12\,\mathrm{dB}$ are used as the negative group (corpus ID 98). The negative group constitutes a critical subgroup of dysphonia, namely waveforms with high levels of short-term modulations (not including beat frequency phenomena) and/or additive noise, or severe dysphonia perceptually. It is shown that the conventional features fail to produce relevant group effects, and that the Diplophonia Diagram is capable of distinguishing diplophonia from the severe non-diplophonic dysphonia.

Figure 5.15 shows the Diplophonia Diagram with green circles for diplophonic intervals, blue circles for dysphonic intervals, a contour plot of the regression's predicted probability (solid lines) and the line of equal synthesis qualities (dotted). The predicted probabilities are obtained from a binary logistic regression model that has been trained on this critical corpus (cf. section 5.2.1). The model coefficients happen to be similar to the ones of the original model ($B_1 = -2.8512$, $B_2 = -1.0332$ and $B_3 = 1.5027$), which suggests that the selected negative group contains all relevant data, or all support vectors. The decision border is tilted with respect to the line of equal synthesis quality, similarly to the tilt of the original regression model. The sensitivity estimate is 80.0 % and the specificity estimate is 92.9 %. Table 5.4 illustrates the performance of the Diplophonia Diagram.

Figures 5.16 to 5.21 summarize the results of the conventional features. Their AUCs are below 0.7, which suggests that their application on the critical corpus is questionable. The DSH achieves the highest sensitivity of all conventional features (69.1 %) and the Jitter achieves the highest specificity (80.3 %). The waveform modeling is superior to the conventional features. Table 5.5 summarizes the sensitivities and specificities of all approaches.

		Diplophonia	
		Present	Absent
P(d)	> 0.5	44	5
	< 0.5	11	65

(a) Category table.

	Estimator	CI
SE	80.0 %	[67.0 %, 89.6 %]
SP	92.9 %	[84.1 %, 97.6 %]
ACC	87.2 %	[80.0 %, 92.5 %]
PR	44.0 %	[35.1 %, 53.2 %]
PPV	89.9 %	[77.8 %, 96.6 %]
PLR	11.3	[4.8, 26.6]
NPV	85.5 %	[75.6 %, 92.5 %]
NLR	4.6	[2.7, 7.9]

(b) Performance measures.

Table 5.4: Performance of the Diplophonia Diagram on the critical corpus. 95 % confidence intervals (CI).

	SE, CI [%]	p-value	SP, CI [%]	p-value
Jitter	50.9, [37.1, 64.6]	n.s.	80.3, [68.7, 89.1]	n.s.
Shimmer	67.3, [53.3, 78.6]	n.s.	69.7, [57.1, 80.4]	0.022
Irr	57.8, [42.2, 72.3]	n.s.	64.0, [42.5, 82.0]	0.025
Noise	60.0, [38.7, 78.9]	n.s.	77.8, [62.9, 88.8]	n.s.
HNR	60.0, [47.6, 71.5]	n.s.	58.2, [44.1, 71.3]	0.002
DSH	69.1, [55.2, 80.9]	n.s.	57.1, [44.8, 68.9]	< 0.001
DD	80.0, [67.0, 89.6]	-	92.9, [84.1, 97.6]	-

Table 5.5: Comparison of all tested features on the critical corpus. The p-values are obtained by a Chi^2-test for equal proportions, six-fold Bonferroni corrected for adjustment of the multiple testing effect [55]. Göttingen irregularity feature (Irr), Göttingen noise feature (Noise), Harmonics-to-Noise Ratio (HNR), Degree of subharmonics (DSH), Diplophonia Diagram (DD), 95 % confidence intervals (CI).

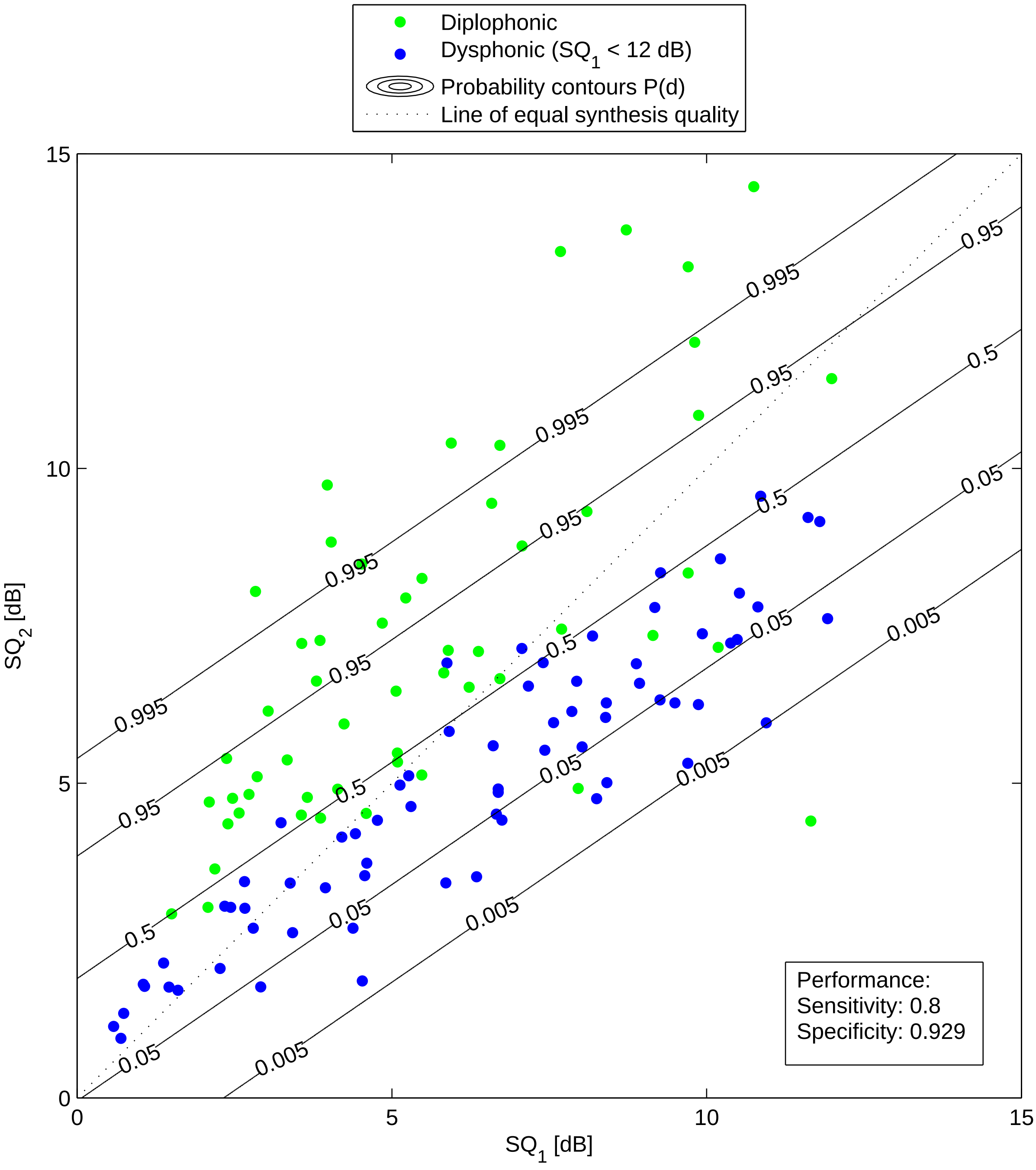
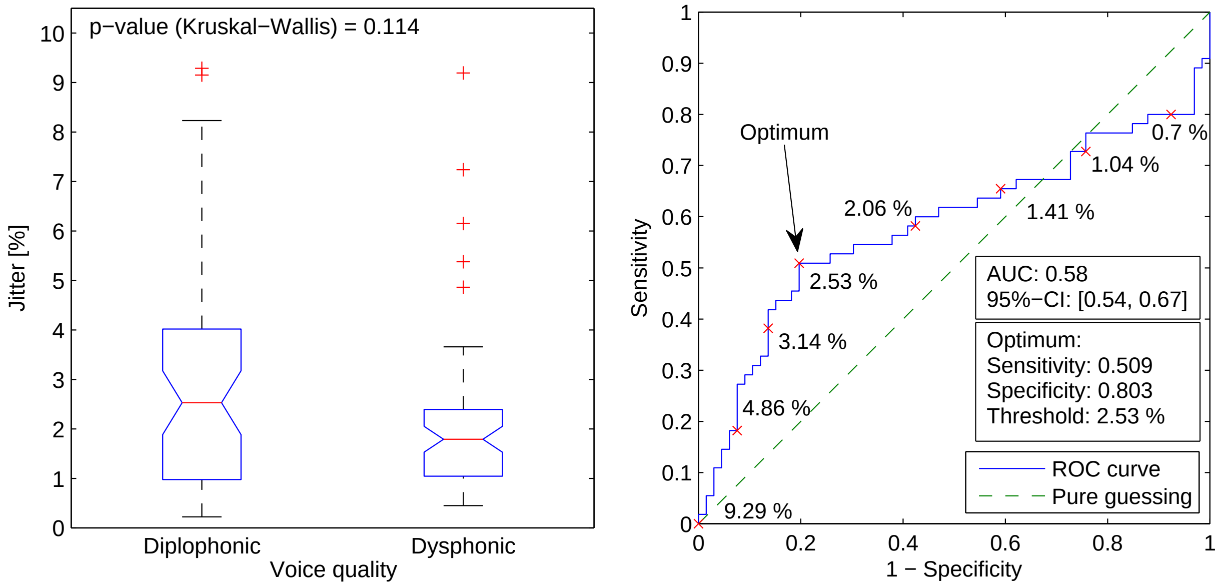
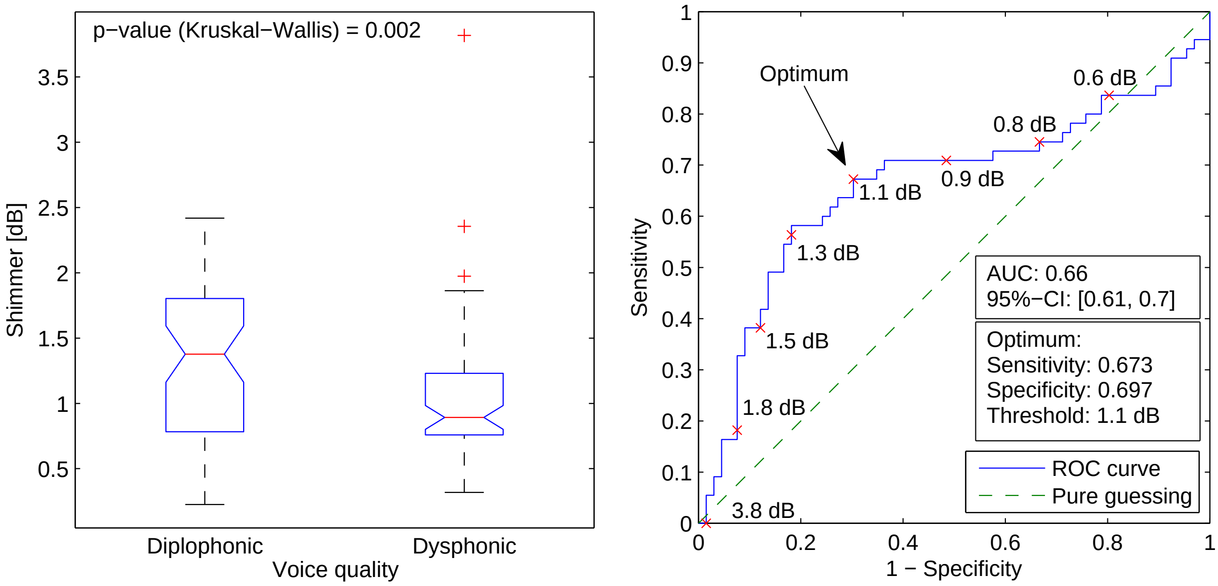


Figure 5.15: The Diplophonia Diagram, critical corpus.



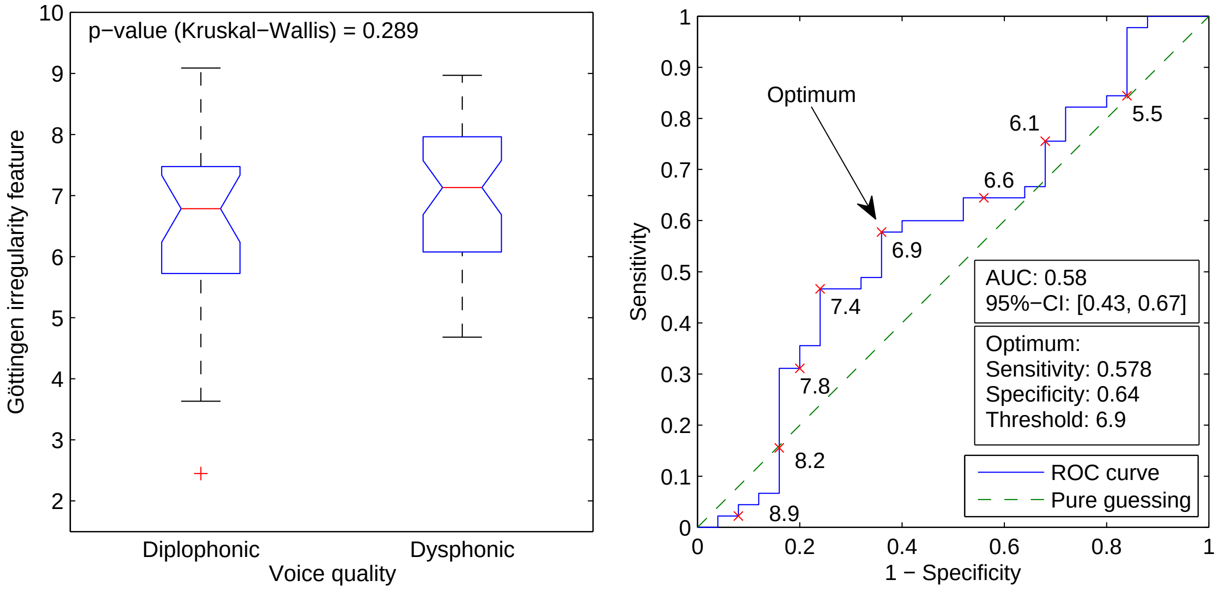
(a) Boxplots. *(b)* ROC curve.

Figure 5.16: Evaluation of jitter versus voice quality, critical corpus.



(a) Boxplots. *(b)* ROC curve.

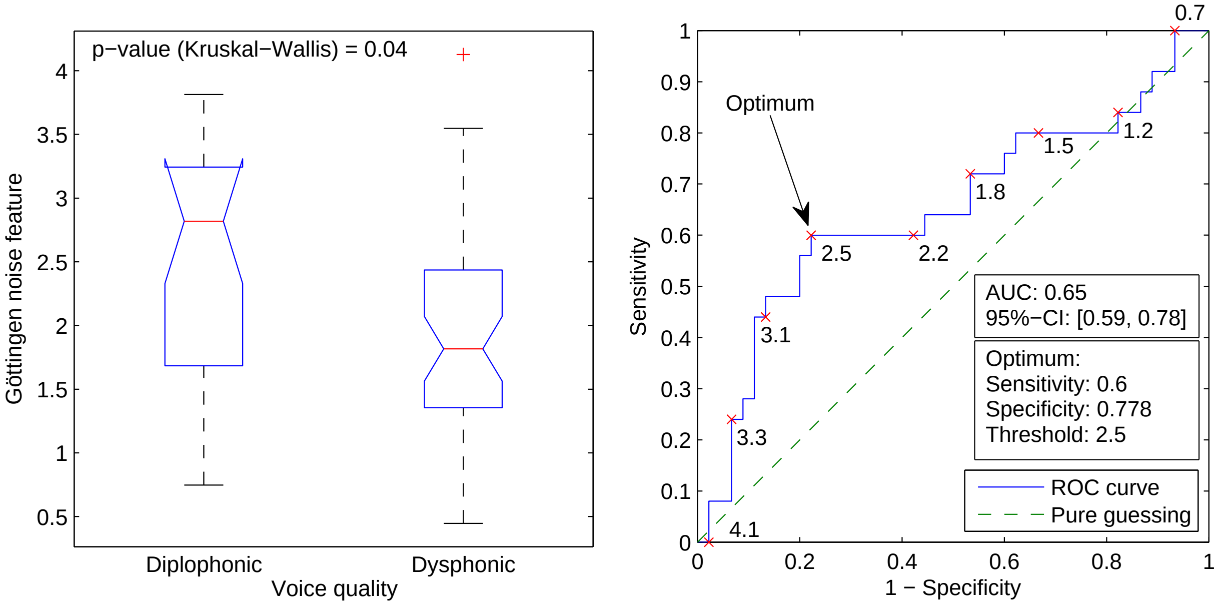
Figure 5.17: Evaluation of shimmer versus voice quality, critical corpus.



(a) Boxplots.

(b) ROC curve.

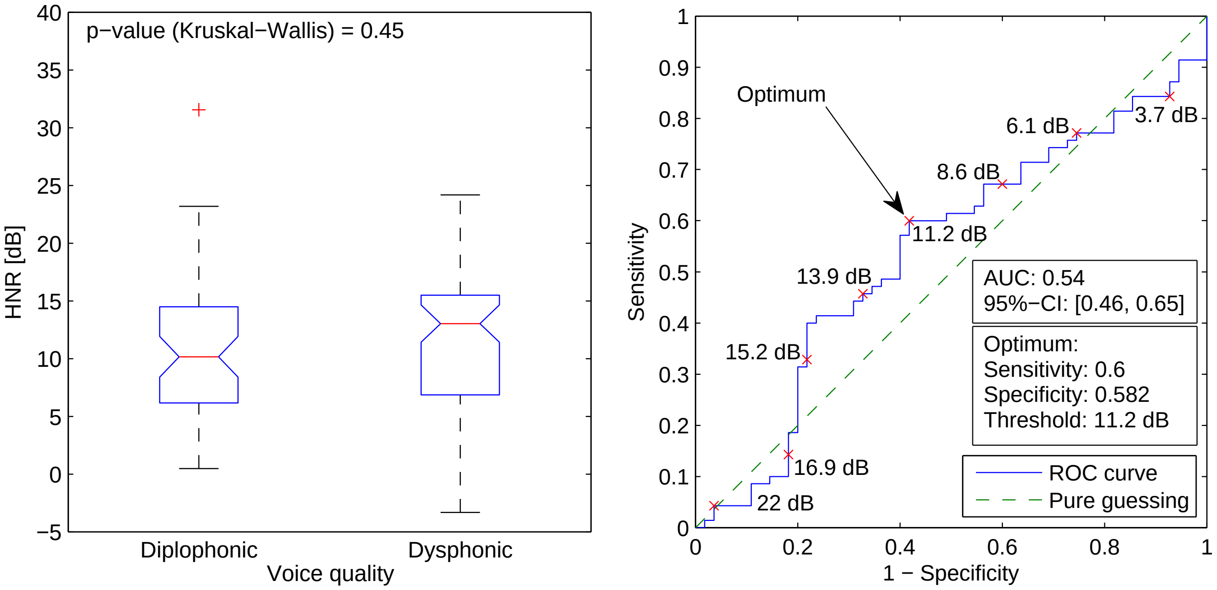
Figure 5.18: Evaluation of the Göttingen irregularity measure versus voice quality, critical corpus.



(a) Boxplots.

(b) ROC curve.

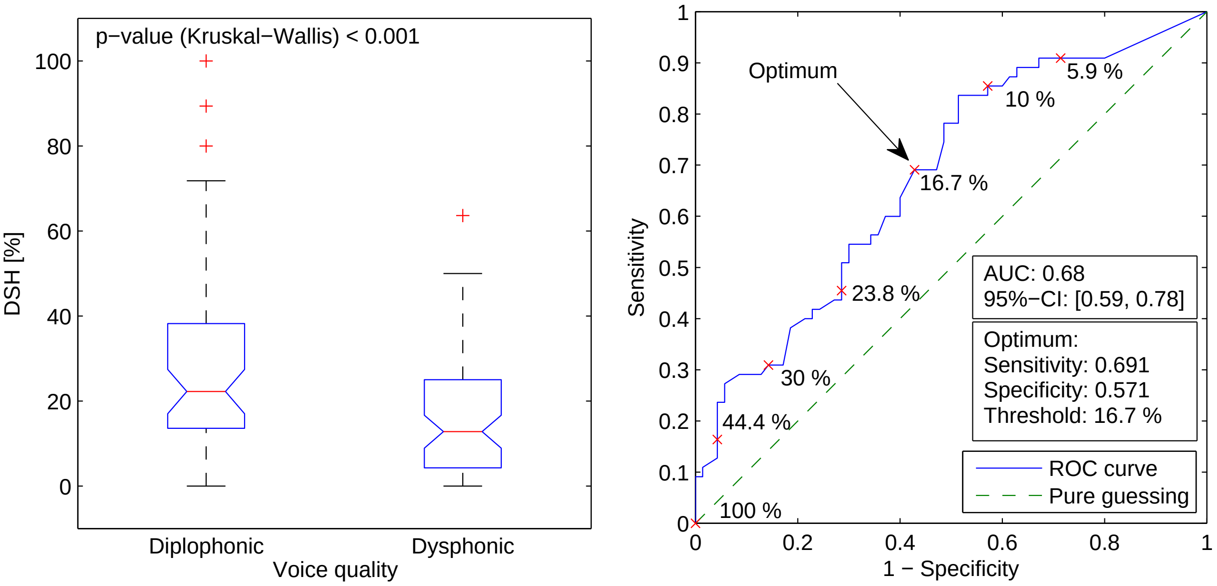
Figure 5.19: Evaluation of the Göttingen noise measure versus voice quality, critical corpus.



(a) Boxplots.

(b) ROC curve.

Figure 5.20: Evaluation of Praat's Harmonics-to-Noise Ratio (HNR) versus voice quality, critical corpus.



(a) Boxplots.

(b) ROC curve.

Figure 5.21: Evaluation of the Degree of subharmonics (DSH) versus voice quality, critical corpus.

5.3 Latent class analysis

One aim of this thesis is to provide a characterization of diplophonia that enables automated detection. Several tests that report diplophonia have been evaluated so far. They use voice quality annotation of phonation intervals with homogeneous voice quality as a gold standard or ground truth, which may be further questioned because the rating is subjective. One may ask now: "Does auditively detectable diplophonia exist, that cannot be detected acoustically?" and "Does diplophonia exist that can be detected acoustically, but not auditively?". Another question that is addressed is: "Having applied several tests to a signal interval, what test outcome combinations do reliably indicate the presence of diplophonia?"

Scientific guidelines exist for conducting diagnostic accuracy studies if no gold standard is available [122]. The test that is expected to provide the most accurate decision with regard to some target condition is referred to as the "gold standard" or the "reference standard". The reference standard may be an expensive procedure, which can mean that 1) the needed devices are costly, 2) highly specialized experts are needed for conducting the test, 3) the procedure is time-demanding, 4) the procedure is displeasing for the patients, or 5) the procedure exposes the patients to some additional risk. The index test in contrast should be less expensive with respect to the enumerated aspects but still provide results that are comparable to the reference standard with some acceptable limit of agreement.

In the past, it was hypothesized that laryngeal high-speed videos would enable the development of a reference standard that could be used to evaluate less expensive index tests. It turned out that the analysis of laryngeal high-speed videos is difficult because of the high amount of user intervention and error proneness. Another possibility for the development of a new reference standard is the recruiting of an expert panel. The experts judge the presence of the target condition, which may be based on several sources of information. Standardized procedures for acquiring expert panel decisions exist, but are currently under discussion [122]. As described in chapter 2, four expert discussions on the definition of diplophonia were organized. The voice quality annotation that serves as ground truth throughout the thesis is based on the outcome of these discussions. Its possible imperfections are addressed in this section.

This section explains how the results of several imperfect index tests can be used to construct a reference standard. Latent class analysis estimates the prevalence of the target condition within the given corpus, as well as the sensitivities and specificities of the individual index tests [122]. Information from the voice quality annotation, the Diplophonia Diagram and the DSH are combined from a probabilistic point of view. It is shown that the most probable interpretation of the observed data distribution is that the voice quality annotation is very close to the truth.

5.3.1 Model description

The model and the parameter estimation has been introduced in [123]. The model parameters are the prevalence, the sensitivities and the specificities. They are estimated by maximizing the log-likelihood function of the model with respect to the observed data Y_{ijk}. Y_{ijk} is 3-dimensional and binary, where i is the phonation interval index, j is the test index and k the test outcome index. Y_{ijk} is 1 if the i^{th} interval put to the j^{th} test produces the outcome k, and 0 otherwise. Let r be the class index of the latent variable, i.e., the presence of diplophonia, and let π_{jrk} denote the class conditional probability, i.e., the probability that the j^{th} test returns outcome k if the interval under test is from class r. π_{jrk} is the matrix of true positive rate (sensitivity), true negative rate (specificity), false positive rate (1-specificity) and false negative rate (1-sensitivity).

Equation 5.8 calculates the probability that interval i of class r produces a certain combination of test outcomes. The equation is equivalent to the $\cap$ operator for the joint probability. For the used corpus and tests, i goes from 1 to 278 (the intervals), r is 1 or 2 (diplophonic or non-diplophonic), J is 3 (the number of tests) and K is 2 (the number of possible outcomes of each test).

$$f(Y_i; \pi_r) = \prod_{j=1}^{J} \prod_{k=1}^{K} (\pi_{jrk})^{Y_{ijk}} \tag{5.8}$$

The mixing proportions p_r, i.e., the prevalence and 1 - prevalence are now introduced, and the probability density function across all classes is derived. The probability density function denotes the probability of observing a certain combination of test outcomes, given the test performance measures in π and the prevalence of diplophonia. Given a certain combination of test outcomes, the probability for the presence of diplophonia is estimated using Bayes' theorem (equation 5.10).

$$P(Y_i|\pi, p) = \sum_{r=1}^{R} p_r \cdot f(Y_i; \pi_r) \tag{5.9}$$

$$\hat{P}(r|Y_i) = \frac{\hat{p}_r \cdot f(Y_i; \hat{\pi}_r)}{\sum_{q=1}^{R} \hat{p}_q \cdot f(Y_i; \hat{\pi}_q)} \tag{5.10}$$

5.3.2 Parameter estimation

The model parameters p_r, and π_{jrk} are estimated by maximizing the log-likelihood function ($\log L$), which is a measure of expectation. It measures the likelihood that the given data distribution has been produced by the modeled random process. The model with the maximal $\log L$ is considered to be optimal.

$$\log L = \sum_{i=1}^{N} P(Y_i|\pi, p) \tag{5.11}$$

The maximization of $\log L$ is achieved using the EM-algorithm [124]. The model parameters are initialized with random numbers between 0 and 1, taking care of the restrictions that $\sum p_r = 1$ and $\sum_k \pi = 1$. In the expectation step, the initial values are inserted into the above equations, to calculate $\log L$ and to estimate the probability for the presence of diplophonia, given a certain combination of test outcomes $\hat{P}(r|Y_i)$. In the maximization step, a new estimate for the mixing proportions and new estimates for the test performance measures are obtained. The new estimates are used to calculate new values for $\log L$ and $\hat{P}(r|Y_i)$. These two steps are iterated until the increase in $\log L$ is smaller than some arbitrary small threshold, which stops the iteration.

$$\hat{p}_r^{\text{new}} = \frac{1}{N} \cdot \sum_{i=1}^{N} \hat{P}(r|Y_i) \tag{5.12}$$

$$\hat{\pi}_{jrk}^{\text{new}} = \frac{\sum_{i=1}^{N} \left\{ Y_{ijk} \cdot \hat{P}(r|Y_i) \right\}}{\sum_{i=1}^{N} \hat{P}(r|Y_i)} \tag{5.13}$$

This procedure for estimating the prevalence of diplophonia and the performance measures of the diagnostic tests is applied to a corpus of 278 phonation intervals, from which 127 are euphonic, 55 are diplophonic and 96 are dysphonic (corpus ID 8). The intervals are tested by means of voice quality annotation (section 2.2.2), a DSH classifier (section 5.1.5) and the Diplophonia Diagram (section 5.2).

The DSH classifier is a cut-off threshold classifier. In section 5.1, the optimal threshold separating diplophonic from non-diplophonic intervals is 11.6 %. The ground truth on which this optimal threshold has been estimated is not necessarily the true ground truth. Therefore numerous models are fitted on data distributions achieved from different DSH thresholds.

Figure 5.22 shows the distance D in percent for the three tests with respect to the DSH threshold. D is $\sqrt{(1-SE)^2 + (1-SP)^2}$, i.e., a classifiers' distance from a perfect test. The threshold at the minimal distance is indeed 11.5 % and practically equal to the previous optimal threshold. This suggests that the previous ground truth was appropriate. At the optimal threshold, the DSH contributes to the new ground truth, but its performance is weaker than the ones of the other two tests. The estimated model parameters are given in table 5.8.

Figure 5.23 shows the log-likelihood function $\log L$ with respect to the number of iterations. The function starts at $-\infty$, and increases monotonically. The function saturates and its increase falls below 0.01 at iteration 18, which stops the optimization. The $\log L$ converges to a value of -327.4.

In the beginning of the optimization, the model parameters were initialized randomly. The optimization iteratively updates the parameter estimates by executing deterministic update rules. The model parameters converge to a local optimum, which is not necessarily the global optimum. To avoid suboptimal parameter estimates at local maxima, the optimization is repeated several times, each with different starting values. All models are globally optimal, because the differences are negligible in practice (table 5.6).

Maximum log-likelihood	Number of occurrences
-327.6	36
-327.4	64

Table 5.6: Outcomes of 100 repetitions of the maximum log likelihood estimation.

The three evaluated tests have two possible outcomes (diplophonic, non-diplophonic). However, the number of latent classes is not necessarily two. A latent variable with less than or more than two classes may explain the observed data better than a latent variable with two classes.

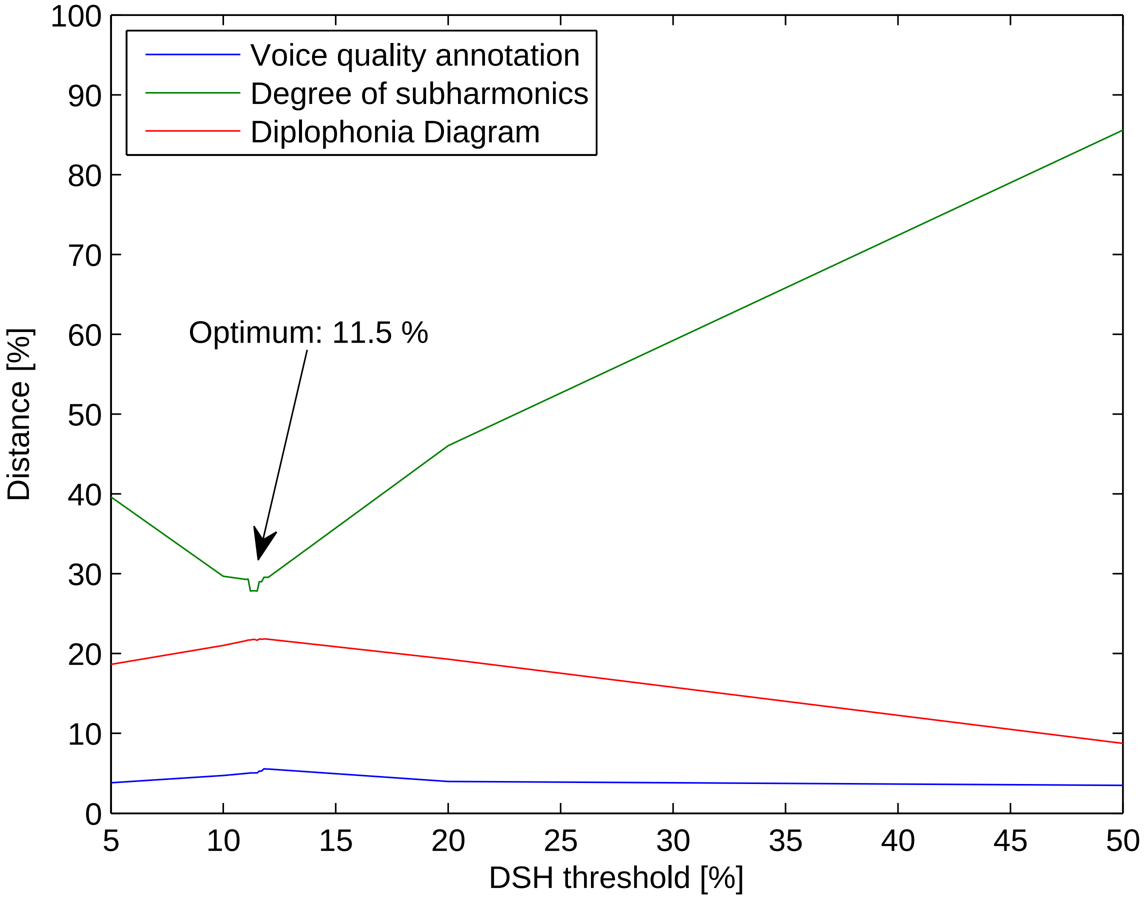


Figure 5.22: Distance of three tests from the virtual optimum in the ROC curve for different Degree of subharmonics (DSH) thresholds.

To test which model is the best for explaining the data, the Bayesian information criterion (BIC) is used as a measure of the goodness of fit [125]. The BIC considers the maximal log likelihood Δ, the number of observations N and the number of estimated parameters Φ, i.e., it receives a penalty for each parameter to estimate. As the number of estimated parameters grows with a growing number of latent classes, models with more latent classes need to produce much better results to survive. Figure 5.24 shows the BIC with respect to the number of latent classes and suggests that two latent classes are optimal.

$$\mathrm{BIC} = -2 \cdot \Delta + \Phi \cdot \ln(N) \tag{5.14}$$

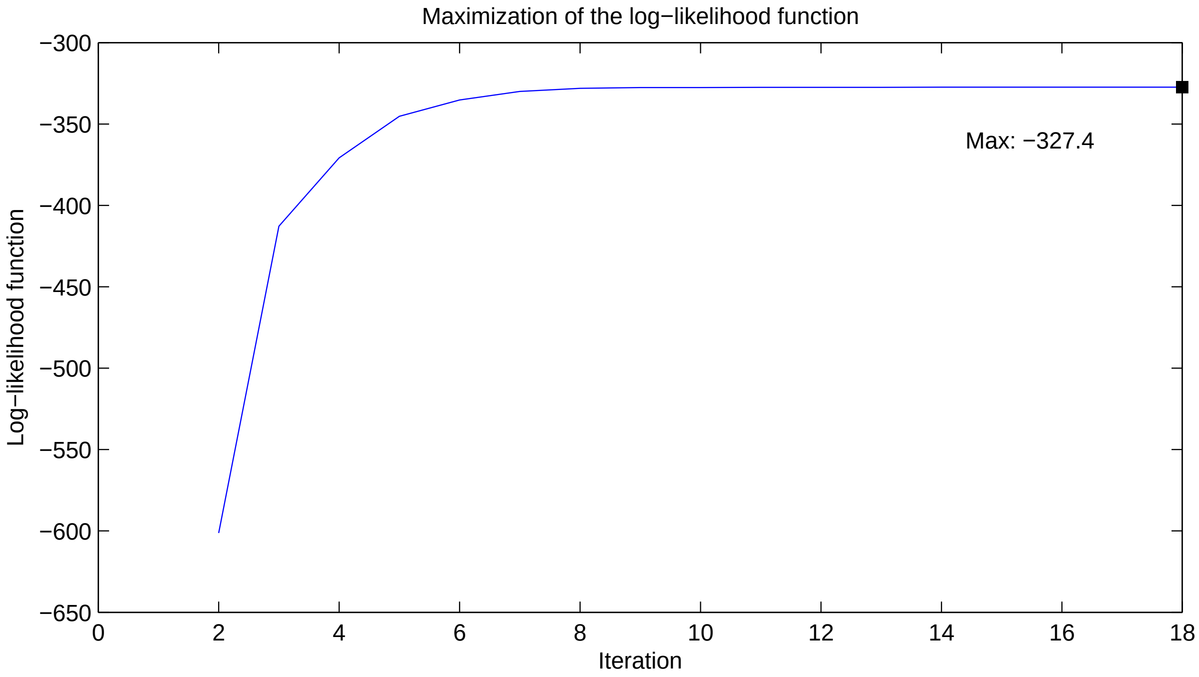


Figure 5.23: Log-likelihood function $\log L$ with respect to the maximization iteration.

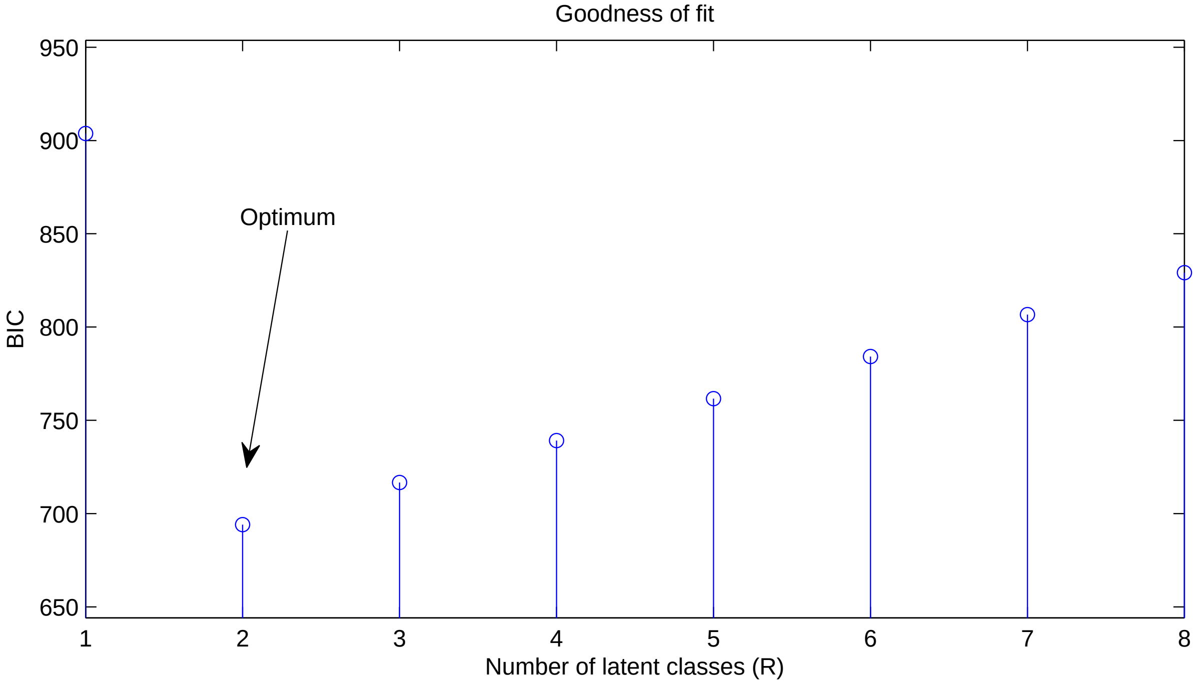


Figure 5.24: Bayesian information criterion (BIC) for different numbers of latent classes.

5.3.3 Evaluation and validation

Table 5.7 shows the data distribution, i.e., the numbers of observed intervals with respect to all possible test outcome combinations. Label 1 denotes non-diplophonic/negative and label 2 denotes diplophonic/positive. The majority of all intervals (170) are tested negative by all tests. The tests agree positively on 36 intervals. All other intervals obtain divergent test results from different tests. The fifth column shows the estimated probability for the presence of diplophonia obtained from latent class analysis, given the respective test outcome combination. Most observations obtain definite probability estimates above 90 % or below 10 %. The sixth column shows the new ground truth, which is derived from $\hat{P}(2)$. Only if $\hat{P}(2)$ is greater than 0.5, the test combination is considered to detect diplophonic phonation. Only two intervals that were annotated as diplophonic are acutally non-diplophonic (test outcome combination 1 - 2 - 2) and all other intervals were correctly annotated.

Test outcomes					
VoQA	DSH	DD	# intervals	$\hat{P}(2)$	New ground truth
1	1	1	170	0.0 %	1
1	1	2	1	41.6 %	1
1	2	1	50	1.1 %	1
1	2	2	2	92.6 %	2
2	1	1	2	94.9 %	2
2	1	2	7	100.0 %	2
2	2	1	10	99.7 %	2
2	2	2	36	100.0 %	2

Table 5.7: The numbers of intervals, their predicted probability for the presence of diplophonia $\hat{P}(2)$ and their new ground truth from latent class analysis with respect to all possible test outcome combinations. VoQA: voice quality annotations, DSH: Degree of subharmonics, DD: Diplophonia Diagram; 1 ... non-diplophonic, 2 ... diplophonic.

Table 5.8 summarizes the estimated sensitivities and specificities for all tests. The voice quality annotation is superior, the Diplophonia Diagram shows an excellent specificity and a fair sensitivity. The DSH has a slightly higher sensitivity, but a lower specificity. The prevalence of diplophonia in the corpus was estimated to be 20.8 %. Tables 5.9 to 5.11 are the diagnostic tables and the performance measure tables for the three tests.

Test	**Sensitivity**	**Specificity**
VoQA	95.0 %	100 %
DD	78.4 %	99.7 %
DSH	83.7 %	77.5 %

Table 5.8: Estimated parameters in latent class analysis. VoQA: voice quality annotations, DSH: Degree of subharmonics, DD: Diplophonia Diagram. The prevalence is estimated to be 20.8 %.

In a situation in which diplophonia must be detected automatically without the aid of expert annotations, only the Diplophonia Diagram and the DSH are available. Table 5.12 shows the number of observed intervals with respect to all possible test outcome combinations and the outcome of an optimal test. The outcome of the Diplophonia Diagram agrees with the optimal test outcome. Thus for the used sample, no need for additionally testing the DSH exists if the Diplophonia Diagram is available.

		New ground truth	
		Diplophonic	Non-diplophonic
VoQA	Diplophonic	55	0
	Non-diplophonic	2	221

(a) Diagnostic table.

	Estimate	CI
SE	96.5 %	[87.9 %, 99.6 %]
SP	100.0 %	[98.3 %, 100.0%]
ACC	99.3 %	[97.4 %, 99.9 %]
PR	20.5 %	[15.9 %, 25.7 %]
PPV	100.0 %	[93.5 %, 100.0%]
PLR	∞	-
NPV	99.1 %	[96.8 %, 99.9 %]
NLR	28.6	[7.3, 111.7]

(b) Performance measures.

Table 5.9: Diagnostic performance of the voice quality annotation (VoQA), ground truth from latent class analysis. 95 % confidence intervals (CI).

		New ground truth	
		Diplophonic	Non-diplophonic
DD	Diplophonic	45	1
	Non-diplophonic	12	220

(a) Diagnostic table.

	Estimate	CI
SE	78.9 %	[66.1 %, 88.6 %]
SP	99.5 %	[97.5 %, 100.0%]
ACC	95.3 %	[92.1 %, 97.5 %]
PR	20.5 %	[15.9 %, 25.7 %]
PPV	97.8 %	[88.5 %, 99.9 %]
PLR	157.8	[24.4, 1019]
NPV	94.8 %	[91.1 %, 97.3 %]
NLR	4.7	[2.9, 7.8]

(b) Performance measures.

Table 5.10: Diagnostic performance of the Diplophonia Diagram (DD), ground truth from latent class analysis. 95 % confidence intervals (CI).

		New ground truth	
		Diplophonic	Non-diplophonic
DSH	Diplophonic	48	50
	Non-diplophonic	9	171

(a) Diagnostic table.

	Estimate	CI
SE	84.2 %	[72.1 %, 92.5 %]
SP	77.4 %	[71.3 %, 82.7 %]
ACC	78.8 %	[73.5 %, 83.4 %]
PR	19.8 %	[15.3 %, 25.0 %]
PPV	49.0 %	[38.7 %, 59.3 %]
PLR	3.7	[2.8, 4.9]
NPV	95.0 %	[90.7 %, 97.7 %]
NLR	4.9	[2.7, 9.0]

(b) Performance measures.

Table 5.11: Diagnostic performance of the Degree of subharmonics (DSH), ground truth from latent class analysis. 95 % confidence intervals (CI).

Test outcome		# intervals		
DSH	DD	Diplophonic	Non-diplophonic	Optimal test
1	1	2	170	1
1	2	7	1	2
2	1	10	50	1
2	2	38	0	2

Table 5.12: Number of observations with respect to test outcome combinations and the presence of diplophonia (new ground truth). The considered tests are the Degree of subharmonics (DSH) and the Diplophonia Diagram (DD). The voice quality annotation is not available, which favors the Diplophonia Diagram.

5.4 Conclusion

Conventional features such as jitter, shimmer, the Göttingen features, the Harmonics-to-Noise Ratio (HNR) and the Degree of subharmonics (DSH) show evidence for group effects when tested on corpora of audio recordings of a broad variety of voice disorders and euphonic phonation. However, only the DSH considers two fundamental frequencies within one analysis block. It has been demonstrated how the conventional one-oscillator features fail when several oscillations are combined in the same waveform, which decreases their relevance. A critical corpus of severely dysphonic phonation on which conventional features fail was defined. The critical corpus consists of diplophonic intervals and dysphonic intervals with high levels of modulation or additive noise. A large corpus may be more representative of clinical applications, but the critical corpus focusses on situations, in which diplophonia must be distinguished from particular dysphonic phonation intervals.

A diagnostic test based on the waveform modeling error measures has been developed. The test relies on binomial logistic regression and is referred to as the ”Diplophonia Diagram”. It has been shown that the Diplophonia Diagram outperforms conventional features when tested on diplophonic and severely dysphonic intervals. The Diplophonia Diagram may be interpreted physiologically because it reports the presence of secondary additive oscillations in waveforms.

In a final step a probabilistic latent class analysis model has been used to check on the ground truth. It has been shown that the voice quality annotation is the most likely reference, if test results from the Diplophonia Diagram and the Degree of subharmonics are considered in addition. One drawback of the latent class model may be that its clinical relevance may be feeble because of the model's abstract nature. Another drawback may be that the model assumes that the tests' false decisions are independent from each other, i.e., there are no ”difficult” cases on which the test is expected to fail systematically. If on the contrary the tests fail on specific phonation intervals, the constructed reference is biased. This also occurs (to a lesser extent) if subgroups of tests tend to fail together. The appropriateness of the assumption of independence cannot be tested explicitly however.

One may argue that it is not surprising that the voice quality annotation and the Diplophonia Diagram do agree and that the DSH performance is lower. This is true because of two reasons. First, the voice quality annotation and the Diplophonia Diagram have been trained by the same person, and second, the selection of the intervals were auditory-based. Regarding the first argument it is indeed true that the external validity of the procedure would have been higher if the voice quality annotation and the development of the Diplophonia Diagram would have been done by different persons. However, it has been proven that diplophonia is a phenomenon that actually exists and that can be reliably determined by showing strong agreement between the subjective voice quality annotation and several (semi-)automatic procedures.

Suggestions for future work are automatic temporal segmentation of phonation intervals, analysis of sustained vowels in the absence of an endoscope, analysis of connected speech and the development of valid interpretation guidelines for the Diplophonia Diagram. Relevant research questions might be: Can and should we distinguish between classes of diplophonia severity? Does a kind of mild diplophonia exist that does not need treatment? Do cases of diplophonia exist that are improved by treatment but not eliminated? What pre-post treatment difference is significant?

6 Discussion, conclusion and outlook

Hoarseness is the main sign of voice disorders. The term summarizes a wide variety of perceptively altered voice sounds and is thus a cover term including voices perturbed by additive noise and modulation. A more distinguished view on voice disorders is elaborated by investigating diplophonia, which focusses on voices involving two oscillators. Diplophonia is described at different levels, which are the level of glottal vibration, the acoustic level and the auditive level.

Each level of description involves its own definition of diplophonia. The auditive level allows for the most intuitive definition, i.e., the simultaneous presence of two pitches. The term describes how a listener perceives the voice and to a lesser extent refers to glottal conditions. Glottal and non-glottal phenomena are pooled, which blurs the focus of investigation. At the level of glottal vibration, diplophonia is defined as spatially distinct vocal fold oscillations at different frequencies and non-glottal vibrations are excluded. Detecting diplophonia at the level of glottal vibration connects with its distal causes, which is desirable in clinical application. However, geometric modeling suggests that available imaging techniques do not enable distinguishing frequency from phase superior-inferior asymmetry. Thus, the interpretation of laryngeal high-speed videos remains difficult and needs further refinement. The acoustic level enables detecting diplophonia automatically. Diplophonia may be characterized by two additively combined pseudo-periodic oscillations, which is exploited by the Diplophonia Diagram. It has been shown that the Diplophonia connects to auditive diplophonia. The exact relations between glottal and acoustic diplophonia should be investigated in the future.

Titze's pulse counting definition are at a waveform level [52]. It can be applied to any type of waveform, such as acoustic, glottal area, electroglottographic or flow waveforms. Strictly speaking, the Diplophonia Diagram detects Titze's "biphonia" instead of diplophonia, which is the combination of two sound sources. Nevertheless, the proposed method is named Diplophonia Diagram, because the term is clinically well established. With regard to pulse counting, biphonia can be diplophonia, triplophonia, quadruplophonia or multiplophonia. The oscillator definition may be favored over the pulse counting definition, because it has a wider applicability.

A high-quality database of laryngeal high-speed videos with synchronous audio recordings has been built for research purposes. It enables interdisciplinary research with regard to glottal vibration, acoustics and perception. Making the database widely available may increase the quality and quantity of published research. Research on the database can be pursued in two directions. First, as yet undocumented vocal fold phenomena may be revealed by investigating distinct perceptual voice qualities. This direction has been pursued in the presented thesis. Second, novel perceptual attributes may be discovered by investigating distinct vocal fold vibration patterns.

Observed limitations of the collected data are the following. First, recording conditions in a hospital are similar to field conditions, because the environment is not perfectly controllable. Second, if future research investigates perceptual cues under realistic clinical conditions, calibrated in-ear recordings and realistic playback setups would be needed. Third, the sound production conditions during telescopic high-speed video laryngoscopy are not natural. One may wish to base future work on vowels or connected speech, recorded under less obstructing conditions. Finally, the prevalence of diplophonia in the used corpora is most likely not representative of the prevalence in the general population.

Further future work regarding data acquisition concerns data pre-selection, audio to video synchronization and image quality. Regarding pre-selection based on data quality, one may wish to investigate what minimal data quality is needed for what kind of analysis. Especially data quality thresholds enabling automatic spatio-temporal segmentation of the glottal gap are needed. Regarding synchronization, clinically certified laryngeal high-speed cameras need to be equipped with hard wired synchronized high-quality audio recording facilities, because manual synchronization is not always feasible and also time-consuming. Regarding image quality, laryngeal high-speed cameras should use light sources with higher luminosity and sensors with higher spatial resolution, sensitivity and dynamics.

Although the FIB and the FPB report glottal diplophonia, they have been evaluated on auditive diplophonia. The detection performance of the features is fair only, but might be better when evaluated on glottal diplophonia. Ground truth annotations for glottal diplophonia should be obtained in the future.

The interpretation of laryngeal high-speed videos is hampered by the projection of a three-dimensional vibration onto a two-dimensional sensor, which possibly visually masks important information. The community may wish to aim at a technique extracting three-dimensional vibrations in the long run. At first vocal fold surface trajectories and later the trajectories of reference points inside the tissue should be extracted. Before having these novel technologies, two-dimensional representations of relevant three-dimensional vibration patterns should be investigated by means of geometrical modeling. Models may be fitted to in vivo video data to identify three-dimensional vibration patterns. They must first be tested on synthetic data, because it is unclear what types of vibrations patterns form well-posed optimization problems. Additional future work with regard to the interpretation of laryngeal high-speed videos may be the following. First, the spatio-temporal segmentation of the glottal gap from laryngeal high-speed videos needs to be investigated with respect to video quality, user intervention and accuracy. Second, optimal sagittal positions of kymogram lines may be determined automatically by using information theory.

The Diplophonia Diagram is based on model structure optimization, audio waveform modeling and analysis-by-synthesis, which enables a more suitable description of diplophonic signals than conventional hoarseness features. Analysis-by-synthesis and waveform modeling had already been carried out in voice research, but systematic investigation of model structure optimization with respect to perceived voice quality is novel. For diplophonia, the switch between one and two oscillators is crucial. Optimal model structure is a qualitative outcome that may be interpreted physiologically and one may conjecture that model structure optimization is also useful for describing other voice phenomena than diplophonia. The obtained descriptors might be more easily accepted by clinicians than the conventional ones.

The Diplophonia Diagram addresses well known problems of joint multiple fundamental frequency estimation and source separation, but the existing approaches cannot be applied directly. Most of them rely on quantitative a priori knowledge about the sources, which is not available in diplophonia, because the sources do not exist separately. Qualitative assumptions that had to be made are oscillator additivity, independence, uncorrelatedness and unmodulatedness, which may in practice be violated. True source separation relies on the correct extraction of the fundamental frequency, which is not achieved in approximately half of the observed diplophonic analysis blocks. A promising direction of joint multiple fundamental frequency estimation and source separation with only qualitative knowledge about the sources is taken in [113]. Future work may also include estimation of the fundamental frequency from kymographic data, handling of correlated oscillators and tracking the evolution of time-variant model parameters. The Diplophonia Diagram may be extended for modulative oscillator combination in a next step, because some types of diplophonic phonation cannot be modeled with available models. A modulative model may inform about coupling, but quantitative comparisons of modulative and additive models are pending.

Waveform modeling connects to perceptual aspects of diplophonia and results from the mutual masked loudness model for multiple sources may be correlated with the perception of two pitches. Model waveforms may in the future be used for ear training of phoniatricians and logopedists. However, the applicability of the waveform model is limited to harmonic signals and the masking model's validity may be decreased for those. Further research may thus focus on computational auditory scene analysis and auditory stream segregation [21].

In contrario to conventional hoarseness features, the Diplophonia Diagram achieves satisfying classification performance, because it relaxes the one-oscillator assumption. If the one-oscillator assumption is violated, the validity of most conventional hoarseness features is lost. Numbers may then be in a normal or abnormal range, but cannot be interpreted validly. In addition, conventional features are not attended by a warning in case of loss of validity. The only conventional feature that accounts for the existence of two oscillators is the Degree of subharmonics (DSH). The Diplophonia Diagram outperforms the DSH in terms of specificity, accuracy, positive likelihood ratio and also validity in the case of two added oscillator waveforms. The Diplophonia Diagram is therefore favored for automated detection of diplophonia at the acoustic level.

Latent class analysis has been used to assess the appropriateness of the ground truth from the perspective of probabilistic modeling. Information from voice quality annotation, the Diplophonia Diagram and the DSH has been combined. It has been shown that the available voice quality annotation is almost perfect. Still, the obtained ground truth is not absolute, because experiments with multiple listeners may increase external validity, and connected speech may shift the labels of scored audio fragments.

More relevant clinical interpretation guidelines for the Diplophonia Diagram should be developed in future clinical studies. It should be investigated how clinical actions should be influenced by the test outcome to enable more efficient clinical care. It is unclear, if the description of diplophonic intervals is necessary in clinical practice or if subject global description is sufficient. Other research questions may be: ”How many classes of diplophonia can be distinguished perceptually?”, ”What changes in voice quality are relevant in treatment effect evaluation?”, ”Does mild diplophonia exist that does not need treatment?” and ”Does inaudible diplophonia exist that needs treatment?”. The ultimate goal is to quantitatively demonstrate that an individual benefits from administering a diplophonia test.

Acronyms

ACC accuracy.

AQA audio quality annotations.

AUC area under the curve.

BIC Bayesian information criterion.

Cam laryngeal high-speed camera.

CI 95 % confidence intervals.

CM cycle mark.

CMOS complementary metal-oxide-semiconductor.

DD Diplophonia Diagram.

DFT discrete Fourier transform.

DSH Degree of subharmonics.

DSI dysphonia severity index.

EGG electroglottograph.

ENT ear, nose and throat.

EPF energy perturbation factor.

EPQ energy perturbation quotient.

F0 fundamental frequency.

FIB Frequency Image Bimodality.

FN false negatives.

FP false positives.

FPB Frequency Plot Bimodality.

fps frames per second.

GAW glottal area waveform.

GNE Gottal-to-Noise Excitation Ratio.

HM headworn microphone.

HNR Harmonics-to-Noise Ratio.

IEC international electrotechnical commission.

IQ intelligence quotient.

LAN local area network.

LF Liljencrants-Fant.

LM lavalier microphone.

MC metacycle.

MD medical doctor.

MDVP Multi-Dimensional Voice Program.

MWC mean waveform matching coefficient.

NHR Noise-to-Harmonics Ratio.

NLR negative likelihood ratio.

NPV negative predictive value.

OS overlapping sinusoids.

PC personal computer.

PCM pulse code modulation.

PDOS phase-delayed overlapping sinusoids.

PF perturbation factor.

PLR positive likelihood ratio.

POP posttest probability.

PPF period perturbation factor.

PPQ period perturbation quotient.

PPV positive predictive value.

PQ perturbation quotient.

PR prevalence.

PR portable audio recorder.

PRP pretest probability.

PSC pre-selection combinations.

relRMSE relative root mean square error.

ROC receiver operating characteristic.

ROI region of interest.

S subject.

SE sensitivity.

SE sound engineer.

SP specificity.

SQ synthesis quality.

STP spatio-temporal plot.

SVA spectral video analysis.

TN true negatives.

TP true positives.

ViQA video quality annotations.

VoQA voice quality annotations.

WAV waveform file format.

Glossary

This section provides definitions of used terms. For some of them various different definitions exist across disciplines and schools, thus universality is not claimed.

auditive diplophonia Auditively determined diplophonia.

cue "a minor stimulus acting as an indication of the nature of the perceived object or situation" [126], *acoustic cue*: "the perceptual effect of an acoustic correlate" [127].

diagnosis "the art or act of identifying a disease from its signs and symptoms" [126].

disease "A definite pathological process having a characteristic set of signs and symptoms. It may affect the whole body or any of its parts, and its etiology, pathology, and prognosis may be known or unknown" [128]; or "an impairment of the normal state of the living animal or plant body or one of its parts that interrupts or modifies the performance of the vital functions, is typically manifested by distinguishing signs and symptoms, and is a response to environmental factors (as malnutrition, industrial hazards, or climate), to specific infective agents (as worms, bacteria, or viruses), to inherent defects of the organism (as genetic anomalies), or to combinations of these factors" [126].

dysphonia "the partial loss of the ability to use the vocal folds to produce phonation" [127].

etiology "Cause or causes of a disease" [126].

evidence-based medicine "the use of evidence from well designed and conducted research in healthcare decision-making" [129].

feature "an individual measurable heuristic property of a phenomenon being observed" [129], or "one of a specified set of phonological primes defined in such a way that every segment in a language, at least at the phonological level, can be exhaustively characterized as some permitted combination, or 'bundle', of features, each with an associated value" [127].

Frequency Image The spatial distribution of dominant frequencies in the laryngeal high-speed video.

Frequency Plot Spatial distribution of dominant frequencies along the vocal fold edges.

fundamental frequency "The fundamental frequency of a periodic sound is the frequency of that sinusoidal component of the sound that has the same period as the periodic sound" [112].

glottal diplophonia Diplophonic vocal fold vibration.

hoarseness The overall deviation from normal voice quality, excluding deviations in pitch, loudness and rhythm [14].

ICD-10 The International Classification of Diseases (ICD) of the World Health Organization. It is the standard diagnostic tool for epidemiology, health management and clinical purposed [130].

irregularity Variation of fundamental frequency, cycle-to-cycle energy or cycle shape, as defined in [44].

label ”One or more characters, within or attached to a set of data, contain information about the set, including its identification” [131].

laryngeal high-speed video A high frame rate video of the larynx.

marker ”something that serves to identify, predict, or characterize” [126].

measure ”The number (real, complex, vector, etcetera) that expresses the ratio of the quantity to the unit used in measuring it” [131].

measurement ”The determination of the magnitude or amount of a quantity by comparison (direct or indirect) with the prototype standards of the system of units employed” [131].

normal ”according with, constituting, or not deviating from a norm, rule, or principle” [126].

parameter ”Any specific quantity or value affecting or describing the theoretical or measurable characteristics of a unit being considered which behaves as an independent variable or which depends upon some functional interaction of other quantities in a theoretically determinable manner” [131].

pathology From Ancient Greek ”pathos” (experience, suffering), and -logia (”an account of”). A synonym of ”disease” [129]; or ”the anatomic and physiological deviations from the normal that constitute disease or characterize a particular disease” [126].

phonation ”the production of vocal sounds and especially speech” [126], or ”the act of producing speech sounds” [127].

phonosurgery ”Phonosurgery is composed of procedures that are intended to maintain or improve the quality of the voice by correcting defects in laryngeal sound production” [132].

pitch ”that attribute of auditory sensation in terms of which sounds may be ordered on a musical scale” [112].

roughness ”Audible impression of irregular glottal pulses, abnormal fluctuations in F0, and separately perceived acoustic impulses (as in vocal fry), including diplophonia and register breaks” [14].

short-time Fourier transform The time-variant Fourier transform that is achieved from windowed signals, i.e., a spectrogram.

sign ”an objective evidence of disease especially as observed and interpreted by the physician rather than by the patient or lay observer” [126].

spectral video analysis Spectral analysis of laryngeal high-speed videos with respect to time. Both pixel intensity time-series and glottal edge trajectories can be Fourier transformed.

symptom ”A departure from normal function or feeling which is noticed by a patient, indicating the presence of disease of abnormality. A symptom is subjective, observed by the patient and cannot be measured directly” [129]; or ”subjective evidence of disease or physical disturbance observed by the patient” [126].

voice ”sound produced by vertebrates by means of lungs, larynx, or syrinx; *especially*: sound so produced by human beings” [126], or ”any phonation type involving vibration of the vocal folds” [127].

Symbols

a Real values of Fourier coefficients.

a' Even polynomial coefficients for waveform modeling.

α Probability of a rejection error in a hypothesis test.

A Spatial distribution of the peak spectral intensity density.

B Model parameter in multinomial logistic regression.

b Imaginary values of Fourier coefficients.

b' Odd polynomial coefficients for waveform modeling.

β The power of a hypothesis test, or the Kaiser window parameter.

D Geometrical distance of the ROC curve from the virtual optimum.

d Glottal edge trajectories or general waveform.

Δ Spectra of the vocal fold trajectories, the phase shift for unit-pulse FIR filtering, or the maximum log likelihood.

δ Mean free and blocked glottal edge trajectories, or unit-pulse function in waveform modeling.

e Error waveform in waveform modeling.

η Noise.

f_s Sampling frequency.

Γ Number of oscillator candidates.

γ Oscillator candidate index.

H Histogram of discrete frequencies, drawn from the spatial distribution of the dominant oscillation frequency.

I Light intensity.

i Block index, or phonation interval index in latent class analysis.

J Time respective spectra of light intensity, or the number of index tests in latent class analysis.

j Index test index in latent class analysis.

K Spatial distribution of the dominant oscillation frequency, or the number of possible index test outcomes in latent class analysis, or the block length parameter in perturbation quotient measurement.

k Discrete frequency, or test outcome index in latent class analysis.

κ Spatial distribution of dominant frequencies along the glottal edges, i.e., the Frequency Plot.

κ_2 Frequency of the second highest peak in the normalized frequency histogram $H(\kappa, i)$, i.e., the diplophonic frequency.

K_p Sgn-coding threshold.

L The likelihood.

l Lag index for cross correlation, or filter coefficient index in unit-pulse FIR filtering.

M Number of distinct frequencies in the spectral video analysis, or the number of oscillators used for waveform modeling.

m Oscillator index used for resynthesis.

M_e Even matrix for coefficient transformation in polynomial waveform modeling.

M_o Odd matrix for coefficient transformation in polynomial waveform modeling.

$M(\omega)$ Continuous Fourier transform of one metacycle.

μ Pulse index.

ν Arbitrary cyclic parameter.

N Block length, period length, number of intervals in latent class analysis, the number of negative intervals or the number of cycles in periods within the analyzed interval in perturbation quotient measurement.

n Discrete time.

n' Relative discrete time index.

ω Angular frequency.

P Number of positive intervals, or probability, or the number of partials.

p Partial index, or the mixing proportions of diplophonic and non-diplophonic intervals, i.e., the prevalence and 1 - prevalence.

$P(d)$ Predicted probability for the presence of diplophonia from multinomial logistic regression.

Φ The number of estimated parameters in latent class analysis.

R The number of latent classes.

π The ratio of a circles circumference to its diameter, or the class conditional probability.

r Normalized autocorrelation function.

r Cross correlation function, pulse shape filter coefficients, or class index in latent class analysis.

τ_{max} Lag of the autocorrelation function's local maximum.

S Oscillator combination in heuristic oscillator selection, or slope for optimal threshold determination in a ROC curve.

thr Threshold.

thr_{opt} Optimal threshold.

u Unit-pulse train.

v Auxiliary vector for coefficient transformation.

w Time-domain window.

x' Lateral position index in the spatio-temporal plot (i.e., left or right vocal fold).

x Lateral position, or Cartesian coordinate.

Y Index test outcomes in latent class analysis.

y' Sagittal position along the main glottal axis.

y Sagittal position, or Cartesian coordinate.

List of Figures

List of Tables

Bibliography

[1] R. J. Ruben, “Redefining the survival of the fittest: Communication disorders in the 21st century,” *The Laryngoscope*, vol. 110, no. 2 Pt 1, pp. 241–245, 2000.

[2] G. Bertino, A. Bellomo, F. E. Ferrero, and A. Ferlito, “Acoustic analysis of voice quality with or without false vocal fold displacement after cordectomy,” *Journal of Voice*, vol. 15, no. 1, pp. 131–140, 2001.

[3] K. Nishiyama, H. Hirose, Y. Iguchi, H. Nagai, J. Yamanaka, and M. Okamoto, “Autologous transplantation of fascia into the vocal fold as a treatment for recurrent nerve paralysis,” *The Laryngoscope*, vol. 112, no. 8 Pt 1, pp. 1420–1425, 2002.

[4] M. Krengli, M. Policarpo, I. Manfredda, P. Aluffi, G. Gambaro, M. Panella, and F. Pia, “Voice quality after treatment for T1a glottic carcinoma,” *Acta Oncologica*, vol. 43, no. 3, pp. 284–289, 2004.

[5] U. Cesari, C. Faggioli, D. Testa, O. Vecchio, and V. Galli, “Montgomery thyroplasty. Case report focusing on endoscopic and functional findings,” *Acta Otorhinolaryngologica Italica*, vol. 24, pp. 226–233, 2004.

[6] K. Tsukahara, R. Tokashiki, H. Hiramatsu, and M. Suzuki, “A case of high-pitched diplophonia that resolved after a direct pull of the lateral cricoarytenoid muscle,” *Acta Oto-Laryngologica*, vol. 125, no. 3, pp. 331–333, 2005.

[7] J.-Y. Lim, S. E. Lim, S. H. Choi, J. H. Kim, K.-M. Kim, and H.-S. Choi, “Clinical characteristics and voice analysis of patients with mutational dysphonia: clinical significance of diplophonia and closed quotients,” *Journal of Voice*, vol. 21, no. 1, pp. 12–19, 2007.

[8] I. Kocak, M. Dogan, E. Tadihan, Z. Alkan Cakir, S. Bengisu, and M. Akpinar, “Window anterior commissure relaxation laryngoplasty in the management of high-pitched voice disorders,” *Archives of Otolaryngology–Head & Neck Surgery*, vol. 134, no. 12, pp. 1263–1269, 2008.

[9] R. Speyer, “Effects of voice therapy: a systematic review,” *Journal of Voice*, vol. 22, no. 5, pp. 565–580, 2008.

[10] J. R. L. Bibby, S. M. Cotton, A. Perry, and J. F. Corry, “Voice outcomes after radiotherapy treatment for early glottic cancer: assessment using multidimensional tools,” *Head & Neck*, vol. 30, no. 5, pp. 600–610, 2008.

[11] J. M. Ulis and E. Yanagisawa, “What’s new in differential diagnosis and treatment of hoarseness?” *Current Opinion in Otolaryngology & Head and Neck Surgery*, vol. 17, no. 3, pp. 209–215, 2009.

[12] M. Kimura, H. Imagawa, T. Nito, K. I. Sakakibara, R. W. Chan, and N. Tayama, “Arytenoid adduction for correcting vocal fold asymmetry: high-speed imaging,” *The Annals of Otology, Rhinology & Laryngology*, vol. 119, no. 7, pp. 439–446, 2010.

[13] G. Molteni, G. Bergamini, A. Ricci-Maccarini, C. Marchese, A. Ghidini, M. Alicandri-Ciufelli, M. P. Luppi, and L. Presutti, "Auto-crosslinked hyaluronan gel injections in phonosurgery," *Otolaryngology-Head and Neck Surgery*, vol. 142, no. 4, pp. 547–553, 2010.

[14] P. H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, and V. Woisard, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques," *European Archives of Oto-Rhino-Laryngology*, vol. 258, no. 2, pp. 77–82, 2001.

[15] D. Michaelis and H. Strube, "The Hoarseness Diagram," Göttingen, 2003. [Online]. Available: www.physik3.gwdg.de/~micha/hd.html [Accessed: July 15, 2014]

[16] S. Vishnubhotla, "Detection of irregular phonation in speech," Master's thesis, University of Maryland, 2007.

[17] B. R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," *Journal of Phonetics*, vol. 29, no. 4, pp. 365–381, 2001.

[18] D. Schacter, D. Gilbert, and D. Wegner, *Psychology.* New York: Worth, 2009.

[19] I. Steinecke and H. Herzel, "Bifurcations in an asymmetric vocal-fold model," *The Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1874–1884, 1995.

[20] P. H. Dejonckere and J. Lebacq, "An analysis of the diplophonia phenomenon," *Speech Communication*, vol. 2, no. 1, pp. 47–56, 1983.

[21] A. Bregman, *Auditory scene analysis.* Cambridge: The MIT Press, 1994.

[22] B. Moore, "Basic auditory processes involved in the analysis of speech sounds," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 363, pp. 947–963, 2008.

[23] J. Müller, *Handbuch der Physiologie des Menschen II.* Coblenz: Hölscher, 1837.

[24] C. L. Merkel, *Anatomie und Physiologie des menschlichen Stimm- und Sprachorgans.* Leipzig: Ambrosius, 1857.

[25] L. Türck, *Klinik der Krankheiten des Kehlkopfes.* Vienna: Braumüller, 1866.

[26] M. J. Rossbach, "Doppeltönigkeit der Stimme (Diphthongie) bei ungleicher Spannung der Stimmbänder," *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin*, vol. 54, no. 3, pp. 571–574, 1872.

[27] P. Grützner, *Handbuch der Physiologie.* Leipzig: Hermann, 1879, vol. 1, no. 2.

[28] I. R. Titze, "Toward standards in acoustic analysis of voice," *Journal of Voice*, vol. 8, no. 1, pp. 1–7, 1994.

[29] O. Dornblüth, *Klinisches Wörterbuch.* Wiesbaden: Dornblüth, 1927.

[30] P. Moore and H. von Leden, "Dynamic variations of the vibratory pattern in the normal larynx," *Folia Phoniatrica et Logopaedica*, vol. 10, pp. 205–238, 1958.

[31] S. Smith, "Diplophonia and air conduction explosions," *Archiv für Ohren-, Nasen- und Kehlkopfheilkunde, vereinigt mit Zeitschrift für Hals-, Nasen- und Ohrenheilkunde*, vol. 173, no. 2, pp. 504–508, 1958.

[32] J. G. Švec, H. K. Schutte, and D. G. Miller, "A subharmonic vibratory pattern in normal vocal folds." *Journal of speech and hearing research*, vol. 39, no. 1, pp. 135–143, 1996.

[33] P. H. Ward, J. W. Sanders, R. Goldman, and G. P. Moore, "Diplophonia," *The Annals of Otology, Rhinology, and Laryngology*, vol. 78, no. 4, pp. 771–777, 1969.

[34] K. Ishizaka, "Computer simulation of pathological vocal cord vibration," *The Journal of the Acoustical Society of America*, vol. 60, no. 5, pp. 1193–1198, 1976.

[35] L. Cavalli and A. Hirson, "Diplophonia reappraised," *Journal of Voice*, vol. 13, no. 4, pp. 542–556, 1999.

[36] C. Bergan and I. R. Titze, "Perception of pitch and roughness in vocal signals with subharmonics," *Journal of Voice*, vol. 15, no. 2, pp. 165–175, 2001.

[37] X. Sun and Y. Xu, "Perceived pitch of synthesized voice with alternate cycles," *Journal of Voice*, vol. 16, no. 4, pp. 443–459, 2002.

[38] S. Granqvist and P. Lindestad, "A method of applying Fourier analysis to high-speed laryngoscopy," *The Journal of the Acoustical Society of America*, vol. 110, no. 6, pp. 3193–3197, 2001.

[39] J. Neubauer, P. Mergell, U. Eysholdt, and H. Herzel, "Spatio-temporal analysis of irregular vocal fold oscillations: biphonation due to desynchronization of spatial modes," *The Journal of the Acoustical Society of America*, vol. 110, no. 6, pp. 3179–3192, 2001.

[40] J. Hanquinet, F. Grenez, and J. Schoentgen, "Synthesis of disordered voices," *Nonlinear Analyses and Algorithms for Speech Processing*, pp. 231–241, 2005.

[41] K. I. Sakakibara, H. Imagawa, H. Yokonishi, M. Kimura, and N. Tayama, "Physiological observations and synthesis of subharmonic voices," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2011, pp. 1079–1085.

[42] J. B. Alonso, M. a. Ferrer, P. Henríquez, K. López-de Ipina, J. Cabrera, and C. M. Travieso, "A study of glottal excitation synthesizers for different voice qualities," *Neurocomputing*, vol. 150, pp. 367–376, 2015.

[43] D. Michaelis, T. Gramss, and H. Strube, "Glottal-to-Noise Excitation Ratio – a new measure for describing pathological voices," *Acta Acustica*, vol. 83, pp. 700–706, 1997.

[44] D. Michaelis, M. Fröhlich, and H. Strube, "Selection and combination of acoustic features for the description of pathologic voices," *Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1628–1639, 1998.

[45] D. Michaelis, "Das Göttinger Heiserkeits-Diagramm - Entwicklung und Prüfung eines akustischen Verfahrens zur objektiven Stimmgütebeurteilung pathologischer Stimmen," Ph.D. dissertation, Georg-August-Universität zu Göttingen, 1999.

[46] D. Mehta, "Impact of human vocal fold vibratory asymmetries on acoustic characteristics of sustained vowel phonation," Ph.D. dissertation, Massachusetts Institute of Technology, 2010.

[47] Y. Shue, "The voice source in speech production: data, analysis and models," Ph.D. dissertation, University of California, Los Angeles, 2010.

[48] H. Kasuya, S. Ogawa, and Y. Kikuchi, “An adaptive comb filtering method as applied to acoustic analyses of pathological voice,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 669–672, 1986.

[49] G. de Krom and G. de Krom, “A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals,” *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.

[50] I. R. Titze, “Nonlinear source-filter coupling in phonation: theory.” *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2733–2749, 2008.

[51] H. S. Bonilha, D. D. Deliyski, and T. Gerlach, “Phase asymmetries in normophonic speakers: visual judgments and objective findings,” *American Journal of Speech-Language Pathology*, vol. 17, no. 4, pp. 367–376, 2008.

[52] I. R. Titze, *Workshop on acoustic voice analysis: Summary statement.* Iowa: National Center for Voice and Speech, 1995.

[53] P. Bossuyt and J. Reitsma, “Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative,” *Clinical Chemistry*, vol. 49, no. 1, pp. 1–6, 2003.

[54] P. Aichinger, F. Feichter, B. Aichstill, W. Bigenzahn, and B. Schneider-Stickler, “Inter-device reliability of DSI measurement,” *Logopedics Phoniatrics Vocology*, vol. 37, no. 4, pp. 167–73, 2012.

[55] M. Bland, *An Introduction to Medical Statistics.* New York: Oxford University Press, 2000.

[56] “MATLAB documentation: perfcurve.” [Online]. Available: www.mathworks.de/de/help/stats/perfcurve.html [Accessed: March 12, 2014]

[57] P. Aichinger, I. Roesner, B. Schneider-Stickler, W. Bigenzahn, F. Feichter, A. K. Fuchs, M. Hagmüller, and G. Kubin, “Spectral analysis of laryngeal high-speed videos: case studies on diplophonic and euphonic phonation,” in *Proceedings of the 8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2013, pp. 81–84.

[58] P. Aichinger, I. Roesner, M. Leonhard, B. Schneider-Stickler, D. M. Denk-Linnert, W. Bigenzahn, A. K. Fuchs, M. Hagmüller, and G. Kubin, “Comparison of an audio-based and a video-based approach for detecting diplophonia,” *Biomedical Signal Processing and Control (in press).*

[59] P. Aichinger, B. Schneider-Stickler, W. Bigenzahn, A. K. Fuchs, B. Geiger, M. Hagmüller, and G. Kubin, “Double pitch marks in diplophonic voice,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 7437–7441.

[60] P. Aichinger, I. Roesner, M. Leonhard, B. Schneider-Stickler, D. M. Denk-Linnert, W. Bigenzahn, A. K. Fuchs, M. Hagmüller, and G. Kubin, “Towards objective voice assessment: the diplophonia diagram,” *Journal of Voice (accepted).*

[61] P. Aichinger, A. Sontacchi, and B. Schneider-Stickler, “Describing the transparency of mixdowns: The Masked-to-Unmasked-Ratio,” in *130th Audio Engineering Society Convention London*, 2011, pp. 1–10.

[62] F. Schenk, M. Urschler, C. Aigner, I. Roesner, P. Aichinger, and H. Bischof, "Automatic glottis segmentation from laryngeal high-speed videos using 3D active contours," in *Medical Image Understanding and Analysis*, 2014, pp. 111–116.

[63] F. Schenk, P. Aichinger, I. Roesner, H. Bischof, and M. Urschler, "Automatic high-speed video glottis segmentation using salient regions and 3D geodesic active contours," *Annals of the British Machine Vision Association (submitted).*

[64] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Döllinger, "Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos," *Medical Image Analysis*, vol. 11, no. 4, pp. 400–413, 2007.

[65] D. D. Deliyski, P. P. Petrushev, H. S. Bonilha, T. T. Gerlach, B. Martin-Harris, and R. E. Hillman, "Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 1, pp. 33–44, 2008.

[66] T. Wurzbacher, M. Döllinger, R. Schwarz, U. Hoppe, U. Eysholdt, and J. Lohscheller, "Spatiotemporal classification of vocal fold dynamics by a multimass model comprising time-dependent parameters." *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2324–2334, 2008.

[67] D. Voigt, M. Döllinger, A. Yang, U. Eysholdt, and J. Lohscheller, "Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods," *Computer Methods and Programs in Biomedicine*, vol. 99, no. 3, pp. 275–288, 2010.

[68] D. Voigt, M. Döllinger, T. Braunschweig, A. Yang, U. Eysholdt, and J. Lohscheller, "Classification of functional voice disorders based on phonovibrograms," *Artificial Intelligence in Medicine*, vol. 49, no. 1, pp. 51–59, 2010.

[69] R. R. Patel, L. Liu, N. Galatsanos, and D. M. Bless, "Differential vibratory characteristics of adductor spasmodic dysphonia and muscle tension dysphonia on high-speed digital imaging." *The Annals of Otology, Rhinology & Laryngology*, vol. 120, no. 1, pp. 21–32, 2011.

[70] E. Inwald, M. Döllinger, M. Schuster, U. Eysholdt, and C. Bohr, "Multiparametric analysis of vocal fold vibrations in healthy and disordered voices in high-speed imaging," *Journal of Voice*, vol. 25, no. 5, pp. 576–590, 2011.

[71] R. Orlikoff, M. E. Golla, and D. D. Deliyski, "Analysis of longitudinal phase differences in vocal-fold vibration using synchronous high-speed videoendoscopy and electroglottography," *Journal of Voice*, vol. 26, no. 6, pp. 816.e13–816.e20, 2012.

[72] A. Krenmayr, T. Wöllner, N. Supper, and P. Zorowka, "Visualizing phase relations of the vocal folds by means of high-speed videoendoscopy." *Journal of Voice*, vol. 26, no. 4, pp. 471–479, 2012.

[73] C. Krausert, Y. Liang, Y. Zhang, A. L. Rieves, K. R. Geurink, and J. J. Jiang, "Spatiotemporal analysis of normal and pathological human vocal fold vibrations," *American Journal of Otolaryngology*, vol. 33, no. 6, pp. 641–649, 2012.

[74] M. Kunduk and A. J. Mcwhorter, "Vocal fold vibratory behavior changes following surgical treatment of polyps investigated with high-speed videoendoscopy and phonovibrography." *The Annals of Otology, Rhinology & Laryngology*, vol. 121, no. 6, pp. 355–363, 2012.

[75] S.-Z. Karakozoglou, N. Henrich, D. C, and Y. Stylianou, “Automatic glottal segmentation using local-based active contours and application to glottovibrography,” *Speech Communication*, vol. 54, no. 5, pp. 641–654, 2012.

[76] M. Döllinger, M. Kunduk, and M. Kaltenbacher, “Analysis of vocal fold function from acoustic data simultaneously recorded with high-speed endoscopy,” *Journal of Voice*, vol. 26, no. 6, pp. 726–733, 2012.

[77] “The Kiel Corpus of Spontaneous Speech, Vol. I-III,” University of Kiel, Germany, 1997. [Online]. Available: www.ipds.uni-kiel.de/forschung/kielcorpus.de.html [Accessed: August 25, 2014]

[78] K. Weilhammer, U. Reichel, and F. Schiel, “Multi-tier annotations in the Verbmobil corpus,” in *Proceedings of European Language Resources Association*, 2002, pp. 912–917.

[79] B. Schuppler, “GRASS: The Graz Corpus of Read and Spontaneous Speech,” in *Proceedings of the European Language Resources Association*, 2014, pp. 1465–1470.

[80] H. Müller, A. Rosset, J.-P. Vallée, F. Terrier, and A. Geissbuhler, “A reference data set for the evaluation of medical image retrieval systems.” *Computerized Medical Imaging and Graphics*, vol. 28, no. 6, pp. 295–305, 2004.

[81] S. Mueller, M. Weiner, and L. Thal, “Ways toward an early diagnosis in Alzheimer’s disease: The Alzheimer’s Disease Neuroimaging Initiative (ADNI),” *Alzheimer’s & Dementia*, vol. 1, no. 1, pp. 55–66, 2005.

[82] G. Langs, H. Müller, B. Menze, and A. Hanbury, “VISCERAL: Towards large data in medical imaging - Challenges and directions,” in *Medical Content-Based Retrieval for Clinical Decision Support Lecture Notes in Computer Science Volume 7723*. Berlin Heidelberg: Springer, 2013, pp. 92–98.

[83] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, S. Della Penna, D. Feinberg, M. F. Glasser, N. Harel, a. C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. E. Petersen, F. Prior, B. L. Schlaggar, S. M. Smith, A. Z. Snyder, J. Xu, E. Yacoub, and W.-M. H. Consortium, “The Human Connectome Project: a data acquisition perspective.” *NeuroImage*, vol. 62, no. 4, pp. 2222–2231, 2012.

[84] “Odds ratio estimation.” [Online]. Available: www.medcalc.org/calc/odds_ratio.php [Accessed: October 7, 2014]

[85] P. Boersma and D. Weenink, “Praat: doing phonetics by computer,” 2014. [Online]. Available: www.praat.org [Accessed: July 14, 2014]

[86] M. Kipp, “Anvil: The video annotation research tool,” 2007. [Online]. Available: www.anvil-software.org [Accessed: April 7, 2014]

[87] “Java applets for power and sample size.” [Online]. Available: homepage.stat.uiowa.edu/~rlenth/Power/ [Accessed: August 24, 2014]

[88] “Diagnostic test evaluation.” [Online]. Available: www.medcalc.org/calc/diagnostic_test.php [Accessed: August 25, 2014]

[89] F. Wuyts and M. Bodt, “The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach,” *Journal of Speech, Language and Hearing Research*, vol. 43, pp. 796–809, 2000.

[90] A. Olszewski, L. Shen, and J. Jiang, "Objective methods of sample selection in acoustic analysis of voice," *The Annals of Otology, Rhinology, and Laryngology*, vol. 120, no. 3, pp. 155–161, 2011.

[91] K. I. Sakakibara, H. Imagawa, M. Kimura, H. Yokonishi, and N. Tayama, "Modal analysis of vocal fold vibrations using laryngotopography," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010, pp. 917–920.

[92] S. Hayashi, H. Hirose, N. Tayama, H. Imagawa, M. Nakayama, Y. Seino, M. Okamoto, M. Kimura, and T. Nito, "High-speed digital imaging laryngoscopy of the neoglottis following supracricoid laryngectomy with cricohyoidoepiglottopexy," *The Journal of Laryngology & Otology*, vol. 124, no. 11, pp. 1234–1238, 2010.

[93] D. Dubrovskiy, "Glottis Analyse Tools," University Hospital Erlangen, Germany, 2011.

[94] J. Lohscheller, U. Eysholdt, H. Toy, and M. Döllinger, "Phonovibrography: mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics," *IEEE Transactions on Medical Imaging*, vol. 27, no. 3, pp. 300–309, 2008.

[95] J. Schoentgen, "Glottal area patterns in numerically simulated diplophonia," Third Viennese Scientific Seminar on the Detection of Diplophonia, pp. 1–60, 2014. [Online]. Available: signaux.ulb.ac.be/~jschoent/pdf/7D30_PRESENTATION.pdf [Accessed: November 27, 2014]

[96] I. R. Titze and F. Alipour, *The myoelastic aerodynamic theory of phonation.* Iowa City: National Center for Voice and Speech, 2006.

[97] N. Malyska and T. F. Quatieri, "Spectral representations of nonmodal phonation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 34–46, 2008. [Online]. Available: www.ll.mit.edu/mission/communications/ist/publications/0801_Quatieri_IEEE.pdf[Accessed:November3,2014]

[98] J. G. Švec and H. K. Schutte, "Videokymography: high-speed line scanning of vocal fold vibration," *Journal of Voice*, vol. 10, no. 2, pp. 201–205, 1996.

[99] M. Tigges, T. Wittenberg, P. Mergell, and U. Eysholdt, "Imaging of vocal fold vibration by digital multi-plane kymography." *Computerized Medical Imaging and Graphics*, vol. 23, no. 6, pp. 323–30, 1999.

[100] F. Bilsen, "Pitch of noise signals: evidence for a "central spectrum"," *The Journal of the Acoustical Society of America*, vol. 61, no. 1, pp. 150–161, 1977.

[101] J. Neubauer, M. Edgerton, and H. Herzel, "Nonlinear phenomena in contemporary vocal music," *Journal of Voice*, vol. 18, no. 1, pp. 1–12, 2004.

[102] M. Hagmüller and G. Kubin, "Poincaré pitch marks," *Speech Communication*, vol. 48, no. 12, pp. 1650–1665, 2006.

[103] S. Fraj, J. Schoentgen, and F. Grenez, "Development and perceptual assessment of a synthesizer of disordered voices," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2603–15, 2012.

[104] G. Fant and J. Liljencrants, "Calculation of true glottal flow and its components," *Quarterly Progress Status Report, Department for Speech, Music and Hearing, KTH Computer Science and Communication*, vol. 26, no. 4, pp. 1–13, 1985.

[105] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” in *Proceedings of the Institute of Phonetic Sciences*, vol. 17. University of Amsterdam, 1993, pp. 97–110.

[106] D. Duffy, *Advanced Engineering Mathematics (Applied Mathematics).* Boca Raton, Florida: CRC Press, 1997.

[107] J. Schoentgen, “Non-linear signal representation and its application to the modelling of the glottal waveform,” *Speech Communication*, vol. 9, pp. 189–201, 1990.

[108] P. Aichinger, “Audio examples of polynomial modeling of a diplophonic waveform.” [Online]. Available: www.meduniwien.ac.at/phon/public/aichinger/thesis/chapter4.zip [Accessed: October 8, 2014]

[109] B. Moore, B. Glasberg, and T. Baer, “A model for the prediction of thresholds, loudness, and partial loudness,” *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997.

[110] B. Glasberg and B. Moore, “A model of loudness applicable to time-varying sounds,” *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[111] M. Wu, D. Wang, and G. Brown, “A multipitch tracking algorithm for noisy speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[112] B. Moore, *An introduction to the psychology of hearing.* San Diego: Academic Press, 2012.

[113] H. Kameoka, “Statistical approach to multipitch analysis,” Ph.D. dissertation, University of Tokyo, 2007.

[114] P. Lieberman, “Perturbations in vocal pitch,” *The Journal of the Acoustical Society of America*, vol. 33, no. 5, pp. 597–603, 1961.

[115] R. W. Wendahl, “Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness,” *Folia Phoniatrica et Logopaedica*, vol. 18, pp. 98–108, 1966.

[116] R. W. Wendahl, “Some parameters of auditory roughness,” *Folia phoniatrica et logopaedica*, vol. 18, pp. 26–32, 1966.

[117] D. D. Deliyski, “Acoustic model and evaluation of pathological voice production,” in *Third European Conference on Speech Communication and Technology*, 1993.

[118] P. Milenkovic, “Least mean square measures of voice perturbation,” *Journal of Speech, Language, and Hearing Research*, vol. 30, no. 4, pp. 529–538, 1987.

[119] I. R. Titze and H. Liang, “Comparison of F0 extraction methods for high-precision voice perturbation measurements,” *Journal of Speech, Language, and Hearing Research*, vol. 36, pp. 1120–1133, 1993.

[120] P. McCullagh and J. Nelder, *Generalized Linear Models.* New York: Chapman & Hall, 1990.

[121] A. Dobson, *An introduction to generalized linear models.* Boca Raton, Florida: CRC Press, 2002.

[122] A. Rutjes, J. Reitsma, and A. Coomarasamy, “Evaluation of diagnostic tests when there is no gold standard: A review of methods,” *Health Technology Assessment*, vol. 11, no. 50, pp. iii, ix–51, 2007.

[123] D. A. Linzer and J. B. Lewis, “Journal of Statistical Software poLCA : An R Package for Polytomous Variable,” *Journal of Statistical Software*, vol. 42, no. 10, pp. 1–29, 2001.

[124] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[125] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[126] “MedlinePlus Medical Dictionary.” [Online]. Available: www.nlm.nih.gov/medlineplus/mplusdictionary.html [Accessed: November 5, 2014]

[127] R. L. Trask, *A dictionary of phonetics and phonology.* London and New York: Routledge, 2006.

[128] “Dorland’s online dictionary.” [Online]. Available: www.dorlands.com/wsearch.jsp [Accessed: November 24, 2014]

[129] “Wikipedia.” [Online]. Available: en.wikipedia.org/wiki [Accessed: November 5, 2014]

[130] “ICD-10 (WHO).” [Online]. Available: www.who.int/classifications/icd/en/ [Accessed: December 3, 2014]

[131] *IEEE Standards Dictionary: Glossary of Terms & Definitions.* New York: John Wiley & Sons, IEEE Press, 1984.

[132] D. A. Kieff and S. M. Zeitels, “Phonosurgery,” *Comprehensive Therapy*, vol. 22, no. 4, pp. 222–230, 1996.