Aus der Abteilung für Phoniatrie und Pädaudiologie
der Universität zu Lübeck
Leitender Arzt: Prof. Dr. med. R. Schönweiler

Predicting perceptual voice quality from objective voice parameters in dysphonic patients

Inauguraldissertation
zur
Erlangung der Doktorwürde
der Universität zu Lübeck
-Aus der Medizinischen Fakultät-

vorgelegt von
Elena Kramer
aus Nowosibirsk

Lübeck 2011

1. Berichterstatter/Berichterstatterin:	Prof. Dr. med. R. Schönweiler
2. Berichterstatter/Berichterstatterin:	Prof. Dr. med. W. Eichler
Tag der mündlichen Prüfung:	13.9.2012
Zum Druck genehmigt. Lübeck, den	13.9.2012

Contents

List of Abbreviations

A	astenicity
AI	aperiodicity index
ANN	artificial neural networks
Arate	articulation rate
B	breathiness, Behauchtheit
CQ	closed/contact quotient
CT	cricothyroid muscle
DEGG	first derivative of the EGG signal
DLP	deep lamina propria
DSI	dysphonia severity index
DVB	degree of voice breaks
EGG	electroglottography, electroglottographic
ELS	European Laryngological Society
F	formant
F0	fundamental frequency
FMF	frequency modulation factor
FNN	feed forward neural networks
Fx	instantaneous frequency
G	overall grade of severity
GNE	glottal-to-noise excitation ratio
Gx	unfiltered EGG signal
H	harmonic
H	hoarseness, Heiserkeit
HGG	high-speed glottography
HNR	harmonics-to-noise ratio
IA	interarytenoid muscle
IC	Irregularity component
IFx	irregularity index
ILM	intrinsic laryngeal muscles
ILP	intermediate lamina propria
Int	intensity
Ji	jitter
LCA	lateral cricoarytenoid muscle
LLE	largest Lyapunov exponent
LTAS	long-term average spectrum
Lx	filtered EGG signal
MDVP	Multi-Dimensional Voice Profile
MPT	maximum phonation time
NC	Noise component
Npauses	number of pauses
OQ	open quotient
P(%)	pausing time in % of the total reading time
P/100 syl	number of pauses per 100 syllables
PCA	posterior cricoarytenoid muscle
Plength	mean pause length
PQ	phonation quotient
Ptime	overall pausing time
PVG	phonovibrography, phonovibrogram
R80	80% F0 range

QDA	quadratic discriminant analysis
r	Pearson product-moment correlation coefficient
R	roughness, Rauigkeit
r_s	Spearman's rank correlation coefficient
Rtime	overall reading time
S	strain
S/P	speech-to-pause ratio
SD	standard deviation
SFF	speaking fundamental frequency
Shi	shimmer
SHR	subharmonic-to-harmonic ratio
SLP	superficial lamina propria
Sp	sound pressure
Srate	speech rate
st	semitones
Sylbp	mean number of syllables produced between two pauses
TA	thyroarytenoid muscle
TEP	tracheo-esophageal prosthesis
VC	vital capacity
VFCA	vocal fold contact area
VKG	videokymography
VS	videostroboscopy

Chapter 1: Introduction

The quality and the amount of voice-related research are constantly growing as a result of increased prevalence and incidence of voice disorders in general population. A recent epidemiologic study (n = 1326, 97 % non-teachers) conducted in the USA (Roy et al., 2005) found that almost 30 % (n = 396) of interviewed persons had experienced a voice disorder in the past. 5.9 % (n = 78) of interviewed persons had to seek professional help for voice improvement and 6.6 % were currently experiencing voice problems. Women and persons aged between 40 and 59 complained more often of a chronic voice disorder. Similar studies revealed an even higher lifetime prevalence of voice disorders and a higher risk of developing a chronic voice disorder in the 65+ population and teachers (Roy et al., 2004; Roy et al., 2007).

Voice disorders are often related to vocal abuse, noisy environment, infections or to exposure to substances that irritate and dry out tissues in the throat including tobacco, alcohol and esophageal reflux. Dysphonia is a potential side effect of many a medication. While many factors have contributed to the current state, two main causes like demographic changes with shift to voice and speaking intensive professions and increased awareness of voice/voice problems have prompted voice research questing for objective voice measures.

Most voice disorders originate within the larynx and affect perceptual voice quality. Progressive and persistent hoarseness may be the early symptom of laryngeal cancer. Sometimes, impairment of the voice may be the symptom of a disease or condition that primarily affects organs other than the larynx. In all cases, early screening procedures in patients presenting the symptoms of hoarseness could identify candidates for referral to voice professionals and ensure early identification and treatment of disease.

Changes in perceptual voice quality are the reason why many patients, especially those in speech and voice professions, seek medical attention. Judging a patient's perceptual voice quality has always been subjective and depends on the experience of the voice specialist. The trained ear of the voice professional will probably remain the best instrument in voice quality assessment. But in the view of an increasing need to detect and quantify dysphonia by means of noninvasive methods that could be also used by non-voice specialists, parameterization of signals obtained from voice patients has received a great deal of attention over the last few decades. These are some of the possible advantages of a voice screening tool besides early detection of organic changes in the larynx: testing voice quality prior to surgery to clarify preexisting voice condition and after surgery to document changes in the voice due to tracheal intubation and anesthesia, identification of voice problems and predisposition to professional voice disorders in voice and speaking intensive professions.

In spite of significant progress in pathologic voice research, it has been proved that no clear definition of perceptual voice quality could be solely based on a single objective

voice parameter. However, attempts have not been abandoned to find a method that would be comparable to subjective voice evaluation and could automatically predict the perceptual severity of voice pathology by objective voice parameters. In this thesis, a method of objective voice quality assessment is proposed which is based on a combination of voice parameters that best describe perceptual voice dimensions like breathiness, roughness and hoarseness.

1.1 The outline of the dissertation

The organization of the thesis is as follows:

In the introduction, the focus will be on the basics needed to understand why objective voice parameters succeed or fail to predict perceptual voice quality. A brief review of the basic anatomy of the larynx, the basic concepts inherent to the physics of voice production including those of the theory of nonlinear dynamics and a short description of acoustic and electroglottographic signals and their spectra are presented. Where possible, signal properties will be related to glottal configurations, vocal fold biomechanics and illustrated with patient data. In the penultimate section of Chapter 1, research literature dealing with automatic classification of voice quality by objective voice parameters will be reviewed. Chapter 2 will be centered on methodological issues. Chapter 3 is concerned with experimental results that will be given for each measured voice parameter separately. In section 3.3, two classification methods will be presented and classification results described, followed by a discussion and suggestions for future research in Chapter 4.

1.2 Basic anatomy of the larynx

The larynx is a hollow structure located at the top of the trachea and is composed of three single (thyroid, cricoid, and epiglottis) and three paired (arytenoid, cuneiform and corniculate) cartilages that are fastened together by membranes and muscles. The thyrohyoid muscle and the thyrohyoid membrane connect the larynx to the hyoid bone that in turn is attached to the muscles of the floor of the mouth and the tongue above, so that articulatory movements can be transmitted to the larynx and coordinated with voicing.

Voice is produced in the larynx by vibration of the vocal folds that are brought into a phonatory position and set into motion by airflow, which is the driving force of voice production. The movement from the respiratory to the phonatory position is called adduction and is achieved through a combined activation of several intrinsic laryngeal muscles (ILM), especially interarytenoids. The movement from the phonatory to the respiratory position is called abduction. Whereas the initiation of phonation is a voluntary process, once brought to action, the vibrations of the vocal folds are self-sustained due to physical forces. Vibrations are most

easily sustained in the phonatory position. But if the air supply is sufficient, vocal folds can vibrate in the more open positions.

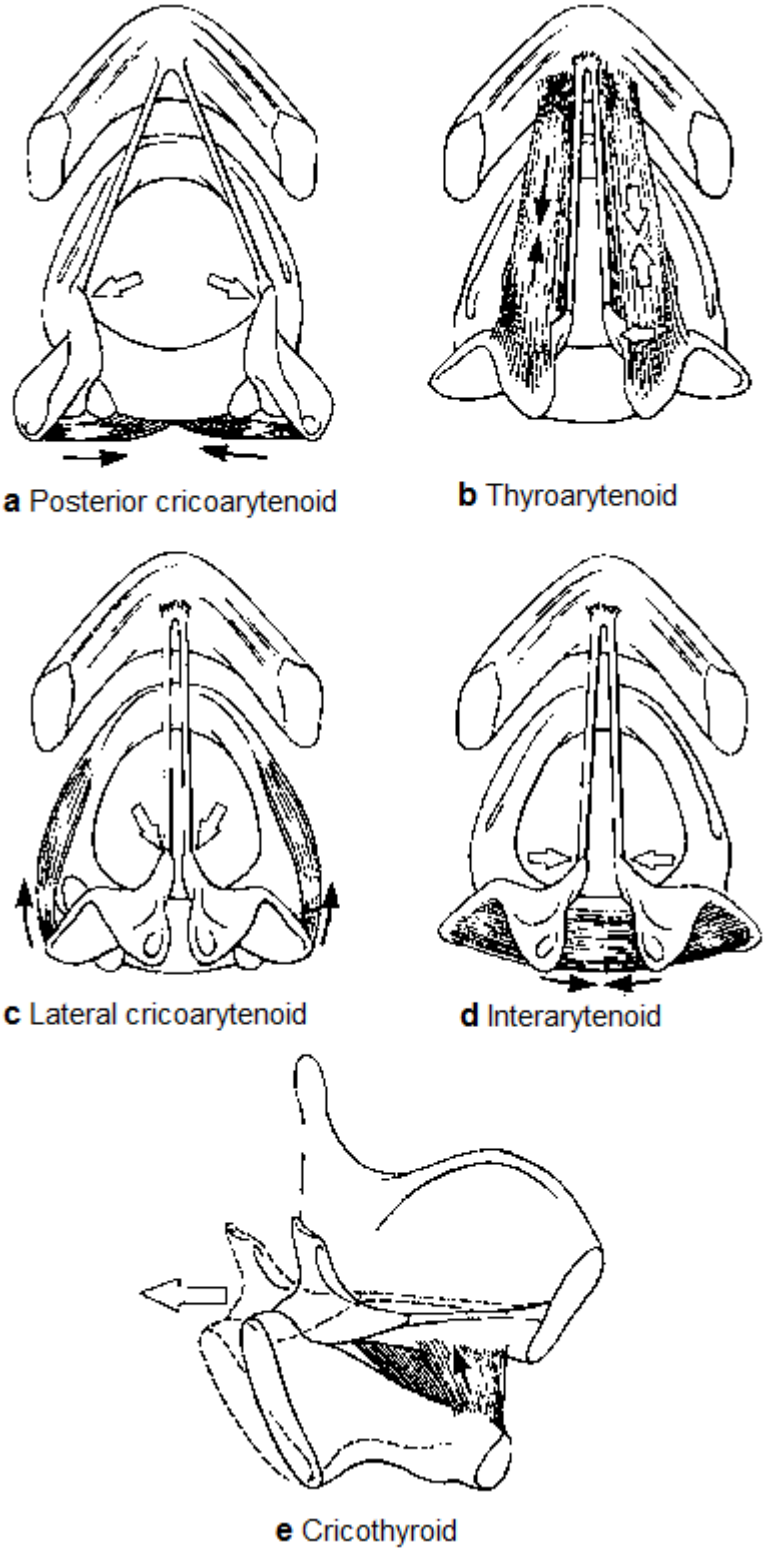
Intrinsic laryngeal muscles (thyroarytenoid (TA), cricothyroid (CT), posterior cricoarytenoid (PCA), lateral cricoarytenoid (LCA), and interarytenoid (IA)) are primarily responsible for changes in the position, shape and tension of the vocal folds (Fig. 1). The body of the vocal folds is formed by the TA, which consists of two parts. They are attached anteriorly to the thyroid cartilage and posteriorly to the arytenoid cartilages. The vocalis muscle is a part of the TA responsible for the active longitudinal tension. The PCA is the only vocal fold separator: outward movements of the arytenoid cartilages around the vertical axis pull the vocal folds apart. While contracting, the LCA pulls the arytenoids in the opposite direction, causing the membranous part of the vocal folds to approximate and increasing the medial compression.

The adductive tension is obtained by contracting the LCA and the IA. The IA closes the posterior part of the glottis. The CT contraction tilts back the lamina of the cricoid cartilage, thereby increasing the vocal fold length and passive longitudinal tension. The same result can be achieved through cricoid or hyoid bone elevation. All three movements result in F0 increase; the latter two being related to articulatory changes. Extrinsic laryngeal muscles regulate primarily the vertical position of the larynx in the neck. Raising of the entire larynx brings about an increase in longitudinal tension.

All ILM are innervated by the recurrent laryngeal branch of the vagus nerve except the CT, which is innervated by the external laryngeal branch of the same nerve. The mean number of motor units as reported in Neto & Marques (2008) was with 268 (1.3) the smallest for IA and the greatest with 431 (1.6) for CT. Of the five examined ILM, TA had the smallest motor unit size and possibly the finest motor control.

At the histological level, vocal folds are composed of three different layers (mucosa, ligament and body) that have different mechanical properties (Hirano, 1974) and therefore react differently to lengthening and tension. The surface of the vocal folds is covered by squamous epithelium that is responsible for shaping, protection and hydration of the vocal folds. Under the epithelium, three further layers can be distinguished differing by the type of fibers. The superficial lamina propria (SLP) has a soft texture similar to gelatin; it consists of fibrous fibers that dampen the impact of vocal fold collision. The epithelium and SLP form the mucosa. The intermediate lamina propria (ILP) consists mostly of elastic fibers, and the deep lamina propria (DLP) of durable collagenous fibers. ILP and DLP form the vocal ligament. The body consists of the muscle fibers.

Fig. 1: Laryngoscopic view of the ILM responsible for activating vocal fold position (from Sasaki CT. Physiology of the larynx. In: English G. ed. *Otolaryngology*. Hagerstown MD, Harper and Row. 1984).



The multi-layered composition of the vocal folds is the prerequisite of the mucosal wave that travels along the surface of the vocal folds once the vocal folds collide. The mucosa is loosely connected to and moves independently from the other layers. The upper and the

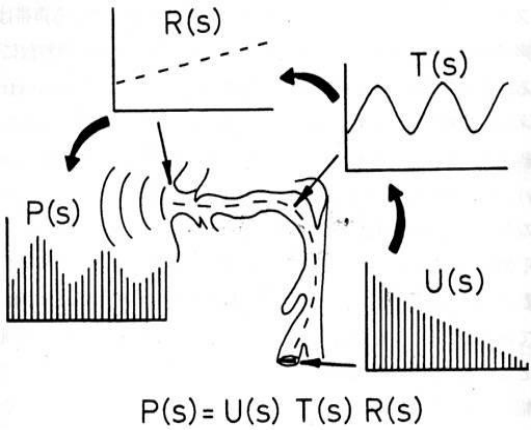
lower margins of the mucosa vibrate with a difference in phase, which effectively co-determines the vocal fold contact area (VFCA), so that opening and closure spreads gradually in horizontal and vertical planes. An increase in the tension of the body or stiffening of the cover reduces the mucosal wave.

Ventricular folds, or false vocal folds, are located above the true vocal folds. Ventricular folds do not contain intrinsic muscles but the distance between them can be changed through activity of external laryngeal muscles. The false vocal folds are usually not involved in voice production since their involvement is known to increase the supraglottal pressure and hinder the vocal fold movement; however, they are sometimes pressed together in hyperfunctional voice or take over the function of the vocal folds when the true vocal folds are lost. In whispered speech, ventricular folds approximate and cause airflow turbulence. Effortful phonation can also be achieved by pressing together the arytenoid cartilages.

1.3 The physics of voice production

In voice healthy subjects, vocal fold dynamics can be concieved of in terms of a pair of coupled ossilators. The air passing into the vocal tract is interrupted at regular intervals by vocal fold vibrations. Their joint effort depends on a number of variables like length, mass, tension, mobility and elasticity. The acoustics of voice production is commonly described in terms of the vocal source-tract filter theory (Fant, 1970) which is a simplified linear model describing normal vowel production.

Fig. 2: The source-tract filter model of vowel production (from Stevens KN, *Acoustic Phonetics*, MIT Press, 1998).



In Fig. 2, the vowel spectrum P(s) is the product of the spectrum of the periodic glottal source U(s), the transfer function of the vocal tract T(s) (i.e., filter), and the radiation characteristics R(s). The spectrum of the glottal source U(s) decays at a rate of 12 dB per octave. The glottal source is modified according to the characteristics of the vocal tract. Peaks in the

transfer function T(s) correspond to resonances of the vocal tract (formants). In the vocal tract, frequencies that coincide with the resonant frequencies are enhanced while others are dampened. At the mouth-opening, the filtered signal is further enhanced at a constant rate of 6 dB/octave which is characteristic for the radiation function R(s).

Many assumptions of the source-tract filter model are invalid in dysphonic voice production. An asymmetry in length, mass, tension, mobility and elasticity may manifest itself as desynchronization of the left and right vocal fold or vibratory modes of the single vocal fold that seems to be the underlying mechanism of nonlinear behavior (Berry et al., 1996; Giovanni et al., 1999a). In particular, there are two kinds of nonlinear behavior: chaos caused by turbulence in incomplete glottic closure and instabilities like bifurcations (sudden transitions to other regimes of vibration), modulations and subharmonics. There has been considerable interest in applying nonlinear dynamic system principles in describing and modelling disordered voices (Titze et al., 1993; Herzel, 1993; Herzel et al., 1994, Svec et al., 1996, Hatzikirou et al., 2006). The results of these studies suggest that some aspects of pathologic voice production are better accounted for in the framework of nonlinear dynamics and that nonlinearities and low-dimentional chaos can be observed even in the simplest models of the vocal folds.

In pathologic voice, the glottal source is not necessarily quasi-periodic. Moreover, vibratory irregularities in pathology are not totally random, but are fairly predictable or organized in patterns that can be captured in nonlinear models. There is substantial evidence that some voice disorders are associated with voice arrests or frequency modulation patterns, others with the presence of subharmonics, short-term modulations in the amplitude or noise component etc. (Isshiki et al., 1966; Koike, 1969; Askenfelt & Hammarberg, 1986; Hirano, 1989; Remacle & Trigaux, 1991; Sapienza et al., 2002).

Sustained phonations can be generated by the ventricular folds and the aryepiglottic sphincter, as well. Whereas ventricular fold vibrations can be considered more or less periodic, aryepiglottic vibrations may be completely aperiodic (Sakakibara et al., 2007). Additional sources of vibration beside the vocal folds can contribute to emergence of additional frequencies (subharmonics and biphonation).

Furthermore, subglottal resonances have a pronounced effect in pathologic voice production. In normal voicing, the influence of the subglottal system is effectively small: when the glottis is closed, the resonances of the vocal tract alone determine the shape of the output. However, even in normal phonation, the glottis is open half the time. The effect of subglottal resonances cannot be neglected in phonations with a permanent glottal leak. Glottal configurations where the glottis is never fully closed during the glottal cycle (breathy voice quality) are known to have a damping effect on the amplitude of the radiated sound. Indicative of the coupling between the supra- and subglottal systems are loss of energy in high-frequency

bands, replacement of harmonic structure by noise at higher frequencies above 2.5 kHz and spectral dips at unusual places (Hanson, 1997; Barney et al., 2007). Södersten & Lindestad (1990) and Hanson (1997) reported that female speakers have in general a more open glottal configuration than male speakers, the acoustic consequences of which are noisier signals with stronger low frequency and weaker high-frequency components found in female voices. A permanent opening at the glottis throughout the glottal cycle, pathologic or habitual, was observed to cause changes in formant positions, formant bandwidths, and the appearance of subglottal formants outside the normal formant pattern as well (Fant et al., 1972). Whereas supraglottal formants are variable depending on a vowel, subglottal formants are relatively fixed and vary a little with larynx height.

Even in normal symmetrical vibration, source-tract interaction may induce voice instabilities (Hatzikirou et al., 2006). In pathology, the source often has to be adjusted to the properties of the filter. For example, changes in the resonating characteristics of the vocal tract due to oronasal coupling cause vowel amplitude to drop by 5 to 10 dB (Hamlet, 1973). Consequently, patients with velopharyngeal insufficiency need a greater vocal effort to speak at normal loudness level than normal subjects. Thus, vowels produced with a high vocal effort have a shallow glottal spectrum slope U(s): the decay in energy to higher frequencies does not occur as fast as in vowels produced with a low vocal effort (Pickett, 1991). Similarly, in esophageal speech, which is notably produced with a high effort, the energy in the high frequencies is stronger than in normal speakers (Lu et al., 1999).

1.3.1 Voice modes and vocal fold biomechanics in norm and pathology

Several major voice modes are possible in normal subjects depending on muscular tension and how closely together the vocal folds are held (see Ladefoged & Maddieson, 1996). In the context of voice pathology, the following voice modes are also highly relevant: breathy, slack, modal, pressed and creaky.

In creaky voice, the arytenoids are strongly adducted, the longitudinal tension is weak meaning thick vibrating mass; vocal folds do not vibrate as a whole. The ligamental and arytenoid parts vibrate separately which leads to pulses of alternating amplitudes. The frequency of vibration is very low.

Pressed (also tense or hyperfunctional) voice has a pronounced contraction of the vocal muscle. Both medial compression and adductive tension are high. Sometimes ventricular folds are involved. Vibration pattern has irregular cycle duration and amplitude. The consequence of the pressed voice on the acoustic signal is the increase in the amplitude of the higher harmonics, similar to increase in subglottal pressure. In pressed voice, the vocal folds are

closed for the most part of the vibratory cycle (Lindqvist-Gauffin, 1972). Hyperfunctional use of voice is a causative factor in organic pathologies like edema, vocal fold thickening, nodules, polyps and contact ulcers. Signs of hyperfunction are described in Hillman et al. (1989) and Sama et al. (2001).

Slack, also lax or hypofunctional, voice is produced with low laryngeal effort, decreased muscle tension and weak adduction. Breathy voice is characterized by either an incomplete glottal closure, a high flow rate, a looser form of vibration, or a combination of these characteristics. Adductive tension, medial compression and longitudinal tension are low. A small degree of glottal adduction increases the time of the glottal cycle spent in glottal opening, reduces the amplitude of the higher harmonics and makes the signal noisier (Lindqvist-Gauffin, 1972).

Besides laryngeal setting, voice pathology may also affect symmetry in length, tension, mass, elasticity and mobility of the vocal folds. Irregular or asymmetric vibrations were observed in many a pathology including laryngitis, granuloma, polyps, Reinke's edema, cysts, hyperplasia, carcinoma, papillomas, and vocal fold paralysis. The loss of looseness of the mucosa or of the mobility interferes with the Bernoulli's effect and contributes to noise production (Isshiki et al., 1969).

A mass lesion normally changes the mode of vibration of the vocal fold in question by affecting mass and stiffness of the impaired side, especially when only one side is impaired, but sometimes may prevent the complete closure. The effect that a mass lesion will exert on glottic closure depends on the location and size of the lesion. For example, vocal cysts located subglotically may be transported upwards with the airflow. Polyps may be squeezed between the vocal folds preventing complete glottal closure or may be very mobile during phonation, especially when pedunculated (Dikkers & Nikkels, 1999). When the lesion is mobile, voice quality is usually very unstable. Incomplete glottal closure can be sometimes observed in laryngitis and Reinke's edema.

The prominence of the mucosal wave depends on the elasticity constants of body (muscle tension) and cover (the degree of stiffness of the mucosa). Some lesions like cysts are known to increase stiffness in the vocal folds, whereas others slacken the mucosa. According to Shohet et al. (1996), the disruption of the mucosal wave, which was absent in 100 % of cysts, was the most reliable criterion in differentiating cysts from polyps and nodules videostroboscopically. The mucosal wave is pronounced in Reinke's edema – a swelling of the mucosa filled with fluid which is very mobile; it can be disrupted at affected portions in smaller lesions like granuloma, nodules, polyps and reduced or absent in advanced carcinoma, sulcus vocalis, epithelial hyperplasia, recurrent laryngeal nerve paralysis, cysts and laryngitis (Hirano, 1974; Hirano, 1989).

In unilateral vocal fold paralysis, the biomechanics and the degree of voice change seems to depend substantially on several factors like position of the paralized vocal fold in both the horizontal and the vertical plane, the degree of bowing of the paralized vocal fold, compensatory glottal maneuvers (Inagi et al., 1997). Neuromuscular pathology affects the tone of the vocal folds and the closure. But its effect on tension should be different than in mass lesions.

The relationship between vocal fold properties and perceived voice quality has been defined as follows: using the method of semantic decomposition of hoarseness, Isshiki et al. (1969) suggested that hoarseness has at least two aspects: noise and irregularity. The perceived roughness is related to irregularity of the fundamental frequency which is produced by the asymmetry of the vocal folds not involving much tissue hardening. Irregularity in periodicity has been reported as characteristic for pressed and creaky voice modes as well and is believed to give the voice a rough character. By contrast, the perceived breathiness is associated with noise component that arises in hard, rough, non-elastic and incompletely approximated vocal folds suppressing the Bernoulli's effect. In many voice disorders both perceptual aspects are prominent. Another consideration of importance is that the more vocal fold properties deviate from the norm, the more salient should be the change in voice quality. In support of this claim, Dejonckere et al. (1993) found that mean scores for R, B and G were higher in organic than in functional voice disorders.

Objective analysis of voice is based on the assumption that vocal fold properties are reflected in quantitative measures of voice. However, phonatory behavior and voice quality in normal subjects show huge variation overlapping with pathology: voice disorders do not have to cause perceptible changes in the acoustic voice signal and normal subjects may sometimes exhibit abnormal voice measures. Sama et al. (2001) observed that videostroboscopically there was no significant difference between patients afflicted with functional dysphonia involving hyperfunction and nondysphonic controls. 60 % of the control population demonstrated features of hyperfunction like anteriorposterior compression, approximation of arytenoids, false vocal folds involvement etc. Many nondysphonic people use slack and breathy voice habitually.

Acoustic measures in turn must be consistent with perceptual impression of the voice. However, voice quality can be normal even if some vibratory irregularities were observed videostroboscopically. Compensatory glottal maneuvers can significantly influence the quality of the voice. Perception of a particular voice quality can be associated with multiple acoustic patterns and multiple underlying vocal fold properties.

Given these inconsistencies, the effect of laryngeal disorder on laryngeal mechanics, acoustic measures and perceptual voice quality may be more complex than previously believed. From literature reviewed in section 1.4, it appears that vocal fold properties and quan-

titative measures relate to perceptual voice quality in a closer and simpler way than to specific diagnosis. It is obvious that distinct pathologies may result in the same biomechanical characteristics of the vocal folds and the same tone of voice; the same diagnosis in different biomechanics and different perceptual voice impressions. This must be the reason why efforts to relate acoustic measures to specific diagnosis have been futile so far.

More promising seems to be the research centered on grouping diagnoses with similar biomechanical properties. Naturally, voice disorders are classified on the basis of mechanical properties of the vocal folds into disorders that primarily modify the vibratory pattern of the vocal folds and those of neurologic-psychogenic nature that affect vocal fold closure pattern. Tanaka et al. (1991) classified voice disorders into three groups according to biomechanically different types of dysphonia: mass lesion group (laryngitis, nodules, polyp, Reinke's edema, cysts, and granuloma), high stiffness group (benign and malignant neoplasms, laryngeal trauma and hyperfunctional dysphonia) and glottic incompetence group (sulcus vocalis, vocal fold paralysis, hypofunctional dysphonia). The problem with this division is that mass and elasticity changes also affect the glottal valving. Therefore, the effect of laryngeal disorder on laryngeal mechanics and control is not always predictable. Michaelis (2000) used a six-fold voice disorders classification: malignant tumors, disorders involving restricted mobility of the vocal folds, benign lesions, functional dysphonia, central dysphonia and others.

Despite some progress, discrimination between different groups of dysphonia remains an extremely complicated task. In Iwata & von Leden (1970), Hecker & Kreul (1971), Murry & Doherty (1980), acoustic measures could not reliably discriminate between different types vocal fold lesions. Remacle & Trigaux (1991) showed that high-resolution frequency analysis despite observed differences between different types of small lesions is not specific and cannot provide the diagnosis. There have been several studies reporting on difficulties in differentiating between different mass lesions videostroboscopically, thereby questioning the results of studies in which the diagnosis was not confirmed histologically (Shohet et al., 1996; Dikkers & Nikkels, 1999). It has been known that the tissue in laryngeal cancer is harder and less elastic than in other lesions. However, criteria to reliably discriminate between cancer and other lesions could not be found (Isshiki et al., 1969; Hirano et al., 1986). Classification attempts seem to be more successful in detecting functional dysphonia. Callan et al. (1999) reported a high success rate in differentiating between normal voices, spasmodic dysphonia, pre-treatment and post-treatment functional disorders. Using self-organizing maps that were trained with 6 acoustic measures they could classify 75.8 % of the voices correctly. The diagnostic value of acoustic voice measures including perturbation measures, DSI and subjective perceived hoarseness was also confirmed in Werth et al. (2010). However, the problem with the automatic diagnosis of voice disorders was that despite significant differences between the clinical groups, individual diagnostics was difficult resulting in misclassification errors in up

to 25 % of cases due to the vast spread of the parameter values within the groups.

Another approach that effectively improved classification rates was to discriminate between normal and pathological voices disregarding the specific pathologic condition. In Fraile et al. (2009), automatic detection of laryngeal pathology by means of artificial neural networks (ANN) with 16 mel-frequency cepstral coefficients as inputs succeeded in 192 (85 %) of 226 study subjects consisting of 173 dysphonic subjects and 53 nondysphonic controls. The fact that cepstral coefficients cannot be related to vocal fold physiology is one of the limitations of this study. In Lin et al. (1998), jitter and shimmer from both EGG and acoustic signals were significantly different in normal and pathologic voices, but failed to distinguish between the ”mass” and ”neuromuscular” group. Alonso et al. (2005) achieved a 92 % success rate in classifying 100 normal and 68 dysphonic subjects using ANN with classical and nonlinear parameters. All in all, there is little evidence of the feasibility of differential diagnosis by objective voice measures.

1.3.2 Modern glottographic techniques

Much of what is known about voice production today has been obtained with the help of glottographic techniques. Glottography is a general term for methods to monitor the vibrations of the vocal folds. Five glottographic techniques are currently in use: electroglottography (EGG), videostroboscopy (VS), high-speed glottography (HGG) and two other techniques based on digital high-speed imaging, namely videokymography (VKG) and phonovibrography (PVG). This section gives a short overview of glottographic techniques. Their full decription as well as relative advantages and disadvantages have been detailed in the literature cited below.

Videostroboscopy is a standard laryngoscopic technique. Although it supplies only an illusion of vocal fold motion, it is superior to other methods in regard to the ability to detect organic findings (Olthoff et al., 2007). Precise assessment of vocal fold functionality regarding duration of the glottal closure, the vocal fold amplitude or the mucosal wave is not possible. Since the F0 triggers the stroboscopic light, VS depends on voice signal quality and often fails in very irregular voices (Schönhärl, 1960; Hirano, 1981; Wendler, 2005).

High-speed glottography visualizes vocal fold movements in real time, which gives HGG an advantage over VS in studying irregular voices and detecting functional findings (Olthoff et al., 2007). However, this advantage is lost in the view of the fact that both VS and HGG need perceptive evaluation of the video recordings (Dejonckere et al., 2001). With inter-rater variability being almost equally high (Olthoff et al., 2007; Lohscheller, 2008), visual assessment of slowered high-speed films is more time-consuming (ca. 60 min). In order to fully exploit the advantage of real-time imaging, there is a substantial and very

obvious need for analysis methods to extract the relevant information from HGG in a more efficient way. Although HGG has been in practice since 1930s, the clinical acceptance of the high-speed technique is still evolving. This is partially due to the lack of a standardized procedure to reconstruct vocal fold vibrations from high-speed videos and insufficient clinical validation (Eysholdt et al., 1996; Schutte et al., 1998; Lohscheller et al., 2007).

Digital kymography and phonovibrography are techniques that reduce the information contained in high-speed films by extracting the motion of the vocal folds. They plot data acquired from different locations of each vocal cord separately. Problems due to rotation of the endoscope or relative movements of patient or examiner can be solved by image processing algorithms.

Digital kymograms arise out of concatenation of single lines from digital high-speed sequences of laryngoscopical examinations. Videokymography previously used to be restricted to a medio-lateral plane and therefore unable to capture irregilarities in anterio-posterio (ap) plane[1]. This problem was resolved by multi-line kymography in Tigges et al. (1999) and Wittenberg et al. (2000). Ca. 40 lines are considered to be sufficient to assess the vocal fold behavior (Neubauer et al., 2001). In this way it is possible to visualize different vibratory modes and irregularities in different sections of the glottis.

The principles of PVG computation is explained in Eysholdt & Lohscheller (2008), Lohscheller (2008), Lohscheller et al. (2008). The phonovibrogram is an image-processing algorithm that is able to identify such landmarks as anterior and posterior comissure and uses the line connecting them as a reference in order to extract the motion of the free edges of the vocal folds for subsequent compression into a single 2D-image. Image interpretation is done on the basis of its geometric form and information contained in colour.

All of the known objective parameters of the voice that can be obtained with EGG like open quotient, speed quotient, jitter, shimmer, symmetry factors etc. can be derived from high-speed glottograms (Eysholdt et al., 1996). Recently, several studies have been published on diagnostic value of digital kymography and phonovibrography. Since it is possible to reconstruct vocal fold motion in different sections of the glottis from one video sequence, it is obvious that the extraction of parameters might depend on location of the motion curves. In particular, Döllinger et al. (2003) found that in normal voices, most stable results were acquired such voice parameter as the degree of symmetry when applied to the medial motion plane. For dorsal and ventral motion curves, correct performance of the algorithm is reduced to 85%. In Voigt et al. (2010a), 81% of functional disorders were classified correctly with

[1] Disordered voices exhibit two types of asymmetry, the ”horizontal” and ”vertical” asymmetry, that can be separated. Asymmetry in the medio-lateral direction is expressed as different fundamental frequencies on each side, while asymmetry in the anterior-posterior direction leads to ap-mode vibrations (Eysholdt et al., 2003a, Eysholdt et al., 2003b).

PVG features. In a similar study involving subjects with vocal fold paresis (Voigt et al., 2010b), the rate of correct classifications was estimated at 93% for 2-class, 73 % for 3-class discrimination.

Advances in quantitative analysis of vocal fold motion using high-speed imaging go hand in hand with vocal fold modeling. Modeled normal and pathological phonation helps to better understand vocal fold dynamics and explain the observations made when viewing high-speed films. Once the model has been verified, real-time imaging techniques serve best to validate them. Successful modeling could allow extrapolation of parameters that are not directly accessible to observation or measurement like subglottal pressure, vocal fold velocity, vertical deflection, elasticity etc. Recent achievements in vocal fold modeling are presented with models that differ in complexity and are designed to accomodate both symmetric and asymmetric vocal fold vibrations. Hier again, a valuable contribution has been made by scientists from Germany. To simulate biphonation, Mergel et al. (1996), successfully tested a two-mass model of vocal folds with 7 parameters including subglottal pressure, elasticity and masses on the right and the left side. The current progress in model-based classification of vocal fold vibrations was reported in Döllinger et al. (2002), Schwarz et al. (2006) and Wurzbacher et al. (2006). Recently, a 3D analysis of vocal fold vibrations has been performed in Döllinger et al. (2008). The use of 3D analysis has been necessitated by the need for more accuracy. Yang et al. (2010) developed and tested a five mass model which can predict the 3D vibrations of the entire medial surface of the vocal fold, the most critical region of mucosal wave propagation.

Although EGG is the only glottographic technique that does not allow for the visualization of the larynx during phonatory activity, it has a number of advantages that outweigh the above mentioned limitation. The EGG signal is more easily analyzed and can be acquired in parallel to other imaging techniques. Most authors agree that EGG is useful in providing estimate of vocal fold contact during the glottal cycle, at least in normal voices. Compared to EGG, high-speed imaging and VS are not only more difficult and expensive, they cannot be conducted under natural conditions. Visual examination of the larynx is performed with protruded tongue under local anaesthesia of the oral cavity and cause a strong foreign body sensation in the throat. All the mentioned factors might affect phonation. Moreover, visualization of the larynx during phonatory activity is sometimes impossible to perform due to individual oral-pharyngeal anatomy, certain hyperfunctional postures and the gag reflex. Beside being invasive to some degree, all visual glottographic methods have a limited duration for recording (several seconds). Even if tongue protrusion, which is supposed to have the most detrimental effect on sustained phonation, can be discounted as a potential

problem[2], visual glottographic methods have the disadvantage of not being able to capture speech and singing samples. For these reasons, it is desireable to use less difficult and less invasive glottographic techniques with potential to be applied to speech samples in the place of HGG.

1.3.3 Comparison of electroglottographic and acoustic signals

Electroglottographic (EGG) and acoustic recording are noninvasive methods of instrumental voice analysis. EGG and microphone signals can be considered complementary since they provide different information with regards to vocal and articulatory phenomena. Whereas the majority of voice professionals all over the world routinely use acoustic voice analysis in clinical practice, only ¼ of participants of the survey conducted by Hirano (1989) reported to use EGG for voice evaluation.

The microphone signal (also called Sp signal) registers variation in sound pressure at the mouth and contains the effects of both the vocal folds and the vocal tract; EGG signals are free from the effect of the vocal tract filter. Although such methods as linear predictive coding and inverse filtering can be used to reconstruct the original signal from the microphone waveform, there is no guarantee that these methods are successful in cancelling the effect of the vocal tract in highly abnormal voices.

Electroglottography is an impedance method which is based on the property of body tissues to conduct electricity: skin and fat are less conductive than fluids and tissues, and air is a poor conductor. The laryngographic method to register variation in vocal fold contact area was first introduced by Fabre in 1957. For a detailed overview of electroglottography, see Colton & Conture (1990), Baken (1992), Rothenberg (1992) and Baken & Orlikoff (2000).

EGG signals contain information on the change in conductivity between the vocal folds and are recorded by positioning two electrodes at the level of the vocal folds. When the vocal folds are fully closed, the current flows through the vocal folds from one electrode to the other and the impedance is the lowest. When the vocal folds are separated, the current has to circumvent the glottis and electrical impedance is the highest. The amplitude of the output signal shows variation in electrical conductance that is proportional to the amount of vocal fold contact.

Increasing and decreasing vocal fold contact accounts for only 1–2 % of the impedance change in EGG signal (Baken, 1992). Structures near the larynx may interact with the current in a way that every change in geometry or orientation of those structures in relation to each

[2] Jilek et al. (2003) used EGG signal to prove that tongue protrusion which cannot be avoided in videostroboscopy and high-speed glottography does not affect vocal fold vibrations.

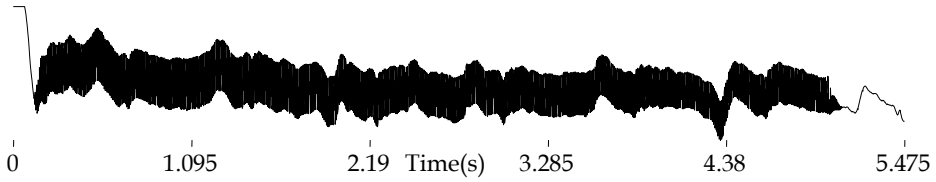
other induces change in impedance. Among sources of transcervical impedance change are muscle activity, vertical movements of the larynx and changes due to blood pulsation. Anatomic factors like position of the glottis within the neck and kind of tissues around the larynx are factors that may also affect impedance. Women are supposed to have a smaller vocal fold contact area as their vocal folds are thicker and shorter, so the EGG signal from female voices has smaller amplitude.

The EGG output waveform is called Gx if it also registers the vertical movements of the larynx like those in articulation or swallowing. The Lx signal emerges after removing the non-phonatory impedance changes via high-pass filtering and amplifying the vocal fold contact area contribution. Sp signals do not provide information on vertical larynx position.

Experiments with X-ray photography showed that during quiet inspiration and expiration the vertical larynx position remains relatively fixed (Mitchinson & Yoffey, 1947; Andrew, 1955). The lowering of the trachea on inspiration is not transmitted to the larynx since larynx elevator muscles normally contract to counteract the forces depressing the larynx.

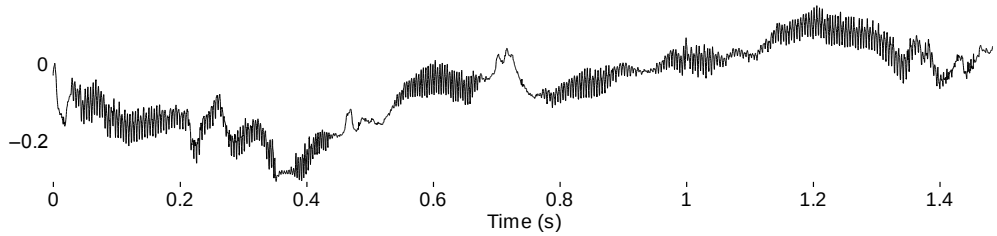
In phonatory position, the larynx is raised by 1 cm as compared to the respiratory position (Fujimura, 1976). The shape of the vocal tract is relatively stable during sustained phonations, but the larynx may move up and down slightly. The vertical height of the larynx is reflected in the baseline of the EGG signal, which is an imaginary line showing the tendency of the larynx to move (Fig. 3). Extreme baseline shifting in sustained phonation is a sign of unstable larynx and is often observed in vertical laryngeal tremor.

Fig. 3: Electroglottographic trace of a sustained vowel /a/ produced by a healthy subject.



In connected speech, the vocal tract shape and the larynx position are changed continuously (Fig. 4). Larger baseline shifts in EGG signals correspond to vertical movements of the larynx associated with articulation (changes in tongue position) and intonation. Baseline shifting is typical in sounds involving high pressure in the pharynx like stops and fricatives. Larynx is raised during production of voiceless stops like /p, t, k/. There appears to be a positive correlation between larynx height and pitch of voice (Shipp & Haller, 1972; Kakita & Hiki, 1976). Larynx rising by 1 mm can result in up to 8–10 Hz F0 increase (Hamlet, 1980).

Fig. 4: Electroglottographic trace of a phrase “Der Nordwind und die Sonne” spoken by a female patient.



Vertical movements of the larynx can be registered but not reliably measured in electroglottographic signals. The vertical excursions of the larynx are more prominent in women than in men, which poses a problem of undesired signal clipping. A habitual high vertical larynx position was observed in hyperfunctional dysphonia and strained voices (Aronson, 1990; Boone & McFarlane, 1993).

Electroglottography has the further advantage that EGG signals can be taken in rooms with a high ambient noise. Microphone signals require an acoustically treated room. In the present study, Lx and Sp signals were acquired almost simultaneously. A minor delay arises due to the distance from the glottis to the mouth and signal processing.

Electroglottographic signals are simpler than microphone waveforms unless they are contaminated with noise which is not related to changes in VFCA: artefacts and noise inherent to the equipment and vibrations of the body surfaces. A small amount of random noise from the equipment can cause cycle-to-cycle variations similar to those caused by irregular vocal fold movements. Voice-synchronous noise is supposed to be generated by anatomical structures around vocal folds: skin, pharyngeal wall, tongue, false vocal folds (Rothenberg, 1992). Böckler & Hacki (1999) found that neck soft tissue vibrations might be important in the interpretation of the EGG signal.

In Sp signals, noise is mainly related to articulatory phenomena during production of stops and fricatives as the oral pressure rises behind the articulatory closure or constriction.

1.3.3.1 Qualitative behavior of Sp signals

Microphone signals show a greater variation in patterns than EGG signals. Laryngographic waveforms of all voiced sounds are more or less the same. It is not possible to identify a voiced sound from an Lx signal alone.

Fig. 5 to Fig. 8 show several consecutive periods of sustainable speech sounds /i/, /e/, /o/, /u/, /a/, /n/, /m/ and /s/ produced by the same nondysphonic speaker. The lower trace represents the Lx signal.

Fig. 5: Waveforms of Sp (up) and Lx signals (down) of /i/ sustained at 151 Hz (left) and /e/ sustained at 145 Hz (right).

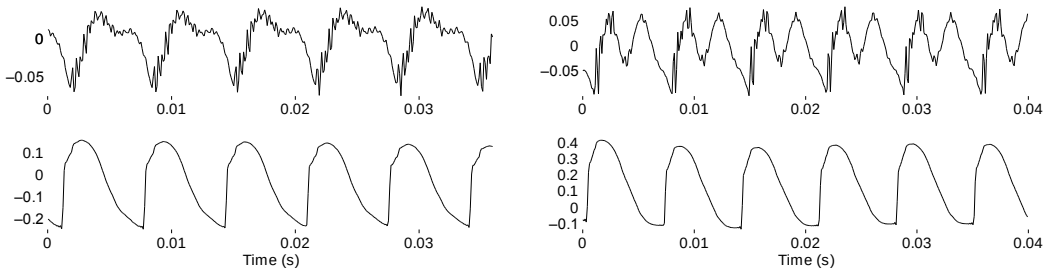


Fig. 6: Waveforms of Sp (up) and Lx signals (down) of /o/ sustained at 152 Hz (left) and /u/ sustained at 155 Hz (right).

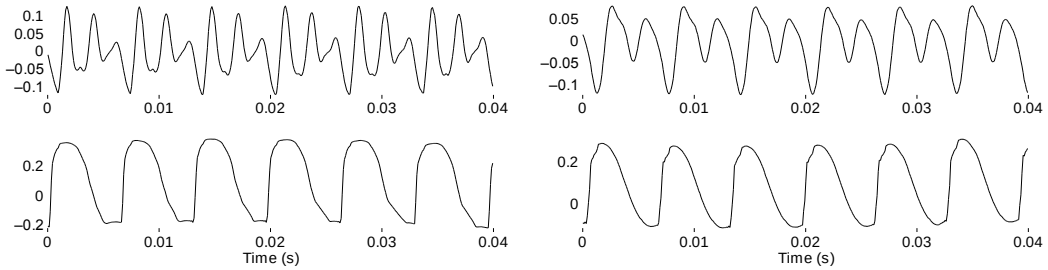


Fig. 7: Waveforms of Sp (up) and Lx signals (down) of /a/ sustained at 146 Hz (left) and /n/ sustained at 119 Hz (right).

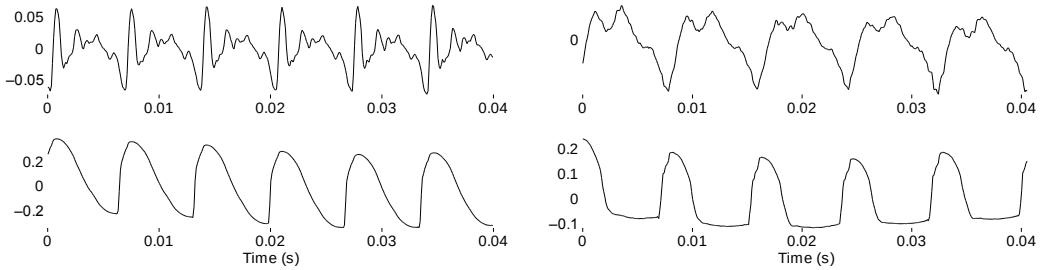
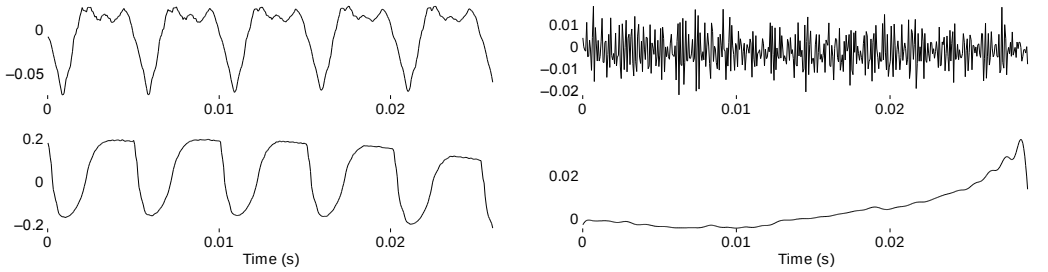


Fig. 8: Waveforms of Sp (up) and Lx signals (down) of /m/ sustained at 200 Hz (left) and /s/ (right).



In microphone signals, cycles of different vowels have different waveform shapes differing in the number of peaks per cycle. These peaks represent the harmonic frequencies in the signal. They seem to be more prominent in vowels than in nasals. On comparing vowels,

it is evident that the largest peak in each glottal cycle is sharper and steeper in /a/ than in other vowels. This difference remains even when the signal is overlaid by noise. This is what makes /a/ less prone to F0 extraction errors with methods based on zero-crossing (Vieira et al., 1996; Vieira et al., 2002). Low-pass filtering is a means to simplify the waveform to an almost sinusoidal shape by removing these peaks.

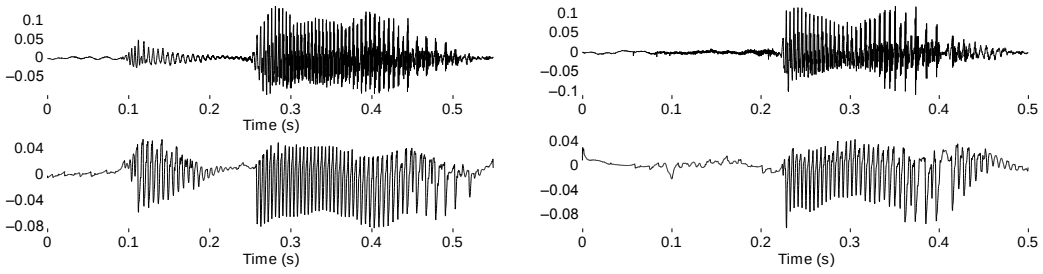
The number of peaks per vowel cycle in acoustic signals of different vowels may differ from individual to individual. To our knowledge, the intersubject and intrasubject variability in speech sound waveforms have never been systematically studied. It is uncertain if there is more variation in wave shapes among different talkers than among different sounds of the same speaker. In pathology, the waveform shapes are likely to be more complex.

During the production of voiceless /s/, the acoustic waveform does not exhibit a recognizable rhythmic pattern and the Lx signal consists of a baseline moving upwards since vocal fold vibrations are absent.

EGG signal is believed to be better suited for the detection of voiced/unvoiced segments with simple methods like zero-crossing. As illustrated by the examples above, strong noise-free Lx signals from nondysphonic voices have just two zero-crossings per period, whereas Sp signals may have more than two. Electroglottographic signals of voice healthy subjects are often used as reference to prove the efficiency of pitch detecting algorithms. In pathologic voices, though, both signals are susceptible to pitch measurement errors. Electroglottographic signals may contain a considerable amount of electrical noise and irregularities, especially in pathologic voices. Therefore, they need smoothing or filtering prior to voice parameter extraction. If a clear EGG signal can be obtained, no filtering is required.

The next two examples are meant to illustrate changes in the Lx and Sp signal due to articulation. In voiceless consonants, the glottis can be opened widely enough to prevent vibration (Fig. 9).

Fig. 9: Microphone (up) and electroglottographic (down) traces of /za/ (left) and /sa/ (right) spoken by a healthy subject.



Alternatively, the vocal folds may remain in the phonatory position but vibrations cease due to lack of effort to sustain them. This is the reason why sometimes voiced conso-

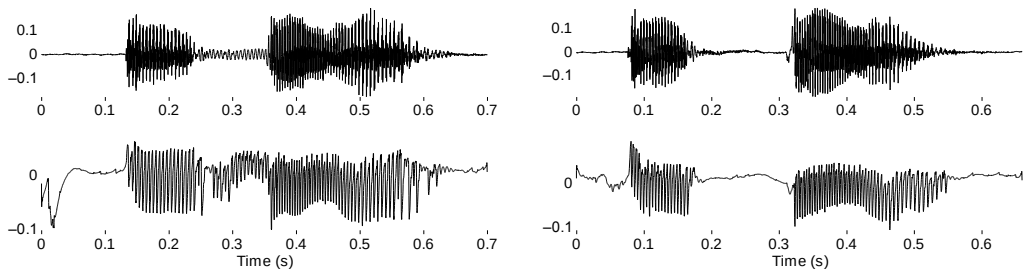
nants are realized as devoiced. In both Lx and Sp signals, vibrations are ceasing at transition between /z/ and /a/ (Fig. 9). Vibration in voiceless consonants can be inhibited by an increase in stiffness in M. vocalis (hence an increase in F0).

Sustaining voicing during voiced consonants requires additional effort (e.g., lowering the F0 via downward movement of the larynx or other cavity enlarging movements to rarefy the air) because vibration is naturally prevented as the supraglottal pressure builds up behind the articulatory closure or constriction (Stevens, 1977; Ladefoged & Maddieson, 1996; Boucher & Lamontagne, 2001).

Systematic F0 downward shifts have been frequently observed in rapid transitions between vowels and voiced consonants. They are associated with vocal tract constrictions and appear to be partially responsible for the instability of the speech signal.

In Fig. 10, voicing is interrupted in /apa/, but is persistent throughout /aba/. The overall amplitude of both the Sp and Lx signal in /b/ is less than that of adjacent vowels but constant throughout the duration of /b/. At normal-to-fast rates of speech, the amplitude of the Lx signal is influenced by intraoral pressure (Boucher & Lamontagne, 2001).

Fig. 10: Microphone (up) and electroglottographic (down) traces of /aba/ (left) and /apa/ (right) spoken by a healthy subject.



1.3.3.2 Qualitative behavior of Lx Signals

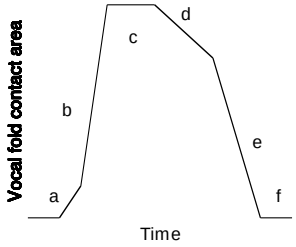
In this section, the ideal Lx waveform is discussed and typical Lx signal forms found in pathologic vowels are shown. Signals were band-pass filtered (30–1000 Hz) and normalized to a standard amplitude.

1.3.3.2.1 The ideal Lx waveform

The idealized cycle of vocal fold vibration is shown in Fig. 11. This model was first introduced by Baken. Each cycle ideally has four phases: closing, closed, opening and open phases. A point of contact break and contact initiation between the upper and lower vocal fold margins are indicated by change in the slopes. These indentations in the rising and falling slopes help to make distinctions between the phases. The phases of the glottal cycle are as

follows: The lower margins of the vocal folds make initial contact and begin to close in *a*. The upper margins are closing in *b* until the maximum contact is reached. The contact is maximal in *c*. The upper margins of the vocal folds begin to separate in *d*. The lower margins continue to separate in *e*. The glottis is open in *f*.

Fig. 11: A single cycle of vocal fold vibration depicting relative vocal fold contact area. The increase in vocal fold contact is plotted upwards (modified from Baken, 1992).



The relationship between videostroboscopic images and EGG curves were successfully validated for normal subjects. The onset of closure was observed to be signaled with a knee in the closing slope (Baken & Orlikoff, 2000). The beginning of vocal fold upper-edge separation in videostroboscopic images was found to correspond to a knee in the EGG opening slope (Anastaplo & Karnell, 1988). Identification of events related to opening and closing of the glottis is not possible in waveforms lacking these characteristic edges. This is why the scientific validity of EGG analysis has long been in dispute.

1.3.3.2.2 Methods for determining the glottal closure and opening in the Lx signal

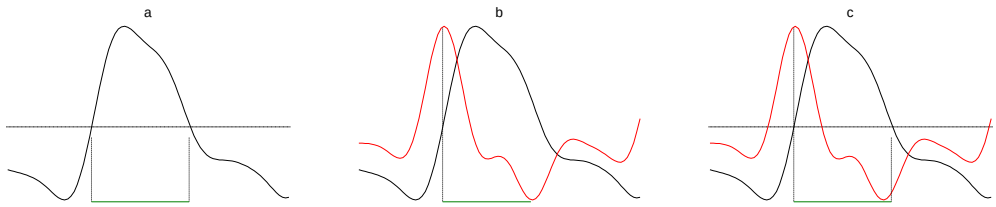
There is no proven method to estimate the timing of glottal opening and closure from the EGG waveforms alone. When the rising and falling slopes lack discontinuities, the beginning and end of each of the four phases is not that clearly defined. In this case, the vocal folds probably do not make complete contact of the upper and lower margins. For practical reasons, to obtain a measure of vocal fold abduction/adduction the vibratory cycles are divided just into two phases: an open phase during which the vocal folds are open and a closed phase during which air flow is blocked by vocal fold closure.

By convention, the temporal positions of glottal closure and glottal opening are determined using one of the two available methods or their combination. As EGG signal contains more information on the closed phase, all these methods aim at estimating the contact quotient (CQ). The open quotient (OQ) is complementary to the CQ and is obtained as *1 - CQ*.

The criterion-level method (Rothenberg & Mahshie, 1988) introduces an arbitrary threshold value defined as a percentage of the peak-to- peak amplitude (between 25–50 %) to

identify the points of glottal closure and opening. The closed phase is the interval between these two points. The open phase equals the difference between the total cycle length and the duration of the closed phase. To obtain the contact quotient (CQ), the closed phase is set in relation to the total cycle length. In norm, the glottis is closed for approximately one half of the glottal cycle.

Fig. 12: a) Calculation of the EGG contact quotient by a criterion-level method. The criterion threshold is set at 42 % of the peak-to-peak amplitude; b) calculation of the EGG contact quotient by a DEGG method. The interval between the positive and the negative peak in the DEGG signal corresponds to the closed phase; c) calculation of the EGG contact quotient by a hybrid method. The red line represents the DEGG. The threshold level is illustrated by the black dashed line. The corresponding closed phase is indicated by the plain green line. The increase in VFCA is plotted upwards.



The DEGG method (Henrich, 2004; Herbst & Ternström, 2006) uses the maximum and the minimum of the EGG first derivative to find the events of glottal closure and opening. A special solution is needed when there are more than one maximum and/or minimum[3].

The hybrid method (Howard, 1995) is a combination of the two, where the time of glottal closure is defined by the maximum of the DEGG signal and glottal opening is assumed to be the point in time when the signal amplitude falls below 3/7 (42 %) of the peak-to-peak amplitude.

All methods introduce a certain degree of arbitrariness in the calculation of the EGG contact quotient. As shown in Herbst & Ternström (2006), the results of the different methods to calculate the contact quotient vary considerably.

1.3.3.2.3 Typical Lx waveforms found in patients

Considerable research has been devoted to establish the relationship between EGG waveform shapes and mechanical vocal fold properties. This relationship is still poorly understood in pathologic voices. According to Hanson et al. (1988), variation in EGG signal patterns only partially reflects the degree of approximation of the vocal folds. The waveform shapes vary so greatly as a consequence of factors like vocal fold mass, tension, medial compression etc.

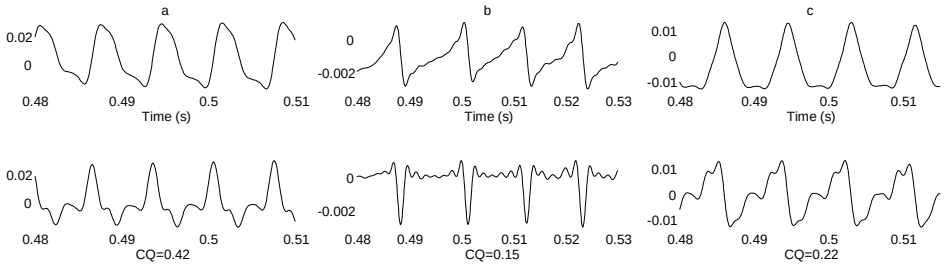
[3] In Henrich et al. (2004), the occurence of double peaks was explained by the manner in which opening/closing takes place in the vertical (over the thickness of the vocal folds) or horizontal (anterior-posterior parts of the glottis) direction. No explanation was offered for the occurence of imprecise peaks. Obviously, this is a research area that calls for an extensive study.

that it is not possible to give an overview of possible waveforms without running an interpretative risk. Due to diversity of possible waveforms and obscurity of their meanings for voice pathology, it seems more convenient to describe their composite features.

In the following, the most common Lx waveforms in prolonged /a/ are shown. One should keep in mind that differences in the shape of waveforms are important only if they result in perceptually different vocal qualities. For each waveform, the contact quotient (CQ) was calculated using the DEGG method. The closed phase was determined as the distance between the strongest local maximum and the strongest local minimum.

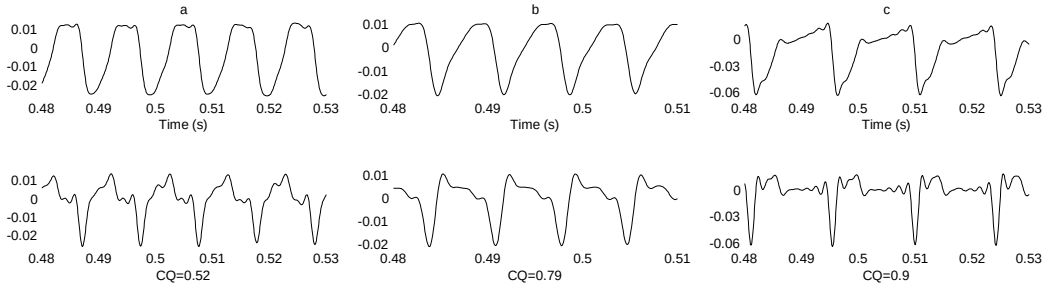
Efficient voice production involves a rapid and sharp closing of the vocal folds. In modal voice, the EGG curve should be slightly slanted to the left, as the vocal folds close faster than open (Fig. 13a). The closing slope becomes shorter, almost vertical, with increasing intensity. Both unilateral and bilateral lesions of the vocal folds (nodules, polyps, edema, tumors etc.) induce poor vocal fold contact which can increase the closing time. This is particularly apparent in the slanting of the cycles to the right. Waveforms as shown in Fig. 13b are typical for breathy voices, especially in combination with low overall amplitude of the signal. Note the 10-fold difference in peak-to-peak amplitude corresponding to VCFA between Fig. 13b and Fig. 13a. The Lx signal retains its periodic structure even if the closure is partial. Many pathologic voices were observed to have symmetric slopes as in Fig. 13c. If, in addition to symmetry, the maximum contact phase is short, proper vocal fold contact was probably not achieved.

Fig. 13: EGG waveforms (up) and the corresponding DEGG (down) obtained from a) patient 8, female, 60, after surgical removal of bilateral leukoplakia; b) patient 38, female 83, diagnosed with recurrent laryngeal nerve paresis; c) patient 141, male, 57, with glottal carcinoma.



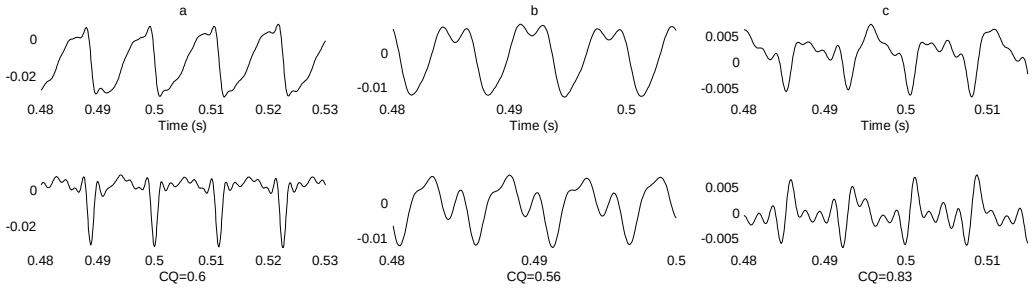
The basic difference between the waveshapes is the number and location of major slope discontinuities. The waveshapes may have no slope discontinuities at all, one at closing or opening, at both opening and closing, one or more at peak. Lx waveforms differ in the forms of the plateau: plateaus can be flat (Fig. 14a), rounded (Fig. 14b), slanted (Fig. 14c), indented (Fig. 15b), with additional peak (Fig 15a), irregular (Fig. 15c), sharp (Fig. 13c) or broad (Fig. 14c).

Fig. 14: EGG waveforms (up) and the corresponding DEGG (down) obtained from a) patient 85, male, 77, diagnosed with a polyp; b) patient 12, male, 63, complaining of dysphonia without apparent organic cause; c) patient 19, male, 53, after surgical removal of leukoplakia.



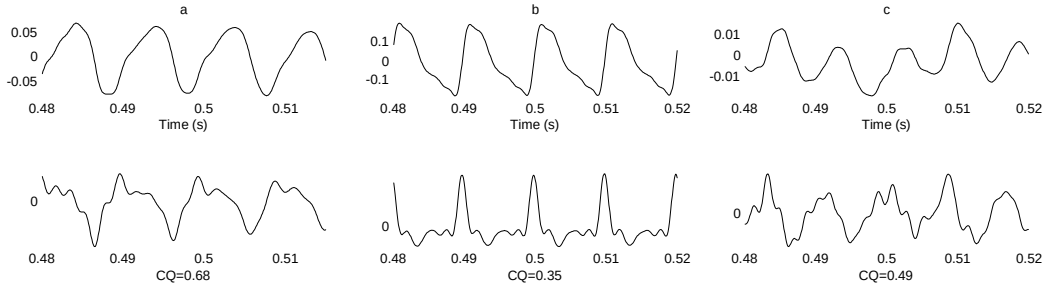
Following clinical implications can be formulated for the form of the EGG plateau. Wider plateaus mean that the closed phase is longer than the open phase. Prominent changes in the plateau VCFA indicate instability of maximum contact phase. An extra peak in the plateau, as shown in Fig. 15a, arises according to Scherer et al. (1988) due to additional tissue collisions or fluid interactions.

Fig. 15: EGG waveforms (up) and the corresponding DEGG (down) obtained from a) patient 37, male, 60, diagnosed with Reinke's edema; b) patient 107, female, 57, functional dysphonia; c) patient 106, male, 53, larynx carcinoma.



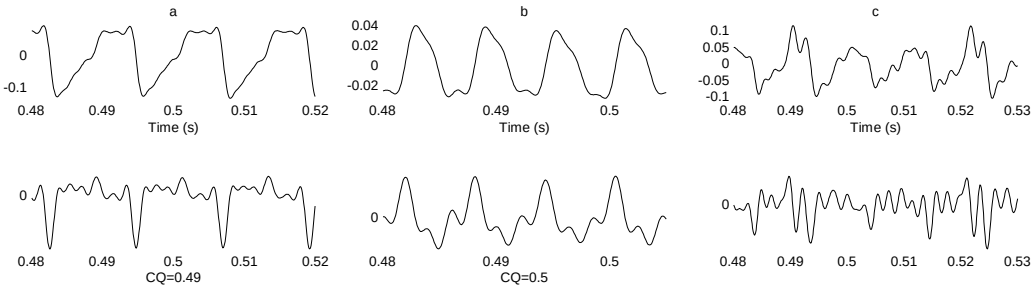
Observation of patient data enables us to conclude that the width of the plateau and, to a lesser extent, the slanting of Lx waveforms seem to have a major effect on CQ estimates derived with the DEGG method. Thus, waveforms with a long rising slope and a prominent convex knee tend to have higher CQ values than those with a long falling slope and a prominent concave knee (Fig. 16a, Fig. 16b), slopes with convex knees (Fig 16a, Fig. 17b) tend to give in general a higher CQ value than slopes with concave knees (Fig. 16b, Fig. 17a).

Fig. 16: EGG waveforms (up) and the corresponding DEGG (down) obtained from a) patient 34, male, 77, diagnosed with vocal fold paresis; b) patient 55, male, 74, diagnosed with spasmodic dysphonia; c) patient 48, male, 29, recurrent laryngeal nerve paresis.



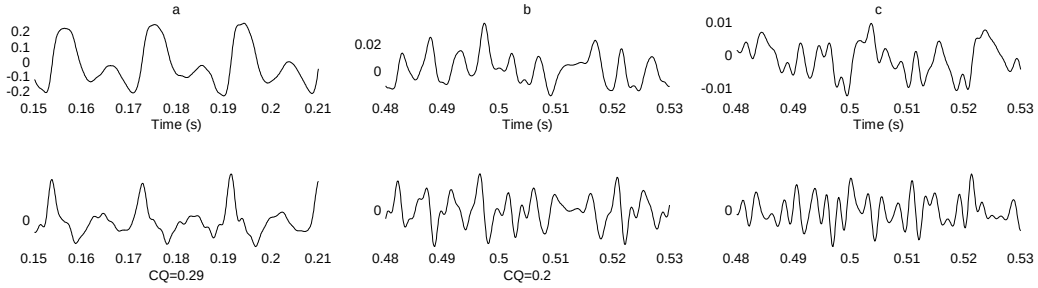
The waveform in Fig. 16c is an example of the signal with unstable baseline of the glottal closure which might be attributed to irregular reduction of electrical contact that takes place on separation of vocal folds when there is a mucous bridge or tissue between them. In Fig. 17c, the slopes are overlaid with noise that may stem from secretions in the larynx. The waveform shows a repetitive pattern of three cycles which makes the calculation of the CQ problematic. In this particular case, one bigger peak is followed by two smaller peaks. Double, triple and quadruple cycles are common in dysphonic voices.

Fig. 17: EGG waveforms (up) and the corresponding DEGG (down) obtained from a) patient 67, male, 63 after vocal fold granuloma removal; b) patient 150, female, 40, after surgery on bilateral Reinke's edema; c) patient 137, male, 67, leukoplakia involving epiglottis.



Mechanical abnormalities may alter the EGG waveform shape in a completely aberrant way. This may be reflected in the irregularity of the cycle-to-cycle shapes: inconsistent vertical amplitude (Fig. 18b), aperiodicity (Fig. 18c) of the signal, noisy waveforms and abrupt transitions to other regimes of vibration (bifurcations). Bifurcations that can be easily detected in time waveforms are transitions to a new frequency or new amplitude, transitions to regimes with alternating period or amplitude and transitions to completely aperiodic vibrations. Noisy and irregular waveforms render automatic CQ estimates less reliable.

Fig. 18: EGG waveforms (up) and the corresponding DEGG (down) obtained from a) patient 84, male, 57, diagnosed with carcinoma of squamous epithelium; b) patient 109, male, 72, recurrent laryngeal nerve paresis; c) patient 145, female, 66, diagnosed with Reinke's edema.



Alternate cycles in the time domain, as shown in Fig. 18a, correspond to subharmonics or creaky voice in the frequency domain. In creaky voice, short and long pitch periods alternate. Subharmonics, can involve both: a pattern with two different amplitudes or two different periods. Alternate cycles are often found in voices characterized as perceptually rough.

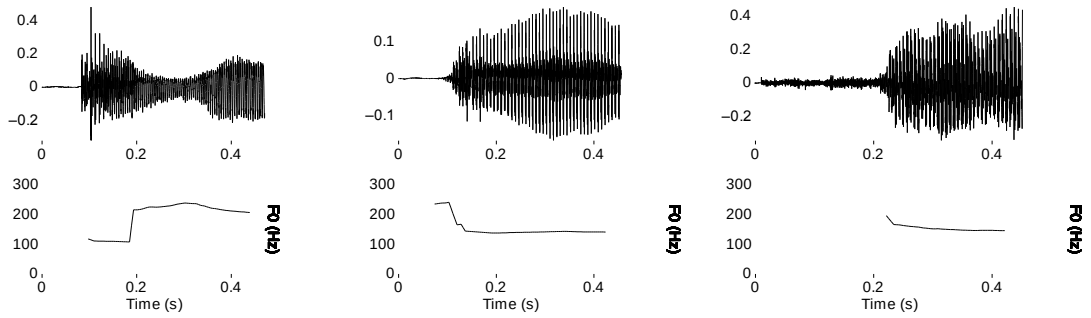
1.3.3.3 Onset transients in Lx and Sp signals

Voice onset is about how phonation is initiated. This involves coordination of timing between the onset and completion of adduction gesture, the onset of airflow and the onset of phonatory vibrations. Three types of voice onset are relevant for voice assessment in clinical practice: normal, breathy and hard.

In breathy voice onset, vocal folds begin to oscillate during the adduction gesture. Adduction is slow. A release of air precedes the vocal fold closure. The completion of adduction comes before the onset of phonatory vibrations in hard voice onsets. The adduction duration is least for hard onsets. Hard voice onsets involve greatest subglottal pressure and muscle tension. The closure is burst open forcefully that can easily result in vocal fold injury. The pressure behind the closure is high and often leads to clipping in the acoustic signal. Ventricular fold contact has also been observed in hard voice onset (Moore, 1938).

These three types can be clearly distinguished in the microphone signal just by looking at periodicity and run of the signal amplitude. The normal voice onset is characterized by a gradual growth of the amplitude before it gets steady (Fig. 19b). The breathy voice onset is started with a low-amplitude aperiodic expiration noise (Fig. 19c). In the hard voice onset the amplitude tends to overshoot before it gets stabilized: that is, the signal amplitude initially exceeds its steady state value (Fig. 19a). An interesting finding was reported by Cooke et al. (1997): the adduction gesture may be truncated and reinitiated; however they did not provide an explanation how this would affect the Sp signal.

Fig. 19: A composite graph showing an oscillogram and a corresponding F0 trajectory in three types of voice onset: hard (left), normal (middle) and breathy (right).



Braunschweig et al. (1997) found that voice onset classification is also possible on the basis of instantaneous fundamental frequency values. In hard voice onset, fundamental frequency increases to a constant value (Fig. 19a). On the contrary, in signals with gradually growing amplitude, fundamental frequency decreases to a constant value (Fig. 19b, Fig. 19c). This finding should be treated with caution as F0 values at on- and offsets are often unreliable.

It has been suggested that the time that the Sp signal needs to reach a stationary state can identify the type of voice onset. In technical terms, voice onset has been defined as an interval between the onset of sound pressure to the point at which the signal amplitude reaches the mean amplitude of the steady portion of the phonation (definition by Koike cited in Orlikoff et al., 2009). This interval is longest (247 ms) in normal voice onset, followed by breathy (121 ms) and hard (29 ms) voice onsets. In clinical setting, this finding implies that at least 250–300 ms at both ends of the vowel should be left out being unsuitable for voice parameter extraction.

Likewise, conclusions on the type of voice onset can be made by inspecting the synchronized sound pressure and electroglottographic waveforms (Fig. 20, Fig. 21). Simultaneous recording of both signals offer complementary information on voice onset and increase the usefulness of both techniques. The sound pressure rises when vocal folds begin to oscillate. This may take place in the pre-phonatory phase during gross adduction and tension adjustment. In the phonatory phase the full contact is achieved and the amplitude of vibration reaches a steady value. The Lx signal does not begin until the vocal folds are in contact. Orlikoff et al. (2009) observed that the time delay between the rise of the signals corresponds to vocal attack characteristics. So breathy voice onset measures a time delay of 7.6 ms to 38 ms, normal voice onset –1.4 ms to 9.6 ms, and hard voice onset –9.5 ms to –1.7 ms. A special algorithm was used to quantify the exact start and end points for measurements. Negative values arise when the Lx signal begins before the Sp signal. These findings were validated

with kymographic analysis[4] on normal subjects. Data has still to be validated on pathologic voices.

Fig. 20: Delayed (a) and almost simultaneous (b) onset of Sp and Lx signals.

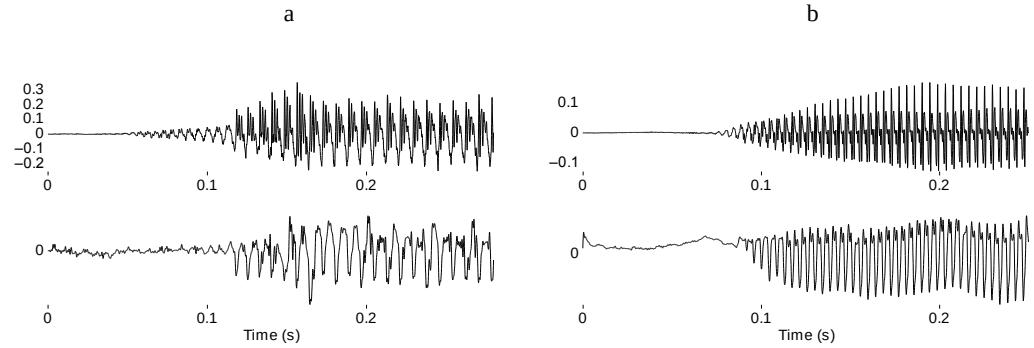
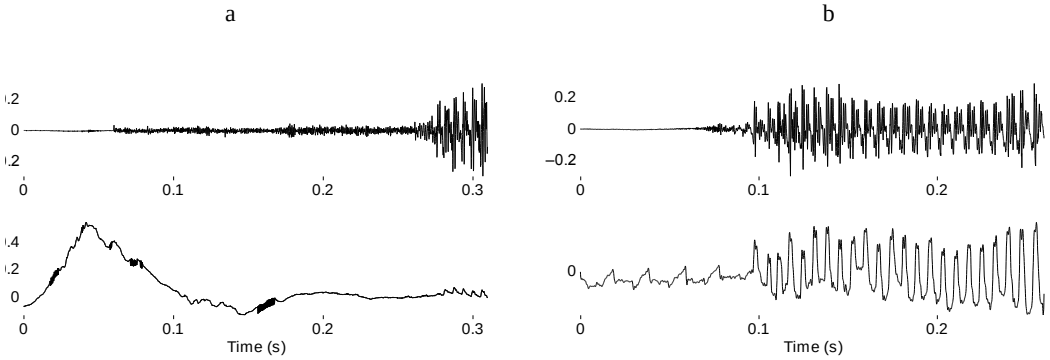


Fig. 21: A composite graph of Sp and Lx signals corresponding to breathy (a) and hard (b) voice onsets. The phonation in (a) is started with the expiration noise characterized by aperiodic fluctuations of low amplitude in the SP signal. Note extremely low amplitude of the Lx signal in breathy phonation starting around 0.3 s. The Lx signal precedes the Sp signal in hard voice onset (b).



Voice onset is believed to determine the subsequent perceptual voice quality. According to de Krom (1994a) and de Krom (1994b), the onset seems to carry more acoustic information for the perception of roughness than the mid-vowel segment or offset. The results of his research suggest that roughness ratings were more reliable when the stimuli contained vowel onsets. No such effect was observed for breathiness. These findings pose a question of what vowel fragments should be used in voice research to predict perceptual voice quality. Voicing control in connected speech is an even bigger challenge in the presence of voice pathology. Its influence on perceptual ratings of dysphonia severity grade has never been systematically studied.

For the purpose of the present study, it was not necessary to identify the type of voice

[4] Digital kymography provides probably the best parameter that allows to classify the type of phonation onset. A method to extract the phonation onset time from digital high-speed videos is described in Mergell et al. (1998). Here, the phonation onset time is defined as duration of the vocal fold amplitude growth from 32.2 % to 67.8 % of the saturation amplitude. The authors report good agreement between theory and measurements. Some examples are also given in Wittenberg et al. (2000).

onset in every single case. Instead, voice on- and offsets were indirectly considered in the calculation of the instrumental measures derived from connected speech.

1.3.3.4 Capabilities and limitations of electroglottographic and acoustic analysis

The relative advantages and disadvantages of EGG and acoustic analysis can be summarized as follows. Both are noninvasive low-cost techniques. Both analyses can be used to obtain measurements of vocal characteristics. They are especially helpful in patients with voice disorders that are not laryngostroboscopically visible or disorders without any apparent organic cause. When the view into the glottis is obscured by swollen false vocal folds or arytenoids, the examiner has to rely on EGG and acoustic analysis.

Acoustic and electroglottographic analyses are superior to many other instruments available for assisting in voice assessment in the sense that they can be used to record both connected speech and sustained phonations. Visual techniques like videolaryngostroboscopy are limited to the observation of vocal function during sustained phonations. However, it has to be emphasized that research on voice measurements made on running speech is still in its infancy.

As to clinical relevance of acoustic and EGG analysis, there are many subjects for whom EGG does not produce valid waveforms even when the electrodes are placed optimally. Suboptimal signals are those small in amplitude or noisy waveforms dominated by random noise from equipment and voice synchronous noise from tissues around vocal folds. In some subjects, with or without vocal pathology, anatomical properties of the neck prevent a clear EGG signal. In other subjects, acoustic and EGG analysis do not produce valid measurements because waveforms obtained from highly disturbed voices are too irregular and therefore not analyzable. In Hill et al. (1990), acoustic analysis was not successful in 35 % of voice patients; electroglottography failed in 46 % of patients. In most cases, either the patient was not able to sustain phonation for at least several seconds or the signal was too aperiodic including aphonic cases. One reason in favor of the limited validity of the numbers reported in this study is that the study sample was too small (n = 26) to allow general conclusions on the success rate of electroglottographic and acoustic analysis in voice patients.

When the software analysis is unsatisfactory, in many cases, the researchers are advised to resort to visual examination of the waveforms and the spectrum of the signal. Visual inspection of EGG and acoustic waveforms allow general judgment of periodicity of vocal fold vibration. Some irregularities in vibration may be detected already in the waveforms. Visual inspection of the spectral characteristics of the signal will be dealt with in the next section.

EGG waveforms help to make judgments concerning laryngeal adduction. On the basis of the obtained CQ and OQ values, it is possible to conclude normal adduction, hyperadduction or hypoadduction[5]. Hacki (1989) and Hacki (1996) were able to discriminate between hyper- and hypofunctional dysphonia on the basis of crescendo task. The range of OQ variation was found to be signficantly restricted in hyper- and hypofunctional dysphonia as compared to normal speakers. In normal speakers, OQ varied from 0.75 (at low vocal intensity) to 0.40 (at high intensity). Besides, OQ behaved differently in different types of dysphonia: OQ decreased with increasing intensity in normal and hyperfunctional speakers, but increased in hypofunction. Jilek et al. (2004) confirmed the results of the previous research conducted by Hacki by reporting successful differentiation of hypertonic and hypotonic voices with high sensitivity and specificity on the basis of the sum of the amplitude perturbation and the quasi-open-quotient perturbation measured before voice loading. Similar results were obtained in Verdolini et al. (1998b): subjects with hypoadduction disorders like nodules, paralysis and vocal fold bowing have smaller CQs than normal subjects without regard to voice quality that they are asked to produce.

An approximation of CQ and OQ can be derived from the inverse filtered microphone signals. With the exception of some noise parameters, many voice parameters that are routinely applied to Sp signals can be computed from the Lx waveforms. However, it is not possible to draw conclusions on mechanical properties of vocal folds by means of objective measurements from acoustic and EGG signals.

EGG and acoustic signals are useless in judging symmetry, amplitude and mucosal wave. Unlike photoglottography, electroglottography cannot differentiate between different types of paralysis (unilateral recurrent, superior and idiopathic) (Hanson et al., 1988). EGG waveforms do not allow conclusions as to which side is impaired, or where in the glottis contact is changing: high, low, back or front. No absolute measure of contact area is possible. Likewise, the completeness of the closure cannot be assumed from the EGG waveforms alone as EGG waveforms retain their characteristic pattern with one maximum and one minimum per period even if the closure is partial.

As shown in Scherer & Titze (1982), there seems to exist a relationship between CQ and impact stress[6]. In Verdolini et al. (1998a), the use of EGG CQ (35 % threshold level method) as a noninvasive indicator of the impact stress between the membranous vocal folds

[5] In hyperadduction, more than 60 % of the cycle is spent in glottal closure; in hypoadduction less than 40 %.

[6] Some lesions of the vocal fold tissue are believed to be a reaction to excessive mechanical stress. Different types of mechanical stress that can damage vocal folds are discussed in Titze (1994). Among the stresses that harbor a high risk of tissue damage are the tensile stress in connection with the action of the CT at high pitches, vocal fold impact (collision) stress and arytenoid contact stress. The maximum impact stress was positively related in Jiang & Titze (1994) to high subglottal pressure, excessive elongation and adduction. It is hypothesized that hyperadduction, loud voice and high pitch substantially contribute to voice abuse and vocal fold lesions, although methods to quantify different types of mechanical stress in human larynges are still in development.

was discussed. The authors found a strong correlation (r = 0.81) between CQ values and measures of impact stress (force per area) in canine larynges. An increase of 0.15 in CQ corresponded to approximately 1 kPa increase in impact stress. For the reasons stated in the following, the EGG CQ is probably not suitable to measure the impact stress between the arytenoids in human larynges.

Obviously, EGG signals contain more information on the membranous part of the glottis which significantly contributes to discrepancies between videostroboscopic and laryngographic findings. The cartilaginous part of the glottis has less impact on the EGG curve since cartilage has lower conductivity than wet muscle tissue and the density of the field lines is weaker in the back of the glottis, and so is the contribution to the signal-noise ratio (Titze, 1990).

Other limitations like ineffectiveness in detecting gross structural and tissue changes restrict EGG and acoustic analysis to the detection of mild and moderate voice dysfunction when videolaryngostroboscopy shows normal anatomic structures with normal-appearing periodicity and amplitude.

Scientifically, EGG's best success has been in voice research involving normal subjects and singing registers. It has been found that the CQ and OQ change as one varies loudness and height of the tone (Howard, 1995; Henrich et al. 2005; Mooshammer, 2010). Following general rules were proved to be valid for the OQ. OQ is greater for high tones than low ones. OQ decreases with increasing intensity. OQ tends to increase in age (Winkler & Sendlmeier, 2006) and in voices with glottal incompetence. Hanson & Chuang (1999) proved that male speakers have lower OQ values than female speakers due to more complete glottal configuration that hinders energy loss during phonation.

Electroglottography has helped to show that voice qualities other than modal are not always a sign of pathology; they are frequently employed to express the emotional and attitudinal state of the speaker and that modal, breathy, whispery, creaky, harsh and pressed voice qualities produced by the same nondysphonic subject have different waveform shapes and contact quotients (Verdolini et al., 1998b). The greatest CQs were found in pressed; the lowest in breathy voice. However, even if normal speakers can produce phonations that come close perceptually to dysphonic voice production, normal subjects are not able to copy the exact biomechanics of impaired voice production that is involved in pathology. This is why one should be cautious to generalize the results of these studies to dysphonic population.

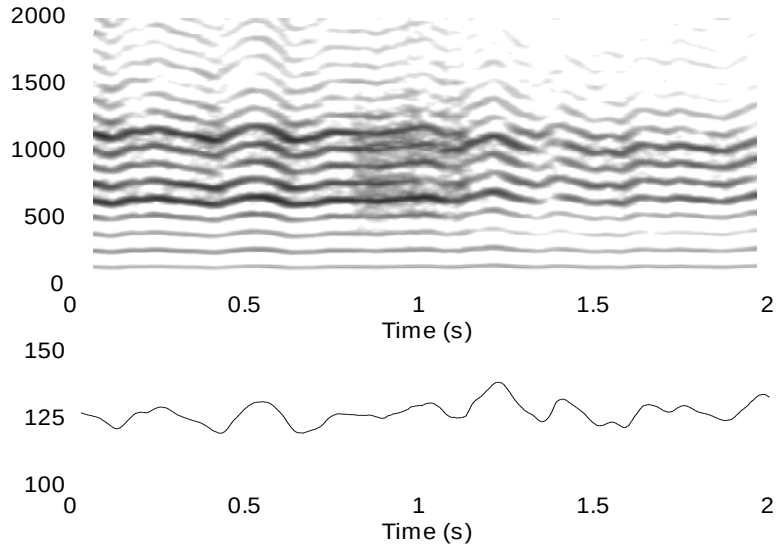
1.3.4 Spectral characteristics of normal and pathological vowels

Both Lx and Sp signals can be visualized via the spectrographic method. Spectrographic

analysis carried out on Sp signals is, however, superior since it allows both narrow-band and wide-band visual inspection of speech sounds. EGG signals are suited for a narrow-band analysis only which is useful in assessing the presence or absence of harmonic structure in the signal. For an accurate description of articulatory, acoustic and distinctive features of speech sounds in the wide-band spectrum the reader is referred to Jacobson et al. (1952) and Ladefoged & Maddieson (1996). In voice research, narrow-band spectrography is more important and has received a great deal of attention since invention of the method.

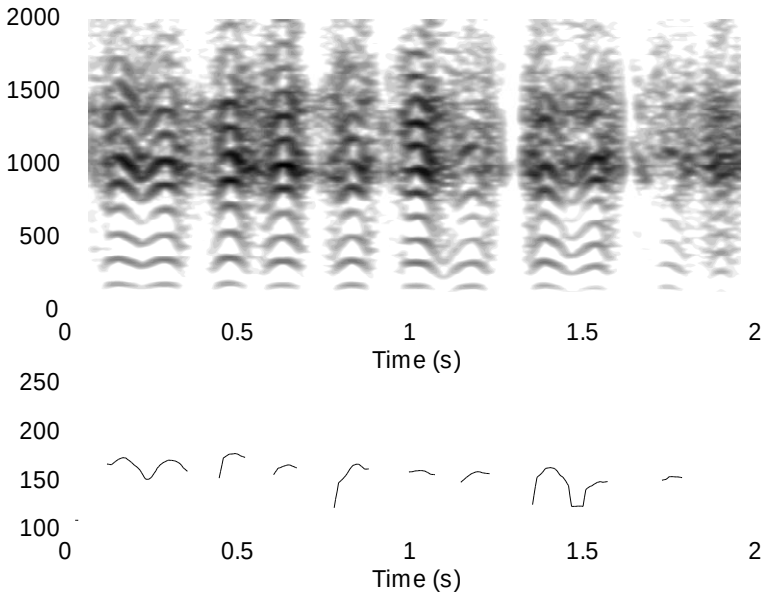
In sustained phonation of a healthy voice, the harmonic traces are expected to remain level. Waviness in partials (often referred to as vibrato in sung vowels) is a sign of pathology in sustained vowels as wavy partials reflect inability to control the larynx position (Fig. 22).

Fig. 22: Spectrogram showing a mid-vowel segment of a sustained /a/ produced by patient 17, male, 81, diagnosed with vocal cord dysfunction, and the corresponding F0 trajectory below. Tremulous phonation is evident as waviness in partials and fluctuations in F0.



Strong frequency and amplitude modulations in sustained vowels are typical for tremulous voices. Both vibrato and tremor normally arise from subtle movements of the larynx against the airflow; in the latter case involuntary movements that are transmitted to other supraglottal structures. Other locations of tremor besides vertical larynx movements are true vocal folds, arythenoid cartilages, pharyngeal walls, epiglottis, tongue and strap muscles (Perez et al., 1996). Independently of tremor location, voice tremulation involves excessive pitch fluctuations with a fairly constant frequency (Lebrun et al., 1982) that can be confused with the simultaneous presence of two distinct pitches. Cavalli & Hirson (1999) observed that vibrato can be perceived as diplophonia. Fig. 23 shows another example of tremulous voice with sudden interruptions in voice production (voice arrests) at irregular intervals.

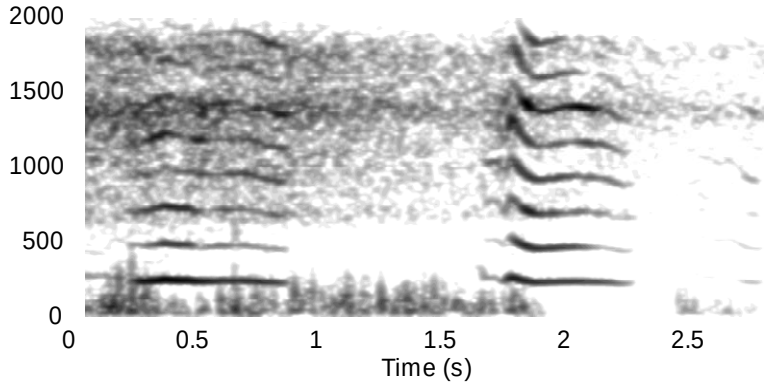
Fig. 23: Spectrogram of a mid-vowel segment of a sustained /a/ showing tremulous phonation interrupted by voice arrests and the corresponding F0 trajectory below. Patient 51 (female, 65) is diagnosed with spasmodic dysphonia.



Voice stoppages can be equally caused by either spasmodic closures of the glottis or by a sudden slackening of the vocal muscles (Lebrun et al., 1982). Inability to maintain constant subglottal pressure may also result in involuntary and intermittent episodes of aphonia. Wavy harmonics can be observed in the EGG signal, as well.

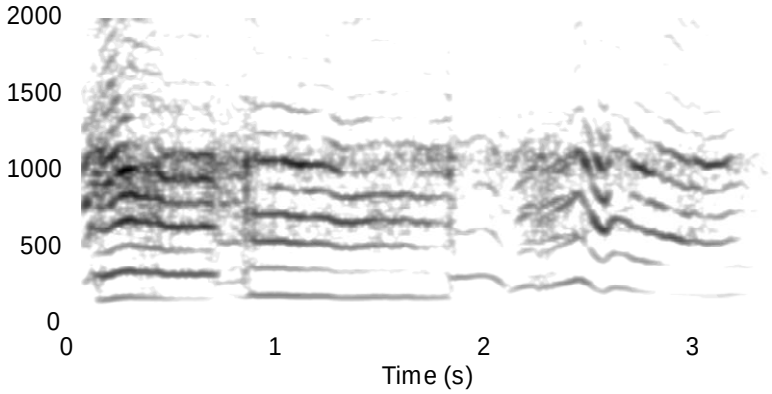
Loss of harmonic structure is another abnormality typical for pathological vowels. In suboptimal phonation, higher harmonics are often weak and replaced by noise. However, destruction of harmonic structure below 2000 Hz is strongly suggestive of a highly disturbed voice. Noise source is characterized by the presence of random energy in a narrow-band spectrum (Fig. 24).

Fig. 24: Spectrogram showing a mid-vowel segment of sustained /a/ from a patient with psychogenic aphonia and extremely unstable voice quality, female, 15, illustrating intermittent transitions from normal to completely irregular mode of vocal fold vibration.



Bifurcations like pitch jumps, splits in F0, emergence of two independent frequencies or subharmonics can be seen in spectrograms of pathological voices. Sudden jumps to a new frequency were interpreted in Berry et al. (1996) as cases of transition from single vocal fold vibrations to synchronized vocal fold vibrations. At large subglottal pressure, instabilities may occur in symmetric vocal fold vibration as well. Several bifurcation types can be present in one phonation. In Fig. 25, the patient starts phonation at 157 Hz. In the course of phonation he switches the register twice by moving up to 298 Hz and back. In an interval between 2.1 s and 2.5 s, two independent frequencies can be observed resulting in bitonal voice quality. The patient corrects the pitch by gradually moving down to the original frequency of phonation.

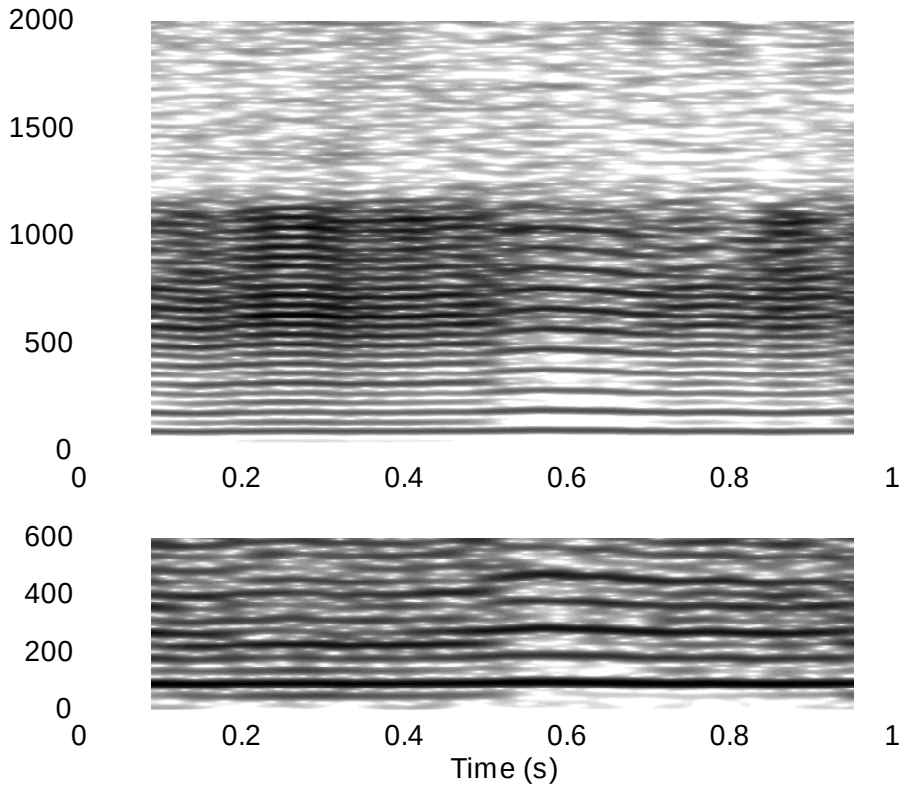
Fig. 25: Spectrogram of a mid-vowel segment of a sustained /a/ from patient 27, male, 72, diagnosed with a polyp.



Variation in number of harmonics is a further acoustic abnormality that sometimes appears in asymmetric vocal fold vibration. Essential for the visual detection of the subharmonics in the spectrogram or power spectrum is a clear defined harmonic structure. Subharmonics are whole-number divisors of the fundamental frequency and can be seen as additional traces between the harmonics (Fig. 26).

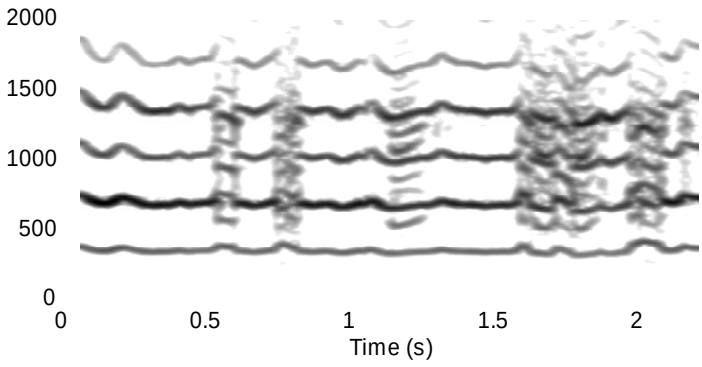
Subharmonics are usually weaker than harmonics and do not persist throughout the duration of the vowel but emerge and disappear from the spectrum. As the high-frequency harmonics in pathological voices are often obscured by noise, the inspection of the lower part of the spectrum up to 2000 Hz is sufficient for screening vowels for subharmonics. Considering a wide variety of pathologies that cause asymmetric vocal fold vibration or desynchronization of the vibratory modes of a single vocal fold, subharmonics are expected to be common in disordered voices.

Fig. 26: Narrow-band spectrogram obtained from synchronized acoustic (top) and EGG signals (bottom) illustrating subharmonic frequencies during phonation of /a/ sustained at 90 Hz with a subharmonic element at 45 Hz in patient 115, male, 37, diagnosed with a polyp.



The number of subharmonics that can be resolved in the spectrum depends on the resolution of the spectrum. Up to three subharmonics can be distinguished between two consecutive harmonic traces. Multiple peaks between two consecutive harmonics are resolved as noise in the signal.

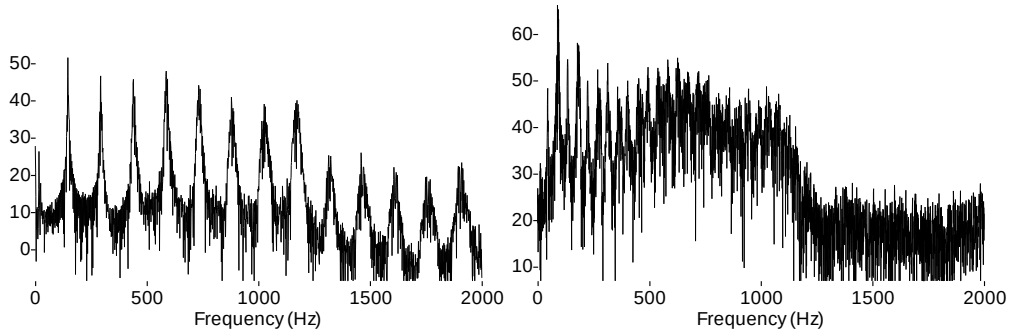
Fig. 27: Spectrogram of a mid-vowel segment of a sustained /a/ with two interharmonics between the harmonics in the interval between 1 s and 1.5 s produced by patient 29, m, 77, diagnosed with vocal fold carcinoma.



True subharmonics stand in a harmonic relationship to the fundamental frequency and contribute to the perceived pitch. Contrariwise, the sensation of biphonation or diplophonia can be evoked by superimposition of two frequencies that are not harmonically related; at least, in theory. By visual examination it is not always possible to distinguish between true subharmonics of the fundamental from separate additional frequencies. The accuracy of the measurement is plus/minus 9 Hz. That is why researchers often do not distinguish between subharmonics and independent frequencies: any interharmonic between the main harmonics qualifies as a subharmonic. Dejonckere & Lebacq (1983) were the first to relate subharmonics to the perception of diplophonia. Wong et al. (1991) made the same observation. In Cavalli & Hirson (1999), 90 % of vowels perceived as diplophonic had subharmonics. Klatt & Klatt (1990) related subharmonics with perceived creak.

Power spectra are helpful in looking for subharmonics and noise. The power spectra of normal voices show relatively clearly defined maxima at regular intervals that correspond to harmonics (Fig. 28).

Fig. 28: Power spectrum of a sustained vowel /a/ by a healthy subject (left). Power spectrum of a sustained vowel /a/ by patient 115 (right). Note the smaller alternate peaks between the main harmonics on the power spectrum that correspond to subharmonic frequencies.



Power spectra of pathologic voices show additional spectral peaks between the harmonics, the peaks are usually widened and notched. In the upper part of the spectrum, the peaks and troughs are badly defined. This characteristic can be exploited in voice analysis. Sasaki et al. (1991) proposed an objective method to quantify the noise (*N*) and total acoustic energy (*V*) using power spectra of vowels and relate it to the perceptual degree of hoarseness. *V* was calculated as the area enclosed between the baseline and the line connecting the peaks; *N* as the area enclosed between the baseline and the line connecting the troughs. The measured parameter *N/V* strongly correlated with hoarseness ratings with a r_s of 0.79 in male and 0.81 in female voices.

Power spectra can help in assessing the relative strength of harmonics. The strength of partials normally decreases at a rate of 6 dB/octave. However, some harmonics, especially

those that fall into formant frequencies, have more energy than others. There is empirical evidence that vowels produced with a stiff or creaky voice have more energy in the harmonics in the region of F1 and F2, whereas vowels produced with slack or breathy voice quality have more energy in F0 and more random energy in the higher frequencies (Hammarberg et al., 1980; Kitzing, 1986).

Spectrograms reveal some but by no means all properties of the signal that can be related to acoustic parameters. The more the spectrum deviates from the norm, the more disturbed is the voice. Whereas gross structural irregularities or noise are hard to overlook, finer changes cannot be assessed with the spectrographic method. Such parameters as jitter and shimmer measure very fine changes in the period or amplitude of the signal that cannot be visually captured. The spectrographic analysis as such is based on the overall impression of the abnormality of the spectrum that is hard to quantify.

To assess the severity of voice disorder in a population with voice problems, it is common to classify pathologic voices with respect to signal type. The most common classification is that proposed by Titze (1995). However, there are other possible classifications (Yanagihara, 1967). This classification is still in use as a subjective index of the degree of hoarseness. There are 4 hoarseness types in this classification that uses the amount of noise and loss of harmonic structure in spectrograms as classification criteria. Three vowels /a/, /e/ and /i/ are needed to assess the spectrographic hoarseness type.

If perceptual voice quality is related to spectral voice characteristics, it deems reasonable to assume that spectrographic analysis can contribute to perceptual voice quality judgements. There have been multiple studies on this subject. Thus, roughness seems to relate to F0 perturbations, low-frequency modulations, the presence of subharmonics and chaos (Omori et al., 1997; Herzel et al., 1994). Similarly, breathiness seems to be connected to amplitude pertubations and the presence of noise. Though, simple spectrographic analysis may fail to detect perceptual breathiness.

The impact of spectrographic analysis on perceptual voice ratings was studied in Martens et al. (2007). They showed that perceptual ratings in different perceptual categories were differently affected by spectrographic analysis of voice: the average hoarseness rating increased and the average breathiness rating decreased in retest tasks done with spectrographic data; judgements on roughness remained, however, unchanged. The inclusion of spectrograms in perceptual rating procedure resulted in greater interrater agreement between experienced judges that was expressed by Fleiss' kappa. Interestingly, although perceptual ratings changed after the inclusion of spectrograms, the correlations between the perceptual ratings and acoustic parameters did not change significantly.

1.3.5 Signal typing

In Titze (1995) acoustic signals were classified into Type 1 (nearly periodic), Type 2 (signals with voice breaks, subharmonics, modulations) and Type 3 (chaotic, without apparent harmonic structure) signals. Behrman et al. (1998) noted potential problems with this classification: whereas the identification of Type 1 signals (Fig. 29) was fairly straightforward, the categories Type 2 and Type 3 did not withstand a critical analysis in clinical application. Fig. 30 and Fig. 31 show typical Type 2 and Type 3 signals, respectively.

Fig. 29: An example of a typical Type 1 signal. Traces a and b show a close-up of some of the cycles extracted at the points that correspond to the solid lines drawn vertically through the EGG and microphone signals. The power spectra are extracted from the center of the signal. The frequency axis of the spectrograms is log-scaled (from Behrman et al., 1998).

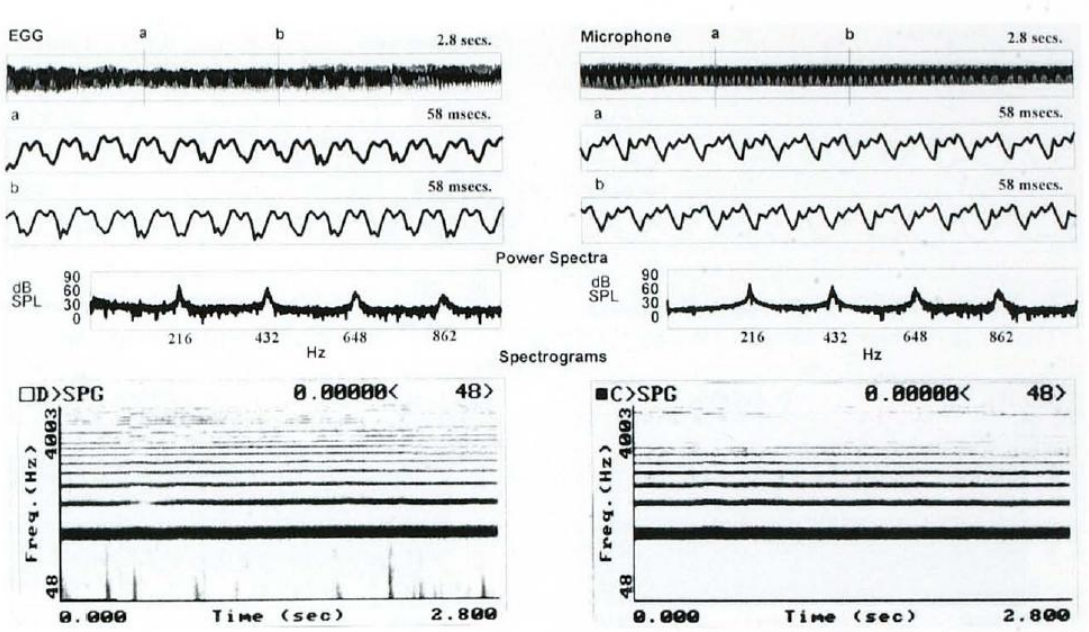


Fig. 30: An example of a typical Type 2 signal (from Behrman et al., 1998).

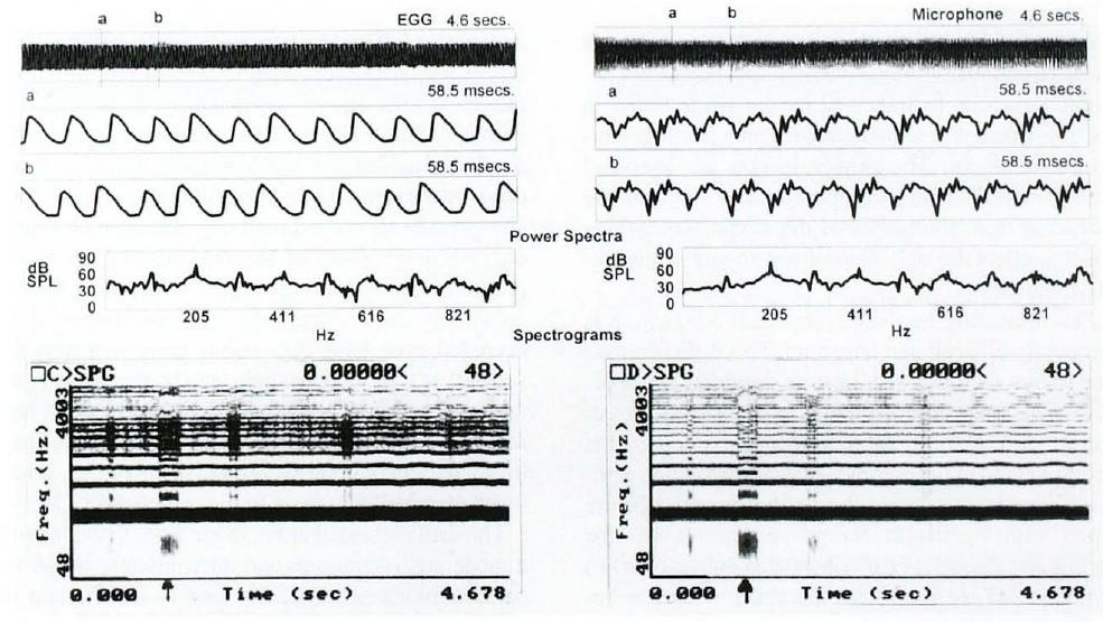
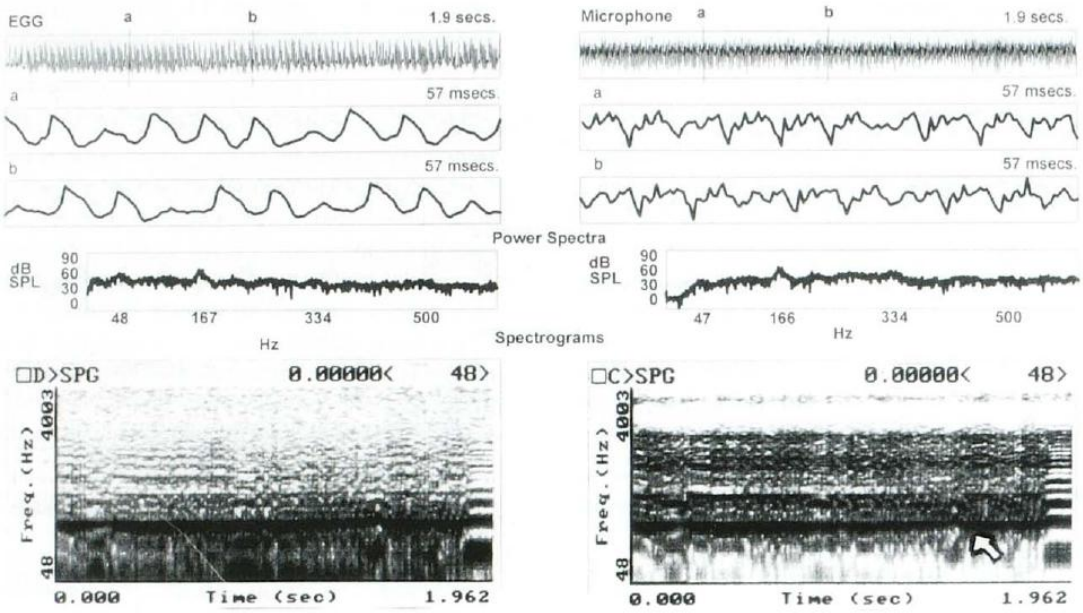


Fig. 31: An example of a typical Type 3 signal (from Behrman et al., 1998).



Ambiguous cases included predominantly nearly periodic signals with a short segment with structural or chaotic irregularities or predominantly chaotic signals that still retained some harmonic structure. They had problems with assigning 40 % of dysphonic voices to either Type 2 or Type 3 suggesting that this tripartite division is not always adequate for dysphonic voices. By way of dealing with the signal type classification problems reported in Behrmann et al. (1998), we introduced two additional signal types.

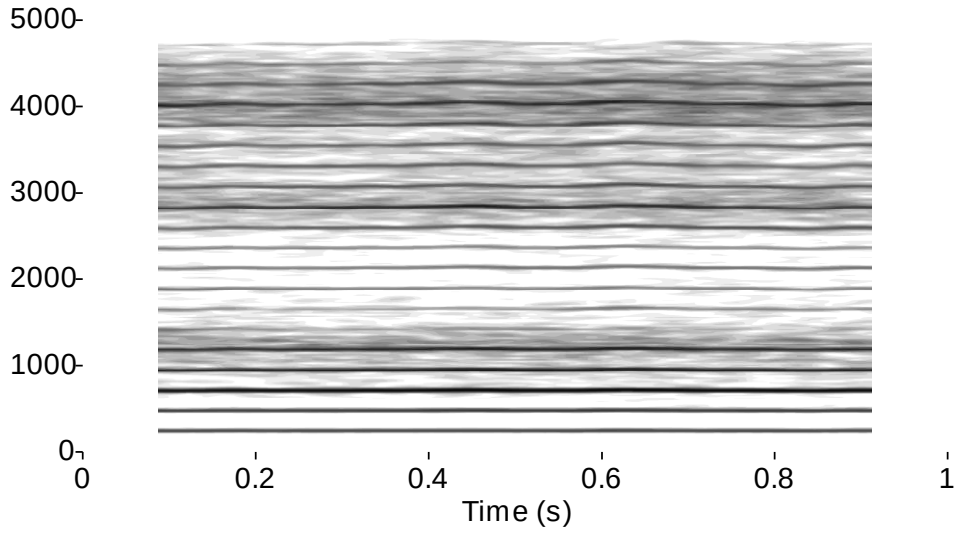
In voice research, voice parameter extraction has often been limited to Type 1 signals as they ensured reliable results. However, the exclusion of non-Type 1 signals from research statistics raises a question of utility in quantifying vocal quality, especially in pathological

voices, as most pathological voices are non-Type 1 signals. According to Behrman et al. (1998), the incidence of non-Type 1 signals in clinical population amounts to 58 %. Acoustic signals that are rejected as unreliable for cycle-to-cycle frequency or amplitude perturbation measurements may be perfectly suitable for noise measurements. The quantification of noise, be it produced by irregularities in the vocal fold vibration or incomplete glottal closure, is another important target in voice quality assessment.

In this section we propose a slightly modified classification of signal types that accounts for both structural irregularities and random-appearing phenomena. Signal typing was carried out on acoustic waveforms on the spectrum ranging from 0 to 5000 Hz. Signal spectra were classified in 5 classes based on the assessment of the harmonic structure, the run of the F0 contour and the amount of noise in the spectrum of a vowel. The same classification cannot be applied to EGG signals. In EGG signals, noise that masks harmonic energy above 1000 Hz is of electrical origin. Therefore, electroglottographic signals can only be screened for the presence of structural irregularities in the lower part of the spectrum.

The following classification of spectrographic vowel types is suggested in the present paper: Type 1 signals are defined as nearly periodic. They may contain a certain amount of low-energy noise between the harmonics as long as the harmonic structure remains preserved in the range up to 4–5 kHz (Fig. 32).

Fig. 32: Example of Type 1 signal: Spectrogram of /a/ by patient 5, female, 29, diagnosed with vocal nodules.

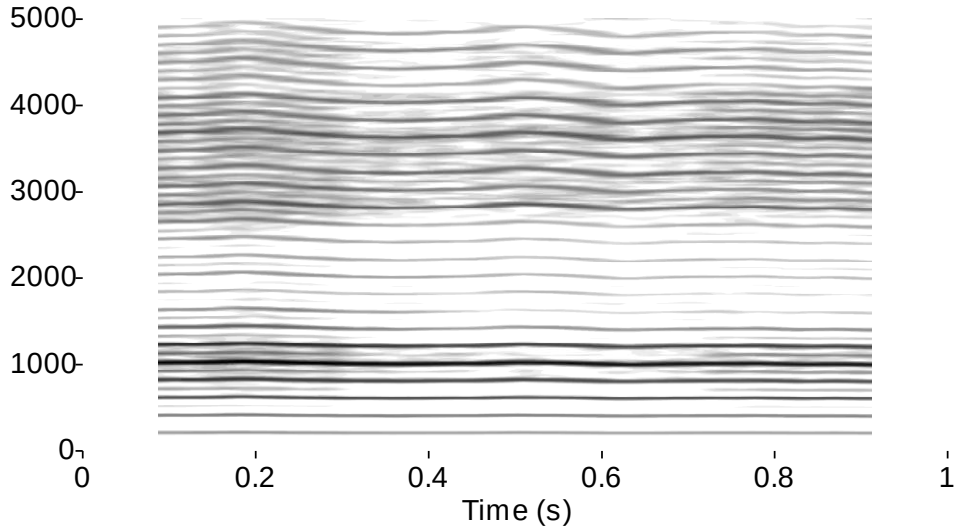


Type 1 signals are either characterized by a flat F0 contour or gradual insignificant F0 changes. Loss of harmonic structure due to excessive signal damping in the upper part of the spectrum was treated as low-energy noise. High-energy noise is characteristic for other

spectrographic types. Type 1 signals correspond to Type 1 signals in traditional classification by Titze.

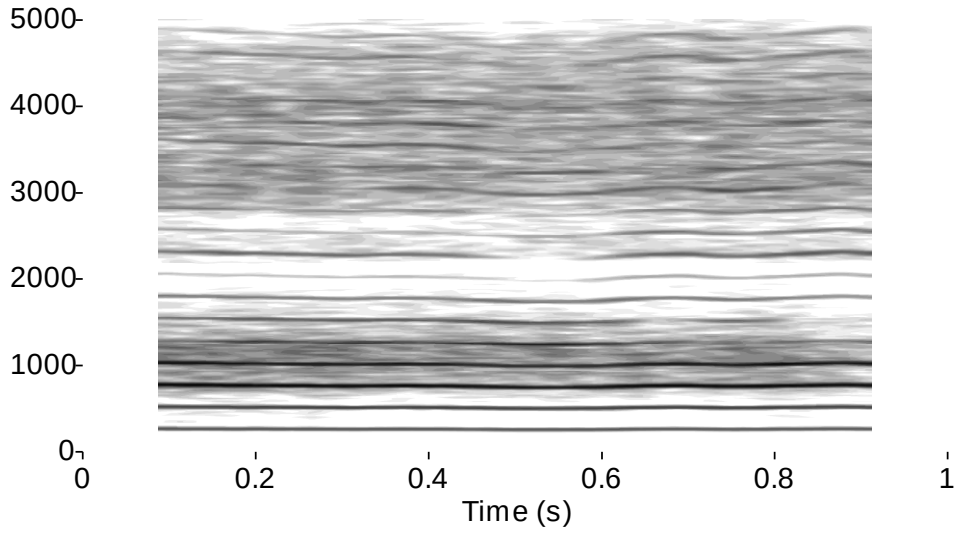
Type 2 signals show some structural irregularities (Fig. 33). Type 2 signals contain subharmonics, strong frequency modulations (evident as excessive waviness in partials), strong amplitude modulations (lack of uniform colour in partials: partials' alternate between darker and lighter parts) and bifurcations including those caused by voice arrests, voice interruptions through noise, the presence of two unrelated frequencies and sudden pitch jumps or F0 splits during phonation. The harmonic structure is largely preserved. Type 2 signals may contain entirely chaotic segments of short duration that are treated as voice interruptions as long as the periodic structure dominates the spectrum. The F0 contour normally shows a bimodal distribution, large variation in F0 values or abrupt changes with pitch breaks and voice interruptions. Type 2 signals correspond to Type 2 signals in traditional classification. The reliability of voice measures derived from Type 2 signals will depend on the presence or absence of structural irregularities in the segment to which parameter extraction is applied.

Fig. 33: Example of Type 2 signal with subharmonics: Spectrogram of /a/ by patient 9, male, 57, diagnosed with unilateral vocal fold paralysis.



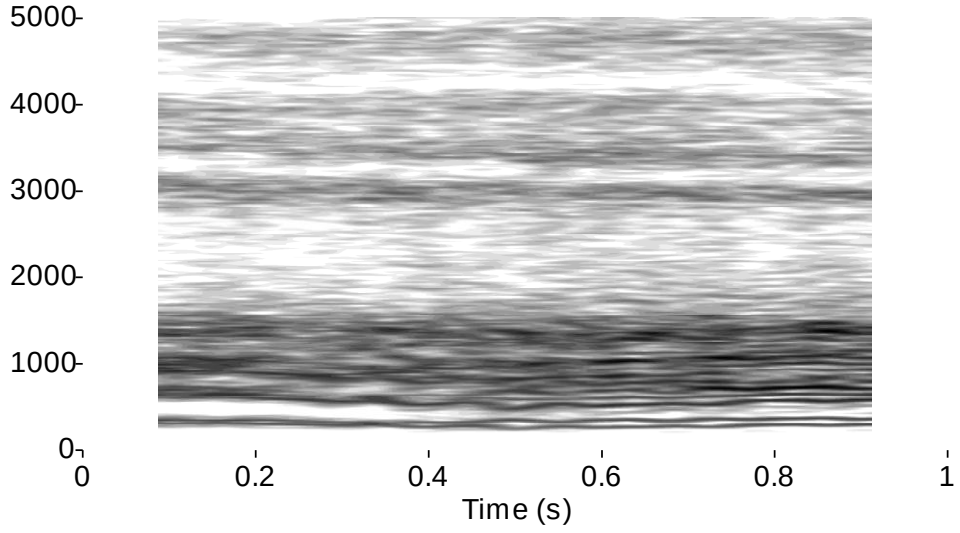
Type 3 signals differ from Type 1 signals by noise in the spectrum. Type 3 signals are defined by the presence of regular harmonic structure in the lower part of the spectrum, with harmonics in the upper part of the spectrum being replaced by noise (Fig. 34). The F0 contour is flat or shows gradual changes. Type 3 signals may correspond to Type 1 signals in traditional classification. Noise in the upper harmonics does not interfere with F0 extraction and may not affect measures based on F0 extraction.

Fig. 34: Example of Type 3 signal: Spectrogram of /e/ by patient 65, female, 82, larynx carcinoma T1 on the right vocal cord, laryngeal epitheleal carcinoma after frontolateral partial larynx resection.



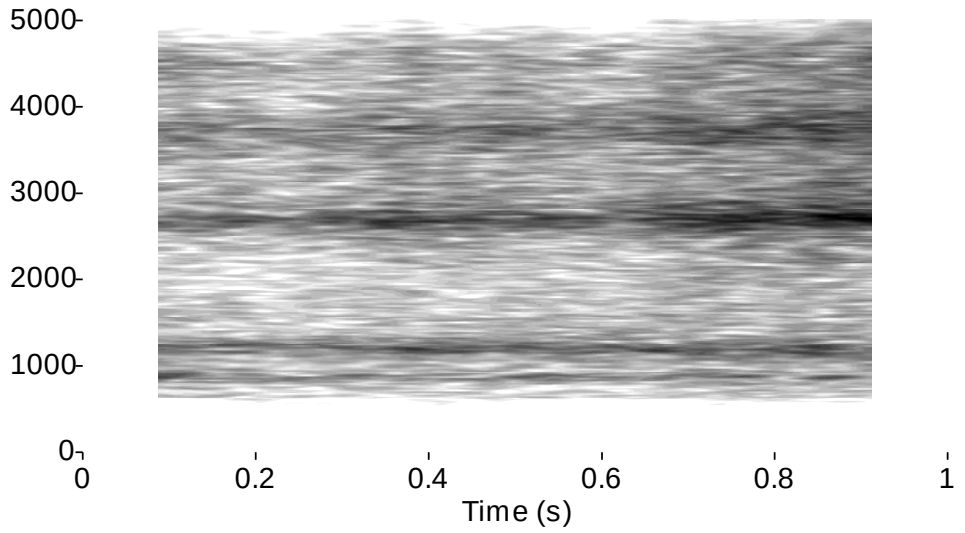
Type 4 signals are characterized by both structural irregularities and a large amount of noise (Fig. 35). They contain strong subharmonics, frequency and amplitude modulations and bifurcations including the presence of two unrelated frequencies, interruptions in voicing and pitch jumps in the lower part of the spectrum. The F0 contour shows a bimodal or multimodal distribution and abrupt changes. The harmonic structure in the upper part, and to some extent in the lower part, of the spectrum is replaced by noise or lost as a consequence of excessive damping. Type 4 signals correspond to Type 2 or Type 3 signals.

Fig. 35: Example of Type 4 signal with a diplophonic voice quality: Spectrogram of /a/ by patient 3, female, 76, leucoplakia. A split in F0 occurs at 0.4 s.



Type 5 signals are pure noise and show no apparent or very little harmonic structure (Fig. 36). The F0 contour is highly irregular or non-existent. The darker horizontal bands in Fig. 36 are noise formants. Type 5 signals correspond to Type 3 signals in traditional classification. Voice measures derived from Type 4 and Type 5 signals are unreliable.

Fig. 36: Example of Type 5 signal: Spectrogram of /e/ by patient 146, male, 68, after partial larynx resection.



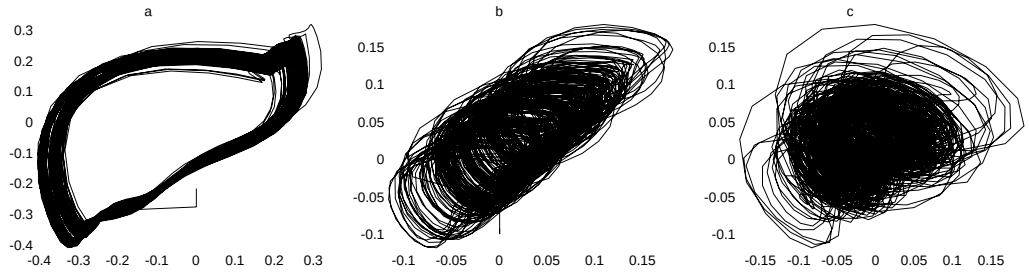
1.3.6 Normal and pathological vowels in the phase space

Another way to visualize voice signals is phase space reconstruction. Time-delay embedding is the most commonly used technique to build phase portraits from time waveforms. In an *m*-dimensional phase space, the original signal is plotted against *m–1* delayed copies of itself.

The choice of the delay time τ and the embedding dimension *m* is somewhat arbitrary and often set by trial and error. When the coordinates of the *m* trajectories are projected into the phase space, a geometrical figure arises which is termed an attractor. Each point of the attractor has the coordinates $\{x(t), x(t + \tau) \ldots\ldots x(t + [m - 1]\tau)\}$. Attractors give preliminary information about the characteristics of the signal. Two-dimensional plots are sufficient to visually assess the attractor type (Herzel et al., 1994).

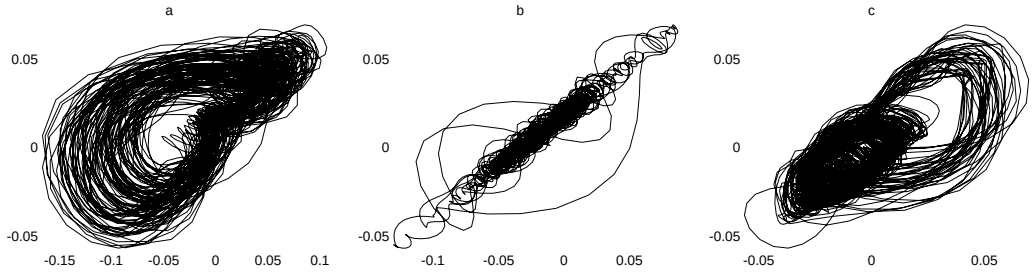
Attractors can be limit cycles, tori and chaotic attractors (Fig. 37). Quasiperiodic signals form attractors that come close to limit cycles. Signals with a linear combination of two frequencies form a torus. The difference between the limit cycle and the torus is the direction in which the signal is spiralling (longitudinal vs. transversal). Strange attractors are formed by noisy and unstable signals, and are characterized by the absence of any structure.

Fig. 37: Examples of the three basic attractor shapes: a) limit cycle, b) torus, c) chaotic attractor. 2D phase plots with $m = 3$, $\tau = 8$ and $N = 5000$ were derived from EGG signals: a) subject 18 without vocal pathology, male, 25; b) subject 24, female, 51, diagnosed with vocal fold edema; c) subject 23, female, 71, diagnosed with vocal fold paresis following strumectomy. The EGG signal in c) contained both frequency modulations stretching over several cycles and an abrupt transition into a subharmonic regime of vibration.



In the phase space, stable signals form a tightly wound ring with the radius depending on the amplitude of the signal (Fig. 38a). Voicing stretches phase portraits along the diagonal plane. Fig. 38 gives further examples of attractors from pathological voices. Smaller deviations from periodicity do not change the shape of the attractor. But they make the orbits appear more disperced (Fig. 38a). The orbits of natural vowels never exactly repeat themselves. On the contrary, synthetic vowels are reported to show no dispersive behavior (Narayanan & Alwan, 1995). Larger deviations from periodicity like period doublings, or subharmonics, may result in two coexisting rings (Fig. 38c). The more unstable is the signal, the more rapidly the orbits will tend to diverge and affect the attractor shape. Herzel et al. (1994) showed that dysphonic patients had more complex attractors than normal controls.

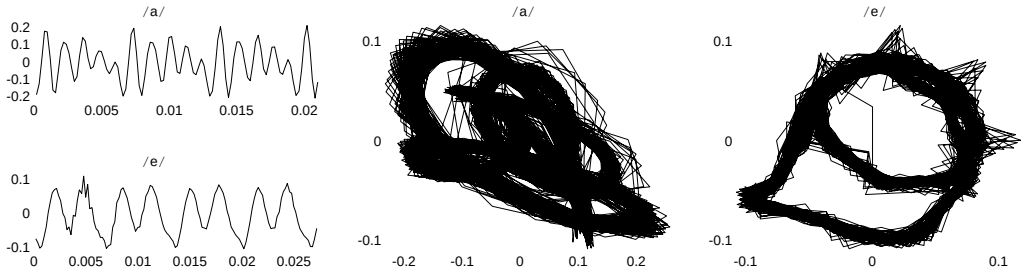
Fig. 38: 2D phase plots with $m = 3$, $\tau = 8$ and $N = 5000$ derived from EGG signals: a) subject 10, male, 44, diagnosed with a contact granuloma, postoperative condition; b) subject 134, male, 69, using ventricular fold phonation presented as very irregular changes in VFCA after partial larynx resection; c) subject 44, male, 66, experiencing hoarseness without apparent organic cause.



Acoustic and EGG signals differ in information content of phase plots. Narayanan & Alwan (1995) and Kumar & Mullick (1996) pointed out the correspondence between time waveforms and phase plots. The number of loops in the phase plot reflects the number of significant peaks per pitch period in the time waveform. The greater the amplitude of the peaks, the larger the relative size of the corresponding loop. If an acoustic waveform of a vowel has 5 peaks per period, as is the case with sustained /a/, the resulting attractor will have up to 5 loops. For this reason, electroglottographic signals are better suited for visualising

vowels in the phase space as they normally have one prominent peak per period. More than one loop in attractors derived from EGG signals can be interpreted as a sign of pathology. The phase portraits from acoustic signals /a/ and /e/ are shown in Fig. 39.

Fig. 39: 2D–phase plots and corresponding waveforms of sustained /a/ and /e/ elicited from subject 11, male, 66, diagnosed with recurrent laryngeal nerve paresis (m = 3, τ = 4, N = 5000).



The complexity of phase portraits obtained from vowels can be quantified by the largest Lyapunov exponent (LLE) that will be discussed in the sections to follow. Giovanni et al. (1999b) reported that LLE from patients with unilateral vocal fold paralysis was significantly different from normal controls. They could successfully predict the degree of hoarseness in 71 % of cases when LLE was included. Further, they suggested that LLE could assist in distingiushing between Type 1 and Type 2 signals in traditional classification. The relationship between signal types and attractor shapes has not been well studied so far.

1.4 Review of literature on automatic voice quality classification by objective parameters

Studies on automatic voice quality classification differ in many aspects, which may significantly determine the research outcome. They differ in the composition of voice parameters that they use. They may concentrate on different perceptual dimensions or use different perceptual rating scales, classification techniques and cross-validation methods. Besides, factors like choice of the signal to extract voice parameters from, choice of speech materal, length and composition of data regarding sex, type and severity of pathology, exclusion of outliers and extremes from statistics seem to play a crucial role.

It is common to rate perceptual voice dimensions on a discrete four-point scale 0–3, ranging from normal (0) through mildly (1) and moderately (2) to severely (3) disturbed, although other scales including discrete 7-point scales, dichotomous (normal vs. pathologic) and continuous scales are also in use. The most popular scales are the GRBAS scale with 5 perceptual dimensions and its lighter version, the RBH scale, with 3 dimensions. Recent

research has been concentrated on the overall severity of dysphonia G that corresponds to H in the RBH scale. Studies to predict other perceptual dimensions like breathiness, roughness, astenicity and strain are less frequent. Among the techniques that have been proposed to relate objective voice parameters and perceptual voice quality, the most popular were discriminant and regression analysis, self-organizing maps and artificial neural networks in the sense of multilayered feedforward classifiers. Since many objective paramters appear to be sex sensitive, some studies that address automatic voice classification avoid mixed gender study populations either by focussing on either male or female voices or by analizing them separately (Yu et al., 2001; Yu et al., 2002). Most studies attempted to predict voice quality from acoustic signals and sustained vowels alone.

Quantative measures can be applied for evaluation of therapy and to support diagnostics only if they carry important information on voice function. Many acoustic voice measures have been in the focus of voice research to prove their significance in the perception of pathologic voice quality. If the relationship between voice measures and perceptual scoring were linear, the prognostic values of measured parameters could be reliably assessed by correlation coefficients. However, there is a large body of evidence that objective measures do not correlate well with perceptual voice quality. Correlations between perceptual ratings of expert raters and measured voice parameters have been a subject of discussion in many studies dealing with dysphonic voices. The strength of association between any single measure and perceptual ratings were mostly low or moderate (Kreiman et al., 1990; Dejonckere et al., 1996; Yu et al., 2001; Bhuta et al., 2004), which compelled researchers to look for a combination of measures to predict perceptual voice quality. Up to now, it is not defined which acoustic and aerodynamic parameters should be used in the multiparametric protocol.

Probably, the most well-known multiparametric method to assess the overall severity of dysphonia G is a linear regression equation called the dysphonia severity index (DSI) proposed by Wuyts et al. (2000). The four parameters included highest frequency, lowest intensity, maximum phonation time and jitter. A negative value of DSI is associated with severe dysphonia; a positive DSI with mild dysphonia or normal voice. The average classification rate by DSI as estimated on 319 patients with dysphonia and 68 normal voices was 50 %. Subsequent studies on normal population have shown that DSI seems to be sex independent since sex-dependent parameters frequency and MPT counterbalance each other but not age-independent (Hakkesteegt et al., 2006).

Giovanni et al. (1996) used 4 acoustic and aerodynamic parameters including jitter, spectrum variable (defined as number of harmonics above noise level/ F0), glottal leakage (oral airflow/intensity) and voice onset time to predict the perceptual dimension G with discriminatory factorial analysis. The average classification rate that was achieved in this study

with 239 subjects (88 controls and 157 dysphonic subjects) approximated 67 %.

Using quadratic discriminant analysis, Yu et al. (2001) achieved with 6 objective parameters a 86 % concordance with perceptual ratings of G. The misclassified cases were mostly G1 and G2. The study involved 84 male subjects of whom 21 presented normal voices. The pertinent parameters were F0 range, largest Lyapunov exponent, subglottic pressure, maximum phonation time, signal-to-noise ratio, and F0. In a similar study (Yu et al., 2002) involving 74 females (68 dysphonic subjects and 6 controls), the authors were able to predict the G ratings in only 65 % of all cases. A larger mixed-sex study involving 391 patients (270 females and 121 males and 58 controls) classified the same set of measurements with a mean success rate of 80.5 % (Yu et al., 2007).

In Linder et al. (2008), using an approach based on artificial neural networks (ANN), 80 % of voices were classified correctly as either healthy (H0–H1) or hoarse (H2–H3). The four-point scale was used for the perceptual dimensions R and B. The mean success rate for R and B after cross-validation was estimated at 58 % and 64 %, respectively; the poorest results being achieved in R0, R3 and B2. The sample size was 120 study subjects (male = 48, female = 72, normals = 8). The objective voice features included jitter, shimmer, GNE and mean period correlation. The same data set was previously used in Schönweiler et al. (2001) to predict R and B by MDVP voice parameters. Classification by regression trees matched the perceived voice quality in 65 % and 63 % for R and B dimensions, respectively. The mean classification accuracy of FNN after cross-validation was lower with 57 % for R and 41 % for B. ANN performed better in mildly and moderately disturbed voices.

In Ritchings & Berry (2006), one of the few studies that makes use of both EGG and acoustic signals to predict vocal quality G, ANN performed badly with the mid-ranking voices irrespective of the choice of the signal. The study assessed voice quality in 178 subjects recovering from radiotherapy for laryngeal cancer on a three-point scale. The percentage of correctly classified good voices was 64 % and 63 % for impedance and acoustic signals, respectively. Bad voices were classified correctly in 83 % and 91 % of cases. Medium-quality voices had the lowest classification accuracy with 26 % and 32 % in impedance and acoustic signals, respectively. This study is outstanding in the sense that comparable match rates were achieved with the same 19 short-term and 3 long-term parameters that were extracted from different signals, which speaks in favour of the hypothesis that ANN seems to be insensitive to the choice of the signal.

Maryn et al. (2010) have recently presented the results of the study that combined both acoustic measures from connected speech and sustained vowels to predict the overall voice quality. Stepwise multiple regression analysis yielded a six-variable model, cepstral measure being the most important variable. The model correlated strongly with the overall

voice quality (r_s = 0.78) and identified healthy vs. pathologic voices with high sensitivity and specificity (ROC area = 0.895).

A direct comparison between the studies is doomed to fail due to differences in technical implementation as one can hardly compare measurement outcomes of different data acquisition hardware and software, different computer systems and measurement algorithms; similarly, it is difficult to compare the measurements of the same voice parameter across different studies. In spite of an enormous research effort, the classification of voice quality by acoustic parameters is still not quite satisfactory. Nevertheless, it appears that predicting voice quality by objective voice parameters is a promising avenue of development.

1.5 Aims of the dissertation

In the present dissertation, a representative sample of 145 dysphonic and 5 normal voices was used for perceptual rating and parameter extraction. The relevant voice parameters were classified according to one of the three examined perceptual dimensions: roughness, breathiness and hoarseness. Two methods were employed to derive classification results: discriminant analysis and ANN. The aims of the present dissertation were set at:

1) Improving classification success rates by means of combining different parameters extracted from both Sp and Lx signals.

2) Testing a wider range of parameters including aerodynamic and prosodic ones to find a combination of sufficiently diverse parameters. Ideally, in each perceptual dimension they should be motivated by voice physiology.

3) Finding the best set of variables with the best discriminative power.

4) Including measures describing connected speech since it is problematic to infer conclusions on perceptual severity of speech from voice parameters measured on sustained vowels.

5) Introducing some new parameters, especially those not based on F0 detection.

6) Not restricting the range of voice pathologies to Type 1 signals.

7) Addressing the issue as to which signal is most useful for extracting voice parameters for discrimination purposes.

Whereas rating auditory impression in good and bad voice qualities is fairly straightforward and mid-ranking voices are more difficult to assess, it can be anticipated that classification results will be lower for intermediate grades of dysphonia. We expect further that inclusion of poor signals that are normally found in severely disturbed voices would not affect classification rates in normal and mildly dysphonic voice categories since parameters measured on poor signals are reportedly unreliable, mostly outliers and extremes.

Chapter 2: Materials and Methods

2.1 Subjects

Study subjects consisted of 145 patients seeking medical advice for voice problems in the period from July to November 2007 and 5 subjects without history of vocal pathology. All types of dysphonia were included. Some of the patients were diagnosed with more than one voice disorder. The distribution of patients by sex, age and diagnosis is shown in Table 1 and Table 2. Organic disorders were predominant in the data.

Table 1: Distribution of study subjects by age decades.

Age	*Patients*		*Healthy Subjects*		*Total*	*Cumulative Distribution in percent*
	Male	*Female*	*Male*	*Female*		
10–20	1	3	0	1	5	3.3
20–30	1	6	2	0	9	9.3
30–40	7	9	0	1	17	20.6
40–50	9	11	0	1	21	34.6
50–60	23	13	0	0	36	58.6
60–70	18	9	0	0	27	76.6
70–80	20	9	0	0	29	96.0
80–90	3	3	0	0	6	100.0

In order to represent organic disorders, patients were included having anatomical and histological alterations of the vocal cords, neuromuscular diseases, or both. Thus, the study subjects presented polyps (9), vocal nodules (3), vocal fold paresis or paralysis (25), leukoplakia (13), T1 and T2 vocal fold cancer (21), laryngeal trauma not connected to strumectomy or intubation (3), granulomas (3), cysts (2), spasmodic dysphonia (3), psychogenic dysphonia (2), monochorditis (1), papillomatosis (1), Reinke's edema (11), vocal cord dysfunction (1), chronic laryngitis (6), acute laryngitis (1). The rest of the cases divide between functional dysphonia (hyper- or hypofunctional) and patients presenting the symptoms of hoarseness in association with reflux, swallowing difficulties or organic changes in structures near the larynx except the vocal cords.

Table 2: Distribution of patients by age, sex and type of diagnosis.

Sex	*Frequency*	*Functional*	*Organic*	*Percent Sex*	*Mean Age*	*SD Age*	*Min Age*	*Max Age*
both	145	35	110	100	56.52	16.31	10	83
male	82	14	68	56.55	60.24	13.55	19	82
female	63	21	42	43.45	51.68	18.32	10	83

No statistical distinction has been made between normal and dysphonic subjects. As normal subjects were not numerous, statistical tests and calculations performed on normal data would have been invalid. At the same time, normal subjects were not excluded from statistics since we did not expect that their inclusion would corrupt the overall statistics in dysphonic data. This approach was justified since sometimes even dysphonic subjects give

normal values and normal subjects can be perceived as having a voice disorder. There is a good deal of overlap between normal and pathological values.

2.2 Speech tasks

Apart from the standard text, the protocol of the European Laryngological Society (ELS) prescribes the use of /a/ sustained at a comfortable pitch and intensity level followed by sustained phonation at a slightly louder than comfortable intensity level (Dejonckere et al., 2001). In deviation from the ELS guidelines, each subject in the present study was asked to sustain the vowels /a/ and /e/ at a comfortable pitch and intensity level for up to 5 seconds and read a standard text passage (Appendix A) of approximately 1-minute duration. The reasons for this deviation from the ELS protocol are stated in the following.

There is a strong tradition in voice research to calculate voice parameters from vowels. Sustained phonations are characterized by a relatively stable vocal tract shape and the air flow has a presumably laminar pattern. The most stable phonations are believed to be elicited from vowels held at a comfortable pitch and intensity level. Combinations of pitch and intensity held at other than comfortable level may introduce additional phonatory control issues in dysphonic subjects.

Vowels in normal subjects differ not only in vocal tract shape, but also in intrinsic fundamental frequency and intensity. High vowels have a higher intrinsic fundamental frequency than low vowels (Ewan, 1975). Open vowels are louder than closed ones. /a/ gives loudest phonations.

Under controlled recording condition, the main difference between /a/ and /e/ is the amount of harmonic energy around 1000 Hz. In /e/ the F2 is far away from F1, the amplitude level of the spectrum between the formants is low (Pickett, 1991) and energy in the lower end of the spectrum is weak (Fig. 40, Fig. 41).

Fig. 40: Wave shape and power spectrum for /a/ sustained at 145 Hz and 64 dB.

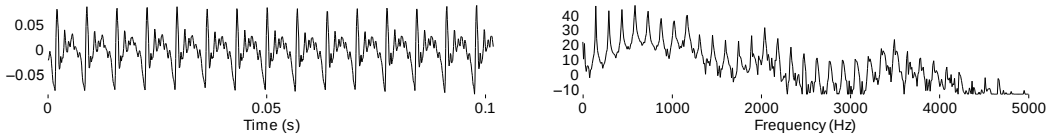
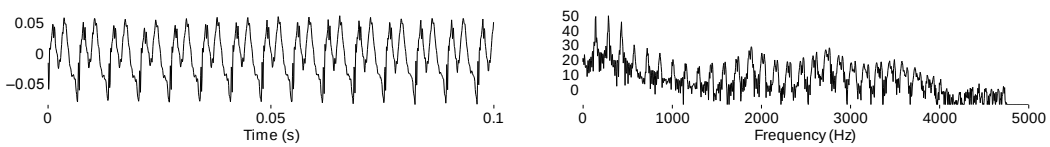


Fig. 41: Wave shape and power spectrum for /e/ sustained at 145 Hz and 65 dB.



Whereas fundamental frequency and intensity can be easily controlled in normal subjects, it is often not feasible to obtain vowels with needed characteristics from voice patients. This is why researchers feel compelled to collect data on several vowels sustained at a comfortable frequency and intensity level.

High vowels /i/ and /e/ are frequently used in acoustic and electroglottographic analysis because they allow comparison with other methods like videostroboscopy and kymography. These vowels garantee a better view into the glottis during visual imaging techniques. This is due to vertical standing of the epiglottis that opens the view to the vocal folds and a higher position of the larynx caused by a higher tongue position.

The vocal function in sustained phonations differs from connected speech in many aspects. It is obvious that if applied to connected speech, voice measures that are normally calculated from vowels would be less reliable, e.g., jitter and shimmer that strongly rely on the fundamental period detection. Aspects of speech like articulatory changes, high level of instability and lack of consistency in voice function, the ability to change the laryngeal tone and intensity according to the requirements of speech, voice breaks, frequent onset and offset of voicing do not manifest in an isolated vowel. However, connected speech is more natural and measures obtained from connected speech are more likely to be generalizable to a patient's everyday speech.

It needs to be explicitly stated that measures made in a clinical setting are not always representative of the dysphonic voice because voice quality is not stable in pathology. It can vary not only within a sentence but also during the course of the day. For this reason, it should be assumed that a voice is as normal or as pathological as it is presented to the examiner.

2.3 Data acquisition

2.3.1 Recording equipment and technique

Recordings were made in a quiet room with an ambient noise below 50 dB(A), verified by repeated measures with a Bruel & Kjaer sound level meter (type 2235) using an omnidirectional microphone (type 4176) at peak hold mode. Recording data was acquired with the Laryngograph software and hardware package through a 16-bit resolution analog to digital converter with a sampling frequency of 50 kHz. The recording equipment consisted of a pair of one-channel electrodes and a high-quality electret microphone with a frequency response from 20 Hz to 20 kHz. Several trial phonations to ensure that an approapriate recording level in both channels had been used preceeded the final recording. The microphone was placed ca. 7 cm from the subject's mouth at an angle of 45–90 degrees from

the line perpendicular to the plane of the lips (Dejonckere et al., 2001; Friedrich & Dejonckere, 2005). Patients were instructed to minimize movement during testing. Recording were made in a sitting position. The text was read from a table positioned at the level of the eyes. As the extent of F0 excursions is known to be influenced by linguistic and paralinguistic factors, to minimize the prosodic effect, the subjects were asked to read a test passage with a detached attitude. All patients had analizable EGG signals.

2.3.2 Pitch detection algorithm settings

In the present study, measures taken by the Laryngograph software program were not used for automatic voice quality classification. With the Laryngograph software, electroglottographic analysis failed in 13.5 % of subjects in /a/ vowels, 8 % in /e/ vowels and 2 % in connected speech samples. Thus, data analysis was made using the Praat editing software program (Boersma & Weenink, 2005). By adjusting the algorithm parameters to pathologic population, it was achieved that the chosen voice parameters applied to all study subjects regardless of the quality of the signal. For the analysis described here, we have modified the default parameters as follows: the silence and voicing thresholds were lowered as recommended, to ensure that the voice detector function is positive during a vowel segment where there is little energy in the microphone signal. A frame was labelled as silent if its energy was less than 10 % of the overall signal energy, a frame was designated as voiced if its energy was more than 25 % of the overall signal energy. This seems legitimate for most pathologic voices characterized by reduced ability to produce loud phonations und unstable voice quality. Another level of the voicing and silence thresholds may have lead to different classification results.

F0-tracking was performed with the autocorrelation method. The autocorrelation method computes the correlation between the signal and a delayed copy of itself between the minimum and maximum expected fundamental period. In periodic signals, the autocorrelation function peaks at a delay that corresponds to the fundamental period. Setting the pitch floor down to 30 Hz was a necessary measure to safeguard against missing vocal breaks due to subharmonics in male voices and low F0 values typical for creaky voice quality. The upper limit of the F0 tracker was 600 Hz.

Obtaining correct values of F0 in noisy and irregular voices is difficult as the poor quality of the signal is the main source of both pitch perception and pitch detection errors. Three phenomena typical for pathological voice can be visually observed in spectrograms of connected speech and tracked by pitch detecting algorithms: subharmonics, birfurcations and modulations caused by asymmetry in the mechanical and geometrical properties of vocal folds, exceptionally low frequencies in creaky voice and voices in which F0 is mainly absent.

In non-F0 dysphonic voices, the pitch detection algorithm appears to give values that

are neither F1 nor F2, since both normally lie above 600 Hz. The non-existing fundamental frequency was replaced in these particular instances by strong noise formant frequencies. Since listeners are able to perceive pitch in normal non-F0-voices, we did not exclude these cases from statistics[7]. However, it is questionable whether the term frequency range could be appplied to voices without F0, even if non-F0-speakers may successfully convey differences in intended pitch.

Extracted vowel samples rarely contained unvoiced analysis frames, so it had little prospect of success to include the degree of voice breaks (DVB) as a predictor variable since the DVB value was different from zero in a few subjects only.

2.3.3 Vowel segmentation

To cover a representative portion of the signal, parameter extraction in this study was applied to 2 possibly non-overlapping 1-second mid-vowel fragments per vowel, voice onset and offset excluded. By choosing only one what appears to be ”representative” /a/ and /e/ vowel per subject, we did not take into account intrasubject variability in phonatory tasks within one session.[8] This was done to imitate conditions under which the automatic classification method might be used in practice.

Parameter extraction was not limited to traditional Type 1 signals or Type 1 segments of the vowel. In many patients, the voice quality was unstable, either improving or deteriorating during phonation, which means that the same measure applied to different segments of the same vowel could be very different. The greatest care was taken to choose the most stable vowel segments for voice parameter extractions. Still, since vowel length was limited to 3–5 seconds, it could not be avoided that one ”good” and one ”bad” segment or two ”bad” samples of a specified length were chosen. The segments were resampled at 10 kHz, thus consisting of 10,000 sampling points. The same cursor positions were applied to Sp and Lx signals.

The two voice parameter estimates made on the same vowel were not averaged, since this would hardly give a more representative estimate than separate measurements. Instead, in classification experiments and statistical calculations, the 2 measurements on the same vowel

[7] Thomas (1969) and Higashikawa & Minifie (1999) argued that the formant frequencies F1 and especially F2 were related to the perceived pitch in such voices.

[8] Higgins et al. (1994) considered nine productions per vowel sufficient to obtain a representative sample. To counteract the intrasubject variability effect, measurements taken from different phonations are normally averaged to obtain one representative measure during the session. Scherer et al. (1995) suggested that in unstable voices up to 15 vowel samples are needed to obtain representative voice measures. In their study each vowel token consisted of 100 consecutive glottal cycles. This requirement seems difficult to enforce in clinical practice as vocal breaks, nonexistent F0 and inability to sustain phonation for more than 1–3 seconds are common in dysphonic voices.

were pooled together resulting in 300 samples per vowel. This approach allowed us to artificially double the number of data sets as if we had 300 subjects.

Problematic with artificial data enlargement is a partial overlap in data. The enlarged data yields less unique information since repeated measurements are correlated but may contribute to noise reduction in the data. According to Harri & Wade Brorsen (2009), many articles in social sciences now use overlapping data to increase prediction accuracy. The reasons for using overlapping data are nonnormality, errors in the explanatory variables and missing data. Here, we deliberately included overlapping data to ensure that our results are not particular to just one vowel segment, but to any randomly chosen vowel segment.

Variables calculated from speech material may be more stable and more representative of a particular voice as they are based on more material than those calculated from one-second vowel segments. The majority of instrumental measurements within one vowel were found to be in more than 95 % of cases inconsistent and varied by more than 10 % (Appendix B)[9]. We found that the magnitude of the correlation coefficient between the first and the second measurement depends on the measure. For example, two measurements of acoustic shimmer correlate with an *r* of 0.63 and 0.51 in /a/ and /e/ vowels, respectively. The difference between the first and second measurement averaged 43 % and 48 % in /a/ and /e/ vowels, respectively.

Despite being different, the first and the second measurements seemed to be more or less identically distributed. It was also observed that pooling the two measurements together did not significantly change correlations with perceptual voice quality in comparison to separate calculations. Similarly, the results of statistical tests performed on a pooled data might have a slightly different *t* or *F* value compared to separate data. However, the magnitude of the effect and the significance level were almost not affected. The advantage of pooling vowel data together was simulating conditions closest to practical use, given that there is a certain degree of arbitrariness in choosing the one-second vowel segment from a longer phonation.

2.3.4 Signal type screening

Signal type screening was applied to whole vowels and texts. As the high-frequency harmonics in pathological voices are often obscured by noise, narrow-band spectrography (15 Hz) in the range from 0 Hz to 2000 Hz was considered sufficient for screening vowels for

[9] Voice measurements on sustained vowels are known to vary with fundamental frequency and intensity of phonation. However, when normal subjects were required to phonate at a comfortable level, intrasubject variability in F0 and intensity within one session and across different sessions were reported to be minimal in all speaker groups (Brown et al., 1996; Brown et al., 1998).

subharmonics. Any additional frequency between the harmonics was treated as a subharmonic. The number of subharmonics was not assessed.

The same investigator made the repeated screening. The rescreening was performed 2 months after initial analysis. A correlation of 0.92 was obtained between the repeated screening sessions.

Screening speech samples for subharmonics is problematic for the following reason: in connected speech, pitch variation is normal and has to be distinguished from unintended pathologic voice breaks. It was assumed that intended lowering of pitch is gradual, causing slanting and rounding of the harmonics. Voices with at least 10 episodes of a sudden pitch halving across vowels in word medial positions were judged as having subharmonics.

2.3.5 Voice parameters

The choice of parameters that underwent a closer examination as candidates for prediction of voice quality included classical measures from vowels like jitter local (Ji) and shimmer local (Shi), irregularity component (IC, Michaelis), frequency modulation factor (FMF), open quotient (OQ, Gendrot) and parameters known to measure the amount of noise or chaos in the signal: glottal-to-noise excitation ratio (GNE, Michaelis), harmonics-to-noise ratio (HNR, Boersma), long-term average spectrum (LTAS, Boersma), aperiodicity index (AI), subharmonics-to-harmonics ratio (SHR, Sun) and largest Lyapunov exponent (LLE, Wolf). All voice parameters were calculated from the same one-second mid-vowel fragments. A short description and evaluation of these parameters are given in the section devoted to results.

When identical parameters were calculated from both signals, the same algorithm was applied. No restrictions were introduced to limit the corresponding EGG and acoustic measures, e.g., jitter or shimmer, to the same range of values.

The parameters extracted from the speech records refer mostly to statistical properties of the distribution of the fundamental frequency during speech: mean, median, and modal F0, standard deviation, 80 % F0 range, percentage of F0 values below specified thresholds. In addition, jitter, shimmer and irregularity index were calculated, as well as average intensity in dB. Estimates of average intensity of speech apply to voiced segments only. All measures were calculated automatically.

Data analyses were performed using SPSS, Matlab and Stata. When means are reported, standard deviations are given in parentheses. The majority of variables were not normally distributed. However, the normality assumption can be violated in data with sample size above 40, as long as data come from similarly shaped distributions (Gardner, 1975; Moore, 1995).

2.3.6 Speech data labeling

Recordings were labelled using both auditory and visual cues. Labelling was performed by the author.

2.3.6.1 Voiced vs. unvoiced frames

The identification of voiced segments in connected speech was done with Praat followed by manual editing of acoustic signals to remove artefacts caused by vowel-consonant transitions and intrusive sounds like groaning and loud breathing. These were excessive F0 downward and upward shifts associated with vocal tract constriction during vowel-consonant transitions. Edited voice segments in acoustic signals contained mainly vowels and nasals. EGG signals were corrected for intrusive sounds only and contained all sounds that were identified as voiced. Signals were resampled at 16 kHz.

F0 extraction was based on instantaneous frequency values. The F0 trajectories were extracted at a rate of 100 values per second by means of an autocorrelation method. Voice quality was found to vary across phrases and sentences. In many dysphonic voices, the speaking intensity was getting weaker at phrase and sentence boundaries, so that despite lowered thresholds for silence and voicing detection no fundamental frequency could be found. Still, most of the subjects had sufficient number of pitch periods from which F0 histograms could be calculated. If F0 values were found in whispered vowels, they were considered typical for dysphonic population and were not eliminated[10].

2.3.6.2 Pausing time

The duration of pauses was measured on oscillographic traces. The major acoustic manifestations of pauses are silent intervals in the signal. For practical reasons, the minimum silent duration to be counted and labeled as a pause was set at 200 ms, long enough not to include silent intervals in stops as pause time. Registration of pauses of less than 200 ms duration was not undertaken as it required enormous manual effort[11]. As shown in Campione & Veronis (2002) pauses of less than 200 ms account for only 4.8 % of all pauses in German and are

[10] In normal whispering, the arytenoids are slightly separated, thereby opening the posterior part of the glottis, while the membranous part of the glottis remains closed.

[11] The auditory threshold for the perception of pauses is reported to lie between 200–250 ms (Goldman-Eisler, 1968; Grosjean & Collins, 1979; Zellner, 1994; Mattys et al., 2005) as this appears to be the pause length that can be most easily detected in perceptual experiments. Recent experiments show that the auditory threshold could be set somewhat lower. In a study of connected German speech, Potapova (2002) found that silent intervals of less than 146 ms duration were not detected as pauses by the raters. Other studies use thresholds of 150 ms (Tsao & Weismer, 1997); 130 ms (Dankovičová, 1997); 100 ms (de Pijper & Sandermann, 1994). Butcher (1981) referred to pauses of less than150 ms duration as unheard pauses.

therefore negligible. The mean and median duration of pauses in their read German data was estimated at 490 ms and 485 ms, respectively.

Other pause manifestations like final lengthening had no influence on pause measurements. Pause duration was measured from the end of periodic voicing, release of the voiceless stop closure or ending of the fricative noise to the beginning of voicing or fricative noise. Closure durations in post-pausal voiceless or devoiced stops were counted as part of the total pausing time, however. The glottal pulse preceding a vowel was taken to be evidence of the vowel onset. In German, initial vowels are frequently preceded with a glottal stop.

2.3.7 Criteria for mode detection

Modes in histograms correspond to strong peaks and represent the most probable values. The detection of statistical modes in distributions of instantaneous frequency data is a complex mathematical problem and requires a fine tuning of the bin width and starting points. Inappropriate choice can lead to either merged peaks or generate false peaks. To overcome these problems with conventional histograms we used univariate kernal density estimation with Epanechnikov kernel and default optimal bandwidth.

The F0 distribution was considered to be bimodal when two local maxima were detected in the histogram, when at least 10 % of Fx values fell into the low register mode and there was a gap between the modes, meaning that some frequencies are seldom or never used.

2.3.8 Syllable count

The number of syllables was taken to be the maximum number of syllables (182) contained in the text in canonical reading pronunciation. If the subjects added or omitted words, the actual number of syllables produced in each reading was counted. Atypical pauses and phrase repetitions were cut out.

Automatic pause detection and syllable count can be reliably implemented using the intensity contour. Local maxima in the intensity contour correspond to syllable nuclei. The performance of a syllable count algorithm implemented in Praat was described in de Jong & Wempe (2007).

2.3.9 Timing measures

Measurements were made for each reading of total speaking time, number of pauses and total pause (including silence and breathing) duration. These were used to determine the speaking rate (number of syllables per second of total time including pauses) and articulation rate (number of syllables per second excluding pauses), mean pause length in ms, speech/pause ratio (articulation time divided by pause time), count of pauses per 100 syllables and average

number of syllables produced between two pauses. These measures can also be referred to as prosodic measures since they are calculated from speech units larger than a phoneme.

As dysfluent reading was assumed to undermine the validity of timing measures and their power to predict perceived breathiness, it was ensured that all study subjects demonstrated adequate reading skills.

2.3.10 Aerodynamic measures

Three aerodynamic measures including maximum phonation time (MPT), vital capacity (VC) and phonation quotient (PQ) were taken. These parameters are part of a routine voice assessment in phoniatric facilities. All measures were carried out in a standing position. MPT and VC were measured twice. The longest values were used for further analysis. VC measurements were taken with a hand-held spirometer. According to Rau & Beckett (1984), hand-held spirometers permit reliable VC measurements.

2.4 Analysis of experimental data

Results are reported without attention to diagnosis. When data was checked for vowel, sex, age and signal effect, the reported *t*-test estimates refer either to paired samples or independent samples *t*-tests. Group means and standard deviations were calculated and are listed for each examined measure separately in Appendix C. Data were averaged to yield group means according to the four levels of the grouping variable. In most cases, data was not broken down by sex even when sex effect was proved to be present since it was not intended to perform classification experiments for male and female subjects separately. The strength of association with perceptual voice categories expressed as Spearman's correlation coefficients and the number of significant contrasts across the four levels of each perceptual category according to Mann-Whitney U-test are given for each examined parameter, as well.

A series of Mann-Whitney U-tests were performed for each of the measured parameters and perceptual category to explore if the chosen parameters can discriminate between contiguous voice-quality grades. For all statistical testing, a significance level of $p = 0.05$ was used. The number of statistically significant contrasts between the voice-quality grades ranges from 0 to 3. When a variable has 3 signficant contrasts, the group medians differ significantly across all four levels of the grouping variable. The test results are tabulated in Appendix D. Variables with poor discrimination ability were not eliminated from further analysis.

2.5 Disordered voice quality rating

Perceptual evaluation was performed by means of the RBH scale, which is the most widely used scaling method in Germany (Nawka et al., 1994). The RBH scale for describing vocal

quality consists of three perceptual categories: roughness (R), breathiness (B) and overall degree of hoarseness (H). Each category has to be judged on a four-point scale with the outcomes ranging from 0 (clear or normal) to 3 (severe). Data was checked for reliability and agreement. We used discrete ratings in statistical calculations.

Perceptual categories are supposed to be rated in reference to what normal voice quality should sound like. As the reference levels vary from listener to listener or with experience, to date, research on voice quality prediction uses data averaged across several raters.

8 experienced speech-language pathologists, all native speakers of German, agreed to participate in the study. The judges were recruited from different centers and were personally contacted via e-mail or phone. Neither have they been working on the same team nor did they receive the same training for voice evaluation. All judges used the same rating protocol which was available online and included voice samples in mp3 format. The access to the website with recordings was granted on request to all specialists and students interested in voice quality rating. 10 % of voice samples (15 recordings) were repeated for assessment of intralistener reliability. These recordings were chosen by chance and were different for each rater. The online protocol was divided into three sections of 55 voice samples, each of which could be accomplished in approximately 30–40 minutes. The judges did not receive any instructions about the RBH rating system and the scoring procedure. All of them were familiar with the RBH scale. Judges were allowed to play recordings more than once and to score at individual pace. Participation was not remunerated.

There are many different measures of interrater agreement and reliability, none of which is universally accepted. All the existing indexes suffer from several limitations with respect to estimating consensus and consistency. For a review of 57 papers published between 1951–1990 on reliability and agreement, see Kreiman et al. (1993).

Two measures of agreement between all the judges averaged together were used in this study: Kendall's *W* and kappa statistic. Kendall's concordance coefficient or Kendall's *W* is based on the mean value of the Spearman's rank correlation coefficients between all pairs of the rankings over which it is calculated. Kendall's W is not an ideal measure of agreement since it makes no assumptions regarding the distribution of data. It might fail to measure the exact agreement as scores can highly correlate with each other without being exactly the same. This is the case when judges agree on ordering but not on magnitude of voice pathology.

Another statistical measure of assessing consensus and consistency across the raters is kappa statistic. Cohen's kappa is used to measure the agreement between each two raters or between scale values on the first and second presentation of the 15 retest voice samples of the

same rater. It accounts for chance agreement[12]. The agreement between any numbers of raters is given by Fleiss' kappa.

One of the limitations of kappa statistics is that if the ratings are different but proportional, reliability can be relatively high even though there is little agreement between the judges and vice versa. Kappa estimates are less useful when ratings are mostly assigned to one or two rating categories while other categories remain underrepresented (Jones et al., 1983). Similarly, kappa is influenced by the relative balance of agreements and disagreements. Whitehurst (1984) argued that kappa yields higher agreement measures when raters disagree on the distribution of ratings in the data and lower measures when they agree.

Although much of the voice analysis in this paper was done on sustained vowels, voice quality judgments were made on speech sample. While the type of voice segment (vowel vs. sentence) presented for evaluation has reportedly no effect on reliability and agreement between the raters (de Krom, 1994a; Revis et al. 1999), the magnitude of the ratings seems to depend on the choice of the voice segment. Askenfelt & Hammarberg (1981) argued that vowels are not representative of voice function status. Only in the case of severe pathology was vocal function found to be consistent between vowels and sentences. Despite strong correlation between vowels and sentences in perceptual quality ratings (r ranging from 0.72 to 0.89), Hanson & Emanuel (1979) found that dysphonic patients occasionally produce vowels that are less severely disturbed than sentences. There was no significant difference found in ratings of complete vowels and connected speech in Revis et al. (1999). On the other hand, Wolfe et al. (1995a) reported that vowels from normal subjects were assigned greater severity ratings than sentences. They measured a relatively high correlation coefficient of 0.78 on dysphonic severity between vowels and sentences. Vowel ratings accounted for 61 % of the variance in the prediction of sentence severity. Only 7.5 % of all vowel-sentence pairs differed by 2 or 3 scale points. The results of Wolfe et al. (1995a) are compromised by the fact that naïve raters recruited for the perceptual evaluation are likely to judge the abnormality of voice as such rather than distinguish between distinct perceptual categories like breathiness or roughness. Up to date, the question remains unresolved as to what extent acoustic measures made on vowels can be expected to predict the perceptual quality of speech.

2.6 Classification scheme

Classification of voice quality by objective voice parameters presents a multiclass classification problem. A combination of predictors is to be found that can reliably

[12] A kappa of 1.0 means perfect agreement. A kappa of zero means that agreement is due to chance alone. A kappa of 0.7 is an acceptable reliability value and would mean that observed agreement is 70 % accounted for by the true agreement between the raters or between the first and the second rating.

discriminate between four different degrees of the examined perceptual voice dimensions. To lower the experimental complexity, redundant variables should be eliminated while maintaining the maximum success rate.

In order to estimate the discriminant function in QDA and weights in ANN, a grouping variable representing actual class membership is required. On the basis of available data, it is possible to predict class membership in cases when it is not known.

The proper assessment of classification success involves an assessment of how the method works on a validation data set. For assessing the classifier performance, we used the leave-one-out cross-validation procedure. Data was split into 2 partitions. 299 samples were used for determining the discriminant function which was subsequently tested on the remaining 1 sample. The total number of misclassifications in the test set was computed and averaged over all partitions.

2.6.1 Quadratic discriminant analysis (QDA)

The choice of relevant variables was based on their discriminating power that was estimated using statistical methods. Variables were first brought into a ranking order in accord with their individual contribution to the classification accuracy measured by the amount of explained variance.

The floating search method (Pudil et al., 1994) was used to determine the reduced variable set that led to the highest success rate after validation by leave-one-out method. The search is started with 2 variables on the top of the ranking list. The inclusion or exclusion of the next variable was to perform depending on the success of this step measured on the classification success rate. After each inclusion, backward elimination was performed in accordance with the following scheme. The result after the first backward step had to be compared with the result that was achieved through inclusion of an additional variable. Should the success rate without additional variable exceed that of with additional variable, the added variable in question has to be eliminated from the variable set. The number of backward steps to try amounts to $n–1$, where n is the number of variables currently used for selection. The result after the second backward step was compared with the result after the first backward step etc. Backward elimination has to be performed as long as there is an improvement in the success rate. Otherwise, one should procede with inclusion of a new variable. If elimination of any variable would lead to a worse result as compared to the result after the last inclusion, no elimination takes place. To prevent infinite loops, it is nessessary to control for variable sets that have already been tested. For this purpose, every single combination has to be stored with the corresponding success rates. On reaching this state, one

should go on to the next step. The search is complete when the variable set with the highest success rate is found.

2.6.2 Artificial neural networks (ANN)

Identical ANNs were employed for each perceptual category. A feedforward network was trained with the "Approximation and Classification of Medical Data" (ACMD) software program. For a detailed review of the used ANN, the reader may refer to Linder & Pöppl (2001). The network typology included one hidden layer with 100 hidden neurons and an output layer comprising 4 neurons according to the 4 classes in question. In the learning phase, the relationship between the input and the known output was determined by adjusting the weights between the computational nodes according to the rules specified in the learning algorithm until the error value is minimized. Learning was stopped after a predefined number of 1000 epochs. During the learning process, 5-fold cross-validation method was used to continuously validate the performance of the ANN in order to recognize the optimal weight set. This method implies that 80 % of the database is used as training material; the remaining 20 % serves for validation purposes. The *winner-takes-all* rule applies in determining the final output.

Redundant variables were eliminated by using the Neural Net Clamping Technique (Wang et al., 1998). This technique is an ANN-based feature selection search that starts with *n* available variables. Upon completing the learning phase, the ANN is tested *n* times, whereby one of the *n* variables in succession is set to its mean value. The feature with the smallest contribution to classification (showing the best classification rate despite being dropped) is omitted. The search continues with *n–1* variables until the best feature set is found.

Chapter 3: Results

In this chapter and the chapter to follow, we present, interpret and discuss the empirical results of our research. The conclusions made here are primarily based on own observations and calculations, but facts and figures found in literature are also considered.

3.1 Subjective analysis of experimental data

Clinicians are advised to use subjective and objective methods of voice quality evaluation. Whereas quantitative analysis reveals many interesting features that characterize a certain voice, subjective analysis of data is important for overall impression of tone quality.

3.1.1 Perceptual voice evaluation

This section summarizes the results of auditory-perceptual evaluation. Patient examination always begins with auditory-perceptual evaluation of voice quality. The perceptual dimensions examined in this study are probably the most popular terms to describe voice quality. In Askenfelt & Hammarberg (1986), breathy voice quality is defined as audible escape of air through the glottis due to incomplete glottal closure. Rough vocal quality is assumed to be perceived as a low-pitched noise caused by irregular vocal fold vibrations. Hoarseness is a complex sensation consisting of a combined sensation of breathiness and roughness, overall severity is even more complex including an evaluation of hoarseness, breathiness, roughness and vocal fry (Eskenazi et al., 1990).

3.1.1.1 Perceptual ratings

Table 3 shows the results of individual ratings across 8 raters. Perceptual voice ratings ranged from normal to severely disordered. The distribution of individual ratings implies that some raters favoured the one or the other end of the scale, whereas others preferred the middle. These differences went lost in averaging over 8 raters.

Table 3: Distribution of individual ratings by the 8 experienced raters for the R, B and H over 150 voices.

Rater	*Roughness*				*Breathiness*				*Hoarseness*			
	0	*1*	*2*	*3*	*0*	*1*	*2*	*3*	*0*	*1*	*2*	*3*
1	38	60	41	11	51	42	37	20	37	43	47	23
2	19	27	46	58	9	46	51	44	12	31	53	54
3	26	46	43	35	33	51	35	31	6	41	47	56
4	31	58	46	15	32	56	43	19	5	31	62	52
5	46	50	39	15	51	62	26	11	17	55	55	23
6	30	61	41	18	27	64	50	9	3	60	63	24
7	50	42	37	21	65	61	18	6	32	49	44	25
8	19	53	44	34	36	59	34	21	16	45	48	41

As indicated in Table 4, the distribution of averaged ratings centered around grades 1 and 2 in all three perceptual categories. Fewer voices were assigned ratings at the lowest and highest end of the scale. Discrete values were obtained through averaging across the ratings and rounding the mean to the nearest integer.

Table 4: Distribution of discrete R, B and H ratings across sex.

Grade	*Roughness women*	*Roughness men*	*Breathiness women*	*Breathiness men*	*Hoarseness women*	*Hoarseness men*
0	11	5	10	8	4	2
1	28	39	42	34	28	21
2	20	29	14	28	24	38
3	7	11	0	14	10	23
∑	66	84	66	84	66	84

The results of the one-way ANOVAs indicated that the difference between sexes was not significant in judging roughness ($F = 2.09$, $p > 0.15$), but highly significant in judging breathiness ($F = 18.83$, $p < 0.01$) and hoarseness ($F = 9.69$, $p < 0.01$). Against expectations, male voices were rated as more breathy. It might be also the case that raters allow for more breathiness in female voices, whereas lower levels of breathiness in male voices are perceived as more abnormal. The mean scores for B had a mean of 1.48 (0.81) in males voices against 0.98 (0.54) in female vioces. The mean scores for H were higher in male voices with a mean of 1.9 (0.73) against 1.52 (0.77) in female voices.

It is noteworthy that a high rating on the R scale tends to go together with at least some degree of perceived breathiness. The same is valid for high ratings on the B scale. When voices were perceived as moderately or severely breathy, they were also judged as at least slightly rough (Table 5). Apparently, perceptual categories R and B are not completely independent.

Table 5: Pairwise distribution of R and B ratings.

	B0	*B1*	*B2*	*B3*
R0	4	10	1	1
R1	13	34	14	6
R2	1	28	17	3
R3	0	4	10	4

In our data, the correlation between R and B ratings was with an r_s of 0.35 rather low, suggesting that R and B are separate perceptual dimensions of voice quality. Both R and B ratings correlated strongly with H ratings ($r_s = 0.75$, $r_s = 0.68$, respectively), which is not surprising as H is understood as a superordinate category for the other two perceptual dimensions.

3.1.1.2 Interrater agreement and reliability

Reliability shows the consistency of judgments over repeated tests by the same rater. The extent to which two or more raters agree when rating the same set of voices is captured in interrater agreement.

Table 6 summarizes the distribution of ratings on the first and second presentation of voice samples across 8 raters. 122 records were presented and judged twice. With 5 ratings missing, 117 ratings were available for R and 122 ratings for B and H dimensions, respectively. On the second presentation, the same roughness rating was assigned in 54 %, the same breathiness rating in 64 % and the same hoarseness rating in 69 % of the records used in the test-retest task. It appears that voice samples were rated as more rough and breathy but less hoarse when heard for the second time. On the average, the first and second ratings differed by 0.37 in roughness and hoarseness dimensions and by 0.45 in breathiness.

Table 6: Frequency table for the first and second rating of R, B and H assessed by all raters.

	Second rating											
First rating	*R*				*B*				*H*			
	0	*1*	*2*	*3*	*0*	*1*	*2*	*3*	*0*	*1*	*2*	*3*
0	12	13	2	0	26	4	3	2	8	2	4	0
1	7	14	8	1	4	26	15	0	7	24	4	0
2	1	7	26	5	0	5	11	3	0	6	31	6
3	1	3	6	11	2	0	6	15	0	2	7	21

One way to assess the interrater agreement and intrarater reliability is to correlate the ratings. Correlations between the first and the second ratings are shown on the diagonal in Table 7, Table 8 and Table 9. Intrarater reliability measured by r_s between the first and second rating ranged between 0.32 and 0.92, with a mean of 0.66 for R, 0.69 for B and 0.73 for H. Similar correlations were found in Dejonckere et al. (1993).

The levels of test-retest reliability were higher than 0.5 in all but one rater in each perceptual category. No rater had a Spearman's correlation coefficient below 0.5 in more than one perceptual dimension. The Wilcoxon signed rank test confirmed that the difference between the first and the second rating was significant ($p = 0.01$) or marginally significant ($p = 0.053$) in 1 out of 8 raters in each perceptual category.

Table 7 – Table 9 show the Bonferroni-adjusted Spearman's correlation coefficients between each pair of raters across R, B and H dimensions. The correlations between the pairs of raters were mostly moderate and good, ranging from 0.45 to 0.87 with a mean of 0.63 (0.11) for R, from 0.36 to 0.74 with a mean of 0.58 (0.1) in B and from 0.48 to 0.81 with a mean of 0.68 (0.09) in H. The best Kendall's W was found for R ($W = 0.82$, $p < 0.01$), followed by moderate Kendall's W estimates for H ($W = 0.65$, $p < 0.01$) and B ($W = 0.56$, $p < 0.01$).

Table 7: Bonferroni-adjusted Spearman's correlation coefficients between pairs of raters for roughness.

Raters	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	0.45*ns							
2	0.45	0.57						
3	0.53	0.58	0.59					
4	0.65	0.55	0.60	0.78				
5	0.57	0.52	0.61	0.8	0.54			
6	0.60	0.48	0.53	0.74	0.76	0.63		
7	0.53	0.56	0.6	0.77	0.87	0.79	0.84	
8	0.60	0.64	0.58	0.65	0.68	0.7	0.72	0.85

*ns not significant *s significant $p < 0.05$

Table 8: Bonferroni-adjusted Spearman's correlation coefficients between pairs of raters for breathiness.

Raters	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	0.92							
2	0.53	0.79						
3	0.48	0.48	0.59					
4	0.72	0.36	0.39	0.57				
5	0.72	0.69	0.57	0.59	0.32*ns			
6	0.63	0.55	0.47	0.66	0.57	0.84		
7	0.58	0.43	0.52	0.56	0.62	0.62	0.78	
8	0.63	0.54	0.50	0.60	0.74	0.71	0.62	0.7

Table 9: Bonferroni-adjusted Spearman's correlation coefficients between pairs of raters for hoarseness.

Raters	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	0.63							
2	0.52	0.82						
3	0.56	0.68	0.87					
4	0.48	0.81	0.8	0.79				
5	0.57	0.67	0.69	0.7	0.75			
6	0.58	0.71	0.67	0.69	0.79	0.77		
7	0.51	0.66	0.72	0.73	0.76	0.74	0.49*ns	
8	0.54	0.71	0.73	0.77	0.81	0.72	0.73	0.74

Table 10–Table 12 show the Cohen's kappa for each two raters, the calculations are based on 150 voice samples. All values given in the tables were highly significant ($p < 0.01$) except cases marked with an asterisk. As data was ordinal, we used a weighted kappa with simple weights: it is better to disagree by one point than by two points or more.

Values on the diagonal measure the test-retest reliability of individual raters. Here again, we found that judges were least reliable in test-retest judgments of roughness. From 8 Cohen's kappas obtained for each speaker 4 values were below 0.5, indicating poor reliability for R. The proportion of judges with poor Cohen's kappa values for B and H was lower. Our findings suggest that hoarseness seems to be the most reliable perceptual category in the RBH scale. Average reliability was moderate with Cohen's kappas of 0.5 for R, 0.6 for B and 0.63

for H, respectively, but higher than intralistener reliability reported in de Bodt et al. (1997) and Revis et al. (1999).

Table 10: Chance-corrected proportional agreement between each two raters (Cohen's kappa) for roughness.

Raters	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	0.27*ns							
2	0.24	0.48						
3	0.36	0.44	0.32*s					
4	0.59	0.30	0.41	0.50				
5	0.52	0.27	0.43	0.72	0.34*s			
6	0.48	0.29	0.40	0.63	0.64	0.74		
7	0.45	0.33	0.45	0.66	0.81	0.70	0.75	
8	0.38	0.49	0.46	0.44	0.43	0.50	0.48	0.59

*ns not significant *s significant $p < 0.05$

Table 11: Chance-corrected proportional agreement between each two raters (Cohen's kappa) for breathiness.

Raters	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	0.72							
2	0.31	0.58						
3	0.33	0.31	0.40*s					
4	0.61	0.23	0.29	0.58				
5	0.60	0.28	0.37	0.45	0.41			
6	0.47	0.32	0.31	0.52	0.42	0.83		
7	0.40	0.15	0.28	0.37	0.53	0.40	0.67	
8	0.50	0.34	0.37	0.43	0.57	0.56	0.41	0.62

Table 12: Chance-corrected proportional agreement between each two raters (Cohen's kappa) for hoarseness.

Raters	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	0.36*s							
2	0.29	0.75						
3	0.29	0.56	0.74					
4	0.26	0.68	0.67	0.79				
5	0.44	0.43	0.43	0.43	0.62			
6	0.36	0.47	0.47	0.45	0.70	0.71		
7	0.39	0.42	0.43	0.38	0.61	0.56	0.45	
8	0.39	0.52	0.55	0.56	0.66	0.55	0.49	0.65*s

Across pairs of raters, the agreement varies considerably in the present study, ranging from 0.24 to 0.81 with a mean of 0.48 (0.14) for R, from 0.15 to 0.61 with a mean of 0.4 (0.12) for B and from 0.26 to 0.7 with a mean of 0.48 (0.12) for H.

We also considered the assigned ratings averaged together. The kappa values in Table 13 show the agreement between the 8 raters listed for each category and each subcategory of R, B and H. The overall agreement reached a kappa of 0.34, 0.25 and 0.33 for R, B and H, respectively. The calculations were based on the frequency tables, where the scores are shown

for each perceptual category rated by all judges. As expected, inadequate reproducibility of ratings implied by Cohen's kappa < 0.7 was paired off with low agreement.

Table 13: Fleiss' kappa table for the four subcategories of R, B and H.

	Roughness				*Breathiness*				*Hoarseness*			
Rating	R0	R1	R2	R3	B0	B1	B2	B3	H0	H1	H2	H3
kappa	0.42	0.28	0.25	0.46	0.30	0.16	0.21	0.43	0.25	0.29	0.25	0.52
kappa	0.34				0.25				0.33			

Our study did not substantiate earlier claims about hoarseness as the category with the best interrater agreement. Instead, we found that the overall agreement for each perceptual category was low (0.2 < kappa < 0.4). Similar results were obtained in Martens et al. (2007) and de Bodt et al. (1997).

One finding was paradoxical: judges were least reliable in test-retest task involving the perceptual category R, but reached a relatively high agreement on R in comparison with other perceptual categories. Our results indicate that judges seem to disagree on B more often than on R and H despite relatively good intralistener reliability. The average degree of disagreement for R and H was almost the same. On comparing kappas for subcategories, the agreement between the raters was found to be somewhat higher for extremes (clear and very disturbed voices) than for intermediate voices. This finding is in accord with results by Kreiman & Gerratt (1998), Rabinov et al. (1995) and Martens et al. (2007).

3.1.2 Visual examination of vowel spectra

3.1.2.1 Signal typing

As shown in Table 14, a large proportion of acoustic signals was not nearly periodic. Of 300 examined vowels, 100 phonations (33 %) were identified as Type 1 signals.

Table 14: Results of signal typing in acoustic signals.

	Microphone /a/	*Microphone /e/*	*Total*
Type 1	49	51	100
Type 2	26	24	50
Type 3	28	36	64
Type 4	43	36	79
Type 5	4	3	7
Total	150	150	300

The results of the Wilcoxon signed rank sum test revealed no significant difference between signal types in paired phonations ($z = 0.644$, $p = 0.52$). In 92 subjects (61 %), both sustained phonations /a/ and /e/ belonged to the same signal type. This finding does not necessarily imply that in clinical practice one sustained phonation would be enough to assess vocal function status.

As expected, signal type correlated with perceptual voice dimensions. The correlations between signal type and voice quality are strongest for hoarseness ratings with an r_s of 0.6 for /a/ and 0.57 for /e/. Correlations with breathiness ratings are comparable to those with roughness ratings with similar correlation coefficients of about 0.53 for /a/ and 0.5 for /e/.

This is a relatively good result considering that vowels are typed independently from each other. For comparison, the correlation coefficient between the spectrographic type based on three different vowels and the perceived degree of hoarseness in classification by Yanagihara (1967) was estimated at 0.65.

3.1.2.2 Subharmonics

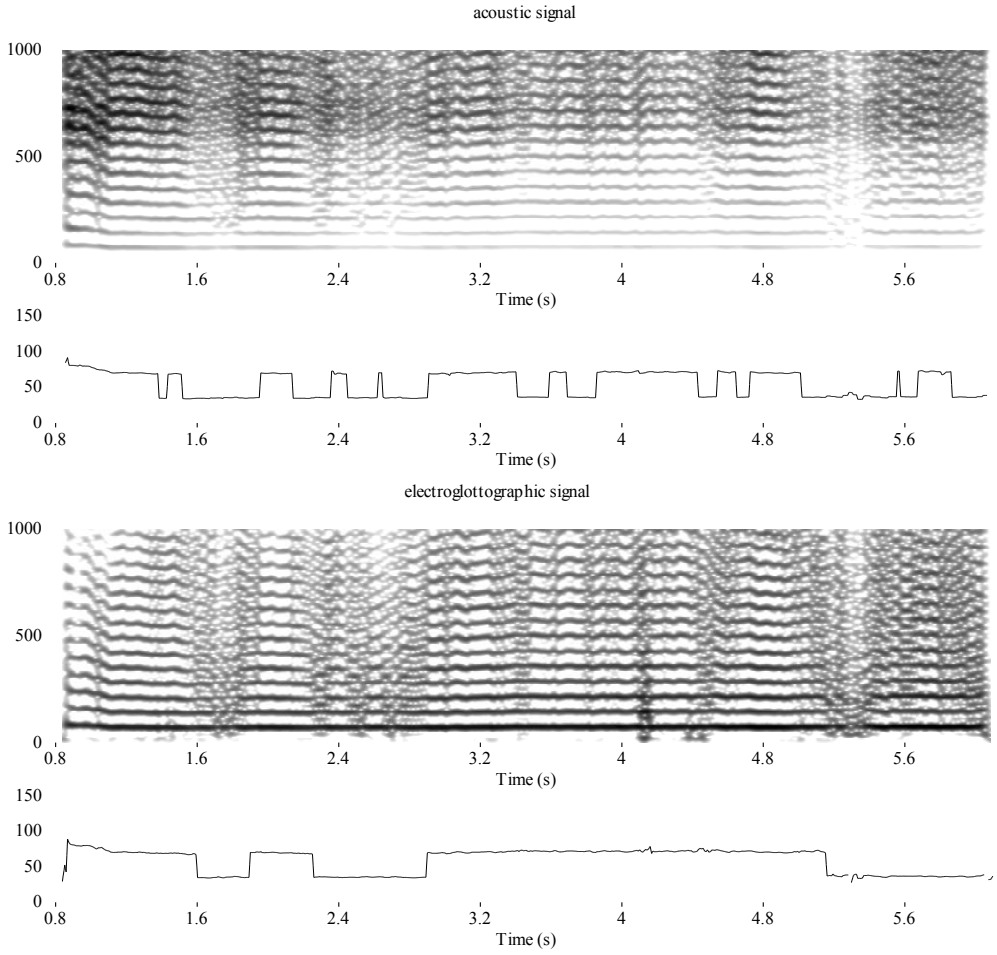
In this section, the incidence of subharmonics in sustained vowels will be examined. The second area of interest is how subharmonics relate to perceived roughness in patient data. Several sources report that subharmonics contribute to perceived roughness (Omori et al., 1997; Dejonckere et al., 1993). In experiments with synthetic signals, the power and frequency of subharmonics seemed to be pivotal for the degree of perceived roughness (Omori et al., 1997). How these findings are applicable to natural voices remains unclear since similar experiments are not possible with natural vowels. Whereas synthesized vowels can be freely manipulated for the purpose of the experiment, it is difficult to elicit natural vowels from subjects that would differ in just one acoustic dimension. In general, findings for the disordered voice with regards to relation between voice quality and acoustic measures are not as definitive as for synthetic stimuli. Observations made in our data are described in the following.

Voices with subharmonic frequences pose a number of problems for objective voice analysis. Subharmonic frequencies normally result in octave errors and are visible as octave jumps in the pitch contour. Although subharmonics above the fundamental are stronger than lower subharmonics, they seem to contribute less to the perceived pitch and detected F0 values than lower subharmonics do. As a rule, whenever a subharmonic frequency below the fundamental exceeded a certain spectral level, the F0 was reduced (Fig. 42). The result was a sudden drop in the F0 contour that usually extended over a number of contiguous periods. In several instances, however, subharmonics were detected visually but were not strong enough to affect the pitch contour. Note that although the spectrograms from acoustic and EGG signals (Fig. 42) look almost identical, the corresponding F0 contours do not overlap.

Substantial difficulties arose in detecting subharmonic structure in voices with a fundamental frequency below 100 Hz as the accuracy of the measurement relies on the resolution of the spectrum. For the same reason, subharmonics and biphonation have been

mostly reported for voices with higher pitches. Needless to say that signals in which subharmonic energy is indistinguishable from noise between the harmonics are treated as not having subharmonics.

Fig. 42: Spectrograms and the corresponding F0 contours of sustained /a/ by patient 6, m, 60, diagnosed with vocal fold paralysis after apopleptic stroke.



In our data, subharmonic energy was usually found in the middle part of the vowel in the range from 200 Hz to 1000 Hz. Lower and higher subharmonics were either too weak to be detected or masked by noise. Few subjects displayed subharmonics consistently across a vowel. Most signals showed intermittent subharmonic structure in the narrow-band spectrogram. The proportion of time that the subharmonic component was present across a vowel varied between 6 % and 100 %. This agrees with Cavalli & Hirson (1999). In 33 patients (23 %), subharmonics were found in both sustained phonations /a/ and /e/ and were mostly detected in the medial portion of the vowel.

In acoustic signals, subharmonics were identified in 60 patients (41 %) in the middle of /a/. In 57 cases, the presence of subharmonics was detected by both the pitch extraction algorithm as judged by the F0 contour and by visual examination of the harmonic structure. In

three remaining cases, subharmonics were detected visually but lacked evidence in the pitch contour. Accordingly, 47 patients (33 %) had subharmonics in the middle of /e/. In 45 /e/ vowels, the evidence was drawn from both breaks in the pitch contour and visual examination of the spectrum. The McNemar's chi-square test statistic suggests that vowels /a/ and /e/ do not differ significantly in the proportion of signals with subharmonics (*McNemar's chi2(1)* = 3.27, $p = 0.07$).

If subharmonics are generated at the level of the vocal folds, evidence of subharmonic structure was naturally expected to be found in both acoustic and electroglottographic signals. The analysis of pathologic vowels has shown that this is not always the case.

In 41 /a/ vowels and 40 /e/ vowels, evidence in favour of the presence of subharmonic frequencies was found in both sound pressure and electroglottographic curve. The analysis of electroglottographic signals showed that segments characterized by octave drops in the pitch-contour exhibit multiple, mostly twofold or threefold, cycles within a single period in a closeup view. This pattern implies periodicity that is achieved every second or third period. However, in voices with pitch breaks in both electroglottographic and microphone signal, the corresponding pitch contours were not identical. In many a pathologic vowel, subharmonic segments in acoustic and electroglottographic signals did not overlap exactly and were often found to be inconsistent with each other with respect to both their duration and location in the signals (Fig. 43).

Fig. 43: Spectrograms of /a/ and the corresponding F0 contours by patient 2, m, 79, diagnosed with vocal fold paresis. Note the difference in the duration and location of segments with subharmonic frequencies in the acoustic (left panel) and EGG (right panel) signals.

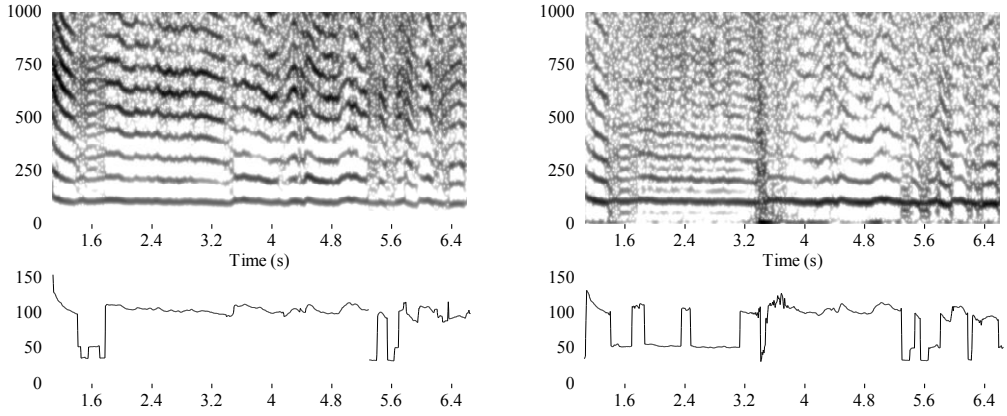
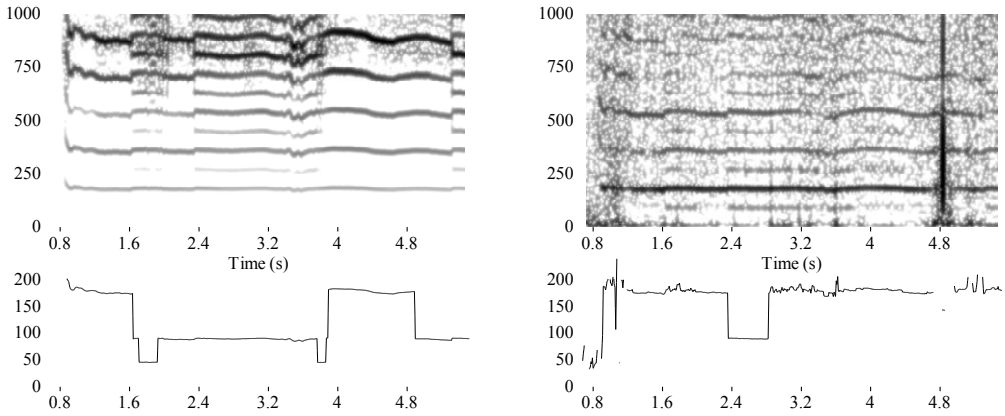


Fig. 44 is an example of inconsistency between acoustic and EGG signals with respect to the number of detected subharmonic frequencies. Whereas the acoustic signal in Fig. 44 exhibits two short segments with 3 subharmonic frequencies between the harmonics,

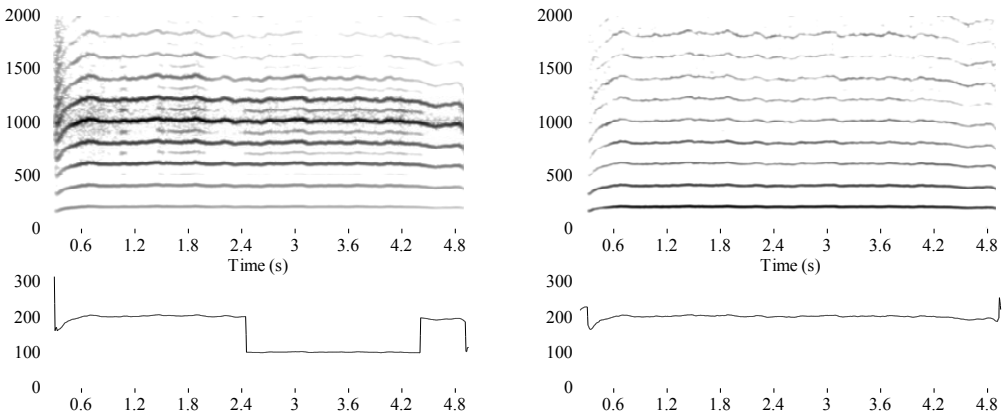
the EGG signal shows only one subharmonic frequency. Note that the presence of 3 subharmonic frequencies is reflected in the corresponding F0 contour.

Fig. 44: Spectrograms of /a/ and the corresponding F0 contours by patient 39, w, 32, diagnosed with functional psychogenic dysphonia. The left panel refers to acoustic signal, the right panel to EGG signal.



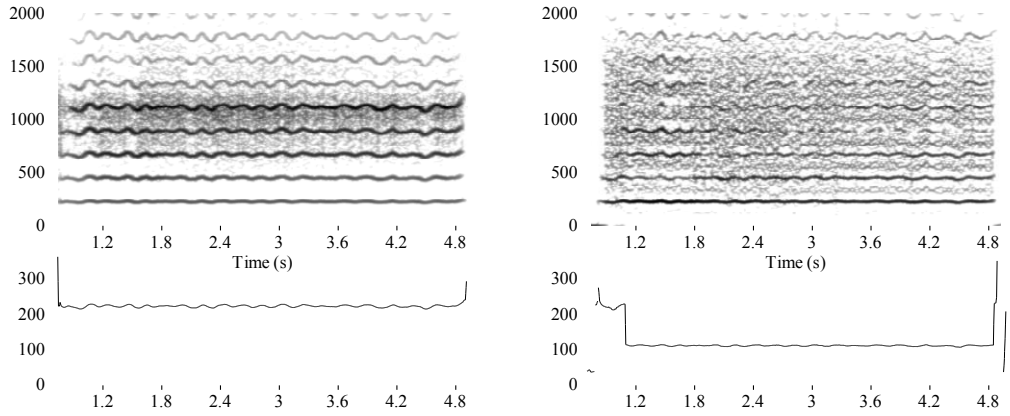
In 15 /a/ and 7 /e/ vowels with subharmonics in the microphone signal, the electroglottographic curve showed no signs of subharmonic component (Fig. 45). In a further 4 cases of /a/, the electroglottographic wave was too irregular to allow conclusions about the harmonic structure. Subharmonics that were not detected in the electroglottographic curve but evident in the sound pressure wave curve alone were assumed not to be originated by the voice source at the level of the vocal folds. In this case, subharmonics should arise from the involvement of anatomical structures other than vocal folds.

Fig. 45: Spectrograms and the corresponding F0 contours by patient 9, m, 57, diagnosed with unilateral vocal fold paralysis. Note the presence and the absence of subharmonic frequencies in the acoustic (left panel) and EGG (right panel) signals, respectively.



In 26 /a/ and 43 /e/ vowels, the subharmonic element was present in the electroglottographic curve, but was lost in the microphone signal (Fig. 46). The subharmonic element seems to be too weak or masked by noise in the microphone signal.

Fig. 46: Spectrograms and the corresponding F0 contours by patient 119, w, 63, diagnosed with unilateral vocal fold paralysis. Note the absence and the presence of subharmonic frequencies in the acoustic (left panel) and EGG (right panel) signals, respectively. Subharmonics are reflected in the EGG F0 contour.



Similar observations were reported by Behrman et al. (1998); however, they concluded that signal typing done on acoustic and EGG signals was relatively consistent between the two signals. In their data, the signal typing between acoustic and EGG signals (traditional method) was different in 6 from 202 subjects.

That EGG and microphone signals may have systematic F0 discrepancies in signals with subharmonics follows from Table 15. According to the paired samples *t*-test, the mean F0 values derived from acoustic and electroglottographic signals are significantly different in /e/ vowels of Type 2 signals (male voices: $t = -3.05$, $p < 0.01$; female voices: $t = -2.22$, $p < 0.05$) and in /a/ vowels of Type 4 signals (male voices: $t = 2.8$, $p < 0.01$; female voices: $t = 4.04$, $p < 0.01$). An example of a typical Type 2 signal is shown in Fig. 33. Accordingly, Fig. 35 depicts a typical Type 4 signal.

It is noteworthy that female voices with subharmonics had exceptionally low mean F0 values. We hypothesized that in male voices harmonics are more powerful than subharmonics, so that low F0 values are less frequently detected by pitch detection algorithms. When comparing the means in male and female voices in acoustic signals, we found that 3 out of 4 independent samples *t*-tests failed to reach significance. Thus, in acoustic signals, men and women did not differ in the mean F0 in /a/ vowels of Type 2 and Type 4 as well as /e/ vowels of Type 4.

Table 15: The means and standard deviations of mean F0 values derived from acoustic signals in patients with subharmonics and corresponding data from electroglottographic signals. Data was stratified by vowel, signal type and sex.

Signal Type	*Variable*	*Male Voices*	*Female Voices*	*Male Voices*	*Female Voices*
		Mean (SD) /e/	Mean (SD) /e/	Mean (SD) /a/	Mean (SD) /a/
		n=16	n=20	n=18	n=24
Type 2	F0 (Hz)	149.4 (71.5)	188.3 (49.1)	144.6 (69.2)	146.9 (64.0)
	F0 EGG (Hz)	125.2 (58.5)	156.7 (52.8)	131.5 (28.9)	164.3 (60.1)
		n=40	n=18	n=50	n=28
	F0 (Hz)	111.3 (46.5)	127.9 (46.8)	112.6 (47.8)	115.3 (44.8)
Type 4	F0 EGG (Hz)	113.6 (32.1)	137.1 (43.3)	131.3 (42.1)	153.4 (43.3)

Contrary to expectations, setting the pitch floor of the pitch detection algorithm to 30 Hz does not result in lowering the mean F0 values to values marginal to or outside the normal frequency range in male voices. The possible explanation that in the spectrum of male voices harmonic frequences are more powerful than subharmonic frequences is not sound enough to explain the observed pitch breaks in the F0 contour in male voices. Only by assuming that the frequency range of male voice in pathology is close to female frequency range, the resulting mean F0 values in Type 4 signals can be interpreted as lowering due to subharmonic effect.

A total of 75 patients (50 %) had subharmonics in the microphone signal in at least one vowel task. Another consideration of importance is how the presence of subharmonics in vowels relates to subharmonics in speech. A total of 52 % of study subjects habitually used exceptionally low pitch across vowels in reading. The low F0 values come about through subharmonics and creaky voice. This excessive lowering of the pitch in vowels seems to cause the perception of roughness in speech. This question will be resumed in section 3.2.2.

Given that roughness judgements are made on speech samples, it would be of interest to look at how the degree of perceived roughness in speech relates to signal typing in sustained vowels. In the following table, two methods of subjective voice classification are compared. The results are similar for both vowels.

Table 16: Distribution of signal types across 4 roughness grades for the vowels /a/ (left) and /e/ (right). Values in brackets represent voices with subharmonics.

Signal Type	*Roughness*				
	R0	*R1*	*R2*	*R3*	*∑*
Type 1	12	29	8	0	49
Type 2	3 (3)	9 (8)	9 (8)	5 (2)	26
Type 3	2	12	12	2	28
Type 4	0	9 (8)	19 (17)	15 (14)	43
Type 5	0	2	0	2	4
∑	17	61	48	24	150

Signal Type	*Roughness*				
	R0	*R1*	*R2*	*R3*	*∑*
Type 1	12	29	9	1	51
Type 2	3 (3)	10 (8)	10 (6)	1 (1)	24
Type 3	1	16	14	5	36
Type 4	1 (1)	4 (3)	14 (11)	17 (14)	36
Type 5	0	2	1	0	3
∑	17	61	48	24	150

Table 16 shows the number of voices classed into each signal type and roughness grade. For the description of signal types, see section 1.3.5. Spearman's rank correlation

coefficients between roughness ratings and signal type amounted to 0.52 in /a/ and 0.53 in /e/ vowels. Even though sustained vowels do not adequately represent connected speech, it is reasonable to assume that less disturbed signals come from less rough voices.

Vowels from all 5 normal subjects were classed Type 1 signals. Only two of them were rated R0. Voices of three healthy subjects were judged mildly rough (R1). 31 % of patients with voice problems produced Type 1 signals that would give reliable vocal parameters. The majority of voices that were perceived as not rough were presented with Type 1 signals in sustained phonation task. The quality of vowel signal in mildly and moderately rough voices is difficult to predict, with a slight predominance of Type 1 signals in voices judged R1. Severely rough voices tended to produce Type 4 signals. No generalization can be drawn for Type 5 signals on the basis of the available data. In general, the proportion of voices with subharmonics increased with increasing roughness grade. Approximately 50 % of voices that were perceived as moderately and severely rough had subharmonic component in sustained phonations, and the speech of 67 % of subjects with subharmonics in sustained phonations was judged as moderately and severely rough.

The results of the Fisher's exact test with data from both vowels suggest a statistically significant relationship between the presence of subharmonics in the spectrum of sustained vowels and roughness ratings (*Fischer's exact* < 0.01). The strength of association was rather low (r_s = 0.36 for /a/ and 0.27 for /e/). A similar relationship between subharmonics and breathiness ratings could not be confirmed.

This finding has the following implication for the connected speech: the mere presence of subharmonics in vowels might not necessarily explain rough voice quality in speech. As vowel data shows, rough voices tended to produce phonations with subharmonics and noise (Type 4 signals). Thus, spectrographic noise might contribute not only to perceived breathiness but also to perceived roughness.

3.2 Objective analysis of experimental data

In this section, the relation between instrumental measures and perceived voice quality is investigated for both isolated vowel segments and connected speech. We grouped the voice parameters that were measured on vowels into two main categories: perturbation and noise measures. Some sections begin with what is known about these measures from previous research. The summary sections present findings common to all variables of the same type.

3.2.1 Analysis of mid-vowel segments

3.2.1.1 Fundamental frequency and intensity

The analysis of mid-vowel segments begins with the average comfortable pitch and intensity levels in dysphonic patients. Both fundamental frequency and intensity of phonation might be affected in laryngeal pathology in a fairly predictable way. Thus, patients with mass lesions like laryngitis, nodules, polyps, Reinke's edema, cysts and granulomas are expected to have a lower F0 in sustained vowels compared to normals. Patients with glottic incompetence might stand out through lowered intensity of phonation.

As many voice measures are based on reliable F0 extraction, it would be rewarding to study if there are any systematic differences in the mean fundamental frequency derived from acoustic and electroglottographic signals in the first place. For this end, we used our set of dysphonic voices to compare the corresponding acoustic and electroglottographic F0 values.

One would expect that in normal subjects the mean F0 values obtained from acoustic and electroglottographic signals are very similar. In normal subjects, Orlikoff (1995) measured an almost perfect correlation between individual acoustic and EGG-derived mean F0 values. In dysphonic patients, we expected a bigger discrepancy between acoustic and electroglottographic signals: EGG signals may not accurately measure F0 if the vocal folds do not properly contact each other; acoustic signals are prone to measurement errors when dominated by noise.

Our patient data confirm that the mean F0 values derived from microphone and electroglottographic signals do not differ significantly in /a/ vowels (t = –1.09, p = 0.27). However, there was a minor difference in /e/ vowels which was also statistically significant (t = 2.44, $p < 0.01$). The Pearson's r between acoustic and electroglottographic mean F0 values was estimated at 0.76 and 0.68 in /a/ and /e/ vowels, respectively. In /e/ vowels, there was a notable difference in the magnitude of the correlation coefficients across sexes; women having a lower correlation coefficient than men. We attributed this finding to a more open glottal configuration during the production of /e/ and weaker EGG signals in women.

Acoustic signals show significant difference in the mean F0 for vowel type. The average difference in F0 between /a/ and /e/ vowels in our data was 7 and 14 Hz in male and female voices, respectively. /e/ samples had significantly higher F0 estimates than /a/ vowels (t = –4.07, $p < 0.01$). A higher F0 in high vowels that was observed in the present study is explained by an increased laryngeal tension as a consequence of tongue root raising and anterior movement of the hyoid bone during high vowel production (Ewan, 1975; Honda, 1983). In contrast to our data, no statistical differences among the vowels were noted in Orlikoff (1995) when a single-factor repeated-measures analysis of variance was applied to the mean F0 data.

Further, we confirmed a significant effect for sex in F0 data (t = –6.4, p < 0.019). In acoustic signals, male subjects used a mean F0 of 127 (66) Hz maintained at 71.5 (5.5) dB in /a/ vowels. /e/ vowels were sustained at a mean F0 of 134 (61) Hz. Female subjects used a mean F0 of 175 (61) Hz maintained at 71.2 (4.7) dB in /a/ vowels. As expected, female subjects had a higher mean F0 of 189 (49) Hz in /e/ vowels.

The mean F0 extracted from Lx signals in male voices equals 119 (43) Hz and 128 (42) Hz in /a/ and /e/ vowels, respectively. Women measured a mean of 169 (54) Hz and 177 (54) Hz in /a/ and /e/ vowels, respectively. The difference in the mean F0 was significant across vowel types (t = –2.68, p < 0.01). Here again in accord with expectations, /e/ vowels give higher F0 estimates than /a/ vowels.

Table 17 summarizes the mean F0 data in both signals broken down by spectrographic vowel type. In Type 2, Type 4 and Type 5 signals, the mean F0 was found to be significantly reduced in female voices as compared to signals without F0 irregularities in the lower part of the spectrum. This effect proved to be pronounced in /a/ vowels and was present in both acoustic and EGG signals. The mean F0 values lay at the lowest level or outside the normal frequency range for female voices. Extremely high F0 values in 3 male voices were detected in the absence of F0 in Type 5 signals.

The present data point to the need to examine whether F0 affects the perception of voice quality. Wolfe & Ratusnik (1988) found that voices with high F0 were perceived as less rough than voices with a low F0. We found little evidence for this effect in our data.

Table 17: The means and standard deviations of mean F0 values derived from acoustic and electroglottographic signals stratified by vowel type, signal type and sex. The data is based on 300 observations per vowel.

Signal Type	*Variable*	*Male Voices*	*Female Voices*	*Male Voices*	*Female Voices*
		Mean (SD) /e/	*Mean (SD) /e/*	*Mean (SD) /a/*	*Mean (SD) /a/*
Type1		n=44	n=58	n=44	n=54
	F0 (Hz)	133.0 (36.2)	213.9 (29.8)	113.8 (28.9)	215.7 (37.1)
	F0 EGG (Hz)	126.9 (40.0)	184.7 (54.5)	110.2 (30.6)	203.2 (46.4)
Type2		n=22	n=26	n=20	n=32
	F0 (Hz)	151.8 (67.2)	197.7 (53.5)	147.3 (65.9)	154.4 (57.5)
	F0 EGG (Hz)	132.0 (40.0)	200.8 (48.8)	141.2 (55.6)	146.3 (52.8)
Type3		n=50	n=22	n=42	n=14
	F0 (Hz)	129.8 (43.8)	173.9 (41.3)	115.8 (21.3)	188.9 (23.2)
	F0 EGG (Hz)	126.6 (40.2)	152.3 (57.6)	109.3 (23.3)	176.4 (39.3)
Type4		n=46	n=26	n=56	n=30
	F0 (Hz)	114.5 (46.3)	140.4 (48.4)	111.9 (48.4)	125.4 (57.9)
	F0 EGG (Hz)	124.8 (34.4)	160.3 (44.2)	126.2 (38.8)	135.1 (37.4)
Type5		n=6		n=6	n=2
	F0 (Hz)	266.1 (179.9)		381.3 (86.8)	53.8 (7.6)
	F0 EGG (Hz)	194.4 (112.5)		176.8 (129.2)	70.7 (3.6)

Table 18 shows the mean F0 values in different roughness grades. Relation between roughness and pitch in acoustic signals seems to hold only for female voices (r_s = –0.35); no such effect being observed for male voices (r_s = 0.02).

Table 18: The means and standard deviations of the mean F0 values from acoustic and electroglottographic signals stratified by vowel type, roughness grade and sex.

Roughness Grade	*Variable*	*Male Voices*	*Female Voices*	*Male Voices*	*Female Voices*
		Mean (SD) */e/*	*Mean (SD)* */e/*	*Mean (SD)* */a/*	*Mean (SD)* */a/*
0		n=12	n=22	n=12	n=22
	F0 (Hz)	125.8 (30.3)	214.4 (38.8)	115.9 (28.1)	213.9 (33.9)
	F0 EGG (Hz)	131.7 (32.7)	185.2 (60.9)	124.4 (37.0)	187.7 (43.2)
1		n=64	n=58	n=64	n=58
	F0 (Hz)	140.5 (72.6)	197.7 (40.2)	132.3 (81.9)	182.9 (62.5)
	F0 EGG (Hz)	131.1 (57.8)	175.2 (56.1)	125.7 (50.8)	176.2 (60.3)
2		n=62	n=34	n=62	n=34
	F0 (Hz)	128.3 (39.1)	181.6 (53.8)	118.9 (34.8)	151.3 (55.2)
	F0 EGG (Hz)	130.7 (33.0)	187.8 (48.9)	117.5 (33.0)	166.3 (47.9)
3		n=30	n=18	n=30	n=18
	F0 (Hz)	136.1 (81.6)	147.9 (54.8)	138.1 (84.8)	146.8 (64.3)
	F0 EGG (Hz)	121.7 (34.3)	157.3 (47.0)	118.8 (52.5)	128.7 (43.9)

Across roughness grades, the mean F0 in both vowels systematically decreased with increasing roughness grade in female voices. In EGG signals, this effect was observed in /a/ vowels produced by female subjects only (r_s = –0.29). In other data, there was no tendency for the mean F0 to decrease with increasing roughness grade.

All in all, we found little evidence that roughness ratings were strongly associated with lower F0 estimates in vowels. This finding is in agreement with Murry & Doherty (1980). They found that lower F0 values in dysphonic subjects were valid in sentence data only: normals and subjects diagnosed with cancer did not differ in F0 for sustained vowels.

Although mean F0 may have a possible effect upon the perception of voice quality, especially in female voices, it was not included as a predictor variable. Separate classification schemes for male and female voices would have led to a smaller sample size.

In the view of interaction between pitch and intensity[13], data on the comfortable frequency and intensity level are dealt with in one section. The ability to control intensity is

[13] Isshiki (1964) defined two mechanisms of intensity regulation, respiratory and laryngeal, depending on pitch. Voice intensity is proportional to subglottic pressure, which is a function of glottal resistance and airflow rate. At low and medium pitches, intensity can be increased by increasing the glottal resistance through activity of laryngeal muscles. The larynx affects intensity by increasing adduction and adjusting the length and stiffness of the vocal folds. At high pitches, the glottal resistance cannot be further increased without affecting the pitch, so exhaling muscles are more important in varying the intensity. The existence of these two mechanisms has never been questioned by other researchers; however, the contribution of laryngeal control in varying the intensity level has been seen as rather marginal (Finnegan et al., 2000). This finding has an important implication for breathy voices. When the mechanism of increasing glottal resistance through laryngeal control is not available, the speakers have to increase the flow rate to maintain a medium intensity level. The flow rate is controlled by the

an essential charateristic of a healthy voice. Habitually weak intensity levels or uncontrollable bursts of intensity are typical signs of pathology; the extreme case being intensity levels in voices with the absent vocal fold activity.

In our data, the mean intensity of /a/ was 71.4 (0.3) dB; the mean intensity of /e/ — 71.6 (0.3) dB[14]. The mean intensity estimates in vowels ranged between 57 dB and 86 dB. According to a paired *t*-test statistic, the difference in intensity between vowels /a/ and /e/ was not significant ($t = –0.91$, $p = 0.36$). This is a surprising finding in the light of the fact that the area of the mouth opening is related to the radiation impedance: when all other variables are held constant, the larger the mouth opening, the greater the intensity (Isshiki, 1964). In accordance with this rule, open vowels should be pronounced with a higher intensity than closed vowels.

It is noteworthy that dysphonic subjects were not able to maintain a stable intensity level for 3–5 seconds. In both vowels, the second sample had a significantly lower intensity than the first one ($t = 11.08$; $t = 10.8$, $p < 0.01$). Intensity data in vowels remained largely uncorrelated with perceived voice quality and can only with difficulty be used to descriminate between contiguous voice-quality grades (Appendix C).

3.2.1.2 Perturbation measures

Perturbations are cycle-to-cycle variations in the waveform. Perturbations in the signal are thought to be caused by multiple factors like tissue abnormalities, vibration asymmetry and neuromuscular fluctuations. 5 perturbation measures were computed for all patients. Pitch and amplitude perturbation measures indicate instabilities in the fundamental frequency and amplitude. OQ reflects variations in the open phase relative to the pitch period. Traditionally, perturbation parameters are believed to measure the ability of the subject to exert phonatory control over a sustained vowel. A healthy voice will have little difficulty in maintaining stable fundamental frequency and amplitude across a sustained vowel segment. A pathological voice can be expected to show more variation. Large perturbations are associated with severe laryngeal dysfunction. A low degree of perturbation is, however, physiological.

3.2.1.2.1 Jitter and shimmer in Lx and Sp signals

The most popular perturbation measures are jitter and shimmer. Jitter measures fluctuations in cycle length (frequency perturbations), shimmer (amplitude perturbations) – in cycle amplitude. Jitter has always been considered one of the most useful acoustic measures of pathology

---

activity of the exhaling muscles. From the point of view of air economy, the second mechanism of increasing intensity is less efficient since the glottis remains largely open and causes the sensation of frication noise.

[14] Intensity data in Praat refers to the intensity in air expressed in dB relative to the auditory threshold.

for clinical application.

The ”jitter (local)” that was used in this study is defined as the average absolute difference between consecutive periods, divided by the average period. Similarly, the ”shimmer (local)” is defined as the average absolute difference between the amplitudes of the consecutive periods divided by the average amplitude.

Most methods for calculating jitter and shimmer are sensitive to pitch measurement errors and use averaging over several cycles to minimize the variance. Parameters like jitter and shimmer cannot discriminate between subtle instabilities and gross modulations of the signal over several cycles like vibrato or F0 movements in intonation. This is why measurements have to be made on stable portions of sustained vowels. However, some researchers used nonstationary speech to estimate perturbation measures (Askenfelt & Hammarberg, 1981; Vasilakis & Stylianou, 2009). EGG signals can also be used to obtain perturbation measures like jitter and shimmer, though it is commonly used to draw abduction/adduction measures.

We found that jitter and shimmer are correlated. Strong correlations between jitter and shimmer extracted from acoustic and electroglottographic signals are well known from many publications. In our data, compared to observations made by Michaelis (2000), correlations between jitter and shimmer were stronger in EGG signals than in acoustic signals. Correlation coefficients between jitter and shimmer in acoustic and electroglottographic /a/ signals were estimated at 0.47 and 0.73, respectively. The strength of association between jitter and shimmer was found to be somewhat lower in /e/ vowels with a Pearson’s r of 0.41 and 0.61, respectively.

In our data, the vowel effect was significant ($t = 3.19$, $p < 0.01$; $t = 10.14$, $p < 0.01$) in acoustic jitter and shimmer, respectively. It was found that /a/ vowels had higher jitter and shimmer values in acoustic signals. This agrees with Sussman & Sapienza (1994) and Orlikoff (1995). No vowel effect was present in electroglottographic signals ($t = 1.05$, $p = 0.14$).

We could confirm that acoustic jitter is weakly influenced by the mean F0. The strength of linear association between the mean F0 in sustained phonations and jitter in /a/ and /e/ vowels in acoustic data was estimated at –0.27 and –0.20, respectively. Acoustic shimmer was uncorrelated with the mean F0. This finding harmonizes with Horii (1979) who found that in male voices increasing F0 in the range from 98 Hz to 210 Hz is coupled with decreasing jitter and that higher vowels have less jitter.

Further, acoustic jitter was observed to be uncorrelated with the mean intensity of phonation. Acoustic shimmer weakly correlated with intensity data ($r = –0.16$ in /a/ and –0.18 in /e/). These findings do not agree well with Orlikoff & Kahane (1991).

Against expectations, we found that jitter estimates from Sp and Lx signals despite

being measured on the same vowel fragment were very different. The same is valid for shimmer measures. The difference between the means in all the four *t*-tests was very significant ($p < 0.01$). Apparently, EGG signals in pathological voices are much noisier than microphone signals; so that jitter and shimmer values extracted from Lx signals are significantly higher than those in the microphone signals. Acoustic jitter was lower than or equal to EGG jitter in 68 % (203/300) of /a/ and 72 % (215/300) of /e/ tokens. Acoustic shimmer was lower than or equal to EGG shimmer in 75 % (226/300) of /a/ and 82 % (247/300) of /e/ tokens.

Unlike Vieira et al. (2002) who reported a very strong correlation between acoustic and EGG jitter in dysphonic /a/ vowels with jitter values up to 2.7 %, we found that acoustic and EGG jitter were uncorrelated with an r of 0.05 in /a/ and 0.08 in /e/ fragments. Acoustic shimmer was weakly but significantly correlated with EGG shimmer in both examined vowels ($r = 0.24$ in /a/ and $r = 0.30$ in /e/).

Further on, we tested data on jitter and shimmer for sex effect. An independent samples *t*-test revealed that acoustic and electroglottographic jitter values obtained from vowels did not differ significantly among men and women. Sex effect was not confirmed in EGG shimmer, either. In both vowels, acoustic shimmer, however, was significantly different across male and female subjects ($t = 2.19$, $p = 0.03$; $t = 2.85$ $p < 0.01$). Male subjects were found to have more shimmer than female subjects.

The pathology threshold for acoustic jitter as it is implemented in Praat is reported to be 1 %. In the present study, jitter estimates ranged between 0.03 % and 8.67 %. /a/ vowels had a mean jitter of 0.56 (0.87) %. In /e/ vowels we measured a lower mean of 0.38 (0.76) %. Acoustic jitter did not prove to be sufficient for pathology detection since the majority of study subjects had an acoustic jitter below 1 %. When jitter was measured on acoustic signals, about 85 % (255/300) of samples had a jitter below 1 % in /a/ vowels and 92 % (276/300) in /e/ vowels. Our data suggests that the pathology threshold for jitter was set to high.

The pathology threshold for shimmer in Sp signals was given at 3.81 %. We found that 14.5 % (44/300) of /a/ samples and 39 % (118/300) of /e/ samples have an acoustic shimmer below the specified threshold. Shimmer measured a mean of 8.68 (5.07) % within the range between 1.22 % and 24.73 % in /a/ vowels and 6.03 (4.58) % ranging from 1.19 % and 25.74 % in /e/ vowels.

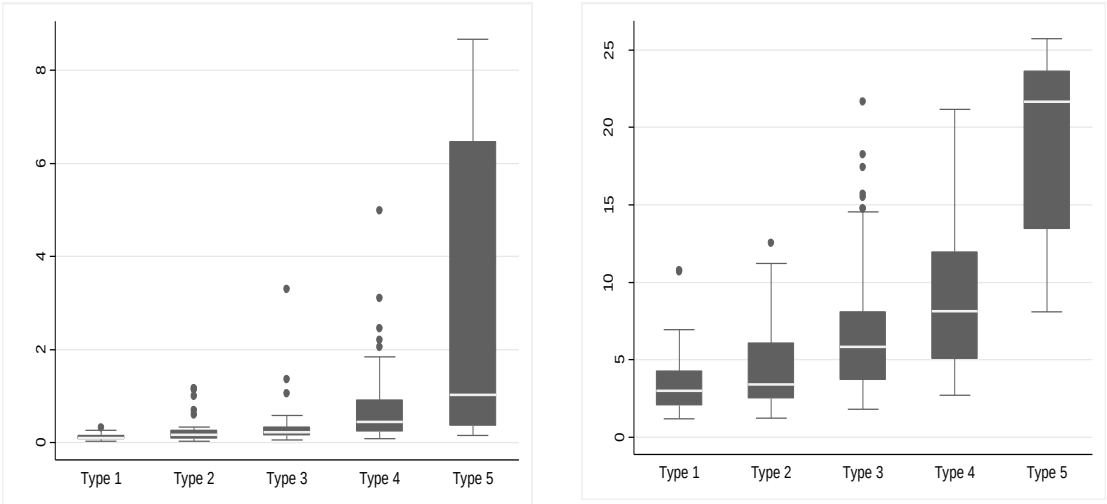
Voices with subharmonics had a higher proportion of higher shimmer values but a lower proportion of higher jitter values compared to average data. Of 60 patients who had subharmonics in /a/, only 28 % of samples (34/120) had abnormal jitter and 91 % (109/120) had abnormal shimmer. The jitter and shimmer analysis in /e/ segments revealed abnormal jitter in 23 samples (24 %) and abnormal shimmer in 81 samples (84 %). This finding would suggest that subharmonics is a phenomenon that arises in the first place due to excessive amplitude modulation.

Pathology thresholds for the EGG jitter and shimmer were not specified. However, 48 % (145/300) of /a/ samples and 50 % (151/300) of /e/ samples had an EGG jitter below 1 %. In approximately 47 % (140/300) of /a/ samples and 50 % (149/300) of /e/ samples both acoustic and EGG jitter lay below the 1 %-threshold. If we restricted the range of possible jitter values to values below the 1 %-threshold, the Pearson's *r* between acoustic and EGG jitter would be no higher than 0.47 in /a/ and 0.42 in /e/.

We found that acoustic jitter was associated with roughness (rather than with breathiness) and with hoarseness. Correlations with hoarseness were almost of the same order as correlations with roughness. On the contrary, acoustic shimmer was stronger correlated with breathiness (rather than with roughness) and with hoarseness. Correlations with hoarseness were even higher than correlations with breathiness. Acoustic jitter and shimmer outperformed EGG jitter and shimmer in discriminative effectiveness between adjacent roughness and hoarseness grades. Acoustic parameters exhibit higher correlations and a greater number of significant contrasts. Acoustic jitter seems to be more discriminative than acoustic shimmer.

Acoustic jitter was correlated with spectrographic vowel type with an r_s of 0.61 and 0.65 in /a/ and /e/ vowels, respectively. In acoustic shimmer data, correlations with spectrographic vowels type measured an r_s of 0.60 and 0.59 in /a/ and /e/ vowels. Fig. 47 shows boxplots with jitter and shimmer data by spectrographic vowel type. All types have a similar skewed distribution. Jitter seems to be highest in most irregular signals (Type 4 and Type 5). In noisy signals (Type 3, Type 4 and Type 5) acoustic shimmer seems to be higher than in signals without noise.

Fig. 47: Boxplots showing acoustic jitter (left) and acoustic shimmer (right) data in /e/ samples grouped by spectrographic vowel type.



3.2.1.2.2 Frequency modulation factor (FMF)

Standard deviation of F0 is a measure of variability of F0 and reflects laryngeal stability across duration of the vowel. Zwirner et al. (1999) were able to document post-therapeutic improvement in voice quality using F0 SD in vowels.

To facilitate comparison between subjects with very different F0 means, standard deviation of F0 had to be expressed as a frequency modulation factor. FMF is defined as the ratio of the standard deviation to the mean F0.

Table 19 summarizes results on the FMF data by spectrographic vowel type. Vowels /a/ and /e/ did not differ significantly in the FMF values. No sex effect was apparent in the FMF data, either. Correlations with spectrographic vowel type amounted to 0.60 and 0.57 in /a/ and /e/ vowels, respectively.

FMF had lowest values in Type 1 and Type 3 signals – signals without irregularities in the lower part of the spectrum. In signals with irregularities in the lower part of the spectrum without noise (Type 2), the FMF estimates were lower than in signals dominated with noise (Type 4 and Type 5).

The presence of subharmonics in the spectrum was expected to have a strong impact on the standard deviation of the fundamental frequency. As expected, voices with subharmonics had higher FMF values (Kruskal-Wallis equality-of-populations rank test, *Chi-squared* = 52.0 in /a/ vowels, *Chi-squared* = 27.7 in /e/ vowels, $p < 0.01$).

Table 19: The means and standard deviations of the FMF values derived from acoustic signals stratified by vowel type, signal type and sex. Data is based on 300 observations per vowel.

Signal Type	*Male Voices*	*Female Voices*	*Male Voices*	*Female Voices*
	Mean (SD) /e/	*Mean (SD)* /e/	*Mean (SD)* /a/	*Mean (SD)* /a/
Type1	n=44 .9 (.3)	n=58 .88 (.43)	n=44 1.7 (2.9)	n=54 1.4 (2.3)
Type2	n=22 8.2 (12.8)	n=26 8.3 (12.5)	n=20 8.1 (9.1)	n=32 9.2 (11.8)
Type3	n=50 5.4 (17.8)	n=22 1.6 (1.3)	n=42 1.9 (2.19)	n=14 1.1 (.4)
Type4	n=46 12.9 (18.5)	n=26 27.5 (22.3)	n=56 20.2 (24.2)	n=30 28.3 (23.6)
Type5	n=6 49.8 (51.1)		n=6 26.6 (19.7)	n=2 26.9 (5.0)

Among voices with high FMF values that had no subharmonics were voices showing the following spectral characteristics: voices with voice arrests in the presence or in the absence of laryngeal tremor, tremoulous voices without voice arrests, voices with a highly

disturbed harmonic structure below 2000 Hz and voices without identifiable harmonic structure.

FMF correlated moderately with perceived roughness and hoarseness. Correlation coefficients were higher in /a/ samples. In both sexes, roughness and hoarseness ratings tend to be higher with increasing FMF.

3.2.1.2.3 Irregularity component (IC)

Irregularity component is an aggregate measure consisting of three measurements: jitter averaged over 11 periods, shimmer averaged over 15 periods and mean period correlation. For the exact formula to calculate IC, the reader is referred to Michaelis (2000) and Fröhlich et al. (2000). IC is defined for all voices regardless of periodicity requirements and therefore can be computed from highly disturbed, even aphonic, voices. The waveform-matching method is used to determine the fundamental frequency. The pathology threshold for IC was set at ca. 4.7, which was assessed over 92 normal subjects.

By applying this threshold to our data, we found that 197 (66 %) samples had IC values above the pathology threshold in /a/ vowels and 111 (37 %) samples in /e/ vowels. A serious vowel selection effect was confirmed with a paired samples *t*-test. Thus, parameter extraction done on /e/ gives less pathological values ($t = 10.82$, $p < 0.01$). Further, we found that male subjects had higher IC values than female subjects in both vowels ($t = 2.85$, $p < 0.01$; $t = 4.01$, $p > 0.01$). IC showed a reasonably high correlation with hoarseness for both male and female voices. Correlations with breathiness and roughness were somewhat lower. Further, it had the highest discrimination number in all three perceptual categories.

IC correlated strongly with spectrographic vowel type. We measured an r_s of 0.71 and 0.75 in /a/ and /e/ vowels, respectively.

3.2.1.2.4 Open quotient (OQ)

OQ was calculated using EGG signals[15]. The mean OQ in our data was estimated at 0.62 (0.16). Some of the subjects had an OQ of 1,0 or near 1,0.

OQ values did not differ significantly across male and female subjects. This finding disagrees with Hanson & Chuang (1999) who reported smaller OQs in nondysphonic male speakers. Obviously, the sex effect is lost in dysphonic speakers. Vowel effect was not significant, either.

Correlations with spectrographic vowel type were found to be low and insignificant. In

[15] The Praat file that we used to calculate the OQ using the DEGG method was provided by C. Gendrot. It is available in: Henrich N, Gendrot C, Michaud A. Tools for electroglottographic analysis: Software, documentation and databases. http://voiceresearch.free.fr/egg/. (last visited: 30th September 2010).

Type 1 signals, we measured a mean OQ of 0.58 (0.14) in /a/ and 0.61 (0.14) in /e/. 56 subjects received OQ values above 0.75 in /a/ and 52 subjects in /e/. Vowels with subharmonics had a mean OQ of 0.64 (0.16). 16 /a/ vowels and 14 /e/ vowels with subharmonics were characterized by a large OQ (0.75 or higher).

There was little evidence in our data that the OQ interacts with fundamental frequency and intensity. The correlations with the mean F0 of phonation was estimated at –0.20 and –0.15 in /a/ and /e/ vowels, respectively. This finding can be indirectly related to Södersten & Lindestad (1990) who found that pitch does not affect the degree of glottal closure. Correlations with the mean intensity of phonation were even lower.

Södersten & Lindestad (1990) confirmed the relationship between incomplete glottal closure in the horizontal plane and breathiness. According to this, a lower degree of vocal fold closure should correspond to a higher degree of breathiness. We expected that OQ estimates would correlate strongly with perceived breathiness. However, OQ was less informative on voice quality than other perturbation measures. It did not correlate significantly with any of the examined voice-quality dimensions. It showed inferior performance in dicriminating between contiguous voice-quality grades.

3.2.1.3 Noise parameters

Noise parameters give some indication of the noise content in the voice. They are widely applied in voice-quality evaluation. Noise measures are highly valuable as they normally do not rely on precise F0 extraction. Some noise measures decompose the waveform into signal and noise and relate the energies of these two components to each other.

3.2.1.3.1 Glottal-to-noise excitation ratio (GNE) and noise component (NC)

Glottal-to-noise excitation ratio is a voice parameter that measures the ratio between excitation due to vocal fold vibrations and excitation due to turbulent noise in vowel signals. GNE is based on correlations between the Hilbert envelopes in several frequency bands. In the presence of pathology, frequency bands normally contain a large amount of noise so that the correlations between the Hilbert envelopes are low. The lower the highest correlation, the lower the GNE estimate. The best results are reported to be achieved with window size of 500 ms shifted by 250 ms, 3kHz frequency bands and 100 Hz frequency shift (Michaelis et al., 1998; Michaelis, 2000; Fröhlich et al., 2000). GNE does not require reliable F0 estimates, so it can be used to measure severely disturbed voices. Olthoff et al. (2003) showed that using GNE it is possible to discriminate between very irregular voices after total and partial laryngectomy.

GNE was originally related to perceived breathiness. However, a recent study by Godino-Llorente et al. (2010) found that GNE has a high discrimination capability to classify between normal and pathological vowels in general. In their study, the efficiency for screening reached 90 %.

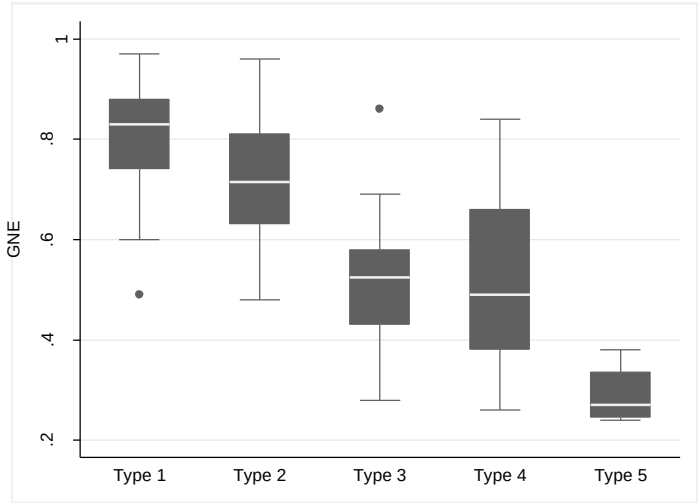
The pathology threshold for GNE was not formulated in the original work as GNE was used to calculate the noise component (NC) for application in the Goettinger Hoarseness Diagram (GHD). For the description of how to calculate NC using GNE the reader is referred to the original manuscript. The pathology threshold for NC was calculated over 92 normal subjects and was set at ca. 2.5. Applying this threshold to our data, only 90 samples of vowel /a/ (30 %) and 72 samples of vowel /e/ (24 %) would be declared pathologic.

We used both measures in our study. NC was calculated with the original algorithm which was available online. GNE was calculated with the algorithm that was implemented in the Praat software.

With increasing grade, GNE was observed to decrease. As expected, we measured the highest correlations with breathiness ratings. Though, GNE correlated moderately with hoarseness, as well. The GNE values differed significantly across all four breathiness grades.

Measures made within one sustained vowel production were not significantly different. However, we observed a minor vowel effect. /e/ vowels had significantly higher GNE values than /a/ vowels ($t = –4.9$, $p < 0.01$). Male subjects had lower GNE values than female subjects in both vowels ($t = –3.35$, $p < 0.01$; $t = –3.41$, $p < 0.01$), which does not agree well with the fact that women reportedly tend to have more breathy voices.

Fig. 48: Boxplots of the GNE values by spectrographic signal type in /a/ vowels.



Spearman's correlation coefficients between GNE and spectrographic vowel type amounted to –0.70 in /a/ and –0.49 in /e/ vowels. Noisy signals of Type 3, Type 4 and Type 5

had lower GNE values than signals without noise. The results are visualised in Fig. 48. The discriminative efficiency of GNE was poorest between Type 3 and Type 4 vowels as the signals of these types may contain on average the same amount of noise.

NC correlated positively with R, B and H. The calculated correlation coefficients were of the same magnitude as in GNE. NC was identical with GNE in the number of significant contrasts. /a/ vowels gave higher NC values than /e/ vowels ($t = 5.3$, $p < 0.01$). Male subjects had significantly higher NC values in both vowels than women ($t = 3.36$; $t = 3.49$, $p < 0.01$). NC differed from GNE in the magnitude and direction of correlations with spectrographic vowel type. They equaled 0.66 and 0.51 in /a/ and /e/ vowels, respectively. The Pearson's *r* between GNE and NC was estimated at –0.99. In this respect, the algorithm implemented in Praat for calculation of the GNE measure may be considered to work with the same precision as the original algorithm.

3.2.1.3.2 Harmonics-to noise ratio (HNR)

HNR was developed as an objective measure of the degree of hoarsenesss obtained by spectrographic method by Yumoto et al. (1982). HNR can be measured in different frequency bands. For an overview of different methods to calculate HNR, see Severin et al. (2005).

There is a great discrepancy in the normative HNR values reported in the literature. Yumoto et al. (1982) found HNR values between 7.0 and 17.0 dB in normal subjects. Other researchers reported higher or lower HNR values in normal subjects (Horii & Fuller, 1990; Ferrand, 2002). Ferrand, (2002) found that HNR tends to decrease with age in women and proposed HNR as an index of vocal aging.

To calculate HNR, we used the method implemented in the Praat software. The method proposed by Boersma (1993) works under assumption that additive noise is white t.i. it contains all frequencies. Since the method in question is not based on F0 estimation, it is suitable to quantify noise in severely disturbed voices without any periodic component.

We expected that HNR would prove a good predictor of all three voice qualities. These expectations were based on the findings by de Krom (1994b) who reported that in mid-vowel fragments, a decrease in harmonic energy irrespective of the examined frequency band was associated with both roughness and breathiness. However, values that he measured on voice onsets indicated that a low HNR value in the band below 2 kHz was associated with roughness, a low HNR in the interval 2–5 kHz – with breathiness. HNR produced a correlation of –0.32 with dysphonic severity in Wolfe et al. (1995b).

In our data, HNR correlated moderately with all three perceptual categories. HNR values decreased with increasing ratings. Correlations were highest with perceived hoarseness. The discrimination efficiency was lower in breathiness as compared to roughness

and hoarseness. HNR was found to be highly correlated with spectrographic vowel type. The correlation coefficients between HNR and spectrographic vowel type was estimated at –0.76 in /a/ and –0.62 in /e/ vowels. We found that male subjects measured less HNR than female subjects in both vowels (t = –4.28, $p < 0.01$; t = –5.72, $p < 0.01$). With significantly higher HNR values in /e/ vowels (t = –12.8, $p < 0.01$), /e/ was less pathologic than /a/. We found a low correlation of –0.27 between HNR and age of the study subjects in /a/ vowels. Age effect was less pronounced in /e/ vowels with a negative correlation coefficient amounting to –0.15.

3.2.1.3.3 Long-term average spectrum

There have been a considerable research interest in relating spectral characteristics to perceived voice quality. Whereas roughness seems to relate to the noise increase throughout the spectrum, breathy voices are normally characterized by high energy in the frequency bands covering the lowest harmonics and low energy above 2 kHz.

Long-term average spectrum (LTAS) describes how energy is distributed over frequencies by averaging individual spectra over time. ”Long-term” implicates a timespan of seconds to minutes, covering entire phrases of speech, at least. Good voices might differ from pathologic voices in terms of the characteristics of LTAS, which are peak locations and spectral tilt[16]. Hammarberg et al. (1980) found that all LTAS frequency bands were significantly correlated with rated breathiness.

The LTAS measure in Praat presents the average power in a sound during a certain time and frequency range relative to $2*10^{-5}$ Pa (which is considered to be a threshold of human hearing at 1 KHz) and is expressed in dB/Hz. LTAS was measured in the frequency range between 0 and 2 kHz. This frequency band covers the first two formants. This part of the spectrum carries much linguistic and non-linguistic information, so it might be also useful for the discrimination of voice quality. It is possible to measure LTAS in connected speech samples. However, speech samples of certain length are required for the LTAS to stabilize in order to give reliable results.

Similar to the mean intensity of phonation data, the first and the second measurements within the same vowel were significantly different. In both vowels, the first measurement gave a higher LTAS value than the second one (t = 11.19, t = 10.86, $p < 0.01$). Correlations with the mean intensity of phonation were estimated at 0.99 in both vowels. No sex or vowel effect was proved for the LTAS data. Correlations with perceived vowel quality

[16] Spectral tilt measures energy decay along the frequency axis. It is normally defined as the difference in the amplitude between the first and the second harmonic (H1 and H2) or between the first harmonic and the harmonic closest to the first formant in the frequency domain. Other landmarks are possible. If the harmonic structure is disturbed or nonexistent, it is the energy in the bandwidths, e.g., from 0 to 200 Hz and from 200 to 1000 Hz that goes into the calculation of spectral tilt: for this is where H1 and F1 are expected to be found. The cutoffs for the used frequency bands are chosen arbitrarily.

were either low or insignificant; the same is valid for correlation with spectrographic vowel type which measured –0.11 and –0.13 in /a/ and /e/ vowels, respectively. LTAS did not differ significantly across contiguous voice-quality grades meaning that information on voice status associated with this measure is poor.

3.2.1.3.4 Largest Lyapunov exponent (LLE)

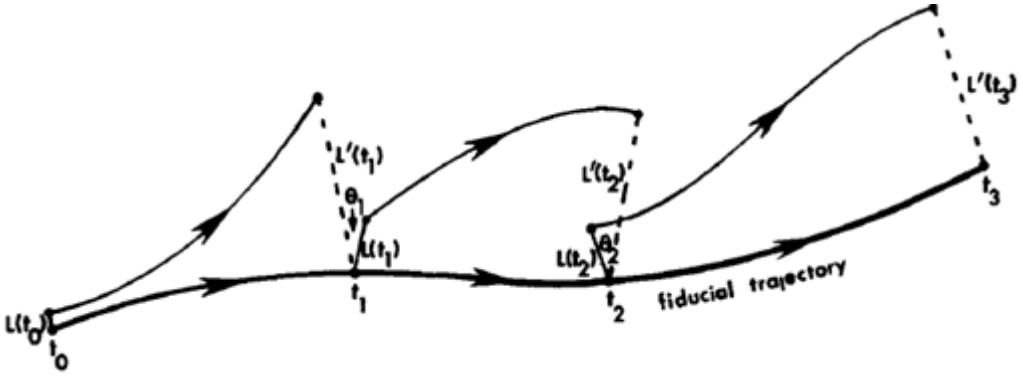
LLE is a measure that quantitatively describes attractors by measuring the rate at which the *m* trajectories of the phase space diverge and is supposed to measure irregularity that manifests itself in the amplitude, phase and periodicity of the signal. The algorithm to determine LLE (bit/s) from empirical data was originally proposed by Wolf et al. (1985). The method deals with data when equations of motion are not expicitely known. It analyses the separation of trajectories along the most unstable direction within a certain radius. A slightly modifed version of the algorithm was described and tested on pathological voices in Giovanni et al. (1999b). In this version, LLE estimates are dimensionless units.

The calculation of the LLE from time series data is based on the assumption that the reconstructed attractor (see section 1.3.6), though defined as a single trajectory, can provide points that may be considered to lie on different trajectories. Factually, these points lie on nearby orbits. Given the time series $x(t)$ and an m-dimensional phase space, the initial (fiducial) point is by definition a point with coordinates $\{x(t_0),\ldots,x(t_0+[m-1]\tau)\}$ and $L(t_0)$ is the distance between the initial point and its nearest neighbour in the Euclidian sense. This pair of data points is propagated a fixed number of steps through the attractor, after which the log of the ratio of final to initial separation between these points is computed and a replacement is attempted. Fig. 49 shows a schematic representation of the evolution and replacement procedure used in Wolf et al. (1985) to calculate the LLE. In Fig. 49, the points t_0–t_3 correspond to replacement points. The time step between replacements is constant $\Delta = t_{k+1} - t_k$. At replacement points, the non-fiducial data point is replaced with a point closer to the evolved fiducial point. The point closest to the evolved fiducial point is ideally confined between the smaller and the bigger length scales. If more than one point is found, the candidate with the smaller angular change is chosen. Points closer than the smaller length scale or further away than the bigger length scale are normally omitted. However, if no point could be found that satisfy the replacement criteria, the large distance criterion and then the angular acceptance criterion are stepwise relaxed. Continued failure results in propagating the same pair of points until the next replacement point is attained. The procedure is repeated until the fiducial trajectory reaches the end of data file. The LLE is calculated as follows:

$$LLE = \frac{1}{t_M - t_0} \sum_{k=1}^{M} \log_2 \frac{L'(t_k)}{L(t_{k-1})},$$

where M is the total number of replacement steps.

Fig. 49: Schematic drawing showing how the LLE is computated from the growth of length elements. When the length of the vector L between two points becomes large, a new point has to be chosen near the reference trajectory. Thereby the replacement length L and the orientation change Θ are to be minimized.



Although both algorithms and various combinations of algorithm parameters were tested, the results of only one of the tested combinations (LLE_{Wolf}) are tabulated in Appendix C. The choice was arbitrary since all combinations were characterized by poor performance with regards to association with perceived voice quality and statistically significant contrasts across the four levels of the grouping variable. The time series data that was used to calculate LLE were acoustic waveforms. The data length used in the present study was ca. 10,000 points. Like Giovanni et al. (1999b), we did not determine attractor dimension but applied the consistency test that ensures that different combination of parameters lead to little change in the same estimation of the LLE value. When embedding dimension m is chosen sufficiently large, LLE values do not change significantly with changing time delay τ and m (Herzel, 1993; Giovanni et al., 1999b). Still there remains a certain degree of arbitrariness in the choice of both length scales.

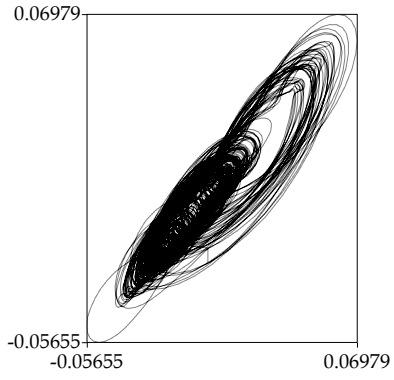
We encountered several problems in calculating LLE from pathologic vowels. It appears that in noisy signals, the selection of neighbouring points is problematic. LLE did not converge in all instances producing missing values in the data matrix. We attribute this effect to short data length and high-dimensionality of attractors in pathologic voices. Not a single combination of parameters could be found to achieve convergence in all 600 vowel samples. Electroglottographic signals were less suited for LLE calculation as they were noisier than acoustic signals. This agrees well with observations made by Giovanni et al. (1999b) and Wolf et al. (1985) that LLE is sensitive to noise.

Despite claims that LLE would quantify the degree of disturbance in voice signals, we found that this measure often fails to characterize dysphonic data. Analysis of normal vowels showed in many cases nonchaotic behavior with negative or slightly positive LLEs

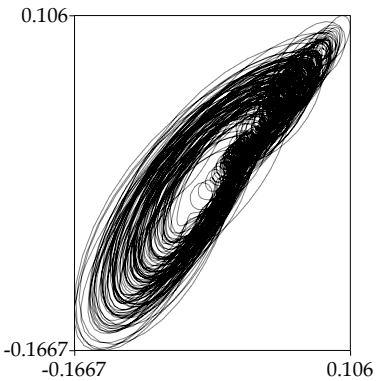
and regular limit cycles. Values close to zero or negative values (obtained with the algorithm by Giovanni) are supposed to indicate stable phonations. Normal vowels had lower LLE estimates than pathologic vowels when calculated with the algorithm by Wolf. In many dysphonic voices, we found indications of irregularities in the attractor shape. Observation of phase portraits revealed that poorer signals formed more complex attractors. Noisy vowels had very complex attractor shapes indicative of high-dimentionality and nonstationarity.

Fig. 50 shows 8 typical attractor shapes that were found in patient data. The corresponding LLE estimates are given for both algorithms. We found that LLE, especially $LLE_{Giovanni}$, poorly reflects the complexity of attractor shapes. It is obvious that the most irregular attractors (Fig. 50d, Fig. 50g, Fig. 50h) have lower LLE estimates than the most regular attractor in Fig. 50f. Further, variation of the algorithm parameters, affected the stability of results in an aberrant manner. In vowels of Type 4 and Type 5 (as in Fig. 50h), $LLE_{Giovanni}$ had either a high positive or a low (sometimes even negative) value suggesting that it is probably not applicable to these voices. Similarly, LLE estimates obtained with the algorithm by Wolf gave either very low or very high LLE estimates when applied to these vowel segments.

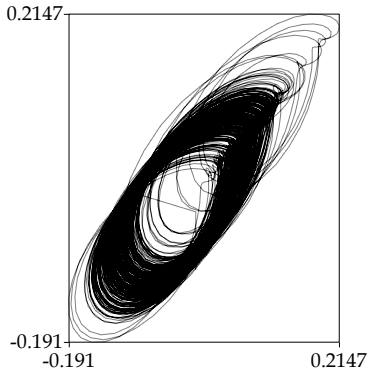
Fig. 50: Typical attractor shapes found in patients. 2D phase plots (m = 3, τ = 8 and N = 5000) of sustained /a/ vowels were calculated from EGG signals. The corresponding LLE values were calculated from acoustic signals using both algorithms and are given after information on each subject.



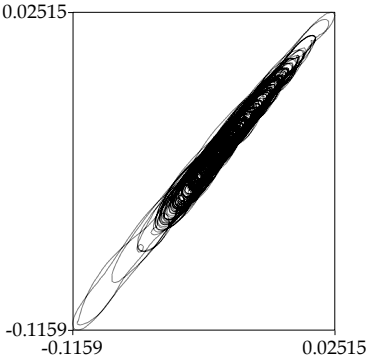
a) subject 44, male, 66, experiencing hoarseness without apparent organic cause
$LLE_{Giovanni}$ = 0.33
LLE_{Wolf} = 295 bit/s



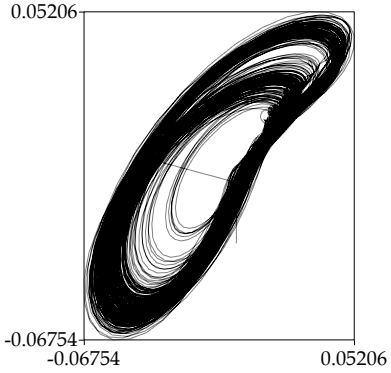
b) subject 10, male, 44, diagnosed with a contact granuloma, postoperative condition
$LLE_{Giovanni}$ = 0.32
LLE_{Wolf} = 411 bit/s



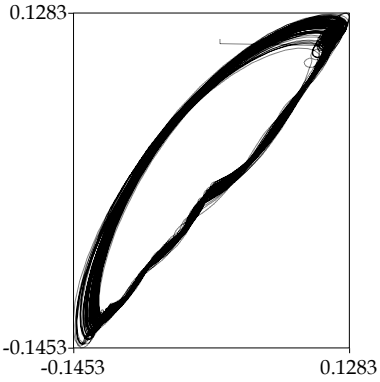
c) subject 26, female, 57, diagnosed with functional dysphonia
$LLE_{Giovanni}$ = 0.35
LLE_{Wolf} = 308 bit/s



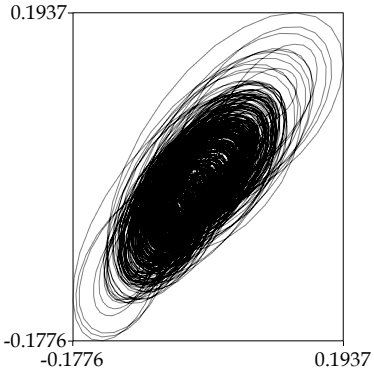
d) subject 50, male, 77, diagnosed with dysarthrophonia of central origin
$LLE_{Giovanni}$ = 0.37
LLE_{Wolf} = 126 bit/s



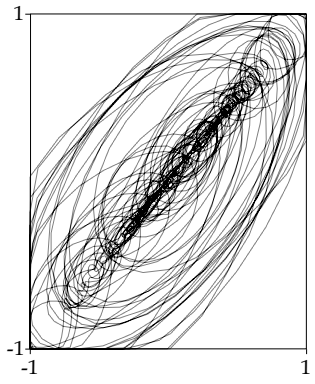
e) subject 87, female, 72, hyper-functional dysphonia, acute laryngitis
$LLE_{Giovanni}$ = 0.18
LLE_{Wolf} = 561 bit/s



f) subject 90, female, 72, diagnosed with unilateral vocal fold paresis
$LLE_{Giovanni}$ = 0.35
LLE_{Wolf} = 221 bit/s



g) subject 23, female, 71, diagnosed with vocal fold paresis following strumectomy
$LLE_{Giovanni}$ = 0.19
LLE_{Wolf} = 1018 bit/s



h) subject 61, male, 59, diagnosed with glottic cancer, post-operative condition
$LLE_{Giovanni}$ = 0.25
LLE_{Wolf} = 930 bit/s

Although we calculated the LLE with twice the amount of data points that were used in Giovanni et al. (1999b), the results were not conclusive as far as perceptual voice quality is concerned. We measured low correlations with perceptual voice quality. In the previous source, pathologic voices that were judged G1 and G2 did not seem to have significantly different group mean LLE estimates from those measured in controls, either. Merely G3 voices had a large mean LLE in the order of ca. 1,0.

It was not possible to distinguish voices with subharmonics from those without, either on the basis of the attractor shape or the numeric value of LLE. The EGG signals of 5 out 8 subjects whose attractors are shown in Fig. 50 had subharmonics. Only in Fig. 50e, the two frequencies correspond to two loops. The attractor in Fig. 50a shows two loops that belong to EGG signal without subharmonics. As expected, LLE was weakly correlated with spectrographic vowel type. A minor vowel effect was found in the LLE measurements ($t = -3.04$, $p < 0.01$). Men and women did not differ signficantly in the LLE estimates.

It is quite possible that better results could be achieved by calculating the exact attractor dimension in every single case or by applying the consistency test to every individual vowel sample instead of the entire vowel inventory. To analyse this amount of information would require immense computing time.

3.2.1.3.5 Aperiodicity index (AI)

Aperiodicity index (AI) is an experimental measure proposed by the author. It is based on power spectrum observations in dysphonic voices: dysphonic signals lack clear-defined harmonic structure, especially in the mid and upper frequencies and many pathologic voices have spectral peaks that appear to be randomly distributed and do not relate harmonically to the fundamental or to each other (Fig. 28). It was hypothesized that when the ratios of the frequencies that contribute to the spectrum of a vowel are not as simple as expected, the vowel is perceived as rough. In contrast, frequencies that stand in a harmonic relationship may be perceived as one entity and do not envoke the perception of roughness. Inharmonic partials make the spectrum denser than the perceived pitch allows and interfere with the purity of the perceived vowel.

AI was calculated in several steps:

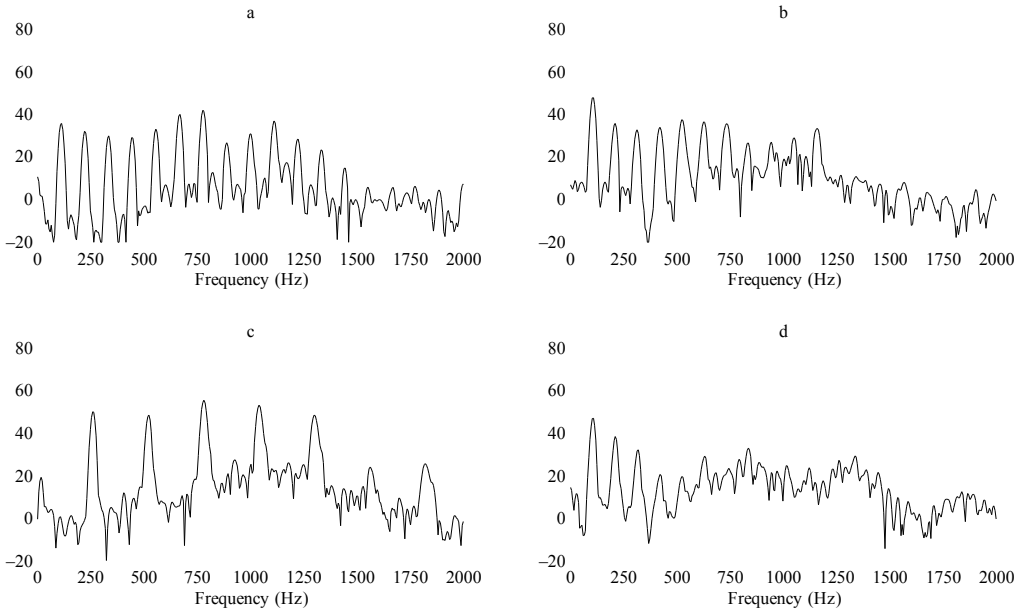
1. Extraction is applied to 1000 ms vowel segments.
2. The signal was downsampled to a sampling frequency of 10 kHz.
3. An 1024-point FFT is applied from which the power spectrum is calculated. The result consists of the amplitude spectrum of points spaced 8 Hz on a linear frequency scale.

4. All local maxima in the range from 30 Hz to 2000 Hz were determined. Experiments were carried out with different frequency ranges equal to 2 kHz, 3 kHz, 4 kHz, and 5 kHz. The best result was achieved with the 2 kHz frequency range, which is in perfect agreement with the results of Remacle & Trigaux (1991) who found that frequency range up to 5 kHz is less rewarding in detecting voice irregularities. They used the frequency range between 0 and 1 kHz in high-resolution frequency analysis of voices with small vocal lesions.
5. Local maxima were further enhanced by removing all frequences 40 dB below the power of the strongest spectral peak and by setting all values around the peaks that are less than 1 FFT point (8 Hz) separated from the local maxima equal to 0 since F0 movements in the power spectrum may be presented as broad peaks.
6. Values other than local maxima were discarded. This measure resulted in a simplified power spectrum.
7. Starting with the second peak, the frequency of every peak, which was previously weighted with its amplitude, is set in relation to the frequency of all preceding peaks in turn. In this way, AI reflects both the occurence and the magnitude of additional frequencies. The absolute values of the difference between every single ratio to the nearest integer are summed up and divided by the number of the peaks found in the power spectrum.
8. The sex effect was supposed to be eliminated by dividing the sum by the number of peaks. However, it is important to note that the calculation was fully automatic; no attempt was made to verify that the first peak corresponds to the fundamental period. Since the mean F0 in voices with detected subharmonics is lower than in voices without subharmonics, the AI estimate would be greater in voices with subharmonics. AI is not supposed to identify subharmonics as such, but evaluate the amount of inharmonic component in a vowel. AI estimates would be even higher in very noisy voices as more peaks contribute to the index.

In fact, we found that simplified spectra of pathologic voices had a greater number of peaks than normal voices. The more irregular the spectrum (harmonics replaced by noise), the higher the AI estimates (Fig. 51a, Fig, 51b and Fig. 51d). AI seems to be insensitive to the type of noise t.i. noise between the harmonics and noise replacing the harmonics (a female voice in Fig. 51c has a higher AI value than a male voice in Fig. 51d). Especially rough voices were found to have a strong inharmonic component. An increase in the level of inharmonic noise seemed to be related to perceived roughness rather to perceived breathiness. An example of the power spectrum of a purely breathy voice with a clearly defined harmonic

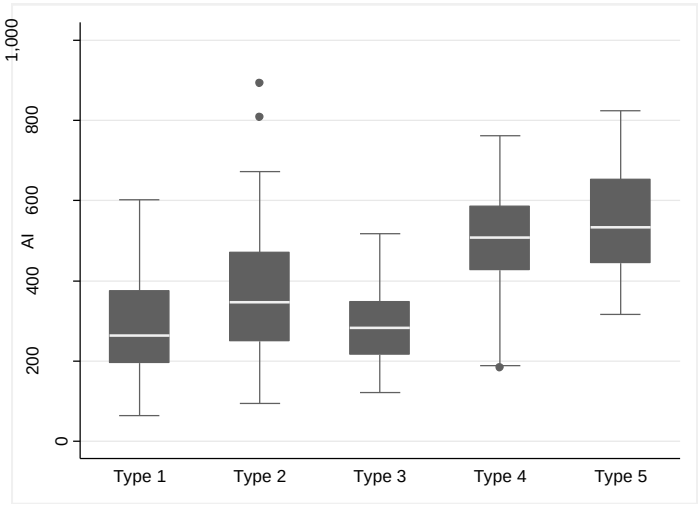
structure and a few disharmonic peaks (voice with subharmonics judged R0 B3, AI = 148) is shown in Fig. 53e. Additive noise between the harmonics was evenly distributed over the whole spectrum, had very low energy and contributed little to the AI estimate.

Fig. 51: Power spectra of sustained /a/ and the corresponding AI estimates produced by a) subject 44, male, 66, experiencing hoarseness without apparent organic cause (AI = 190); b) subject 10, male, 44, diagnosed with a contact granuloma, postoperative condition (AI = 259); c) subject 26, female, 57, diagnosed with functional dysphonia (AI = 475); d) subject 50, male, 77, diagnosed with dysarthrophonia of central origin (AI = 392).



Correlations with spectrographic vowel type were moderate with an r_s of 0.49 and 0.41 in /a/ and /e/ vowels, respectively. AI estimates were lower in signals of Type 1 and Type 3 without irregularities in the lower part of the spectrum (Fig. 52).

Fig. 52: Boxplots of the AI estimates measured on /a/ vowels by spectrographic vowel type.



A minor sex effect was still found to exist. Men had higher AI values in both vowels than women ($t = 3.0$, $p < 0.01$; $t = 7.11$, $p < 0.01$). There was a minor difference between the AI estimates calculated from the first and the second sample within the same vowel ($t = 3.19$, $t = 3.15$) which was significant ($p < 0.01$). /a/ vowels had significantly higher AI values than /e/ vowels ($t = 26.16$, $p < 0.01$). This is not surprising since the spectral peaks of /e/ are weaker than those of /a/ and the harmonic structure of /e/ vanishes in spectral dips between the first two formants.

AI correlated moderately with roughness and hoarseness. Correlations with breathiness were much lower. The discriminative efficiency was highest in /a/ vowels that were used to predict roughness.

AI seems to work well with vowels with rich harmonic structure and vowels without harmonic structure. Its derivation can be related back to the physics of voice production. Still, a number of improvements and extensions to this implementation are possible.

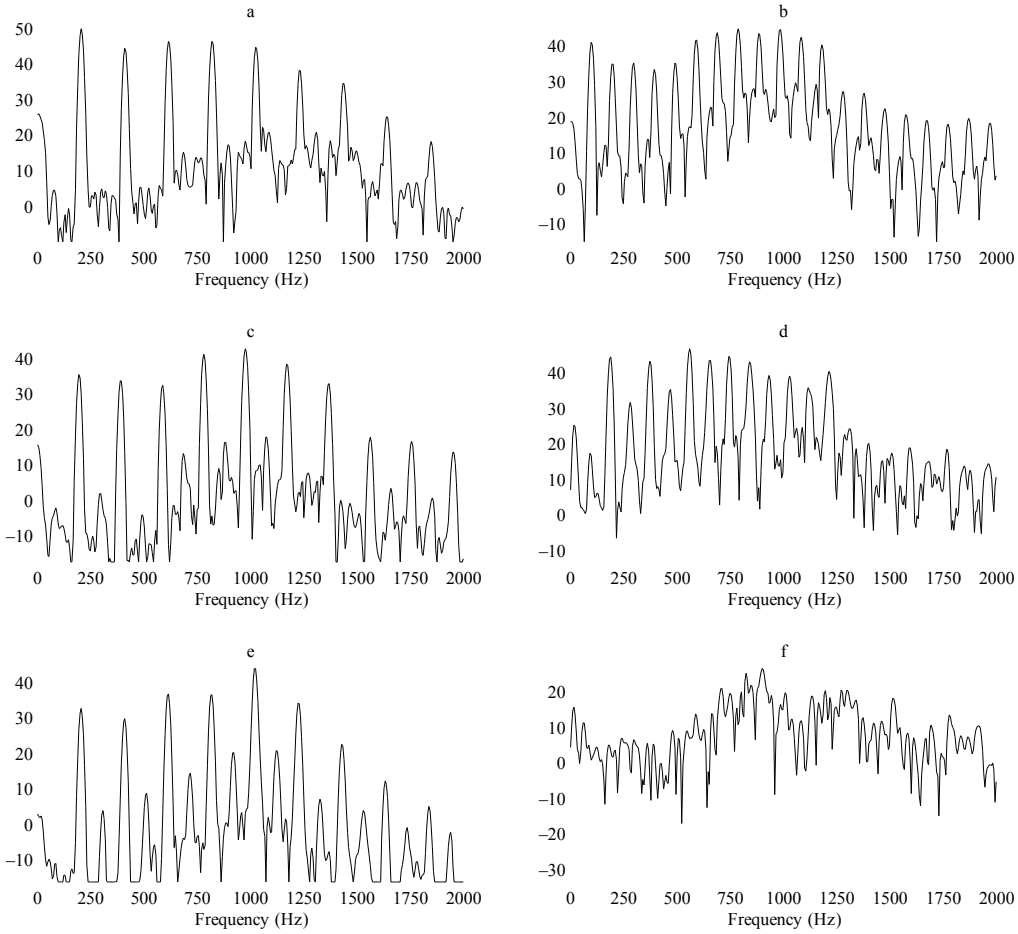
3.2.1.3.6 Subharmonics-to-harmonics ratio (SHR)

SHR was primarily invented as a pitch detection method based on the idea of spectrum compression that involves summation of a sequence of compressed spectra. Compressed spectra are shifted along the logarithmic frequency scale, which makes the F0 peak more prominent. To reduce computational cost, only part of the spectrum (max. F0 = 1250 Hz) is used.

SHR describes the amplitude ratio between subharmonics and harmonics using pitch estimation by considering subharmonic effects. The greater the ratio the greater is the degree of deviation from the normal voice. Thus, an SHR value above 0.4 suggests a strong subharmonic component, values between 0.2 and 0.4 are to be considered ambiguous. The basic algorithm to calculate SHR is described in Sun (2000). The Matlab script that implements the algorithm was obtained from Sun (2008).

We tested the SHR algorithm on our database of pathologic voices and came to conclusion that this measure tells little about voice function status. Although many voices have plausible SHR values (normal voice in Fig. 53a, voices with subharmonics in Fig. 53c and Fig. 53d, non-F0 voice in Fig. 53f), there exists quite a number of counterexamples that do not conform to the proposed interpretation of SHR data. A normal voice in Fig. 53b was assigned a high SHR value. Accordingly, a pathologic voice in Fig. 53e with clearly defined subharmonics became a low SHR value.

Fig. 53: Power spectra of sustained /a/ vowels and the corresponding SHR values produced by a) subject 57, female, 32, normal (SHR = 0.001); b) subject 138, male, 29, normal (SHR = 0.45); c) subject 87, female, 72, diagnosed with hyperfunctional dysphonia and acute laryngitis (SHR = 0.27); d) subject 90, female, 72, diagnosed with unilateral vocal fold paresis (SHR = 0.72); e) subject 9, male, 57, diagnosed with unilateral vocal fold paralysis (SHR = 0.03); f) subject 146, male, 68, after partial larynx resection (SHR = 0.31).



In our analysis, SHR did not correlate with spectrographic vowel type. We cannot confirm the tendency for voices with subharmonics or voices with impaired harmonic structure to have a high SHR value. Merely voices without F0 (Type 5 signals) were clearly identified as ambiguous with regards to F0. Sex effect was found in /e/ vowels only. Male subjects had slightly higher SHR values than female subjects ($t = –4.65$, $p < 0.01$). No vowel effect could be detected.

Correlations with perceived voice quality were found to be low or insingnificant. SHR estimates do not appear to differ significantly across voice-quality grades except in /a/ vowels, where correlation with roughness ratings and the number of significant contrasts were highest.

3.2.1.4 Summary

Measures with both at least moderate correlation coefficients with and high discriminative ability across different voice-quality ratings were considered promising candidates for clinically applicable voice measures. Using these criteria, several measures like acoustic jitter and shimmer, FMF, IC, GNE and HNR measures could be especially useful in predicting perceptual voice quality.

It has been observed that some predictor variables were strongly correlated with each other. When variables highly correlate, some of them may be perceptually redundant since correlated parameters probably explain the same part of the variance in perceptual voice-quality ratings. Correlations between the examined parameters are presented in Table 20. To mention just the most important observations, LTAS correlated highly with the mean intensity of phonation; OQ correlated moderately with EGG jitter; perturbation measures like acoustic jitter, acoustic shimmer, IC and FMF did not only correlate strongly with each other but they also correlated highly with noise measures like HNR and GNE.

Table 20: Correlation matrix presenting the Pearson's *r* between voice parameters measured on mid-vowel segments.

Variables	*Ji*	*Ji E*	*Int*	*Shi*	*Shi E*	*LTAS*	*HNR*	*GNE*	*FMF*	*IC*	*LLE*	*AI*	*OQ*	*SHR*
Ji	1.0													
Ji EGG	0.04	1.0												
Int	-0.08	0.01	1.0											
Shi	0.46	0.25	-0.16	1.0										
Shi EGG	0.09	0.73	-0.01	0.24	1.0									
LTAS	-0.08	0.01	0.99	-0.17	-0.01	1.0								
HNR	-0.63	-0.22	0.19	-0.78	-0.24	0.20	1.0							
GNE	-0.21	-0.18	0.18	-0.54	-0.19	0.19	0.57	1.0						
FMF	0.66	0.08	-0.02	0.52	0.12	-0.02	-0.58	-0.26	1.0					
IC	0.67	0.24	-0.16	0.75	0.26	-0.17	-0.93	-0.53	0.64	1.0				
LLE	0.09	0.15	-0.15	0.24	0.06	-0.16	-0.32	-0.15	0.13	0.29	1.0			
AI	0.50	0.11	0.25	0.50	0.15	0.24	-0.71	-0.16	0.50	0.66	0.19	1.0		
OQ	0.01	-0.57	0.04	-0.17	-0.49	0.04	0.09	0.16	0.04	-0.12	-0.04	0.04	1.0	
SHR	0.15	0.01	0.08	0.11	-0.03	0.07	-0.18	0.11	0.22	0.15	0.04	0.33	0.04	1.0

In the next step, the percentage of explained variance was calculated for each of the predictor variables. The results for logtransformed or squared data are given in Appendix E. Only measures with at least 10 % of explained rating variance are displayed. The upper range of the percentage of variance explained by a single variable did not exceed 40 %.

The best predictors of roughness were IC and AI followed by acoustic jitter and FMF. Measures taken from /a/ vowels explained more variance than those from /e/ vowels. Noise measures explained little roughness rating variance with exception of AI, which explained 33 % of the rating variance. The perceptual irrelevance of AI for predicting breathiness was confirmed by the data.

GNE proved the best single predictor of rated breathiness followed by NC and acoustic

shimmer. As might be expected, little breathiness rating variance was explained by perturbation measures. This was not surprising since irregularities in the waveform are more related to roughness than to breathiness. Acoustic shimmer was a less useful predictor of roughness than of breathiness.

Similarly, hoarseness ratings were related to many voice parameters. The parameter with the highest explained variance was the IC measure.

3.2.2 F0 statistics in connected speech

Speaking fundamental frequency (SFF) carries important information on voice quality that could be relevant in predicting perceptual roughness in addition to conventional voice parameters. In this section, F0 distributional characteristics and measures based on instantaneous fundamental frequency estimates (Fx) in connected speech will be addressed.

In contrast to sustained phonations, pitch variation is normal in speech. It is expected to be found at voice on- and offsets, at paragraph or sentence boundaries, at placement of focal and contrastive stress. The lowest possible pitch is normally realized at the endpoint of sentences with a falling F0 contour. Nevertheless, we hypothesized that even greater cycle-to-cycle changes in SFF should be reflected in the extracted instrumental measures from dysphonic population.

3.2.2.1 F0 means, medians and standard deviations

This section focusses on the mean and median F0 in 150 study subjects. By itself, mean F0 is a poor predictor of voice pathology (Murry, 1978), unless the perceived pitch lies outside the range that would be considered normal for a particular speaker i.e. if a speaker stands out from the average by an unusually high or low voice. F0 properties like F0 standard deviation, lower and upper limits of the F0 range have proved so far to be more valuable features for detecting voice pathology.

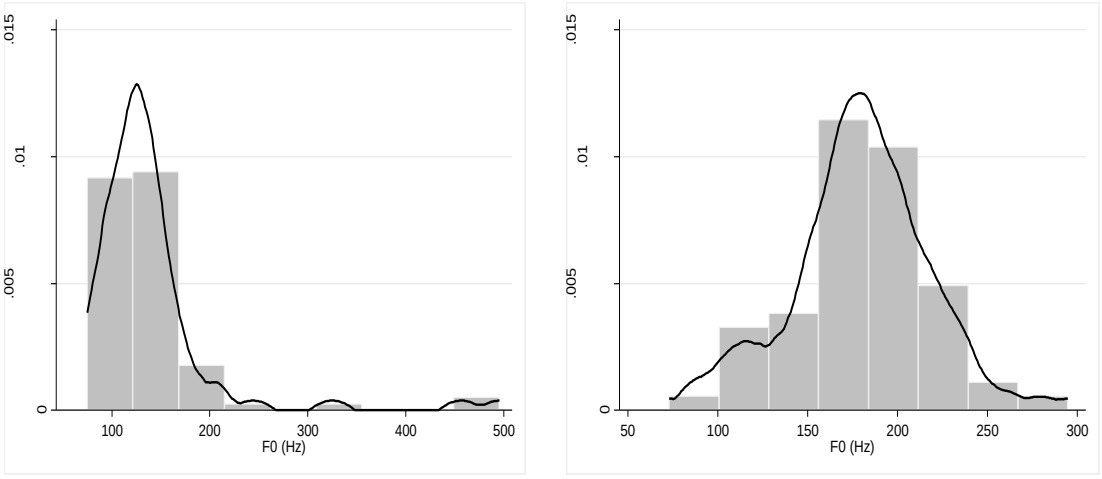
Fig. 42 provides the distribution of the means in male voices. The solid black line presents the kernel density plot. Data is calculated from acoustic signals. Approximately 35 % of the male speakers had a mean F0 lying between 98 Hz and 131 Hz, considered the normal mean F0 range for male voices (Wendler & Seidner, 1997). The mean of the means is 138 (65) Hz. There is a positive skewing (3.62) with 15 means below 100 Hz and 18 means above 150 Hz.

The means in the frequency range between 300 Hz and 500 Hz are attributable to 2 outliers with a strained near to whispered voice quality in which F0 was absent. These cases are definitely pitch detection errors.

The distribution of the medians is very similar to the distribution of the means. According to a paired samples *t*-test, the F0 medians in male voices differ significantly from the F0 means ($t = 2.59$, $p = 0.011$). The distribution of the medians has a positive skewing (3.81), but the mean of the medians has moved down to 134 (66) Hz.

In female voices, the mean of the means was estimated at 179 (39) Hz. The distribution of the means had a slightly negative skewing (–0.06). 21 subjects (ca. 32 %) had a mean F0 between 195 Hz and 262 Hz, the normal mean F0 range according to Wendler & Seidner (1997). 14 female voices (21 %) have a low F0 mean below 160 Hz (Fig. 54).

Fig. 54: Distribution of the mean F0 in acoustic signals in male (left, bin width = 35.25) and female (right, bin width = 46.77) voices.



No significant difference was found between the means and the medians in female voices ($t = 0.64$, $p > 0.52$). The mean of the medians amounted to 180 (41) Hz and had a slight positive skewing (0.09).

In EGG signals, the difference between the means and the medians was not significant in both sexes. The mean of the means was 126 (36) Hz in male voices. The mean of the medians amounted to 126 (37) Hz. In female voices, we measured a mean of the means of 176 (40) Hz and a mean of the medians of 178 (42) Hz.

The F0 means obtained from acoustic and electroglottographic signals were found to be marginally different ($t = 2.36$, $p = 0.02$). When data is broken down by sex, a minor difference in the mean F0 between the signals was found to be statistically significant in male voices only ($t = 2.11$, $p < 0.01$).

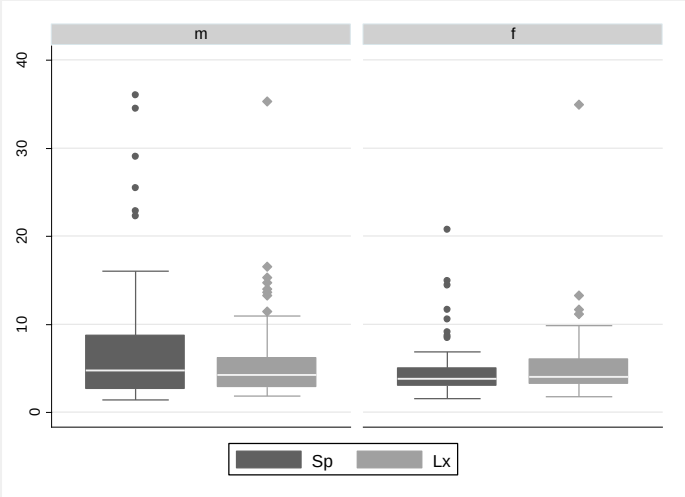
The comparison of the F0 distributions across sexes shows that in clinical population the distribution of the means and the medians in both sexes overlap to a certain extent, the means and the medians of 86 male voices shifting closer to higher values typical for female voices and the means and the medians of 64 female voices moving down to lower values

outside the sex-specific normal frequency range. Vocal pathology is definitely responsible for this shift in prefered fundamental frequency as compaired to normal voices.

The second factor to consider are tissue changes in the ageing larynx since clinical population we studied was on average 56 years of age, changes in male voices having earlier onset and a greater impact on voice characteristics (Kahane, 1987). The correlation coefficients between the mean/median F0 and age was estimated at 0.16/0.17 and –0.33/–0.28 in male and female voices, respectively. Judging by the sign of the correlation coefficients, the shift in the mean and median F0 occured in the expected direction.

Fig. 55 shows the boxplots of the F0 standard deviation estimates in male and female voices given for Sp and EGG signals. We found that sex effect was significant in acoustic signals with men having a higher SD than women (t = 3.7, p < 0.01), which does not agree well with studies claiming that women show more variability in F0. However, the magnitude of the effect is rather small.

Fig. 55: Boxplots of the F0 SD (st) given for acoustic and EGG signals stratified by sex.



In both sexes, the F0 SD measure in semitones increased with increasing voice-quality grade. Both signals showed the highest correlation with perceived roughness. However, judging by the number of significant contrasts F0 SD calculated from EGG signals could be more useful in predicting hoarseness. The same conclusions can be drawn, when F0 variablity is expressed as frequency modulation factor.

Interestingly, the mean F0 estimates obtained from sustained phonations and speech samples did not differ significantly (p < 0.05). The same effect was found in EGG signals, as well. This finding may be indicative of the fact that comfortable pitch levels in sustained phonations and speech are very likely to lie very close together. This finding seems to be in conflict with another finding reported in section 3.2.1.1. As stated earlier there is a significant difference in the mean F0 measured in /a/ and /e/ vowels. Upon closer examination of patient

data, it turned out that comfortable pitch level might depend on voice quality since dysphonic voices were found to behave differently with regard to the choice of comfortable pitch level for speaking as compaired to sustained vowels. Less hoarse voices (H0 and H1) used a mean SFF that was closer to the pitch level chosen for /a/ vowels ($p < 0.01$) while more hoarse voices (H2 and H3) spoke with a mean SFF that was not significantly different from the corresponding mean F0 in /e/ vowels ($p < 0.01$). It was assumed that higher pitched speech (relative to sustained phonations) in more hoarse voices is a result of putting more effort into speaking.

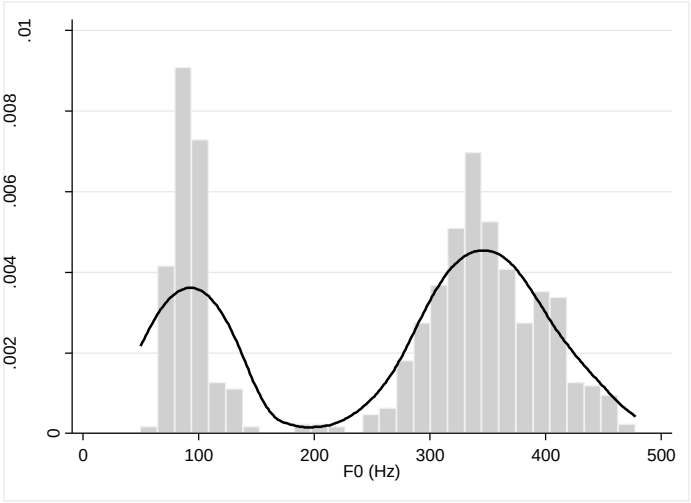
3.2.2.2 Unimodal and bimodal F0 distributions

In perceptual experiments with dysphonic voices, the modal F0 may be more important than the mean and the median F0. Applied to connected speech, pitch is the height of the most frequent tone within an utterance which is equivalent with the tone of the crucial stressed vowel content. Statistically speaking, the perceived pitch of an utterance roughly corresponds to the modal F0.

Histogram analysis provides useful infomation on the distibution of Fx data. A healthy voice shows a comparartively narrow F0 distribution with a single peak and a low standard deviation. In normal F0 distributions, the only mode is supposed to be found in the lower part of the speaker's frequency range and the distribution to have a positive skewing (Jassem et al., 1973; Traunmüller & Eriksson, 1995). A pathological voice is expected to deviate from the normal voice. In fact, we observed two major maxima in the distribution of Fx values in many dysphonic voices. We assumed that the emergence of the second mode in the lower frequencies of dysphonic voices might represent something important.

Fig. 56 shows a histogram of a voice with two subharmonic frequencies that was rated R3. The mode in the lower frequencies is located at 95 Hz; the other mode in the frequencies typical for female voices (normal register) at 340 Hz. The percentage of values below 120 Hz was estimated at 32 %. It is obvious that the mean F0 (260 Hz) falls in the gap with the frequencies that are seldom or never used. The median F0 (320 Hz) is closer to the normal register mode.

Fig. 56: Histogram showing the Fx distribution obtained from patient 3, female, 76, diagnosed with leukoplakia. The solid black line represents the kernel density plot.



Bimodal F0 distributions arise as a result of pitch detection errors due to subharmonics and exceptionally low frequencies in creaky voice. In many studies on F0 distribution of normal voices, subharmonics are treated as pitch detection errors. In dysphonic voices, the quality of the signal itself is the main course of gross pitch errors. Errors related to subharmonic frequencies are errors in the sense that the corresponding Fx estimates found by a pitch detection algorithm were not intended by the speaker and their presence represents a diminished ability of dysphonic speakers to exert control over laryngeal tone. In this case, the lowest subharmonic competes with the fundamental frequency and F0 drops by half or more only if the subharmonic energy exceeds a certain spectral level. Strong subharmonics may not only induce the perception of a lower pitch than that intended by the speaker but also influence perceptual ratings.

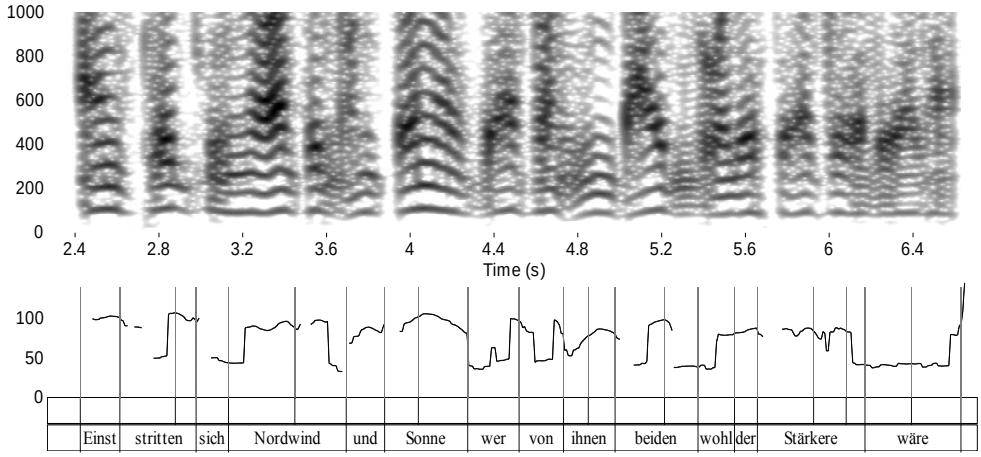
By accounting for the modes in Fx distributions of dysphonic voices, we were looking for the way to use pitch errors for description of dysphonic voices since the problem of pitch errors due to subharmonics cannot be solved by restricting the lowest possible frequency to a certain uniform value, which besides eliminating ”errors” in some but by no means all voices would eliminate evidence of creaky voice as well as voice on- and offsets. So we introduced a new variable that would to a certain extent reflect the modality of the distribution. This measure was the percentage of F0 values below 65 Hz for men and 120 Hz for women. It was termed percentage of low frequency values (PLF). The thresholds lay 11.4 st apart from the mean F0 reported for nondysphonic speakers of German in Rappaport (1958)[17].

Applying the above-mentioned criteria for mode detection to acoustic signals, bimodal distribution was found in 68 speakers (48 %). The distance between the modes comprised

[17] Rappaport (1958), estimated a mean of 129 (17) Hz in male (n = 190) and 238 (26) Hz in female subjects (n = 108).

on average 13.3 semitones. A notable amount of creaky voice and subharmonic frequencies were detected in 78 speakers via spectrographic scanning, of which 84 % (66 speakers) passed the 10 % criterion for a second mode. The F0 distribution was unimodal in 54 male (64 %) and 30 female (45 %) speakers. One healthy voice from 5 had a bimodal F0 distribution. Some distributions had more than one preferred F0 target (smaller humps in the histogram) within a single mode. In this case, there was no gap between the modes with seldom or never used frequencies, the strongest F0 target was identified as the only modal F0. Fig. 57 shows a narrow-band spectrogram of a bimodal male voice with a mean SFF of 83 (43) Hz.

Fig. 57: Narrow-band spectrogram of a phrase ”Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre.” spoken by patient 19, m, 53, after bilateral removal of leukoplakia. The bottom panel shows the corresponding pitch contour. Vertical lines indicate syllable and word boundaries. Segmentation was made using the broadband spectrogram.



This particular voice favours F0 values around 100 Hz. On several occasions within the phrase, the voice changes into subharmonic pattern across the vowels, in which case frequencies around 50 Hz can be detected in the spectrogram and the corresponding pitch contour.

In distributions with two modes, the first maximum was located in the lower frequencies and therefore is referred to as low register mode. The mean of the low register modes in bimodal voices was found at 60 (17) Hz in male and 89 (19) Hz in female speakers. The mean of the normal register modes was 138 (56) Hz and 185 (48) Hz, respectively. In unimodal distributions, the mean of the modal F0 equals 132 (60) Hz and 174 (43) Hz in male and female voices, respectively.

Of 52 male voices with a unimodal F0 distribution, the majority (31 male speakers) was classed R1. The remaining 6 speakers were judged R0, 13 speakers R2 and 2 speakers R3.

In unimodal distributions, the mean of the means and the mean of the medians stratified by roughness grade measured 120/119 Hz in R0, 137/135 Hz in R1, 137/127 Hz in R2 and 337/339 Hz in R3, respectively. Voices judged as R1 and R2 tend to have slightly raised group means; the gap between the group mean and the group median being larger in R2. Voices rated R3 had abnormally high group mean and median values. The mean SD had a tendency to increase with increasing roughness grade and measured 22, 30, 49 and 48 Hz, respectively.

Among 32 male voices with a bimodal F0 distribution, no voice received R0, just 1 voice was typed R1, 18 voices were judged R2 and 13 voices were graded R3.

In bimodal male voices the mean of the means and the mean of the medians measured 129 Hz/129 Hz in R1; 133 Hz/129 Hz in R2; 126 Hz/119 Hz in R3. The mean SD was increasing with increasing roughness: 39 Hz, 46 Hz and 59 Hz. The mean locations of the low register modes for voices grooped by roughness grades R1, R2 and R3 were estimated at 50, 59 and 61 Hz, respectively. The mean normal register modes were found at 139 Hz, 121 Hz and 179 Hz in R1, R2 and R3, respectively.

In male voices, roughness ratings were found to significantly correlate with F0 SD (r_s = 0.52), PLF (r_s = 0.57) and the number of modes (r_s = 0.68), $p < 0.05$. The mean and the median F0 weakly correlated with perceived roughness with an r_s of 0.28 and 0.25, respectively.

The majority of unimodal female speakers were rated either R0 (8 speakers) or R1 (15 speakers). The remaining speakers were judged as follows: 4 speakers were graded R2 and 3 speakers received R3.

In female voices, the mean of the means and the mean of the medians amounted to 201/199 Hz, 178/175 Hz, 218/216 Hz and 99/97 Hz in R0, R1, R2 and R3 respectively. The F0 SD averaged 29 Hz, 29 Hz, 48 Hz and 45 Hz. The modal F0 was found at 196 Hz, 165 Hz, 183 Hz and 128 Hz. Note that speakers judged R3 had exceptionally low mean and modal F0, outside the normal mean SFF range.

Of 36 female voices showing a bimodal distribution, 3 speakers were judged R0, 14 speakers R1, 13 speakers R2 and 6 speakers R3. The means of the means and the means of the medians were found to be consistently decreasing across the first three roughness grades, with 200/199 Hz, 184/187 Hz, 155/155 Hz and 196/210 Hz in R0, R1, R2 and R3, respectively. The corresponding means of the normal register mode were estimated at 190 Hz, 187 Hz, 160 Hz and 253 Hz. The low register modes were located at 87 Hz, 89 Hz, 84 Hz and 107 Hz, respectively. In 6 female voices judged R3, low register modes were nearly as strong or stronger than the normal register mode. The mean SD was consistently increasing across roughness grades: 32 Hz, 38 Hz, 41 Hz and 81 Hz.

In female voices, the correlations are less spectacular, roughness ratings significantly correlated with the F0 mean (r_s = –0.26), F0 SD (r_s = 0.38), PLF (r_s = 0.56) and the number of modes (r_s = 0.31), $p < 0.05$.

Our data suggests that in dysphonic voices the group means and the corresponding medians lie relatively close together disregarding the modality of the distribution. In individual voices with a bimodal distribution, the mean/median and modal F0 may lie very far apart. In extreme cases, the mean and median values may lie somewhere between the modes falling on frequencies which are seldom or never used. Thus, the mean and median F0 may not always be representative of the preferred F0 targets in dysphonic population.

When it comes to the distribution of bimodal voices across roughness grades, there is a clear difference between the two sexes. In both sexes, the proportion of bimodal voices increased with perceived roughness. However, in females, quite a number of bimodal voices were perceived as normal or slightly rough. On the contrary, male voices with bimodal F0 distribution were mainly judged as moderately or severely rough.

The relation between bimodal F0 distribution and subharmonics in read vowels can be regarded as confirmed. However, subharmonic frequencies in read vowels cannot be the only source of roughness, since roughness might be perceived in the absence of subharmonics and the presence of subharmonics do not always result in the perception of roughness. Apparently, there exists an energy threshold below which the presence of subharmonics is not perceived as roughness.

Of 150 acoustic records, only 18 voices did not use low frequencies at all. In 34 subjects PLF exceeded 10 %. PLF tended to rise with increasing roughness grade in both signals. The group means calculated from EGG signals seemed to be more contrastive than those from acoustic signals.

Significant correlations in the order of 56 % and 57 % were found to hold between roughness ratings and acoustic PLF in female and male voices, respectively. EGG PLF correlated moderately with an r_s of 0.42 in male and 0.44 in female voices. When sex effect was neglected, the correlation coefficients dropped to 0.46 in acoustic signals and measured only 0.37 in EGG signals. We hypothesized that perception of roughness may differ in male and female voices. We found that male speakers had significantly more F0 values below the specified threshold than women in acoustic and electroglottographic signals with t = –5.15, $p < 0.01$ and t = –4.27, $p < 0.01$, respectively. In rating roughness, listeners allowed for more low F0 values in male speakers.

3.2.2.3 Subharmonics in vowels and bimodal F0 distributions

This section deals with the association between modality of F0 distribution and subharmonics in connected speech. The following observation was made on comparing the occurence of subharmonics in sustained vowels /a/ and /e/ with modality of the F0 distribution in connected speech: 56 % (38 patients) of the patients with a bimodal F0 distibution had subharmonics in /a/ and 49 % (33 patients) in /e/, 38 % (26 patients) had subharmonics in both sustained vowels and 33% (23 patients) in none. Among the patients with unimodal distribution, 19.5 % (16 speakers) had subharmonics in /a/, 10.9 % (9 speakers) in /e/, 7.3 % (6 speakers) in both and 62 % (51 speakers) in none. This amounts to the conclusion that connected speech and sustained phonation are very different from a physiological point of view and that some patients perform better on sustained vowels than on connected speech and vice versa. However, if a subject has subharmonics in one or both sustained phonations, it might be more probable that he has a bimodal F0 distribution in speech.

3.2.2.4 80 % F0 range (80R)

Speaking frequency range gives some idea of variability of F0. In read speech, the range of F0 values is relatively small in comparison to spontaneous and emotional speech. Women are thought to use a larger F0 range to achieve expressiveness.

In speech samples, the 80 % F0 range was calculated as the difference in semitones between the 90^{th} and 10^{th} percentile of F0 values for each speaker. One more reason to look for the modes in connected speech of dysphonic voices are unrealistic F0-range estimates delivered by automatic systems applied in voice analysis. In bimodal distributions, F0-range calculation based on extreme values is unreliable for the reason that extreme values are probably failure of the analysis software as they were not intended by the speaker.

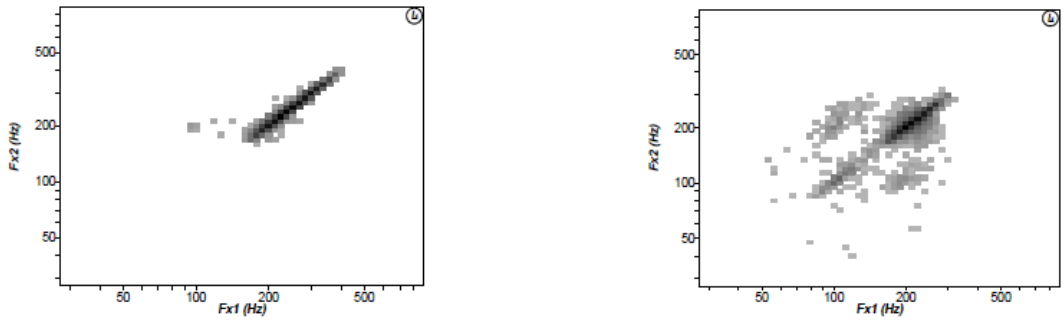
In unimodal F0 distributions, male speakers measured a mean of 6.4 (3.3) st ranging between 2.6 st and 19.1 st. Female speakers used a similar mean value of 6.5 (2.9) st in the range from 3.4 st to 18.1 st. In bimodal distributions, 80R amounted to 10.5 (5.6) st and 12.1 (7.7) st in female and male speakers, respectively; the 80R estimates ranged from 4.1 st to 41.6 st. This finding does not compare well with previous research since the fundamental frequency range is supposed to be severely reduced in pathology.

We could not confirm that men and women differed significantly in 80R. Association with roughness ratings was the highest. 80R correlated moderately with roughness. It differed significantly across 3 roughness grades with a poor discriminative effectiveness between R0 and R1 voices.

3.2.2.5 Irregularity index (IFx)

IFx (Fourcin, 2009) is intended to capture cycle-to-cycle variability in pitch and is obtained by way of measuring the distance between each two consecutive cycles. The measure can be visualized in a crossplot where the frequency of the first vocal fold cycle in pair is plotted against the frequency of the second vocal fold cycle (Fig. 58). When two successive vocal fold instantaneous frequencies are equal, the pair lies exactly on the main diagonal. The central diagonal core of the crossplot contains by definition only cycle pairs that differ in Fx by less than 6 % in both directions. When the separation between consecutive Fx values is larger than 6 %, the pair is located outside the core. IFx measures the number of occurrences of the pairs that fall outside the main diagonal core divided by the total number of the cycle pairs in the speech sample. According to Fourcin (2009), an IFx of 10 % can be conceived of as a pathology threshold.

Fig. 58: Vocal fold period crossplots of normal (left) and dysphonic (right) speakers (from Fourcin et al. (2002)).



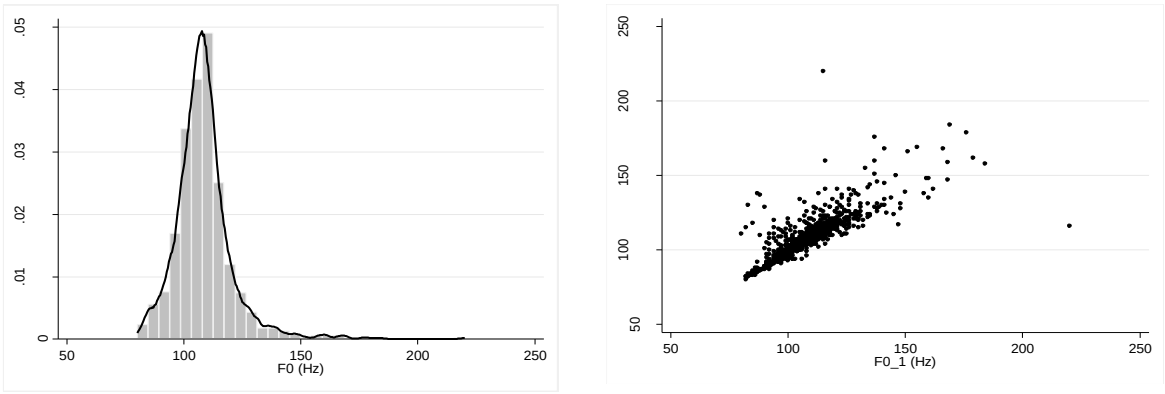
We used both acoustic and electroglottographic signals to calculate IFx. Appendix F shows some examples of Fx histograms and the corresponding IFx crossplots with data calculated from acoustic signals.

The IFx values ranged from 4 % to 38 % in acoustic signals. In EGG signals, the range of IFx values lay between 7.3 % and 87.5 %. This is not surprising, given that IFx was calculated from Sp and Lx signals using different voiced content. Voiced segments in acoustic signals were corrected for artefacts due to coarticulation.

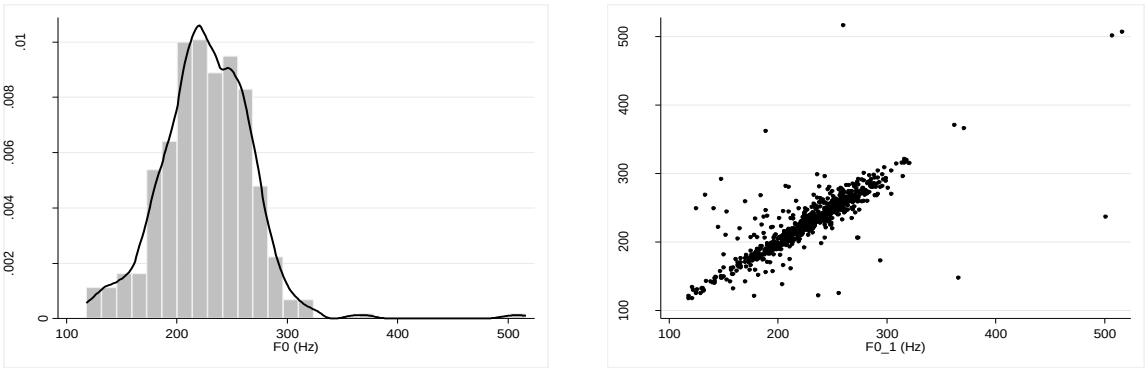
In normal voices, data is supposed to cluster along the main diagonal and form just one lobe. In dysphonic voices, we observed that broader Fx distributions gave greater IFx estimates (Fig. 59, examples (b) and (d)). The broadest distributions were found in non-F0 voices. Two lobes on the diagonal are assumed to represent the two modes in the F0 distribution (example (c)). The second lobe in the lower frequencies indicates a high probability of subharmonic frequencies. Values that fall outside the main diagonal might form side lobes. The presence of the side lobes of either side of the main diagonal might be related to creaky voice or diplophonic frequencies that are not harmonically related to F0. Additional lobes might be responsible for a strong sensation of roughness.

Fig. 59: Examples of F0 distributions and corresponding vocal fold period crossplots.

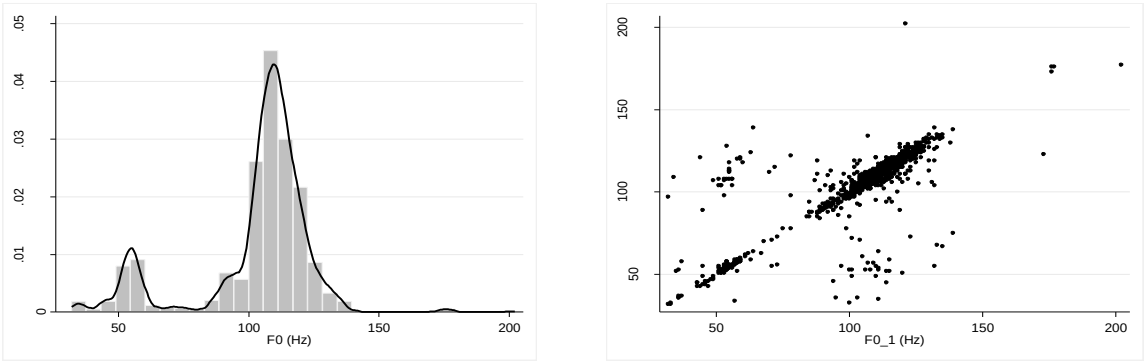
a) Patient 22 (male, 73) diagnosed with vocal fold cancer, IFx = 13 %.



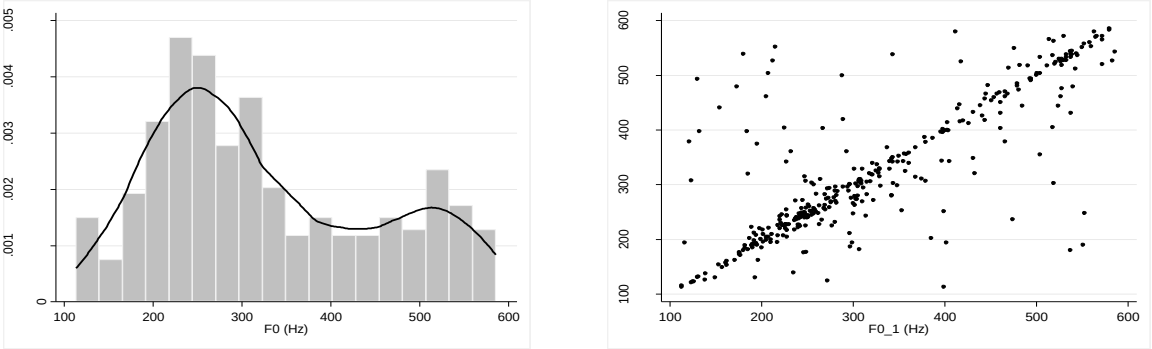
b) Patient 39 (female, 32) diagnosed with functional dysphonia, IFx = 17 %



c) Patient 16 (female, 44) diagnosed with Reinke's edema, IFx = 13 %



d) Patient 134 (male, 69) after partial larynx resection using ventricular fold voice, IFx = 37 %



It is obvious that bimodal voices do not nessessarily have a high IFx value. It seems that IFx accounts stronger for side lobes than for the second lobe on the diagonal since values that form this additional lobe might lie within the 6 %-core of the diagonal.

The IFx values obtained from acoustic and EGG signals were significantly different. EGG signals gave higher IFx estimates (t = –10.03, p < 0.01). The mean IFx in 5 normal subjects was 20.6 % and 23.6 % in Sp and Lx signals, respectively. In acoustic signals, 13 speakers were found to have an IFx value below 10 %, from which 5 subjects received R0, 7 subjects were rated R1 and 1 subject R2. Only 2 speakers had an IFx below 10 % in EGG signals. Both were rated R0. The analysis of scatterplots revealed that it is possible to have a low IFx value and a high R score and vice versa.

Correlations between IFx and roughness were estimated at 0.53 and 0.39 in Sp and Lx signals, respectively. In both signals, IFx increases with increasing roughness but seems to have difficulty in discriminating between R0 and R1. In other voice-quality dimensions, the group means are less conclusive.

Irregularity was uncorrelated with age with a low r of 0.14 and 0.10 in acoustic and EGG signals, respectively. In IFx estimates obtained from acoustic signals, there was a minor but significant difference between sexes (t = 2.21, p = 0.03). Female speakers were found to have lower IFx estimates than male speakers. No such difference was found in EGG signals. When data was stratified by sex, the strength of linear relationship between IFx and roughness in acoustic signals was confirmed by significant correlations amounting to 0.68 and 0.47 in male and female speakers, respectively. The corresponding correlations measured 0.37 and 0.40 in EGG signals.

3.2.2.6 Jitter and shimmer in connected speech

The performance of jitter and shimmer in connected speech was completely different from their performance in sustained phonations. As expected, jitter and shimmer computed from connected speech were significantly higher than those from vowels because of higher variability in F0 due to articulation and intonation. Acoustic and EGG jitter estimates were found to be significantly different (t = –11.10, p < 0.01). Similarly, acoustic and EGG shimmer were significantly different (t = –13.66, p < 0.01).

There was no difference between sexes in acoustic jitter and EGG shimmer obtained from connected speech. EGG jitter and acoustic shimmer were significantly higher in male speakers (t = 2.97, p < 0.01; t = 3.64, p < 0.01).

Both acoustic and EGG shimmer as well as acoustic and EGG jitter increased with increasing roughness. Jitter and shimmer from acoustic signals were superior in their strength of association with perceived roughness and in the number of significant contrasts than EGG

jitter and EGG shimmer. In our data, acoustic shimmer had a higher correlation with perceived roughness than acoustic jitter.

3.2.2.7 Summary

Our results suggest that listeners might use different distributional characteristics of F0 as cues when rating roughness in speech samples. The following observations were made when comparing similar measures calculated from different material. Acoustic jitter and FMF in /a/ vowels performed slightly better than their conterparts calculated from connected speech. Acoustic shimmer performed significantly better in connected speech. Acoustic jitter and shimmer from connected speech were superior to their conterparts in /e/ vowels. EGG jitter and shimmer in both vowels were better than EGG jitter and shimmer in connected speech. Finally, all measures taken from acoustic signals in connected speech outperformed the same measures in EGG signals.

With two exceptions, different measures taken from the same signal were more correlated than identical measures from different signals (Table 21). IFx correlated moderately with its counterpart measured on EGG signals. PLF correlated strongly with EGG PLF.

Table 21: Correlation matrix presenting voice parameters measured in connected speech.

Variables	*SD*	*SD EGG*	*IFx*	*IFx EGG*	*R80*	*Ji*	*Ji EGG*	*Shi*	*Shi EGG*	*OQ*	*PLF*	*PLF EGG*
SD	1.0											
SD EGG	0.14	1.0										
IFx	0.52	0.29	1.0									
IFx EGG	0.21	0.46	0.53	1.0								
R80	0.59	0.19	0.66	0.36	1.0							
Ji	0.75	0.28	0.62	0.37	0.58	1.0						
Ji EGG	0.13	0.45	0.23	0.76	0.15	0.27	1.0					
Shi	0.69	0.45	0.72	0.49	0.60	0.78	0.36	1.0				
Shi EGG	0.06	0.42	0.04	0.52	0.08	0.10	0.69	0.15	1.0			
OQ	0.07	-0.27	-0.12	-0.53	-0.09	-0.10	-0.66	-0.14	-0.66	1.0		
PLF	0.35	0.29	0.18	-0.01	0.34	0.42	-0.06	0.31	0.03	0.07	1.0	
PLF EGG	0.26	0.49	0.18	0.07	0.22	0.36	-0.03	0.33	0.11	0.04	0.85	1.0

As expected, F0 measures from connected speech explained little breathiness ratings variance (Appendix E). Acoustic shimmer obtained from connected speech had the highest percentage of explained roughness ratings variance amounting to 41 %, whereas it explained only 12 % of breathiness ratings variance. Other acoustic measures like IFx, F0 SD, 80 % F0 range and jitter could also be considered stronger predictors of roughness rather than breathiness. Here again, acoustic measures outperformed measures calculated from EGG signals. Acoustic shimmer was also the best parameter explaining the hoarseness ratings variance.

There is only little evidence for interactive effect of perceived roughness and low pitch in connected speech. Subharmonics and creaky voice are related to rough voice quality as they increase the perception of a low-frequency component in the voice. However, PLF and EGG PLF explained little variance in roughness ratings.

3.2.3 Aerodynamic measures

Aerodynamic measures are important part of voice assessment procedure. Numerous studies offer strong evidence for deviant aerodynamic measures in voice disorders like vocal fold paralysis (Franco & Andrus, 2009), laryngopharyngeal reflux (Radish Kumar & Bhat, 2008), glottal carcinoma (Mitrovic, 2003) and others. Three aerodynamic measures were obtained from study subjects: maximum phonation time (MPT), vital capacity (VC) and phonation quotient (PQ).

The relation between the three measures is straightforward. MPT is presumed to provide indication of respiratory support during phonatory function and is defined as the maximum time needed to sustain a tone on one breath. In comparison to tidal volume which is defined as the quantity of air inhaled and exhaled during one cycle of respiration, vital capacity represents the volume of air that can be *maximally* exhaled after a *deep* inhalation. VC reflects the amount of air which can potentially be made available for phonation or speech. In the norm, speech at normal loudness level begins at ca. 60 % of VC (Denny, 2000) and ends at 40 % of VC (Hixon, 1973; Weismer, 1985). Whithin this range, the choice of the lung volume to start speaking with depends on the intended length and intensity of the utterance. Ca. 50 % of VC can be normally used for MPT. Speakers with voice disorders have even less VC available for MPT than normal subjects. Some of the VC is wasted due to incomplete glottal closure or short closed phase resulting in a reduced MPT. A small VC is responsible for voice problems of another kind: it signals a deficient muscular breathing mechanism and therefore inability of the speaker to raise the subglottic pressure needed to produce loud phonations[18]. Short phonations, however, can be produced with just a tidal volume amounting in quiet respiration to 500 ml (Schutte, 1992). The phonation quotient is determined by dividing the vital capacity by the maximum phonation time.

There have been several sources of normative data available for comparison and reference. Hirano (1981) considers MPT values in the range between 25 s and 30 s to be normal for men. Normal values for women lie in the range between 15 s and 25 s. Values

[18] Whereas respiratory disorders do not always result in voice disorders: a serious reduction in lung capacity has to take place in order to affect phonation, since normal conversational speech requires less than 35 % of vital capacity (Aronson, 1990), the reverse may be almost always the case: respiratory behavior is altered to compensate for disordered voice production.

below 10 s can be regarded as abnormal for both sexes (Dejonckere et al., 2001). Solomon et al. (2000) measured an average of 22 s disregarding sex. As no significant differences were found between the sexes, they suggested further that variability in MPT data in normal subjects goes back to changes in airflow, VC and alveolar pressure. However, their data revealed at the same time that the correlation between MPT and VC in normal subjects was not significant.

Hirano (1981) provided normative values for PQ: the normal PQ values are to be found between 120 and 190 ml/s. In voice disorders with short or incomplete glottal closure, PQ, given the formula, has to be strongly dependent on MPT. PQ values around 300 ml/s and above were reported by Hirano (1989) and Mitrovic (2003) in patients with unilateral carcinoma and vocal fold paralysis. Hirano (1968) and Tanaka et al. (1991) examined the relationship between MPT and PQ in normal and pathologic voices. They found high correlations between the two parameters. Similarly, in Hirano (1989) the correlations between MPT and PQ ranged between –0.63 and –0.95 in different pathology groups.

Normative values for VC have been extensively reported in the literature as well. The norm for vital capacity differs significantly between men (5000 ml) and women (3500 ml). The limits of normal generally vary with age. Vital capacity is known to reduce with age in both sexes. Hoit & Hixon, (1987), Hoit et al. (1989), Sperry & Klich (1992) reported age-related reduction in VC in nondysphonic men and women without respiratory problems.

Since aerodynamic measures reflect to a certain extent the degree of vocal fold closure, they are supposed to be primarily associated with perceived breathiness. Ptacek & Sander (1963) were one of the first to demonstrate the inverse relationship between MPT and breathiness. The following sections summarize the findings of the present study with regards to aerodynamic measures.

3.2.3.1 Maximum phonation time (MPT)

MPT values ranged from 3 s to 38 s. At the 95 % level of confidence, the confidence limits lay in the range from 15 s to 18 s in male and from 12 s to 15 s in female patients. The mean MPT amounted to 16.5 (6.9) s in male and 13.4 (5.6) s in female subjects. These values were significantly lower than the test values that were set at 25 s for male and 15 s for female subjects (one sample *t*-test, $t = -11.35$, $p < 0.01$; $t = -2.24$, $p = 0.014$), respectively. Sex difference in MPT was significant (one-way ANOVA with $F = 8.24$, $p < 0.01$). 28 study subjects (ca. 19 %) had an MPT value below 10 s, of which 15 were male subjects. A further 44 subjects (ca. 29 %) were not able to sustain a tone for 15 s or longer. The latter group consisted of 70 % of male subjects. Two outliers correspond to exceptionally long MPT values in subjects diagnosed with functional dysphonia.

Should sex difference be neglected, MPT is weakly associated with all three perceptual categories. The relationship is inversed in all three cases. The mean MPT decreased with increasing voice-quality ratings. Surprisingly, the correlation was strongest with hoarseness. Similarly, the results of Mann-Whitney U-test statistics indicate that MPT is a better predictor of hoarseness rather than breathiness. With only two significant contrasts, MPT data was unsuccessful in discriminating between grades B2 and B3.

When stratified by sex, correlations with roughness were not much different from those given in Appendix C. Correlations with breathiness and hoarseness were higher in male subjects with an r_s of –0.52 and –0.5, respectively. In female voices, MPT correlated more strongly with hoarseness (r_s = –0.47) than with breathiness (r_s = –0.34). In both sexes, MPT correlated weakly with age (r = –0.18).

3.2.3.2 Vital capacity (VC)

In the present study, male subjects had a mean of 3261 (904) ml, female subjects – 2501 (798) ml. The sex difference was significant (one-way ANOVA with F = 28.92, p < 0.01). In both sexes, the measured mean VC was lower than reported norms. The results of the one-sample t-test lead one to conclude that the mean VC values are significantly lower than the normative test values set at 5000 and 3500 ml (t = –17.6, t = –10.16, p < 0.01). This effect was assumed to be caused by the advanced age of many study subjects. In fact, we found relatively high correlations of the order of –0.62 and –0.51 between VC and age in male and female subjects, respectively.

Vital capacity measurements in male and female subjects were found to be weakly but significantly correlated with hoarseness. However, when sex difference is neglected, VC remained uncorrelated with perceptual voice quality. Judging by results of Mann-Whitney U-test statistics, VC should have a weak predictive power. VC correlates moderately with MPT (r = 0.47).

3.2.3.3 Phonation quotient (PQ)

In our data, due to reduced MPT estimates in dysphonic subjects, PQ values were abnormally high with many lesions on vocal folds like nodules, polyps, edema and carcinoma. The range of PQ values extends from 80 ml/s to 860 ml/s with a mean of 221 (109) ml/s. 57 study subjects (38 %) had a PQ value in the normal range between 120 ml/s and 190 ml/s. Low PQ values (< 120 ml/s) were found in 15 study subjects (10 %) with a high MPT. 23 study subjects (ca. 15 %) measured a PQ above 300 ml/s.

The mean PQ tended to increase with increasing voice-quality grade. There was no significant difference between sexes in phonation quotient data. Whereas MPT correlated significantly with all three perceptual categories, PQ correlated significantly albeit to a lesser extent with breathiness and hoarseness. In both voice dimensions, the strength of association was weaker with a lower r_s in comparison to MPT data. In predicting breathiness, PQ allows for 2 significant contrasts according to Mann-Whitney U-test statistics. However, the weakest discriminative power was found to hold between clear (B0) and slightly breathy (B1) voices.

Since PQ is calculated using MPT and VC, the three measures are expected to be correlated. The Pearson's *r* between MPT and PQ in our data was estimated at –0.6. The correlation coefficient was slightly higher in male (r = –0.67) than in female (r = –0.6) subjects. Contrary to expectations, the strength of association between PQ and VC was low with an *r* of 0.28.

3.2.4 "Breathiness measures" from connected speech

Respiration is a critical element of speech production[19]. In speech, expiratory air needs to be sustained through activation of the abdominal and thoracic muscles. Expiratory muscles are responsible for breath support by providing adequate airflow for appropriate intensity level during speech. In the norm, speakers are able to adjust expiration to the needs of speech production, t.i. to regulate subglottal pressure and airflow for voicing and articulation.

Deficiency in maintaining subglottal air pressure in speech may manifest itself as release of air prior to phonation onset (breathy voice onset), increased respiratory effort, decreased intensity, reduced ability to coordinate voice on- and offsets and frequent inhalations. To maintain a constant pressure, dysphonic speakers have to compensate for lowered resistance with a greater rate of airflow. Hoit & Lohmeier (2000) hypothesized that breathy voice quality is also linked to excessive ventilation, which is defined as the amount of gas coming in and out of the lungs during speech. They pointed to the need to investigate the relation between breathing and breathiness.

In the previous chapters we argued that breathy voice quality has certain spectral characteristics that can be more or less successfully captured in acoustic voice measures taken from vowels. However, in forming a general impression of breathiness in speech, listeners may be guided by factors like dysfluency and strenuous speech breathing besides spectral characteristics of speech.

[19] Compared to quiet breathing, normal speech breathing is characterized by shorter inspirations, longer expirations, the activity of exhalation muscles to maintain subglottal pressure at a relatively stable levels, larger lung volume excursions and higher expiratory pressure (Hixon & Weismer, 1995; Sapienza et al., 1997).

For the purpose of the present work, it was essential to find appropriate measures that indirectly describe problems with respiration during speech. To identify the possible link between perceptual breathiness and speech breathing, we used parameters that do not require special equipment and can be reliably calculated from spectrographic traces. Beside intensity and OQ measures, several timing measures were examined as predictors of breathiness ratings. We anticipated that speech of individuals with voice disorders will be also impaired with regards to temporal speech characteristics.

It is conceivable that in dysphonic subjects pausing behavior might strongly correlate with inhalation needs. Pausing for breathing needs is subject to individual physiological constraints. Individuals with voice disorders, weak respiratory capacity, low muscular tone and slow articulation rate may need a greater number of pauses and/or longer pauses compared to normal subjects. If one further assumes that inspiratory volume correlates positively with inspiratory duration, which in turn depends on the length and intensity of the following utterance (Denny, 2000), then the pause length should normally reflect the amount of the inhaled air. Keeping this in mind, we hypothesized that pausing behavior in dysphonic subjects would contribute to predicting breathiness ratings.

3.2.4.1 General observations

Analysis of speech characteristics related to breathing in patients with voice disorders revealed following information of significance. Many patients experienced shortness of breath and had to replenish their air supply more frequently which was characterized by an increased number of pauses filled with breathing noise.

Analysis of pause placement showed that in our data pauses mainly occurred at prosodic-syntactic boundaries[20]. As expected, in less dysphonic speakers the difference in boundary prominence was reflected in the pause length: the more important the boundary, the longer the pause (cf. Strangert, 1993; Fant, 2003).

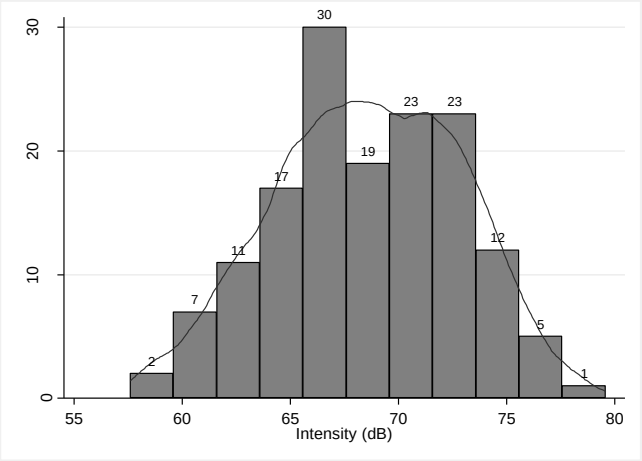
Other abnormalities were observed in connection with decrease in intensity at the end of phrases, frequent changes to whispered speech quality and excessive air release immediately after inhalation and during production of plosives and fricative sounds that often resulted in a sharp increase in sound pressure with signal clipping. Inappropriate pause placement was rarely detected. In our data, we often observed unphonated expirations before speech initiation in dysphonic subjects.

[20] Acoustic signalling of prosodic boundaries is known to be achieved by a combined effect of several variables. Pauses are the most salient indicator of prosodic boundaries. Apart from pausing, slowering speaking rate, final lenghthening of vowels, intonation contours, F0 resets and the prepausal use of creaky voice signal a boundary.

3.2.4.2 Intensity

Inappropriate loudness is one of the most common symptoms in voice impairment. Our data suggested no statistically significant difference in average speaking intensity between sexes (*t*-test, $t = 1.77$, $p = 0.07$). In pathological voices, it seems to be equally affected in both sexes. We measured a lower mean of 68.9 (4.2) dB in male and 68.1 (4.2) dB in female voices compared to figures reported in Orlikoff & Kahane (1991). The lower and upper limits of the speaking intensity range was slightly lower in female voices varying between 57.6 dB and 76.7 dB. In male voices, the speaking intensity ranged between 60.0 dB and 78.1 dB. Distribution of the mean speaking intensity data is shown in Fig. 60.

Fig. 60: Histogram showing the distribution of the mean intensity in dB during reading (bin width = 2 dB). Bar labels give the number of subjects in the bin. The solid black line represents kernel density plot.



Statistical analysis revealed that intensity is uncorrelated with perceptual voice quality. Contrary to expectations, the correlation between the degree of perceived breathiness and speaking intensity was very weak and insignificant ($r_s = 0.12$, $p = 0.14$). The Mann-Whitney U-test statistics confirm that intensity is a poor predictor of perceived voice quality. Speaking intensity data can probably be used to discriminate between B1 and B2.

On comparing speaking intensity data to intensity data in vowels, we found that vowels were significantly louder than speech ($t = 6.91$, $p < 0.01$; $t = 7.28$, $p < 0.01$). This effect could be based on test vowels being shorter and requiring less articulation effort than sentences.

Speaking intensity was uncorrelated with MPT ($r_s = –0.02$). Likewise, the strength of association between intensity in /a/ and /e/ and MPT was low with an r_s of 0.15 and 0.13, respectively. This finding contradicts the results by Max et al. (1996) who found a high degree of association between MPT and intensity of phonation.

3.2.4.3 Open quotient (OQ)

Many study subjects had very high OQ values in speech. This finding can be explained by a high number of patients with no significant vocal fold closure during voiced segments in speech or calculation errors. As shown in the introduction, DEGG method has problems with calculating OQ from irregular signals.

OQ in connected speech was significantly larger than OQ in /a/ and /e/ vowels (t = 13.1, $p < 0.01$; t = 12.7, $p < 0.01$). Phonatory and articulatory changes seem to bring about an increase in measures of OQ. However, we cannot rule out the possibility that this effect may also be attributable to technical limitations of the EGG method giving unreliable measures when the position of the electrodes is changed in relation to the moving larynx.

Whereas sex effect in OQ did not reach significance in sustained vowels, sex difference in OQ data obtained from connected speech was significant with women having greater OQ values (t = 3.64, $p < 0.01$) than men. This finding agrees well with the fact that women have a more open glottal configuration and larger vertical excursions of the larynx during speech.

There was a tendency for the mean OQ to increase with increasing voice-quality ratings. The correlation was highest with breathiness. Note that OQ from vowels did not correlate significantly with voice-quality ratings. Similarly, the number of significant contrasts was highest across breathiness ratings, the weakest discriminative power should be expected between B0 and B1.

Measures of OQ in both vowels and speech were not correlated with aerodynamic measures MPT, VC and PQ or intensity.

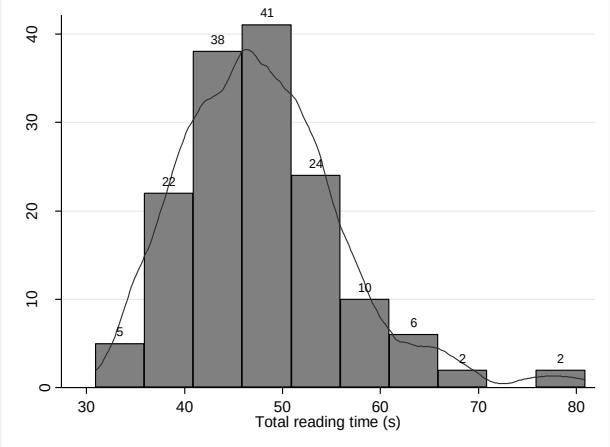
3.2.4.4 Temporal speech characteristics

3.2.4.4.1 Total reading time (Rtime)

It has been observed that B2 and B3 voices needed more time to read the text passage. However, the relation between total reading time and rated breathiness was characterized by a low correlation coefficient.

The difference between sexes was significant (t = 1.8, p = 0.04). Men were slower readers than women. Age-related differences might be partially responsible for the lowered reading rate in dysphonic speakers. Sperry & Klich (1992) found that older subjects needed more time to read the same text passage than younger ones and their reading rates were slower than in young subjects. In accord with findings by Sperry & Klich (1992), we found that Rtime was weakly correlated with age (r_s = 0.36).

Fig. 61: Histogram showing the distribution of the total reading time in seconds over 150 subjects (bin width = 5 s). Bar labels give the percentage of subjects in the bin.



Although the highest correlation was with hoarseness, it was in predicting breathiness that Rtime achieved two significant contrasts. Rtime was weakest to discriminate between B2 and B3.

3.2.4.4.2 Total pausing time (Ptime)

The reading paragraph consisted of 6 sentences ranging in length from 9 to 42 syllables. 6 periods and 11 commas in the standard text were supposed to correspond to 6 major and 11 minor pauses[21]. We did not attempt to analyse and group pauses into classes by position. It is reasonable to assume that the number of pauses produced by normal subjects would lie in the range between 6 and 17, since read speech is less conducive to pauses due to hesitation and cognitive processes involved in spontaneous speech production.

In our data, Ptime ranged from 3.36 s to 34.17 s with a mean of 9.46 (4.42) s. Ptime was also found to moderately correlate with breathiness and hoarseness. The output from the Mann-Whitney U-test provides three significant contrasts in predicting breathiness ratings.

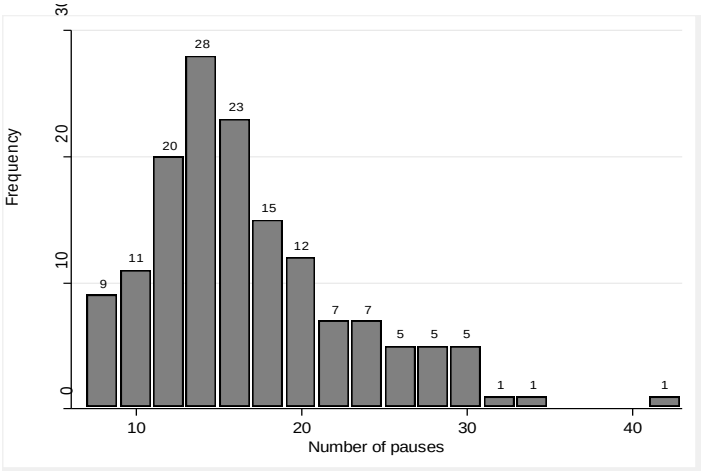
Ptime was weakly correlated with age (r = 0.19). Sex effect was significant with t = 3.06 and $p < 0.01$, although the magnitude of the effect is modest. Male subjects spent more time on pauses than female subjects (10.4 (4.8) s vs. 8.2 (3.5) s). The fact that men have larger lung volumes and need a greater amount of gas for ventilation supplies a reasonable explanation for this effect. This difference was also found to be reflected in the number of pauses and mean pause duration.

[21] The location of pauses for inspiration in read aloud tasks is stable over time and coinsides with major sentence and phrase boundaries (Henderson et al., 1965; Grosjean & Collins, 1979; Winkworth et al., 1994). Syntactic units of less than 16 syllables are normally read without pausing (Fant, 2003).

3.2.4.4.3 Number of pauses (Npauses)

Fig. 62 shows the distribution of pauses across study subjects in the standard text. The number of pauses ranged from 7 to 43. 50 study subjects (m = 34, f = 16) used more than 17 pauses. Male subjects used on average 17.6 (6.6) pauses. Female subjects needed 15.1 (5.2) pauses. Difference between sexes was not great but significant (one-way ANOVA, F = 6.09, p = 0.015).

Fig. 62: Histogram showing the distribution of the number of pauses longer than 200 ms during reading (bin width = 2). Bar labels refer to the number of subjects in the bin.



A significant increase in the number of pauses was observed with increasing breathiness (r_s = 0.44). Those subjects with clear voices made fewer pauses during reading. In B3 voices, the average number of pauses increased to 24.1. Pairwise means comparison indicated significant difference between the adjacent B ratings. However, there is a fair amount of overlap between B0 and B1.

To compare our data with previous research on respiration and counts of breaths per minute, the count of the number of pauses in the text was converted to the count of the number of pauses per minute. Voices rated B0 and B1 were found to make on average 18.7 (4.8) pauses per minute. In B2 and B3 voices, the mean number of pauses per minute amounted to 21.4 (4.1) and 27.6 (5.6) pauses per minute, respectively.

3.2.4.4.4 Number of pauses per 100 syllables (P/100 syl)

This measure is related to the count of the number of syllables. The standard passage consists of 182 syllables. However, when subjects omit or add words or for comparison with data in languages other than German, it is more convenient to calculate the number of pauses per 100 syllables. Here again, we detected a minor sex effect in the data (t = 3.02, p < 0.01). Men made on average 9.7 (3.7) pauses per 100 syllables. Women made fewer pauses than men with a mean of 8.4 (3.3) pauses per 100 syllables.

P/100 syl correlated strongest with breathiness ratings. The Mann-Whitney U-test statistics yielded 2 significant contrasts across voice-quality ratings for each perceived voice category.

3.2.4.4.5 Pauses in percent of the total reading time (P(%))

In the present corpus, pauses occupy on average 19.1 (6.1) % of the overall reading time. Individual measurements range between 9.6 % and 44.1 %. Difference between sexes was significant (t = 4.93, p < 0.01). Men spent more time pausing than women. The means were estimated at 20.6 (6.4) % and 17.2 (5.1) %, respectively.

P(%) correlated positively with all three peceptual categories, especially with breathiness. It differed significantly across all contiguous breathiness grades.

3.2.4.4.6 Mean pause length (Plength)

A significant difference between sexes was found in the mean pause length (one-way ANOVA, F = 8.08, p = 0.005). Male subjects had a mean pause duration of 589 (104) ms. Female subjects had a mean pause duration of 540 (104) ms.

Plength was weakly correlated with perceived voice quality. It did not differ significantly across breathiness ratings. However, it seems to differ significantly across H0, H1 and H2. It appears that subjects with breathy voice quality compensate for air leakage not by longer pauses but by more frequent pausing.

3.2.4.4.7 Speech/pause ratio (S/P)

The speech/pause ratio was calculated as time spent in speaking (total reading time minus total pausing time) set in relation to the total pausing time. Here again, we found that male and females speakers differed significantly in S/P (t = –5.1, p < 0.01), with females having greater values than males.

S/P correlated negatively with all three perceptual categories and had the highest correlation (with the highest number of significant contrasts) with breathiness ratings.

3.2.4.4.8 Mean number of syllables between two pauses (Sylbp)

This measure is an approximation to the number of syllables per breath group used in studies on speech breathing. We could not determine the number of syllables per breath group as breath intakes were not registered.

In our data, a minor sex effect was significant (t = –2.92, p < 0.01). Male subjects produced fewer syllables between two pauses than women (11.8 (4.5) vs. 13.3 (4.4). In agreement with Hoit & Hixon (1987) and Hoit et al. (1989), we found that the mean number

of syllables between two pauses was uncorrelated with age.

The mean Sylbp produced an equal number of significant contrasts in each voice category. As expected, Sylbp decreased with increasing breathiness. We measured a modest correlation of 0.30 between breathiness and Sylbp. In our data, a reduced MPT was not necessarily related to a decrease in the number of syllables spoken between two pauses. Dysphonic patients in the present study produced between 4 and 26 syllables between two pauses with a mean of 12.5.

3.2.4.4.9 Speech and articulation rate

In the present study, individual values for speech and articulation rate at normal reading speed ranged from 2.2 syl/s to 6.2 syl/s and from 3.0 syl/s to 7.2 syl/s, respectively. The mean speech rate was estimated at 3.9 (0.7) syl/s; the mean articulation rate measured 4.8 (0.6) syl/s. Since the number of syllables in the test passage was known, there was no need to determine the number of syllables automatically by means of a syllable count algorithm.

We found a moderate correlation (r = –0.42) between speaking rate and age. Older people tended to have a slower speaking rate. Speaking rate was found to differ across sex with male speakers being slower speakers than women (t = –2.67, p < 0.01). The mean speech rate in male and female speakers equaled 3.8 (0.6) syl/s and 4.0 (0.7) syl/s, respectively. Articulation rate was not sex-specific.

In both speech and articulation rate, correlation coefficients were highest with hoarseness. Speech rate correlated modestly with perceived breathiness. It differed significantly across all four hoarseness ratings in dysphonic population. Voices rated B2 and B3 did not differ significantly in the speech rate.

Articulation rate was weakly correlated with roughness and hoarseness. Interestingly, although correlation with breathiness ratings was not significant, the output of the Mann-Whitney U-test suggested that articulation rate differs significantly across all breathiness grades.

3.2.4.5 Summary

The ”breathiness measures” examined in this section correlated moderately with breathiness ratings. Correlations with roughness were found to be lower than correlations with breathiness and hoarseness ratings or insignificant. It is interesting to note that correlations between aerodynamic and prosodic measures did not exceed the value of 39 % (Table 22). Correlations between different prosodic measures ranged from 12 % to 94 %.

Table 22: Correlation matrix presenting Pearson's *r* between prosodic and aerodynamic measures.

Variables	*MPT*	*PQ*	*Int*	*OQ*	*Rtime*	*Ptime*	*P/100 syl*	*P(%)*	*Plength*	*S/P*	*Sylbp*	*Srate*	*Arate*
MPT	1.00												
PQ	-0.60	1.00											
Int	-0.02	0.10	1.00										
OQ	0.06	-0.02	-0.01	1.00									
Rtime	-0.28	0.14	0.10	-0.01	1.00								
Ptime	-0.35	0.33	0.11	-0.06	0.81	1.00							
P/100 syl	-0.36	0.33	0.10	-0.05	0.70	0.90	1.00						
P(%)	-0.33	0.39	0.11	-0.10	0.62	0.94	0.89	1.00					
Plength	-0.12	0.14	0.03	-0.02	0.51	0.52	0.18	0.47	1.00				
S/P	0.27	-0.30	-0.10	0.08	-0.60	-0.84	-0.81	-0.92	-0.45	1.00			
Sylbp	0.29	-0.24	-0.09	0.07	-0.66	-0.78	-0.88	-0.81	-0.12	0.87	1.00		
Srate	0.28	-0.15	-0.08	0.03	-0.94	-0.78	-0.71	-0.63	-0.54	0.64	0.70	1.00	
Arate	0.18	0.02	-0.05	-0.02	-0.85	-0.47	-0.42	-0.26	-0.42	0.29	0.43	0.90	1.00

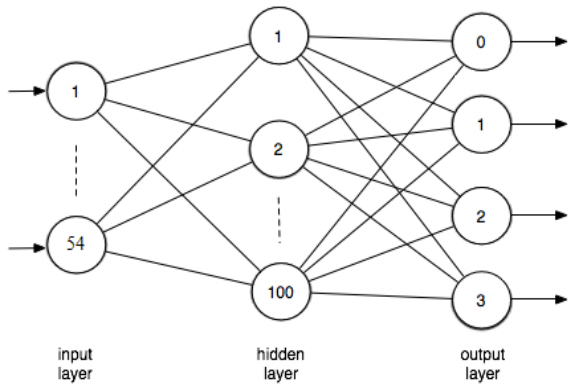
For each of the predictor variables, the percentage of the explained variance is shown in Appendix E. Results are given for B and H. Roughness was excluded since no parameter explained more than 10 % of roughness ratings variance. The single predictor analysis showed that no parameter can be considered an outstanding predictor of breathiness or hoarseness, either. Only two parameters explain more than 20 % of the breathiness ratings variance. S/P ratio and P(%) yielded the largest amount of breathiness variance explained. OQ and intensity did not contribute to the prediction of perceived breathiness. Against expectations, measures dealt with in this section could be less successful predictors of breathiness than measures taken on vowel data. MPT and Srate were the best predictors of hoarseness.

3.3 Classification results

The last section of Chapter 3 covers classification results. Although single model regression was used to determine the amount of explained variance, the reason we did not use regression analysis to predict voice-quality ratings was that most of the variables were not normally distributed. We utilised quadratic discrimination method and ANN which did not require either normal distribution or a strong linear relationship between independent variables and perceived voice quality.

QDA and ANN performed equally well on the present set of pathologic voices. Fig. 63 shows the principle of a feedforward network. 54 variables were used to predict voice quality ratings. Table 23 summarizes the results using QDA and ANN, respectively.

Fig. 63: An example of a 2-layered network with from left to right 54 input units, a hidden layer with 100 units and an output layer with 4 units corresponding to the four voice quality ratings.



Classification results by means of QDA yielded the highest percentage of correct classifications in predicting breathiness. Here, the overall accuracy of classification was found to be approximately 80 %. Roughness and hoarseness were predicted correctly in 74 % and 71 % of all cases, respectively. In all classes with exception of H0, the rate of correct classifications exceeded the chance value of 0.25.

Table 23: Classification results by QDA (left) and ANN (right), Leave-One-Out.

	R	*B*	*H*
0	0.72	0.77	0.00
1	0.86	0.89	0.81
2	0.59	0.73	0.72
3	0.72	0.64	0.68
	0.74	0.80	0.71

	R	*B*	*H*
0	0.78	0.67	0.92
1	0.77	0.73	0.81
2	0.76	0.79	0.67
3	0.78	0.86	0.79
	0.77	0.75	0.75

15 variables were selected to predict perceived roughness, of which 11 predictors were calculated from acoustic signals: jitter, HNR, IC and AI measured in /a/ vowels, SHR and FMF measured in /e/ vowels; measures taken from connected speech included jitter, IFx, 80 % F0 range, F0 SD (st) and PLF. Relevant measures from EGG signals were IFx, shimmer, OQ and F0 SD (st). The worst hit rate estimated was R2 with 59 %. The analysis of errors revealed that in 8.3 % of all cases roughness was overestimated by 1 point, in another 16 % roughness was underestimated by 1 point. The remaining 1.7 % of data sets were misclassified by two or more points. Three variables stand out from the list: OQ, EGG shimmer and HNR. Especially, OQ and EGG shimmer which reflect the amount of vocal fold contact, are normally associated with perceived breathiness.

15 variables contributed to prediction of perceived breathiness. These were shimmer, GNE, intensity, HNR and LTAS measured in /e/ vowels, aerodynamic measures MPT and PQ and measures taken from connected speech including OQ, pause length, S/P ratio, percentage of pauses, frequency of pauses per 100 syllables, speaking rate, speaking intensity and

shimmer. Note that all selected variables were expected to be predictors of breathiness and are in agreement with vocal fold physiology. The lowest classification rate was estimated in B3 with 64 %. Breathiness was overestimated by one point in 8 % of data sets. In another 12 % breathiness was underestimated by 1 point. Misclassification by two or more points occurred in less than 1 % of data sets. Note that only one predictor variable was calculated from EGG signals. Interestingly, for reasons that are not entirely clear, measures obtained from sustained vowels included only measurements on /e/ vowels. In three variables including acoustic shimmer, GNE and HNR, measurements made on /e/ segments were proved to be less pathologic than those taken from /a/ segments.

QDA yielded a discriminant function that included 11 variables for predicting hoarseness. All 11 variables were computed from acoustic signals. The set of predictor variables was comprised of acoustic shimmer and GNE taken from /e/ segments; acoustic jitter, intensity, SHR and HNR measurements from /a/ segments; jitter, IFx, F0 SD (st), frequency of pauses per 100 syllables (all measured on acoustic traces of connected speech) and, finally, MPT. H0 was the only category represented with 12 data sets that did not contribute to the overall success rate. This low level of contribution was attributed to a small size of the H0 class. Hoarseness was misclassified by one point in 26 % of data sets. The remaining 2 % of data sets were misclassified by two points.

Three types of predictor variables were covered in predicting roughness: perturbation and noise measures, as well as measures based on instantaneous frequency values. Predicting breathiness and hoarseness involved additionally aerodynamic and prosodic variables. In case of breathiness, the inclusion of measures based on instantaneous frequency values and prosodic measures resulted in a 20 % increase of predicting accuracy. This fact clearly demonstrates that diverse measures characterize the same perceptual phenomenon in different ways.

In QDA, when identical measures could be calculated from different vowels, only one sustained vowel, either /a/ or /e/, was selected to calculate the discriminant function. Apparently, same measures made on different vowels were too similar and therefore redundant.

We have demonstrated that QDA yields acceptable results with a few predictors. In ANN, the purging process did not reduce the number of predictors significantly. On average, 44 variables were used to obtain the reported numbers of correct classifications. In this respect, ANN gave results comparable to QDA (Table 23); however, they were achieved at the expense of a higher computational complexity. A large number of predictor variables makes it difficult to determine which variables are more important than others. Whereas the ANN method failed to relate every single predictor variable to just one perceptual category,

either breathiness or roughness, in QDA, the list of variables selected to predict breathiness and roughness was almost mutually exclusive and overlapped only in one variable: OQ measured from connected speech.

The greatest mean rate of correct classifications by ANN was found in rough voice quality. Here, the overall accuracy of classification was found to be approximately 77 %. In predicting breathiness and hoarseness, the number of correct classifications amounted to 75 %. The lowest predicted success rate was 67 % in B1 and H2. The classification rates were more or less evenly distributed over separate voice-quality classes. In this respect, classification results improved significantly as compared to similar studies by Schönweiler et al. (2001) and Linder et al. (2008).

Had we treated a combination of H0 and H1 voices as ”healthy” and a combination of H2 and H3 voices as ”pathologic”, the rate of correct classifications using the ANN method would have amounted to 85 % and 84 %, respectively. There is a slight improvement in the overall success rate as compared to Linder et al. (2008), though compared to the present research they successfully solved the two-class classification problem with only four variables.

The classification accuracy for 2-class disrimination R0/R1 vs. R2/R3 resulted in a mean classification rate of 87 %. Thereby, all voices judged R0 were classified correctly. Similar classification rates for breathiness were estimated at 88 %. Here again, the rate of misclassification in voices judged B0 was zero.

The analysis of variables that were rejected as redundant can be summarized as follows: Roughness was predicted with the smallest number of variables. Many of the variables that are traditionally considered to be predictors of roughness were rejected. The list of rejected variables included acoustic and EGG jitter in both vowels, acoustic shimmer and frequency modulation factor in both vowels, standard deviation of the fundamental frequency derived from connected speech in both signals, LLE and AI measured in /a/ and /e/ vowels, respectively. The remaining variables still contained several strong variables that were used to predict roughness with acceptable accuracy. In predicting breathiness, perturbation measures taken from /a/ vowels including jitter, shimmer and frequency modulation factor were less efficient than the same measures from /e/ vowels. Further, measures of F0 SD (st) from connected speech and such measures like LLE and SHR were useless. Not a single measure that was expected to be associated with breathiness was unselected. The largest number of variables were used to predict hoarseness. Here, no general tendency can be identified regarding the nature of rejected variables. LLE and F0 SD from connected speech derived from acoustic signals were redundant in predicting all three perceptual categories.

Our data suggests that in both QDA and ANN classification methods neither intermediate-grade dysphonia nor extreme dysphonia grades were systematically

misclassified. As rater variability was greater when judging intermediate-grade dysphonia, it was expected that the error rate would be greater in intermediate grades. Similarly, we found no evidence supporting the claim that the rate of correct classification is proportional to the number of voices in each class (Table 4) since underrepresented classes 0 and 3 did not yield poorer classification rates than intermediate-grade dysphonia.

Chapter 4: Discussion

We investigated the ability of diverse objective measures to discriminate between four classes of perceptual ratings applied to the three most pertinent perceptual voice-quality dimensions including roughness, breathiness and hoarseness. The focus of the study was laid on voice measures that can be obtained automatically without high personal and equipment cost. Four groups of measures were examined: 1) measures from sustained vowels, 2) aerodynamic measures, 3) measures based on voiced segments in connected speech and 4) prosodic measures.

4.1 Classification results

The major finding of this study was that a combination of examined measures could correctly predict perceptual voice quality in more than 70 % of cases. The results were presented as a four-class classification problem involving not only hoarseness but also breathiness and roughness. The key factor to enhance classification rates appears to be combining, on the one hand, sufficiently diverse and, on the other hand, similar voice measures.

Our results implicate that despite high correlations between individual variables of the same type, correlated variables are not truly redundant. Evidently, noise reduction and better group separation may be obtained by adding variables that are presumably redundant. Similarly, individual variables with allegedly poor predictive power unexpectedly proved to be useful in classifying voices as long as they provide unique information. It might be the case that a variable which is useless by itself, e.g., SHR, can bring about performance improvement when taken with other variables and that variables that are independently and identically distributed do not contain identical information on voices. In contrast, variables regarded as strong predictors of perceived voice quality may remain unselected, e.g., the IC measure in /a/ segments alone accounted for 37 % of the explained variance in perceived hoarseness. However, it did not contribute to perceived hoarseness in QDA when combined with other variables as a predictor.

Unlike QDA, the ANN method does little to provide objective insight into cause-and-effect relationship since it is assumed to define the relationship between perceptual dimensions and instrumental measures in an intuitive and unexplainable way. It does justice to the fact that perceptual scales might integrate a combination of different effects, whereas objective measures capture only one single aspect of voice quality. In this respect, ANN operate in a manner close to human perception that may use a greater variety of cues to evaluate vocal quality, many of which are still not identified. This is a clear disadvantage of

the ANN method since relating voice measures to vocal fold physiology is necessary to ensure that measured variables are not incidentally related to perceived voice quality.

Two non-invasive techniques acoustic analysis and electroglottography were examined for their usefulness in automatic voice-quality classification. These techniques have an advantage of gathering data from larger populations by persons who are not certified to perform invasive procedures. The usefulness of acoustic signal for parametrization of voice signals has never been questioned. As for EGG signal, it has been shown that EGG signal contributed several measures that helped to improve classification accuracy.

Since severly disordered and normal voices could have been underrepresented, there is a possibility of bias limiting the scope of conclusions. We suggest that a larger population must be studied before applying the classification procedure for screening purposes.

Similarly, the present work suggests that sex differences can be better dealt with in larger studies. We cannot discard the possibility that male and female voices may be perceived differently. In our classification experiments, sex differences in variables were neglected. This was justified since even when sex effect was apparent in the examined measures, the magnitude of the sex effect was not great.

In accord with previous research, our results do not indicate that there is a clear-cut distinction between roughness and breathiness. Vocal pathology is highly complex and voice patients exhibit frequently both breathiness and roughness in their voices. As has been shown in the results section, voice measures do not relate to just one single perceptual category, either, suggesting that different voice-quality dimensions may have similar acoustic properties. According to our statistics, the degree of one voice quality dimension might influence the perception of the other ones. Though, our results indicate that by using statistical methods it is possible to find two non-overlapping sets of variables to predict breathiness and roughness. Since our results leave some space for improvement, we conclude that classification results might suffer from less than optimal validity and reliability of both voice measures and perceptual ratings.

4.2 Subjective voice-quality evaluation

Perceptual evaluation is known to be one of the most controversial subjects in assessment of vocal quality. It has been heavily criticized for poor reliability and agreement across listeners. Kreiman et al. (1990) claimed that even experienced clinicians differ substantially in perceptual behavior and apply different criteria in judging voice quality. Naive listeners, in contrast, used similar perceptual strategies. This is not a surprising finding since voice-quality dimensions are fuzzy semantic concepts (Isshiki et al., 1969). There is little consensus on the terms used for describing voice quality common to all voice specialists even though the

perception of roughness, breathiness and hoarseness was found to be essentially the same across different cultures (Yamaguchi et al., 2003). In this respect, high level of disagreement between raters reflects poor operational definitions of perceptual categories describing voice quality or insufficient training in distinguishing between the relevant perceptual categories.

It has been shown that although reliability and agreement between the raters in our study were not optimal, our results were compatible with other studies on this subject or better (Dejonckere et al. (1993); Rabinov et al. (1995); de Bodt et a. (1997); Kreiman & Gerrat (1998); Revis et al. (1999); Martens et al. (2007)). In the present study, average reliability was higher than average reliability in de Bodt et al. (1997) and Revis et al. (1999). de Bodt et al. (1997) measured kappas of 0.35 for R, 0.38 for B and 0.6 for G. Revis et al. (1999) obtained an average kappa of 0.47, 0.46 and 0.55 for R, B and H, respectively. The agreement between the pairs of raters was somewhat lower than in Revis et al. (1999). However, the overall interrater agreement in the present study was better than in Martens et al. (2007) and de Bodt et al. (1997). Martens et al. (2007) reported a Fleiss' kappa of 0.13 for B, 0.22 for R and 0.31 for G. In their data, there was a substantial increase in interrater agreement after repeated procedure with spectrographic analysis. Even then, the interrater agreement did not exceed a value of 0.4. Data obtained in de Bodt et al. (1997) yielded a Fleiss' kappa of 0.44 for G, 0.17 for R and 0.21 for B.

Further, in our data hoarseness seems to be more determined by roughness than by breathiness. In this respect, our findings do not agree well with Michaelis (2000) and Dejonckere et al. (1993). In Michaelis (2000), the correlation between B and R in expert raters amounted to 0.48. Correlations between other dimensions measured 0.69 for R and H, and 0.82 for B and H. Dejonckere et al. (1993) arrived at a conclusion that severity grade G was mainly determined by B, the strength of association between G and B being estimated at 0.88, whereas correlation between R and G equaled 0.63. Surprisingly, the correlation between R and B in their data was with –0.62 strong and negative, suggesting that R and B were almost mutually exclusive perceptual dimensions.

Spectrographic screening of vowels and speech samples revealed a higher proportion of signals with subharmonics than reported elsewhere. In fact, information on the incidence of subharmonics in sustained phonations of pathologic population is sparse and inconsistent. According to (Núñez Batalla et al., 2000), subharmonics were estimated in 31 % (36 from 115) of examined pathologic voices. The incidence of subharmonics in pathologic population was reported to be significantly lower in Omori et al. (1997). Only 20 patients in 389 (5.1 %) were found to have subharmonics. In Behrman et al. (1998), 9.4 % of subjects (19/202) had subharmonics. This discrepancy may be partly attributable to how subharmonic frequencies are defined and to the fact that many studies exclude poor signals from statistics.

The discrepancies between microphone and electroglottographic signals clearly demonstrate that the microphone and electroglottographic signals represent different physiologic phenomena. Considering the discrepancy between microphone and electroglottographic data, the legitimate question would be where besides the vocal folds subharmonics may come from. Several anatomical structures may be considered as additional vibratory souces to assist in pathologic vowel production. The use of ventricular and aryepiglottic folds is frequently observed in dysphonic subjects. In voice healthy population, subharmonics were reported in connection with different singing styles. Thus, aryepiglottic folds were found to vibrate in growl voice in jazz singing whereas ventricular folds were involved in throat singing (Sakakibara et al., 2007). Among possible anatomical structures besides vocal folds that can contribute to the emergence of subharmonics in the spectrum, probably the involvement of the ventricular folds can be captured in the EGG signal. Using the electroglottographic method, several authors (Saito et al., 2006; Lecluse et al., 1981; Brasnu et al., 1989) succeded in registering vibrations of the pharyngeal wall in laryngectomees and confirmed that EGG waveforms show a relatively arythmic pattern with unstable baseline in phonation of subjects using esophageal voice. The maximum EGG responce was found when the electrodes were placed about 2 cm above the stoma. This effect was ascribed to the activity of the thyropharyngeal muscles that form a new vibratory source by setting the paryngeal wall in motion (Minifie et al., 1968; Ewanowski et al., 1968)[22]. Vibrations of the uvula is another possible source of subharmonics during sustained phonations. It is certain that they cannot be reflected in EGG signals. Since the above-mentioned anatomical structures are thicker and stiffer than the vocal folds, there is a possibility that they seldom generate true subharmonics of the fundamental. They vibrate at their own resonant frequencies producing interharmonics and cause the sensation of roughness and biphonation.

Signal typing revealed that pathologic vowels are unstable in quality. Our results, even though restricted to a relatively small set of pathologic voices, showed that there is a connection between voice measures and spectrographic vowel type. As a rule, correlations with spectrographic vowel type were higher than correlations with perceived voice quality. Thus, spectrographic analysis of the signal prior to acoustic parameter extraction allows to identify unusual spectral characteristics and safeguard against erroneous analyses.

[22] The pharyngeal cross-section area can be changed by a factor of 20 from 0.3 to 6.8cm^2 (Lindqvist-Gauffin & Sundberg, 1971).

4.3 Objective voice analysis

It is obvious that neither subjective evaluation nor instrumental analysis of vowel and speech fragments inform the examiner on the vocal fold properties of the patient since divergent laryngeal conditions may result in similar sounding dysphonias and similar voice parameter measurements. However, the need for objective parameters to quantify voice function seems to be recognized by most experienced voice specialists. The main application of these parameters lies in screening for voice disorders and evaluation of therapy. Compared to subjective voice assessment, objective measures are well-defined since they are based on mathematical formulas and bring out quantitative voice characteristics that cannot be captured by human ear.

Yet, there are several shortcomings which make the use of objective parameters in clinical setting problematic. It is noteworthy, that voice measures based on F0 extraction contained measurement errors as a major component. Our results have shown that many instrumental measures are unreliable in moderately and severely dysphonic subjects and that estimates greately vary even within a single vowel. Some objective measures were not that robust to be applied to connected speech. Further disadvantage of objective measures to consider is that they did not always reflect subjective impressions of the voice. These findings agree well with Rabinov et al. (1995) who also claimed that acoustic analysis is presumably superior to subjective voice evaluation only in discriminating among normal and mildly disturbed voices. The same source has demonstrated that reliability in subjective judgements increased with increasing severity of pathology whereas reliability of objective measurements systematically decreased with increasing severity of pathology and when compared across several automatic systems.

Measurements across different studies are not directly comparable as same parameters may be derived from similar but different mathematical formulas. Automatic systems differ in their F0 output even in normal subjects (Morris & Brown, 1996) as they use different F0-extraction strategies. Aperiodicity and noise in signals from dysphonic patients may further reduce both the agreement between and the reliability of automatic voice analysis systems. For this reason, it is important to have program-specific normative data, which is often not available. In cases when reference values were available, they were only marginally applicable to subjects in our study. Since it is common to obtain normal values from younger speakers, there is a need to modify the expectations with regard to dysphonic subjects who are for the most part middle-aged or aged. Thus, deviations from what should be considered normal could be a manifestation of age, disease, or both.

A close examination of the means and standard deviations of the instrumental measures tabulated by voice-quality grades in Appendix C revealed in some cases high standard deviations of the magnitude of the mean value and little distance between the means

of contiguous groups. High standard deviations of the group means are indicative of great variability of the measure within the group. The greater variability, in turn, may reflect the fact that the mean values obtained from dysphonic subjects are not representative since mean values are sensitive to extreme values. Thus, the group means have to be interpreted with caution and should not be taken as reference without further validation. Undoubtedly, a good predictor variable is supposed to have little variation within one class but allow for sufficient distance between the classes. However, one should remember that we did not exclude poor signals from analysis and that for some speakers meaningful F0-dependent parameters could not be obtained. To reject voices that are difficult to parametrisize may be an attractive way to obtain better classification results. However, it is not a practically useful alternative.

The influence of the type of material used for instrumental analysis is reflected in our results. Correlations with perceptual voice quality tended to be higher in /a/ vowels. Differences in spectral characteristics (section 2.2) seem to be the reason why high vowels gave less pathological values of voice measures. This finding is in accord with previous research. Hanson & Emanuel (1979) showed that not only high vowels of dysphonic subjects were rated as being less rough than low vowels and had a lower spectral noise level in the frequency band from 100 to 2600 Hz, but that the vowel content of the test sentence may have impact on the sentence rating of roughness. Similarly, the vowel content of sentences seemed to determine the perceived voice quality which was found to vary as a function of vowel type and vowel context in Rees (1958): isolated vowels were judged less harsh than vowels in text; high vowels were generally perceived as less harsh than low vowels. This effect was attributed to a higher intrinsic fundamental frequency of high vowels, lower average intensity and shorter duration in comparison to low vowels. Vowels in voiced consonant environment were more harsh than those in voiceless environment. Vowels in fricative environment were more harsh than vowels near stops. It is very likely that roughness, breathiness and hoarseness ratings depend on the vowel content of the speech sample. So, it seems reasonable to acquire and analyse data from both vowels and connected speech. Still, little research has been done to develop procedures for acoustic analysis of connected speech in dysphonic patients.

Contrary to expectations, some measures from sustained vowels demonstrated higher correlations with perceived voice quality than measures from connected speech, from which it can be deduced that features extracted from connected speech may be less sensitive to vocal pathology as they are influenced by segmental and suprasegmental factors occuring in connected speech. Although the validity of voice measures obtained from connected speech may be problematic, analysis of connected speech could increase our knowledge about the voice function and is necessary in order to obtain a performance which would be representative of a patient's voice function. Evidence in favour of using connected speech in clinical voice analysis is provided by Parsa & Jamieson (2001) who found that measures

based on vowels gave more accurate classification results than the same measures from connected speech and that classification accuracy improved when vowel and connected speech measures were combined. In any way, instrumental analysis of sustained vowels should not be replaced by analysis of connected speech in voice evaluation even though voice-quality rating is performed on speech material.

There is a great discrepancy between the performance of variables in single and multiple predictor models. In the present study, correlation coefficients between instrumental measures and perceived voice quality did not exceed a value 0f 0.62. This fact might be partly attributable to the loss of precision due to rounding, which was necessary to obtain discrete voice-quality ratings. Correlations with perceived voice quality would have been higher, if the mean ratings had not been rounded to the next integer. Most measured parameters were correlated with more than one perceptual voice quality. Judging by at best moderate correlation coefficients, there did not appear to be a well-defined linear relationship between any specific parameter and perceptual voice quality. Since the methods we used to assess predictive accuracy were not based on a linear relationship between instrumental measures and perceived voice quality, the results were assumed to be not affected by low correlations. Even when the difference between two contiguous voice-quality grades was statistically significant, low *z*-values (Appendix D) mean that individual parameters are not likely to reliably differentiate between voice-quality grades on their own.

There is little consensus in scientific literature as to which measure is associated with which perceptual voice quality and how successful it is in differentiating between norm and pathology or different pathologies. This seems to be true even for most investigated measures like jitter, shimmer, MPT and HNR (Hecker & Kreul (1971); Ludlow et al. (1987); Eskenazi et al. (1990); Martin et al. (1995)).

4.3.1 Measures obtained from vowels

Measures that expose important characteristics of voice function may vary in information power, reliability and usefulness for clinical purposes. This is particularly obvious when one compares the results of numerous studies on the most popular voice measures like jitter and shimmer.

For instance, we found that acoustic jitter is a poor indicator of pathology. Maryn et al. (2009) came to the same conclusion. They found that in dysphonic subjects acoustic jitter values measured in Praat ranged from 0.17 % to 1.93 % with a mean of 0.79 (0.07) %. Acoustic shimmer values ranged between 1.41 % and 6.82 % with a mean of 3.69 (0.17) %. However, this data is difficult to compare with our data since outliers and extremes were excluded from statistics.

Many studies provide inconsistent information on the usefulness of jitter to discriminate between normal and pathologic voices. In patients with laryngeal diseases, high jitter values were observed more frequently than in normal subjects (Iwata & von Leden, 1970). Increased jitter values were found in voice patients when pathology was present but not yet heard. In turn, normal jitter values were found in 40 % of pathological voices in Zyski et al. (1984). In Ludlow et al. (1987), only 30 % of subjects with laryngeal pathology had jitter values outside the confidence interval predicted by a multiple regression model with 95 normal subjects under consideration of several factors that may influence jitter.

The research by Orlikoff (1995) was devoted to the systematic study of differences between the perturbation measures extracted from acoustic and EGG signals. He believed that jitter must measure the same value in acoustic and electroglottographic signal, as in both signals jitter represents the time dimension or periodicity of vibrations. On the contrary, shimmer derived from acoustic and electroglottographic signals should be very different. This follows from the fact that the signal amplitude represents different physical phenomena: lip-radiated sound energy in acoustic signal, but impedance change in electroglottographic signal. His results confirmed that there is a certain agreement between acoustic and EGG jitter in normal speakers. His jitter values derived from Sp signals were slightly lower than those derived from EGG signals. In normal subjects, Orlikoff (1995) found a strong correlation between acoustic and EGG jitter in the order of 0.80 for male and 0.94 for female subjects, respectively. The shimmer values obtained from EGG signals were substantially lower than those from acoustic signals.

In contradiction to Orlikoff (1995), Michaelis (2000) pointed out that both jitter and shimmer extracted from EGG signals are very different from those obtained from microphone signals. In particular, it has been observed that jitter and shimmer in microphone signals correlate more strongly with each other than jitter and shimmer calculated from EGG signals, which was attributed to the influence of the vocal tract filtering function.

In the present study, we found that acoustic and EGG jitter were very different and uncorrelated. Even though data by Vieira (who claimed the opposite) was admittedly biased by signal selection before parameter extraction and subsequent rejection of signals with jitter values above 10 %, large discrepancies between EGG and acoustic jitter were observed (but ignored) in signals with a jitter above 2.7 %. Earlier research by Vieira et al. (1997) also showed large discrepancies between EGG and acoustic jitter in /i/ and /u/ vowels, independently of the F0 extraction method.

The relation between perturbation measures like jitter and shimmer and perceptual voice quality reported in scientific literature seems to be rather controversial. Hillenbrand (1988) found that a high degree of jitter was perceived as roughness rather than breathiness. Hirano (1976) demonstrated the connection between the degree of irregularity of vocal fold

vibrations and the degree of perceived roughness. In clinically hoarse voices, both jitter and shimmer were related to perceived roughness: the correlation between jitter and roughness ratings was found to be highest in /a/ with the Pearson's *r* of 0.69 in Deal & Emanuel (1978). The degree of linear association between shimmer and roughness ratings was equal for /a/ and /i/ with an r_s of 0.62. Deal & Emanuel (1978) concluded that shimmer was more strongly related to perceived roughness than jitter. Carding et al. (2004) found a strong correlation (0.71) between shimmer and perceived breathiness in the vocal fold paralysis group. Wolfe et al. (1995b) found a moderate correlation (r = 0.54) between shimmer and hoarseness grade but no correlation between jitter and hoarseness. An increase in jitter in patients with acute laryngitis was not well correlated with the perception of hoarseness (Ng et al., 1997). In accord with previous research, we found that jitter was related to roughness. In contrast, shimmer was found to be related to breathiness and both were moderately correlated with hoarseness.

The absence of sex effect in acoustic and EGG jitter in the present study does not agree well with other studies on this subject with normal speakers. According to Nittrouer et al. (1990), Sussman & Sapienza (1994) and Orlikoff (1995), there was a significant sex effect found for acoustic jitter, EGG jitter and acoustic shimmer. We believe that in dysphonic speakers sex differences were lost in the group means due to high proportion of outliers and extremes masking the sex effect. Our data on EGG shimmer suggests that male speakers have stronger EGG signals with a higher signal amplitude indicating more vocal fold contact, a fact acknowledged by many researchers.

According to studies on relation between roughness ratings and pitch, there seems to be an interaction between the perceived pitch and roughness, although the direction of this interaction is less conclusive. In some studies, low pitch was found to induce a sensation of roughness. In Newman & Emanuel (1991), perceived vowel roughness decreased as pitch level was raised over an octave in each four musical voice classifications. Emanuel & Smith (1974) came to the same conclusion by examining sustained vowels. Other sources report that strong subharmonics interfere with the perceived pitch and contribute to perceived roughness (Omori et al., 1997; Bergan & Titze, 2001; Sun & Xu, 2002). Verdonck-de Leeuw & Mahieu (2004) found that the habit of smoking which has a lowering effect on F0 is accompanied by an increase in roughness. Conversely, Wolfe & Ratusnik (1988) found that perceived roughness makes pitch sound lower. Moderately to severe dysphonic vowels received significantly lower pitch match values (r_s = –0.64) than less dysphonic and normal vowels. Another finding of this study relates perceived pitch to spectrographic noise classification in dysphonic vowels (r_s = –0.57). In the present study, fundamental frequency and intensity explained little variance in the B and R ratings, which agrees with de Krom (1995), Södersten & Lindestad (1990) and Klatt & Klatt (1990). There is a possibility that perceived roughness

may depend more on the relative pitch within one's personal range, the percentage of low F0 values or other F0 characteristics than on absolute pitch.

Further, the analysis of the interaction between breathiness ratings and intensity in dysphonic subjects revealed no significant interaction patterns. Here again, our results indicate a possible difference between dysphonic and normal population. According to Södersten & Lindestad (1990), loudness plays a crucial role in breathiness ratings, at least when rating voice healthy subjects. They found that both breathiness ratings and the degree of incomplete glottal closure increased with decreased loudness in nondysphonic women; in nondysphonic men, complete glottal closure was predominant for all loudness levels. The breathiness ratings increased significantly with decreased loudness level, pitch had no effect on the perception of breathiness. The average speaking intensity in nondysphonic female voices was found to be ca. 10 dB lower than in male voices in Brown et al. (1993); women were also judged as having more breathy voices than men. Although voice healthy subjects reportedly exhibit clear differences in speaking and vowel intensity between the sexes, our results with regard to intensity data suggest that sex-specific differences are less likely to be observed in dysphonic speakers. A possible explanation is that dysphonic subjects might factually not differ in comfortable intensity levels and that the difference between sexes or contiguous voice-quality grades may be perceptible only in the loudest possible intensity level that the study subjects are capable of producing.

In the present study, low intensity values were not associated with a large OQ and increased noise, either, although the intensity levels that we measured were compatible with other studies (Orlikoff & Kahane (1991), Max et al. (1996)). This result may be partially explained by how intensity data was acquired and analysed. We measured intensity at normal loudness and habitual pitch, not SPL values in dB(A). Loudness measurements made with a sound-level meter would have been probably very different. Speaking intensity was measured on voiced parts of the signal only. Due to technical limitations, pitch detection in whispered vowels and low-intensity segments is problematic. Thus, sentence fragments below the defined voicing threshold in severely disturbed voices were not included in test statistics with the consequence that across voiced segments severely dysphonic subjects did not differ in intensity levels from less dysphonic subjects. Intensity and other measures related to intensity like LTAS might be additionally influenced by speaker loudness variation and microphone settings. Therefore, they could be less reliable since intensity is not linearly related to loudness.

A possible explanation for the inferior performance of the OQ measure might be the method that was used to calculate the OQ. The DEGG method relies on clear definition of the closing and the opening instants. Therefore, it can work exceptionally well on strong and noise-free signals. The threshold methods are robust and can be applied to noisy and weak

signals, but imprecise, since the threshold value is chosen arbitrarily. The ceiling effect of OQ estimates in subjects who did not achieve glottal closure (OQ around 1.0) might be another explanation why formal statistical evaluation of OQ across diffferent voice quality grades was not conclusive. For the same reason, the usefulness of OQ has been questioned in Hanson et al. (1988). They came to the conclusion that electroglottography cannot provide information necessary to calculate the OQ in severely disordered voices. In their research, although OQ was not useful in differentiating between different types of paralysis, it could nevertheless be used to identify normal vs. pathologic voices.

In accord with findings by Herzel (1993), our results have shown that vocal parameters measured in vowels are of limited value because they are not able to discriminate between different types of irregularities and turbulence. In our data, perturbation measures were correlated with noise measures. Thus, pitch and amplitude perturbation measures may not be independent from additive noise[23]. This fact has been proven in experiments with synthetic signals. In synthetic signals, noise addition was found to lead to increased shimmer and jitter values (Hillenbrand, 1987). In this respect, Wolfe et al. (1995b) spoke of interdependence between jitter, shimmer and noise measures. Similarly, noise parameters are sensitive to both structural and additive noise and cannot be directly related to breathiness. To predict breathiness, pure harmonics-to-additive-noise ratios, parameters insensitive to perturbations are needed.

There is also little consensus on what part of the spectrum is more responsible for perceived breathiness. Klatt and Klatt (1990) concluded that perceived breathiness is controlled by noise in the middle and upper frequencies of the spectrum. De Krom (1994b) and de Krom (1995) reported that the lower part of the spectrum contains information relevant for dicrimination between breathy and clear voices: a relatively high level of the H1 has been associated with rated breathiness. The steepness of the spectral slope seems to be important in the perception of breathiness. In breathy voices the spectral slope exceedes the 12 dB per octave (de Krom, 1995). Hammarberg et al. (1980) found that breathy voices do not always show a decrease in acoustic intensity in the higher frequencies. The spectrum is sometimes counterbalanced by high noise energy at high-frequency regions. Hillenbrand & Houde (1996) pointed to spectral tilt measures as a good breathiness predictor in sentences. Findings by Klatt & Klatt (1990) and Hillenbrand et al. (1994) denied the relative importance of spectral tilt measures as cues to breathiness in vowels. Poor performance of the LTAS

[23] The perception of breathiness is believed to result from the presence of additive noise which is caused by constriction in the glottis or vocal tract leading to turbulent airflow. In contrast, structural noise measured in parameters like jitter and shimmer is attributable to random fluctuations in frequency and amplitude is supposed to induce the perception of roughness.

measure in the present study suggests that low frequencies do not contain information to reliably discriminate between different breathiness grades.

There seems to be little reason to be optimistic regarding applicability of the measures like LLE or SHR in voice analysis. LLE requires substantial data length for the embedding dimension to be reliably calculated. The embedding dimension *m* to estimate Lyapunov exponents should be chosen at least twice the attractor dimension. Given that the amount of data needed to calculate the LLE rises exponentially with the dimension of the attractor, at least 40,000 data points are required when the attractor dimension exceeds 3 (Wolf et al., 1985).

Kumar & Mullick (1996) reported a mean attractor dimension for normal cardinal vowels in the order of 2.89 (0.16). Jiang et al. (2009) found low-dimensional attractors ($D2 < 4$) in signals obtained from patients with polyps and nodules and normals. Similar to fricatives, pathological vowels, though, might have high attractor dimensions. In Narayanan & Alwan (1995), attractor dimension estimates in only 59 % of voiced fricatives and 44 % of voiceless fricatives were low-dimensional ranging from 3.0 to 4.8 and from 4.2 to 7.2, respectively. According to Eckmann & Ruelle (1992), in our data with ca. $n = 10000$ data points, the maximum attractor dimension that could be calculated reliably was $2\log_{10}n = 8$. When this bound is saturated, significantly longer time series are necessary to calculate LLE reliably. Since short time series lead to spurious results, it should be rewarding to repeat the experiment with longer vowel samples, at least 2–3 s, and the highest possible sampling rate. However, it might be the case that some dysphonic patients would not be able to produce phonations long enough to calculate LLE. The accuracy of the estimates reportedly increases with the amount of data available. So does the computational intensity as more points per orbit have to be calculated.

Although the SHR measure was suggested for application in clinical voice research in Sun & Xu (2002), many shortcomings prevent successful quantification of the subharmonic component in dysphonic population. The algorithm used in this thesis accounts only for subharmonics that are half the fundamental frequency although it can be extended to other subharmonic frequencies. Thus, errors are possible with voices containing more than one additional frequency between the harmonics or in cases of not harmonically related frequencies. Further, we expected to find discrepancies between the SHR value and the spectrographic type of a vowel if it had a strong subharmonic component but the analysed one-second samples did not happen to contain subharmonics.

4.3.2 Aerodynamic measures

Aerodynamic measures like MPT and PQ have been in daily clinical use and appear to provide a more sensitive means for monitoring therapeutic effects (Schutte, 1992) than for contributing to diagnoses. In an extensive study on voice disorders, Hirano (1989) found overlapping MPT values in different disease groups, suggesting that MPT cannot be used as a single diagnostic instrument. Despite high variability in aerodynamic measures in normals and dysphonic patients, there have been multiple studies providing typical values for specific voice disorders (Brasnu et al., 1989; Hirano, 1989; Max et al., 1996; Motta et al., 2001; Mitrovic, 2003; Robinson et al., 2005; Radish Kumar & Bhat, 2008; Franco & Andrus, 2009).

In the present study, aerodynamic measures were helpful in predicting breathiness and hoarseness. But taken on their own, they can hardly be considered strong predictors of perceived voice quality. It appears that the relationship between MPT and glottal gap size, a factor responsible for the perception of breathiness, is not as simple as would be expected. As shown in Hirano (1989), both MPT and PQ were negatively correlated with the glottic gap, especially in polyp and paralysis groups, but correlations were not very high.

4.3.3 F0-based measures obtained from connected speech

There is a long tradition of research concerned with fundamental frequency patterns in read and spontaneous speech. Beside clinical voice research, F0-based measures from connected speech have been in the focus of forensic linguistics and prosody research. Although a great deal of effort has been put into studying the properties of F0 in different languages of the world including measures of F0 variability and F0 movement, relatively few studies written in the German language are available for comparison with our data.

In ca. 65 % of study subjects, the mean F0 was found to be located outside the normal range. This finding is compatible with previous research on this subject. Voice disorders affect SFF in a number of ways. Whereas increased vocal effort leads to higher-pitched speech, mass lesions have a lowering effect on SFF. Hirano (1989) found a decreased F0 in patients with Reinke's edema. Murry & Doherty (1980) found that patients diagnosed with laryngeal cancer have a lower SFF and a higher F0 variability. The habit of smoking, the most important cause of laryngeal pathology, has been reported to have a lowering effect on the mean fundamental frequency in women without voice pathology (Gilbert & Weismer, 1974). Thickening of the vocal folds was observed in 87 % of smokers.

According to Rappaport (1958), the mean F0 standard deviation measured 2.3 st and 1.9 st in nondysphonic male and female voices, respectively. We could not use these figures as pathology threshold since the extent of F0 excursions in speech is known to depend on the type of used material. However, it has been observed that only 11 male subjects and 4 female

subjects used an F0 SD below the specified values. Whether this effect is attributable to pathology or age is not clear. As shown in (Pegoraro-Krook, 1988), the extent of F0 excursions increases with age.

Pegoraro-Krook (1988) reported an F0 range of 5.1 st and 5.2 st in nondysphonic female and male subjects reading the standard text, respectively. In accord with Pegoraro-Krook (1988), the difference in the 80R estimates was not significant between the sexes in the present study. In accord with expectations, we observed higher 80R estimates than reported above in 111 (ca. 74 %) voices. Studies that claim that frequency range is larger in female voices as compared to male voices did not happen to use the standard text (Brown et al., 1993; Morris et al., 1995). Obviously, a reduced F0 range and F0 SD that were observed in some pathologies (Murry, 1978; Hirano, 1989; Max et al., 1996) do not apply to reading data but to an F0 range based on the highest and the lowest possible tones that patients are capable of producing.

Our data confirmed the tendency of SFF to increase with age in male voices and to decrease in female voices. Age affects male and female voices differently. F0 begins to rise from age 60+ in male voices but decreases by approximately 15 Hz around age of 70 in female voices (Hollien & Shipp, 1972; Pegoraro-Krook, 1988; Ferrand, 2002). Stoicheff (1981) found a decrease in F0 in non-smoking women aged 50 and older which she attributed to increased vocal fold mass. A decrease in speaking F0 was accompanied by an increased variability in F0 indicating a poorer vocal fold control with age. Hollien & Shipp (1972) attributed the decrease in SFF in male voices between 20 to 40 years of age to the thickening of the vocal fold tissue; the increase in SFF in older age decades was ascribed to senescent changes in the larynx.

Occasionally, we observed that even normal voices may have a high IFx value and that normal healthy voices have irregularities in F0. Numerous studies reported the presence of subharmonics in normal voices as well, to name just a few of them: Svec et al. (1996) described a subharmonic vibratory pattern in a normal larynx, in Wong et al. (1991) subharmonics could be generated in vocal folds of normal stiffness and mass without asymmetry with decreased stress in the longitudinal string tension; the study by Gozalez (2007) yealded a weak positive correlation between subharmonic parameters and height in women without vocal pathology; in Hudson et al. (2007) 3 male subjects among 100 young normal males had creaky voice or subharmonics in connected speech. Given that low frequencies might contribute to rough voice quality, there is a certain degree of arbitrariness in the PLF measure since other thresholds might have led to a different result.

4.3.4 Other reading variables

Our results have shown that the potential relation of breathy voice quality to respiration, pausing behavior and speech tempo seems to offer promising ground for research. In the last three decades, several quantitative studies have been conducted to clarify the interaction between voice disorders and speech breathing. Hixon & Putnam (1987) and Schaeffer et al. (2002) demonstrated on the basis of different types of voice disorders that speech breathing of voice-disordered subjects is different from normal subjects. Bahr et al. (2007) showed that patients with adductor spasmodic dysphonia exhibited disordered breathing. Lowell et al. (2008) reported that mildly dysphonic teachers used different speech breathing strategies than their collegues without voice problems (they initiated and terminated breath groups at a significantly lower lung volumes) whereas no difference was found in measures derived from the EGG signal between the two groups of teachers. Similarly, Sapienza & Stathopoulos (1994) demonstrated that women with vocal nodules compensate for loss of air during voicing by using a greater amount of lung volume and higher rates of glottal airflow. They showed larger volume of air per syllable and per breath group relative to their vital capacity. Poor contact between the vocal folds was reported to affect temporal speech characteristics: Till et al. (1994) observed reduced number of syllables between pauses (speech phrase duration) and increased number of breaths per minute in patients with laryngeal insufficiency.

In the present study, prosodic measures were mainly associated with perceived breathiness. However, it is conceivable that rough voices may also be characterized by deviant prosody since (besides vocal fold closure problems) efficient valving of airflow can be affected by irregularities in the vocal fold vibration (Sapienza & Stathopoulos, 1994).

The absolute number of pauses in B2 and B3 voices in the present study was higher than the numbers previously reported for normal subjects. Recently, the "Northwind and Sun" passage was used in two German studies on pausing behavior in normal subjects. Trouvain (1999) investigated the effect of reading tempo on the number and duration of pauses. At normal reading speed, 3 female subjects needed on average between 8.7 and 13.0 pauses with the mean pause duration between 533 ms and 592 ms. In Siebenhaar (2008), study subjects made at least 9 pauses.

There is little evidence that pauses during read speech in normal subjects correpond to breath intakes. Normal subjects use pauses to parse speech into meaningful segments; they do not use every pause to replenish the air supply. However, data on pauses per minute in our study might be indicative of the fact that dysphonic speakers do use every pause to take in breath. The mean number of pauses per minute in B0 and B1 voices was lower than the mean number of breaths per minute in normal subjects reported in Sperry & Klich (1992) and Hoit & Lohmeier (2000). According to Hoit & Lohmeier (2000), breathing frequency during reading averaged 20 breaths/min in nondysphonic males; per breath group, subjects produced

on the average 16.11 syllables. Similar values were reported for nondysphonic females in Sperry & Klich (1992). In accord with Till et al. (1994), we observed an increased number of pauses per minute in B2 and B3 patients.

Further, a reduced number of syllables between pauses was observed in all breathiness grades. The mean number of syllables between two consecutive pauses in the four breathiness groups was systematically lower than the mean number of syllables per breath group reported for normal subjects and higher than or comparable with the means reported for esophageal and tracheoesophageal speakers in the following studies. Hoit & Hixon (1987) found that the number of syllables per breath group was not significantly different in three age groups including nondysphonic males reading a text passage with the mean values ranging from 16 to 22 syllables per breath group. Similarly, female speakers in three age groups did not significantly differ in the number of syllables per breath group, producing between 15 and 19 syllables one one intake of air (Hoit et al., 1989). As for severely disturbed voices, Max et al. (1996) showed a high correlation (r = 0.87) between MPT and maximal number of syllables between two inspirations in esophageal and tracheoesophageal speakers. On the average, esophageal speakers could maximally produce 7.1 (4.71) syllables on one air intake. These figures suggest that the count of the number of syllables between two consecutive pauses might be a good approximation of the count of the number of syllables on one intake of air.

The mean pause length data in B1, B2 and B3 voices in the present study come close to reading data by Trouvain (1999) obtained from normal subjects. When this data is taken as a reference, subjects with breathy voice quality seem to compensate for air leakage not by longer pauses but by more frequent pausing. Campione & Veronis (2002), however, estimated average pause duration in reading by voice healthy German subjects at 490 ms, which was significantly lower than values reported by Trouvain. On comparing this data with our results, one might conclude that speakers with breathy voice quality also make longer pauses than normal subjects and speakers with clear voice quality do.

The mean speech and articulation rate in H0 voices compare well with the mean speech and articulation rates obtained from nondysphonic native speakers of German in Greisbach (1992), Trouvain (1999) and Siebenhaar (2008). The means in other hoarseness grades were found to be significantly lower than reported above. It should be noted, however, that studies on speech and articulation rate in normal population are extremely difficult to compare and summarize for a number of reasons stated in the following. The results might depend on the used speech material. Experimental results are often unreliable as statistical tests that rely on normality assumption are routinely applied to far from normally distributed data and small study sizes. When research is conducted in one language, the results are often not generalizable to other languages or to a larger population. Not only is the speech tempo different in different languages, but the structure of syllables and words are language-specific.

We would like to point out that a listener's impression of voice quality might be affected by, strictly speaking, non-voice factors like pausing behavior or speech tempo. However, more informaton is necessary on how different voice disorders influence speech characteristics. For instance, the occurence of pauses in read speech is governed by inspiration needs and parsing for content. Sapienza et al. (1997) found that despite loss of air subjects with vocal nodules did not differ from normal subjects in pause placement. They made pauses at logical places within the text. One might hypothesize that the severity of voice disorder might be reflected in dysphonic prosody. However, Finizia et al. (1999) could not confirm differences in speech rate between laryngectomized patients with TEP, radiotherapy-treated patients with preserved larynx and controls. A longer inhalation pause time in laryngectomized was attributed to finger occlusion for speech.

Finally, speech breathing is inherently connected with general respiratory function that was found to deteriorate with age (Hoit & Hixon, 1987; Sperry & Klich, 1992). Older subjects were observed to expend greater volumes of air per syllable and to inhale to a higher lung volume levels, more deeply and longer as the sentences became longer. They were found to use more unphonated expirations following inhalation and preceding phonation. In Hoit & Hixon (1987) and Hoit et al. (1989), a significant difference in the percentage of VC expenditure per syllable across the three age groups was observed. They suggested a reduced economy of valving of the speech air stream as a function of age, which they atrributed to degenerative changes in the larynx. All these facts raise the question of whether speech breathing characteristics are better predictors of age rather than predictors of breathiness ratings.

4.4 Suggestions for future research

This study identified a number of variables important in objective voice-quality classification. Although our results show significant progress in predicting perceptual voice quality from objective voice parameters, there is still a need for improvement. In this chapter, suggestions for future research will be addressed. The list is not intended to be exhaustive but merely highlight promising research directions.

Future research suggestions include increasing study size. In this way, it is possible to avoid underrepresented voice-quality classes.

It can be seen in classification experiments that combining measures of the same type results in a relatively modest improvement in prediction accuracy, which is not surprising given that many measures correlate with each other. There may also be considerable redundancy in identical measures derived from different vowels. A further improvement of classification rates could probably be achieved by adding one more variable type. One

possible way to diversify the list of predictor variables is to add, e.g., a categorical variable type. Thus, spectrographic vowel type, subharmonic content, age decade and sex could be included as variables.

It is somewhat surprising that prosodic measures performed as well as they did in improving prediction accuracy in breathiness ratings since measures that we took were not directly related to speech breathing. It should be noted that these results are pure calculations and do not provide any evidence about the cues used by listeners in judging breathiness. Although we do not have an explanation for this finding, the results are encouraging. A series of perceptual experiments are needed to prove that listeners are influenced by speech breathing and pausing behavior in rating breathiness. Little research has been done so far on respiratory function in specific voice disorders (Hixon & Putnam, 1987; Sapienza et al., 1997). Speech breathing in each type of dysphonia has to be examined independently. Viewed in relation to this kind of application, impaired prosody in dysphonic population is another promising topic for future research.

Further, it is interesting to compare the discriminative power of acoustic features based on vowels with the same set of features extracted from speech. The reason we did not acquire noise measures from connected speech was twofold: first, such measures like GNE and HNR taken from vowels were already relatively strong predictors of perceived voice quality. An increase in information retrieval cost is likely to have limited effect when the results are already satisfactory. Second, estimates of noise measures obtained from voiced segments would depend on the voiced consonant content of the speech sample and cannot be compared across languages. Ideally, noise measures in connected speech should be calculated using isolated vowels. However, vowel detection is a more complex computational problem than voicing detection.

The present study design included both voice healthy and voice-disordered subjects. In this we merely followed an example set by other researchers. The overwhelming majority of studies on automatic classification of voice quality used a mixed group of subjects. In the light of the results of the present study we believe that voice research might gain a lot from comparing many of the examined measures obtained from dysphonic subjects with the same measures obtained from presumably voice healthy population. This issue is worth being pursued in another study.

Bibliography

Alonso JB, Díaz-de-María F, Travieso CM, Ferrer MA. (2005). Using nonlinear features for voice disorder detection. In: Proceedings of the NOLISP: 94–106, Barcelona, Spain.

Anastaplo S, Karnell MP. (1988). Synchronized videostroboscopic and electroglottographic examination of glottal opening. J Acoust Soc Am 83, 1883–1890.

Andrew BL. (1955). The respiratory displacement of the larynx: a study of the innervation of accessory respiratory muscles. J Physiol 130, 474–487.

Aronson AE. (1990). Clinical voice disorders. 3rd ed., New York: Thieme Inc.

Askenfelt A, Hammarberg B. (1981). Speech waveform perturbation analysis revisited. STL–QPSR 22, 49–68.

Askenfelt A, Hammarberg B. (1986). Speech waveform perturbation analysis: a perceptual-acoustical comparison of seven measures. J Speech Hear Res 29, 50–64.

Bahr RH, Biedess K, Ridley M. (2007). Speech breathing in patients with adductor spasmodic dysphonia. In: Proceedings of the ICPhS: http://www.icphs2007.de/conference/Papers/1547/1547.pdf. (last visited: 30th September 2010), Saarbrücken, Germany.

Baken RJ, Orlikoff RF. (2000). Clinical measurement of speech and voice. San Diego, CA: Singular Publishing Group, Thompson Learning.

Baken RJ. (1992). Electroglottography. J Voice 6, 98–110.

Barney A, De Stefano A, Henrich N. (2007). The effect of glottal opening on the acoustic response of the vocal tract. Acta Acust United Acust 93, 1046–1056.

Behrman A, Agresti CJ, Blumstein E, Lee N. (1998). Microphone and electroglottographic data from dysphonic patients: Type 1, 2 and 3 signals. J Voice 12, 249–260.

Bergan C, Titze I. (2001). Perception of pitch and roughness in vocal signals with subharmonics. J Voice 15, 165–175.

Berry DA, Herzel H, Titze IR, Story BH. (1996). Bifurcations in excised larynx experiments. J Voice 10, 129–138.

Bhuta T, Patrick L, Garnett JD. (2004). Perceptual evaluation of voice and its correlation with acoustic measurements. J Voice 18, 299–304.

Böckler R, Hacki T. (1999). Influence of neck soft tissue vibrations on the EGG-Lx signal. Folia Phoniatr Logop 51, 243–249.

Boersma P, Weenink D. (2005). Praat: Doing phonetics by computer. <http://www.praat.org/>.

Boersma P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 17, 97–110.

Boone DR, McFarlane SC. (1993). A critical view of the yawn-sigh as a voice therapy technique. J Voice 7, 75–80.

Boucher V, Lamontagne M. (2001). Effects of speaking rate on the control of vocal fold vibration: Clinical implications of active and passive aspects of devoicing. J Speech Lang Hear Res 44, 1005–1014.

Brasnu D, Strome M, Buchman CL, Pfauwadel MC, Menard M, Laccourreye H. (1989). Voice evaluation of myomucosal shunt after total laryngectomy: comparison with eosophageal speech. Am J Otolaryngol 10, 267–272.

Braunschweig T, Grießbach G, Hanson J. (1997). Anwendung rekursiver Schätzungen bei der Analyse des Einschwingens der Stimmlippen. In: Proceedings of the Internation Scientific Colloquium: Bd. 2, 133–138, TU Ilmenau, Germany.

Brown WS Jr, Morris RJ, DeGroot T, Murry T. (1998). Reliability of single sample experimental designs. J Voice 12, 453–459.

Brown WS Jr, Morris RJ, Murry T. (1996). Confortable effort level revisited. J Voice 10, 299–305.

Brown WS, Morris RJ, Hicks DM, Howell E. (1993). Phonational profiles of female professional singers and nonsingers. J Voice 7, 219–226.

Butcher A. (1981) Aspects of the speech pause: phonetic correlates and communicative function. Arbeitsberichte, 15. Institut für Phonetik, Universität Kiel.

Callan DE, Kent RD, Roy N, Tasko SM. (1999). Self-organizing map for the classification of normal and disordered female voices. J Speech Lang Hear Res 42, 355–366.

Campione E, Veronis J. (2002). A large-scale multilingual study of silent pause duration. In: Proceedings of the ESCA workshop on speech prosody: 199–202, Aix-en-Provence.

Carding PN, Steen IN, Webb A, Mackenzie K, Deary IJ, Wilson JA. (2004). The reliability and sensitivity to change of acoustic measures of voice quality. Clin Otolaryngol 29, 538–544.

Cavalli L, Hirson A. (1999). Diplophonia reappraised. J Voice 13, 542–556.

Colton RH, Conture EG. (1990). Problems and pitfalls of electroglottography. J Voice 4, 10–24.

Cooke A, Ludlow CL, Hallett N, Selbie WS. (1997). Characteristics of vocal fold adduction related to voice onset. J Voice 11, 12–22.

Dankovicová J. (1997). The domain of articulation rate variation in Czech. J Phonetics 25, 287–312.

De Bodt MS, Wuyts FL, Van de Heyning PH, Croux C. (1997). Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. J Voice 11, 74–80.

de Jong NH, Wempe T. (2007). Automatic measurement of speech rate in spoken Dutch. ACLC Working Papers 2, 51– 60.

de Krom G. (1994a). Consistency and reliability of voice quality ratings for different types of speech fragments. J Speech Hear Res 37, 985–1000.

de Krom G. (1994b). Spectral correlates of breathiness and roughness for different types of vowel fragments. In: Proceedings of the ICSLP: 1471–1474, Yokohama, Japan.

de Krom G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. J Speech Hear Res 38, 794–811.

de Pijper JR, Sanderman A. (1994). On the perceptual strength of prosodic boundaries and its relation to supra-segmental cues. J Acoust Soc Am 96, 2037–2047.

Deal RE, Emanuel F. (1978). Some waveform and spectral features of vowel roughness. J Speech Hear Res 21, 250–264.

Dejonckere P, Lebacq J. (1983). An analysis of the diplophonia phenomenon. Speech Commun 2, 47–56.

Dejonckere P, Obbens C, de Moor GM, Wieneke GH. (1993). Perceptual evaluation of dysphonia: Reliability and relevance. Folia Phoniatr 45, 76–83.

Dejonckere PH, Bradley P, Clemente P, Cornut G, Crevier-Buchman L, Friedrich G, Van De Heyning P, Remacle M, Woisard V; Committee on Phoniatrics of the European Laryngological Society (ELS). (2001). A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). Eur Arch Otorhinolaryngol 258, 77–82.

Dejonckere PH, Remacle M, Fresnel-Elbaz E, Woisard V, Crevier-Buchmann L, Millet B. (1996). Differentiated perceptual evaluation of pathological voice quality: reliability and correlation with acoustic measurements. Rev Otolaryngol Otol Rhinol 117, 219–224.

Denny M. (2000). Periodic variation in inspiratory volume characterizes speech as well as quiet breathing. J Voice 14, 34–46.

Dikkers F, Nikkels PGJ. (1999). Lamina propria of the mucosa of benign lesions of the vocal folds. Laryngoscope 109, 1684–1689.

Döllinger M, Braunschweig T, Lohscheller J, Eysholdt U, Hoppe U. (2003). Normal voice production: computation of driving parameters from endoscopic digital high speed images. Method Inform Med 42, 271–276.

Döllinger M, Hoppe U, Hettlich F, Lohscheller J, Schuberth S, Eysholdt U. (2002). Vibration parameter extraction from endoscopic image series of the vocal folds. IEEE T Biomed Eng 49, 773–781.

Döllinger M, Rosanowski F, Eysholdt U, Lohscheller J. (2008). Basic research on vocal fold dynamics: three-dimensional vibration analysis of human and canine larynges. HNO 56, 1213–1220.

Eckmann JP, Ruelle D. (1992). Fundamental limitations for estimating dimensions and Lyapunov exponents in dynamical systems. Physica D 56, 185–187.

Emanuel FW, Smith WF. (1974). Pitch effects on vowel roughness and spectral noise. J Phonetics 2, 247–253.

Eskenazi L, Childers DG, Hicks DM. (1990). Acoustic correlates pf vocal quality. J Speech Hear Res 33, 298–306.

Ewan MG. (1975). Explaining the intrinsic pitch of vowels. J Acoust Soc Am 58 (S1), 40.

Ewanowski SJ, Hixon TJ, Kelsey CA, Minifie FD. (1968). Lateral pharyngeal wall movement during esophageal voice production. J Acoust Soc Am 44, 354–354.

Eysholdt U, Lohscheller J. (2008). Phonovibrogram: vocal fold dynamics integrated within a single image. HNO 56, 1207–1212.

Eysholdt U, Rosanowski F, Hoppe U. (2003a). Measurement and interpretation of irregular vocal fold vibrations. HNO 51, 710–716.

Eysholdt U, Rosanowski F, Hoppe U.(2003b). Vocal fold vibration irregularities caused by different types of laryngeal asymmetry. Eur Arch Oto-Rhino-L 260, 412–417.

Eysholdt U, Tigges M, Wittenberg T, Pröschel U. (1996). Direct evaluation of high-speed recordings of vocal fold vibrations. Folia Phoniatr Logop 48, 163–170.

Fabre P. (1957). Un procédé électrique percutané d'inscription de l'accolement glottique au cours de la phonation. Bull Nat Med 141, 66–99.

Fant G, Ishizaka K, Lindqvist-Gauffin J, Sundberg J. (1972). Subglottal formants. STL–QPSR 13, 1–12.

Fant G, Kruckenberg A, Ferreira JB. (2003). Individual variations in pausing. A study of read speech. Phonum 9, 193–196.

Fant G. (1970). Acoustic theory of speech production. 2nd printing, The Hague and Paris: Mouton.

Ferrand CT. (2002). Harmonics-to-noise ratio: an index of vocal aging. J Voice 16, 480–487.

Finizia C, Dotevall H, Lundström E, Lindström J. (1999). Acoustic and perceptual evaluation of voice and speech quality: A study of patients with laryngeal cancer treated with laryngectomy vs irradiation. Arch Otolaryngol Head Neck Surg 125, 157–163.

Finnegan EM, Luschei ES, Hoffman HT. (2000). Modulations in respiratory and laryngeal activity associated with changes in vocal intensity during speech. J Speech Lang Hear Res 43, 934–950.

Fourcin A, McGlashan J, Blowes (R). (2002). Measuring voice in the clinic - Laryngograph® Speech Studio analyses. In: Proceedings of the 6th Voice Symposium of Australia : Adelaide, Australia. http:// http://www.laryngograph.com/pdfdocs/paper7.pdf. (last visited: 28th Februar 2011).

Fourcin A. (2009). Aspects of voice irregularity measurement in connected speech. Folia Phoniatr Logop 61, 126–136.

Fraile R, Sáenz-Lechón N, Godino-Llorente JI, Osma-Ruiz V. (2009). Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficients. Folia Phoniatr Logop 61, 146–152.

Franco RA, Andrus JG. (2009). Aerodynamic and acoustic characteristics of voice before and after adduction arytenopexy and medialization laryngoplasty with GORE-TEX in patients with unilateral vocal fold immobility. J Voice 23, 261–267.

Friedrich G, Dejonckere PH. (2005). The voice evaluation protocol of the European Laryngological Society (ELS) – first results of a multicenter study. Laryngorhinootologie 84, 744–752.

Fröhlich M, Michaelis D, Strube H W, Kruse E. (2000). Acoustic voice analysis by means of the hoarseness diagram. J Speech Lang Hear Res 43, 706–720.

Fujimura O. (1976). Stereo fiberscope. Seminar on research of the dynamic aspects of speech production. Tokyo, University Press.

Gardner PL. (1975). Scales and statistics. Rev Educ Res 45, 43–57.

Gilbert HR, Weismer GG. (1974). The effects of smoking on the speaking fundamental frequency of adult women, J Psycholinguist Res 3, 225–231.

Giovanni A, Ouaknine M, Guelfucci B, Yu P, Zanaret M, Triglia JM. (1999a). Non-linear behavior of vocal fold vibration: the role of coupling between the vold folds. J Voice 13, 465–476.

Giovanni A, Ouaknine M, Triglia JM. (1999b). Determination of largest Lyapunov exponents of vocal signal: Application to unilateral laryngeal paralysis. J Voice 13, 341–354.

Giovanni A, Robert D, Estublier N, Teston B, Zanaret M, Cannoni M. (1996). Objective evaluation of dysphonia: preliminary results of a device allowing simultaneous acoustic and aerodynamic measurements. Folia Phoniatr Logop 48, 175–185.

Godino-Llorente J, Osma-Ruiz V, Sáenz-Lechón N, Gómez–Vilda P, Blanco-Velasco M, Cruz-Roldán F. (2010). The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders. J Voice 24, 47–56.

Goldman-Eisler F. (1968). Psycholinguistics: Experiments in spontaneous speech. New York: Academic Press.

Gozalez J. (2007). Correlations between speakers' body size and acoustic parameters of voice. J Percept Mot Skills 105, 215–220.

Greisbach R. (1992). Reading aloud at maximal speed. Speech Commun 11, 469–473.

Grosjean F, Collins M. (1979). Breathing, pausing and reading. Phonetica 36, 98–114.

Hacki T. (1989). Classification of glottal dysfunctions on the basis of electroglottography. Folia Phoniatr 41, 43–48.

Hacki T. (1996). Electroglottographic quasi-open quotient and amplitude in crescendo phonation. J Voice 10, 342-347.

Hakkesteegt MM, Brocaar MP, Wieringa MH, Feenstra L. (2006). Influence of age and gender on the dysphonia severity index: a study of normative values. Folia Phoniatr Logop 58, 264–273.

Hamlet S. (1973). Vocal compensation: an ultrasonic study of vocal vibration in normal and nasal vowels. Cleft Palate J 10, 267–285.

Hamlet SL. (1980). Ultrasonic measurements of the larynx height and vocal fold vibratory pattern. J Acoust Soc Am 68, 121–126.

Hammarberg B, Fritzell B, Gauffin J, Sundberg J, Wedin L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. Acta Otolaryngol 90, 441–451.

Hanson DG, Gerratt BR, Karin RR, Berke GS. (1988). Glottographic measures of vocal fold vibration: an examination of laryngeal paralysis. Laryngoscope 98, 541–549.

Hanson HM, Chuang ES. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. J Acoust Soc Am 106, 1064–1077.

Hanson HM. (1997). Glottal characteristics of female speakers: acoustic correlates. J Acoust Soc Am 101, 466–481.

Hanson W, Emanuel FW. (1979). Spectral noise and vocal roughness relationships in adults with laryngeal pathology. J Comm Disord 12, 113–124.

Harri A, Wade Brorsen B. (2009). The overlapping data problem. Quant Qual Analysis Soc Sci 3, 78–115.

Hatzikirou H, Fitch WT, Herzel H. (2006). Voice instabilities due to source-tract interactions. Acta Acust 92, 468–474.

Hecker MHL, Kreul EJ. (1971). Descriptions of the speech of patients with cancer of the vocal folds: Part I: Measures of fundamental frequency. J Acoust Soc Am 49, 1275–1282.

Henderson A, Goldman-Eisler F, Skarbek A. (1965). The common value of pausing time in spontaneous speech. Q J Exp Psychol 17, 343–345.

Henrich N, d'Alessandro C, Doval B, Castellengo M. (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. J Acoust Soc Am 115, 1321–32.

Henrich N, d'Alessandro C, Doval B, Castellengo M. (2005). Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency. J Acoust Soc Am 117, 1417–1430.

Herbst C, Ternström S. (2006). A comparison of different methods to measure EGG contact quotient. LPV 31, 126–138.

Herzel H, Berry D, Titze I, Saleh M. (1994). Analysis of vocal disorders with methods from nonlinear dynamics. J Speech Hear Res 37, 1008–1019.

Herzel H. (1993). Bifurcations and chaos in voice signals. Appl Mech Rev 46, 399–413.

Higashikawa M, Minifie FD. (1999). Acoustical-perceptual correlates of "whisper pitch" in synthetically generated vowels. J Speech Lang Hear Res 42, 583–591.

Higgins MB, Netsell R, Schulte L. (1994). Aerodynamic and electroglottographic measures of normal voice production: Intrasubject variability within and across sessions. J Speech Hear Res 37, 38–45.

Hill DP, Meyers AD, Scherer RC. (1990). A comparison of four clinical techniques in the analysis of phonation. J Voice 4, 198–204.

Hillenbrand J, Cleveland RA, Erickson RL. (1994). Acoustic correlates of breathy vocal quality. J Speech Hear Res 37, 769–778.

Hillenbrand J, Houde RA. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. J Speech Hear Res 39, 311–321.

Hillenbrand J. (1987). A methodological study of perturbation and additive noise in synthetically generated voice signals. J Speech Hear Res 30, 448–461.

Hillenbrand J. (1988). Perception of aperiodicities in synthetically generated voices. J Acoust Soc Am 83, 2361–2371.

Hillman RE, Holmberg EB, Perkell JS, Walsh M, Vaughn C. (1989). Objective assessment of vocal hyperfunction: an experimental framework and initial results. J Speech Hear Res 32, 373–92.

Hirano M, Hibi S, Terasawa S, Fujiu M. (1986). Relationship between aerodynamic, vibratoty, acoustic and psychoacoustic correlates in dysphonia. J Phonetics 14, 445–456.

Hirano M, Koike Y, von Leden H. (1968). Maximum phonation time and air usage during phonation. Folia Phoniatr 20, 185–201.

Hirano M. (1974). Morphological structure of the vocal cord as a vibrator and its variations. Folia Phoniatr 26, 89–94.

Hirano M. (1976). Vocal cord vibration behavior of the larynx-structured vibrator in normal and pathological conditions [Film]. Japan: Kurume University Department of Otolaryngology and Head and Neck Surgery.

Hirano M. (1981). Clinical examination of voice. 52–53, Wien: Springer.

Hirano M. (1981). Clinical examination of voice. Wien and New York: Springer Verlag.

Hirano M. (1989) Objective evaluation of the human voice: Clinical aspects. Folia Phoniatr 41, 89–144.

Hixon TJ, Putnam AB. (1987). Voice disorders in relation to respiratory kinematics. In: Hixon TJ (ed.): Respiratory function in speech and song. 267–270, Boston: College-Hill Press, Little Brown.

Hixon TJ, Weismer G. (1995). Perspectives on the Edinburgh study of speech breathing. J Speech Hear Res 38, 42–60.

Hixon TJ. (1973). Respiratory function in speech. In: Minifie F, Hixon T, Williams F (eds.): Normal aspects of speech, hearing and language. 73–122, Englewood Cliffs: Prentice Hall.

Hoit J, Hixon T, Altman M, Morgan W. (1989). Speech breathing in women, J Speech Hear Res 32, 353–365.

Hoit J, Hixon T. (1987). Age and speech breathing. J Speech Hear Res 30, 351–366.

Hoit JD, Lohmeier HL. (2000). Influence of continuous speaking on ventilation. J Speech Lang Hear Res 43, 1240–1251.

Hollien H, Ship T. (1972). Speaking fundamental frequency and chronological age in males, J Speech Hear Res 15, 155–159.

Honda K. (1983). Relationship between pitch control and vowel articulation. In: Bless DM, Abbs JH (eds.): Vocal fold physiology: contemporary research and clinical issues. 286–297, San Diego: College-Hill.

Horii Y, Fuller BF. (1990). Selected acoustic characteristics of voices before intubation and after extubation. J Speech Lang Hear Res 33, 505–510.

Horii Y. (1979). Fundamental frequency perturbation observed in sustained phonation. J Speech Hear Res 22, 5–19.

Howard DM. (1995). Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers. J Voice 9, 163–172.

Hudson T, de Jong G, McDougall K, Harrison P, Nolan F. (2007). F0 statistics for 100 young male speakers of Standard Southern British English. In: Proceedings of the ICPhS: 1809–1812, Saarbrücken, Germany.

Inagi K, Khidr AA, Ford CN, Bless DM, Heisey DM. (1997). Correlation between vocal functions and glottal measurements in patients with unilateral vocal fold paralysis. Laryngoscope 107, 782–791.

Isshiki N, Okamura H, Tanabe M, Morimoto M. (1969). Differential diagnosis of hoarseness. Folia Phoniatr 21, 9–19.

Isshiki N, Yanagihara N, Morimoto M. (1966). Approach to the objective diagnosis of hoarseness. Folia Phoniatr 18, 393–400.

Isshiki N. (1964). Regulatory mechanism of voice intensity variation. J Speech Hear Res 7, 17–29.

Iwata S, von Leden H. (1970). Pitch perturbations in normal and pathologic voices. Folia Phoniatr 22, 413–424.

Jakobson R, Fant CGM, Halle M. (1952) Preliminaries to speech analysis: The distinctive features and their correlates. Cambridge: MIT Press.

Jassem W, Steffen-Batog S, Czajka M. (1973). Statistical characteristics of short-term average F0 distributions as personal voice features. In: Jassem W (ed.): Speech analysis and synthesis. Vol. 3, 209–225, Warshaw: Polish Academy of Science.

Jiang JJ, Titze IR. (1994). Measurement of vocal fold intraglottal pressure and impact stress. J Voice 8, 132–145.

Jiang JJ, Zhang Y, MacCallum J, Sprecher A, Zhou L. (2009). Objective acoustic analysis of pathological voices from patients with vocal nodules and polyps. Folia Phoniatr Logop 61, 342–349.

Jilek C, Marienhagen J, Hacki T (2003). Auswirkungen der Protrusion der Zunge auf elektroglottographische Parameter. In: Gross M & Kruse E (Hrsg.): Aktuelle phoniatrisch-pädaudiologische Aspekte, Bd. 11, 121–127.

Jilek C., Löhrer H., Marienhagen J., Hacki T. (2004). Vocal stability in functional dysphonic versus healthy voices at different times of voice loading. J Voice 18, 443–453.

Jones AP, Johnson LA, Butler MC, Main DS. (1983). Apples and oranges: An empirical comparison of commonly used indices of interrater agreement. Acad Manage J 26, 507–519.

Kahane JC. (1987). Connective tissue changes in the larynx and their effect on voice. J Voice 1, 27–30.

Kakita Y, Hiki S. (1976). A study of laryngeal control for pitch change by use of thyrometer. J Acoust Soc Am 59, 669–674.

Kitzing P. (1986). LTAS-criteria pertinent to the measurement of voice quality. J Phonetics 14, 477–482.

Klatt DH, Klatt LC. (1990). Analysis, synthesis, perception of voice quality variations among female and male talkers. J Acoust Soc Am 87, 820–857.

Koike Y. (1969). Vowel amplitude modulations in patients with laryngeal diseases. J Acoust Soc Am 45, 839–844.

Kreiman J, Gerratt BR, Kempster GB, Erman A, Berke GS. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. J Speech Hear Res 36, 21–40.

Kreiman J, Gerratt BR, Precoda K. (1990). Listener experience and perception of voice quality. J Speech Hear Res 33, 103–115.

Kreiman J, Gerratt BR. (1998). Validity of rating scale measures of voice quality. J Acoust Soc Am 104, 1598–1608.

Kumar A, Mullick SK. (1996). Nonlinear dynamical analysis of speech. J Acoust Soc Am 100, 615–629.

Ladefoged P, Maddieson I. (1996). The sounds of the world's languages. Oxford: Blackwell.

Laryngoscope 118, 753–758.

Lebrun Y, Devreux F, Rousseau JJ, Darimont P. (1982). Tremulous speech. Folia Phoniatr 34, 134–142.

Lecluse FL, Tiwari RM, Snow GB. (1981). Electroglottographic studies of Staffieri neoglottis. Laryngoscope 91, 971–975.

Lin E, Jiang J, Hanson D. (1998). Glottographic signal perturbation in biomechanically different types of dysphonia. Laryngoscope 108, 18–25.

Linder R, Albers A, Hess M, Pöppl S, Schönweiler R. (2008). Artificial neural network-based classification to screen for dysphonia using psychoacoustic scaling of acoustic voice features. J Voice 22, 155–163.

Linder R, Pöppl SJ. (2001). ACMD: A practical tool for automatic neural net based learning. In: Proceedings of the ISMDA: 168–173, Madrid, Spain.

Lindqvist-Gauffin J, Sundberg J. (1971). Pharyngeal constrictions. STL–QPSR 12, 26–31.

Lindqvist-Gauffin J. (1972). A descriptive model of laryngeal articulation in speech. STL–QPSR 13, 1–14.

Lohscheller J, Eysholdt U, Toy H, Dollinger M. (2008). 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 27, 300–309.

Lohscheller J, Eysholdt U. (2008). Phonovibrogram visualization of entire vocal fold dynamics.

Lohscheller J, Toy H, Rosanowski F, Eysholdt U, Döllinger M. (2007). Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. Med Image Anal 11, 400–413.

Lowell SY, Barkmeier-Kraemer JM, Hoit JD, Story BH. (2008). Respiratory and laryngeal function during spontaneous speaking in teachers with voice disorders. J Speech Lang Hear Res 51, 333–349.

Lu J, Nakamura S, Shikano H. (1999). Spectral analysis of esophageal speech. J Acoust Soc Am 105, 1353–1353.

Ludlow C, Bassich C, Connor N, Coulter D, Lee Y. (1987). The validity of using phonatory jitter and shimmer to detect laryngeal pathology. In: Baer T, Sasaki C, Harris K (eds.): Laryngeal function in phonation and respiration. 492–508, San Diego: College–Hill.

Martens J, Versnel H, Dejonckere PH. (2007). The effect of visible speech in the perceptual rating of pathological voices. Arch Otolaryngol Head Neck Surg 133, 178–185

Martin D, Fitch J, Wolfe V.(1995). Pathologic voice type and the acoustic prediction of severity. J Speech Hear Res 38, 765–771.

Maryn Y, Corthals P, Van Cauwenberge P, Roy N, De Bodt M. (2010). Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. J Voice 24, 540–555.

Maryn Y, Paul Corthals, De Bodt M, Van Cauwenberge P, Deliyski D. (2009). Perturbation measures of voice: A comparative study between Multi-Dimensional Voice Program and Praat. Folia Phoniatr Logop 61, 217–226.

Mattys SL, Pleydell-Pearce CW, Melhorn JF, Whitecross SE. (2005). Detecting silent pauses in speech: A new tool for measuring on-line lexical and semantic processing. Psychol Sci 16, 958–964.

Max L, Steurs W, de Bruyn W. (1996). Vocal capacities in esophageal and tracheoesophageal speakers. Laryngoscope 106, 93–96.

Mergell P, Herzel H, Wittenberg T, Tigges M, Eysholdt U. (1998). Phonation onset: vocal fold modeling and high-speed glottography. J Acoust Soc Am 104, 464–470.

Mergell P, Tigges M, Herzel HP,Wittenberg T, Eysholdt U. (1996). Simulation glottaler Biphonation mit dem 2-Massen-Modell. In: Gross M, Eysholdt U. (Hrsg). Aktuelle phoniatrisch-pädaudiologische Aspekte, Bd 4. Phoniatrie Göttingen, 13–14.

Michaelis D, Fröhlich M, Strube HW. (1998). Selection and combination of acoustic features for the description of pathologic voices. J Acoust Soc Am 103, 1628–1639.

Michaelis D. (2000). Das Göttinger Heiserkeits-Diagramm – Entwicklung und Prüfung eines Verfahrens zur objektiven Stimmgütebeurteilung pathologischer Stimmen. Published in electronic form only: http://webdoc.sub.gwdg.de/diss/2000/michaelis/index.html. (last visited: 30th September 2010).

Minifie FD, Kelsey CA, Hixon TJ. (1968). Lateral pharyngeal wall motion during speech. J Acoust Soc Am 44, 353–354.

Mitchinson AG, Yoffey JM. (1947). Respiratory displacement of the larynx, hyoid bone and tongue. J Anat 81, 118–120.

Mitrovic MS. (2003). Comparative analysis of voice in diagnostics of T1 and T2 vocal cord carcinoma. Arch Oncology 11, 239–242.

Moore DS. (1995). The basic practice of statistics. NY: Freeman and Co.

Moore P. (1938). Motion picture studies of the vocal folds and vocal attack. J Speech Disord 3, 235–238.

Mooshammer C. (2010). Acoustic and laryngographic measures of the laryngeal reflexes of linguistic prominence and vocal effort in German. J Acoust Soc Am 127, 1047–1058.

Morris RJ, Brown WS. (1996). Comparison of various automatic means for measuring mean fundamental frequency. J Voice 10, 159–165.

Morris RJ, Brown, WS, Hicks DM, Howell E. (1995). Phonational profiles of male trained singers and nonsingers. J Voice 9, 142–148.

Motta S, Galli I, Di Rienzo L. (2001). Aerodynamic findings in esophageal voice. Arch Otolaryngol Head Neck Surg 127, 700 – 704.

Murry T, Doherty ET. (1980). Selected acoustic characteristics of pathologic and normal speakers. J Speech Hear Res 23, 361–369.

Murry T. (1978). Speaking fundamental frequency characteristics associated with voice pathologies. J Speech Hear Dis 43, 374–379.

Narayanan S, Alwan A. (1995). A nonlinear dynamical systems analysis of fricative consonants. J Acoust Soc Am 97, 2511–2524.

Nawka T, Anders LC, Wendler J. (1994). Die Beurteilung heiserer Stimmen nach dem RBH-System. Sprache–Stimme–Gehoer 18, 130–133.

Neto HN, Marques MJ. (2008). Estimation of the number and size of motor units in intrinsic laryngeal muscles using morphometric methods. Clin Anat 21, 301–306.

Neubauer J, Mergell P, Eysholdt U, Herzel HP. (2001) Spatiotemporal analysis of irregular vocal fold oscillations: biphonation due to desynchronization of spatial modes. J Acoust Soc Am 110, 3179–3192.

Newman RA, Emanuel FW. (1991). Pitch effects on vowel roughness and spectral noise for subjects in four musical voice classifications. J Speech Hear Res 34, 753–760.

Ng ML, Gilbert HR, Lerman JW. (1997). Some aerodynamic and acoustic characteristics of acute laryngitis. J Voice 11, 356–363.

Nittrouer S, McGowan R, Milenkovic P, Beehler D. (1990). Acoustic measurements of men's and women's voices: a study of context effects and covariation. J Speech Hear Res 33, 761–775.

Núñez Batalla F, Suárez Nieto C, Muñoz Pinto C, Baragaño Río L, Alvarez Zapico MJ, Martínez Ferreras A. (2000). Spectrographic study of voice disorders: subharmonics. Acta Otorrinolaringol Esp 51, 52–56.

Olthoff A, Mrugalla S, Laskawi R, Fröhlich M, Stuermer I, Kruse E, Ambrosch P, Steiner W. (2003). Assessment of irregular voices after total and laser surgical partial laryngectomy. Arch Otolaryngol Head Neck Surg 129, 994–999.

Olthoff A, Woywod C, Kruse E. (2007). Stroboscopy versus high-speed glottography: a comparative study. Laryngoscope 117, 1123–1126.

Omori K, Kojima H, Kakani R, Slavit DH, Blaugrund SM. (1997). Acoustic characteristics of rough voice: Subharmonics. J Voice 11, 40–47.

Orlikoff RF, Deliyski DD, Baken RJ, Watson BC. (2009). Validation of a glottographic measure of vocal attack. J Voice 23, 164–168.

Orlikoff RF, Kahane JC. (1991). Influence of mean sound pressure level on jitter and shimmer measures. J Voice 5, 113–119.

Orlikoff RF. (1995). Vocal stability and vocal tract configuration: an acoustic and electroglottographic investigation. J Voice 9, 173–181.

Parsa V, Jamieson DG. (2001). Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech. J Speech Lang Hear Res 44, 327–339.

Pegoraro-Krook M. (1988). Speaking fundamental frequency characteristics of normal Swedish subjects obtained by glottal frequency analysis. Folia Phoniatr 40, 82–90.

Perez KS, Ramig LO, Smith ME, Dromey C. (1996). The Parkinson larynx: tremor and videostroboscopic findings. J Voice 10, 354–361.

Pickett J. (1991). The spectra of vowels. In: Baken J & Daniloff R (eds): Readings in clinical spectrography of speech. 75–95, San Diego: Singular Publishing Group.

Potapova RK. (2002). An approach to enrich parametric database as applied to German connected speech. In: Issues in Phonetics IV. 225–240, Moscow: Nauka.

Ptacek PH, Sander EK. (1963). Maximum duration of phonation. J Speech Hear Disord 28, 171–182.

Pudil P, Novovicova J, Kittler J. (1994). Floating search methods in feature selection. Pattern Recogn Lett 15, 1119–1125.

Rabinov CR, Kreiman J, Gerrat BR, Bielamowicz S. (1995**)**. Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter. J Speech Hear Res 38, 26–32.

Radish Kumar B, Bhat JS. (2008). Aerodynamic analysis of voice in persons with laryngopharyngeal reflux. Online J Health Allied Scs 7, 5.

Rappaport W. (1958). Über Messungen der Tonhöhenverteilung in der deutschen Sprache. Acustica 8, 220–225.

Rau D, Beckett RL. (1984). Aerodynamic assessment of vocal function using hand-held spirometers. J Speech Hear Disord 49, 183–188.

Rees M. (1958). Some variables affecting perceived harshness. J Speech Hear Res 1, 155–168.

Remacle M, Trigaux I. (1991). Characteristics of nodules through the high resolution frequency analyser. Folia Phoniatr 43, 53–59.

Revis J, Giovanni A, Wuyts F, Triglia JM. (1999). Comparison of different voice samples for perceptual analysis. Folia Phoniatr Logop 51, 108–116.

Ritchings T, Berry C. (2006). A comparative study of impedance and acoustic vowel phonation signals for intelligent voice quality assessment of patients recovering from radiotherapy for cancer of the larynx. Acta Acust United Acust 92, 712–716.

Robinson JL, Mandel S, Sataloff RT. (2005). Objective voice measures in nonsinging patients with unilateral superior laryngeal nerve paresis. J Voice 19, 665–667.

Rothenberg M, Mahshie JJ. (1988). Monitoring vocal fold abduction through vocal fold contact area. J Speech Hear Res 31, 338–351.

Rothenberg M. (1992). A multichannel electroglottograph. J Voice 6, 36–43.

Roy N, Merrill R, Gray S, Smith E. (2005). Voice disorders in the general population: Prevalence, risk factors, and occupational impact. Laryngoscope 115, 1988–1995.

Roy N, Merrill RM, Thibeault S, Parsa RA, Gray SD, Smith EM. (2004). Prevalence of voice disorders in teachers and the general population. J Speech Lang Hear Res 47, 281–293.

Roy N, Stemple J, Merril RM, Thomas L. (2007). Epidemiology of voice disorders in the elderly: Preliminary findings. Laryngoscope 117, 628–633.

Saito M, Imagawa H, Sakakibara K, Tayama N, Nibu K, Amatsu M. (2006). High-speed digital imaging and electroglottography of tracheoesophageal phonation by Amatsu's method. Acta Oto-Laryngologica 126, 521–525.

Sakakibara KI, Imagawa H, Kimura M, Tokuda I, Tayama N. (2007). Observation of subharmonic voices. In: Proceedings of the ICA: http://www.sea-acustica.es/WEB_ICA_07/fchrs/papers/mus-06-009.pdf. (last visited: 30th September 2010), Madrid, Spain.

Sama A, Carding PN, Price S, Kelly P, Wilson JA. (2001). The clinical features of functional dysphonia. Laryngoscope 111, 458–463.

Sapienza CM, Cannito MP, Murry T, Branski R, WoodsonG. (2002). Pre-Botox injection and within early stages of post-Botox injection acoustic variations in reading produced by speakers with spasmodic dysphonia. J Speech Lang Hear Res 45, 830–843.

Sapienza CM, Stathopoulos ET, Brown WS. (1997). Speech breathing during reading in women with vocal nodules. J Voice 11, 195–201.

Sapienza CM, Stathopoulos ET. (1994). Respiratory and laryngeal measures of children and women with bilateral vocal nodules. J Speech Hear Res 37, 1229–1243.

Sasaki Y, Okamura H, Yumoto F. (1991). Quantitative analysis of hoarseness using a digital sound spectrograph. J Voice 5, 36–40.

Schaeffer N, Cavallo S, Wall M, Diakow C. (2002). Speech breathing behavior in normal and moderately to severely dysphonic subjects during connected speech. J Med Speech-Lang Pa 10, 1–18.

Scherer RC, Gould WJ, Titze IR, Meyers AD, Sataloff RT. (1988). Preliminary evaluation of selected acoustic and glottographic measures for clinical phonatory function analysis. J Voice 2, 230–244.

Scherer RC, Titze IR. (1982). Vocal fold contact stress during phonation. J Acoust Soc Am 7l(Suppl)S55(A).

Scherer RC, Vail VJ, Guo CG. (1995). Required number of tokens to determine representative voice perturbation values. J Speech Lang Hear Res 38, 1260–1269.

Schönhärl E. (1960). Die Stroboskopie in der praktischen Laryngologie. 22–24, Stuttgart: Thieme.

Schönweiler R, Wübbelt P, Hess M, Ptok M. (2001). Psychoakustische Skalierung akustischer Stimmparameter durch multizentrisch validierte RBH-Bewertung. Laryngorhinootologie 80, 117–122.

Schutte HK, Svec JG, Sram F. (1998). First results of clinical application of videokymography. Laryngoscope 108, 1206–1210.

Schutte HK. (1992). Integrated aerodynamic measurements. J Voice 6, 127–134.

Schwarz R, Hoppe U, Schuster M, Wurzbacher T, Eysholdt U, Lohscheller J. (2006). Classification of unilateral vocal fold paralysis by the analysis of endoscopic digital high-speed recordings. IEEE T Biomed Eng 53, 1099–1108.

Severin F, Bozkurt B, Dutoit T. (2005). HNR extraction in voiced speech, oriented towards voice quality analysis. In: Proceedings of the EUSIPCO: tcts.fpms.ac.be/publications/.../2005/eusipco05_fsbbtd.pdf. (last visited: 30th September 2010), Antalya, Turkey.

Shipp T, Haller RM. (1972). Vertical larynx movement. J Acoust Soc Am 52, 124–124 (A).

Shohet J, Courey MS, Scott M, Ossoff R. (1996). Value of videostroboscopic parameters in differentiating true vocal fold cysts from polyps. Laryngoscope 106, 19–26.

Siebenhaar B. (2008). Sprechgeschwindigkeit. Pilotstudie. Uni Leipzig. http://www.unileipzig.de/~siebenh/kurse/SS08/s_sprechgeschwindigkeit/02_sprechgeschwindigkeit.pdf. (last visited: 30th September 2010).

Södersten M, Lindestad P-Å. (1990). Glottal closure and perceived breathiness during phonation in normally speaking subjects. J Speech Hear Res 33, 601–611.

Solomon NP, Garlitz SJ, Milbrath RL. (2000). Respiratory and laryngeal contributions to maximum phonation duration. J Voice 14, 331–340.

Sperry EE, Klich RJ. (1992). Speech breathing in senescent and younger women during oral reading. J Speech Hear Res 35, 1246–1255.

Stevens KN. (1977) Physics of laryngeal behavior and larynx modes. Phonetica 34, 264–279.

Stevens KN. (1998). Acoustic phonetics. Cambridge, MA: MIT Press.

Stoicheff ML. (1981). Speaking fundamental frequency characteristics of nonsmoking female adults. J Speech Hear Res 24, 437–441.

Strangert E. (1993). Speaking style and pausing. PHONUM 2, 121–137.

Sun X, Xu Y. (2002). Perceived pitch of synthesized voice with alternate cycles. J Voice 16, 443–459.

Sun X. (2000). A pitch determination algorithm based on subharmonic-to-harmonic ratio. In: Proceedings of the ICSLP: 676–679, Beijing, China.

Sun X. (2008). http://www.mathworks.com/matlabcentral/fileexchange/1230–pitch–determination–algorithm. (last visited: 30th September 2010).

Sussman JE, Sapienza C. (1994). Articulatory, developmental, and gender effects on measures of fundamental frequency and jitter. J Voice 8, 145–156.

Svec JG, Schutte HK, Miller DG. (1996). A subharmonic vibratory pattern in normal vocal folds. J Speech Hear Res 39, 135–143.

Tanaka S, Hirano M, Terasawa R. (1991). Examination of air usage during phonation: Correlations among test parameters. J Voice 5, 106–112.

Thomas IB. (1969). Perceived pitch of whispered vowels. J Acoust Soc Am 46, 468–470.

Tigges M, Wittenberg T, Mergell P, Eysholdt U. (1999). Imaging of vocal fold vibration by digital multi-plane kymography. Comput Med Imag Grap 23, 323–330.

Till JA, Crumley RL, Jafari M, Law-Till C. (1994). Aerodynamic and temporal disruptions of speech in laryngeal insufficiency. Arch Otolaryngol Head Neck Surg 120, 317–325.

Titze IR, Baken RJ, Herzel H. (1993). Evidence of chaos in vocal fold vibration. In: Titze IR (ed.): Vocal fold physiology: New frontiers in basic science. 143–188, San Diego: Singular Publishing Group.

Titze IR. (1990). Interpretation of the electroglottographic signal. J Voice 4, 1–9.

Titze IR. (1994). Mechanical stress in phonation. J Voice 8, 99–105.

Titze IR. (1995). Workshop on acoustic voice analysis: Summary statement. National Center for Voice and Speech, Iowa City.

Traunmüller H, Eriksson A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. Unpublished manuscript. http://www.ling.su.se/staff/hartmut/aktupub.htm. (last visited: 30th September 2010).

Trouvain J. (1999). Phonological aspects of reading task strategies. Phonus 4, 15–35.

Tsao YC, Weismer G. (1997). Interspeaker variation in habitual speaking rate: evidence for a neuromuscular component. J Speech Lang Hear Res 40, 858–866.

Vasilakis M, Stylianou Y. (2009). Voice pathology detection based on short-term jitter estimations in running speech. Folia Phoniatr Logop 61, 153–170.

Verdolini K, Chan R, Titze IR, Hess M, Bierhals W. (1998a). Correspondence of electroglottographic closed quotient to vocal fold impact stress in excised canine larynges. J Voice 12, 415–423.

Verdolini K, Druker DG, Palmer PM, Samawi H. (1998b). Laryngeal adduction in resonant voice. J Voice 12, 315–327.

Verdonck-de Leeuw I, Mahieu H. (2004). Vocal aging and the impact on daily life: A longitudinal study. J Voice 18, 193–202.

Vieira MN, McInnes FR, Jack MA. (1996). Robust F0 and jitter estimation in pathological voices. In: Proceedings of the ICSLP: 745–748, Philadelphia, USA.

Vieira MN, McInnes FR, Jack MA. (1997). Comparative assessment of electroglottographic and acoustic measures of jitter in pathological voices. J Speech Lang Hear Res 40, 170–182.

Vieira MN, McInnes FR, Jack MA. (2002). On the influence of laryngeal pathologies on acoustic and electroglottographic jitter measures. J Acoust Soc Am 111, 1045–1055.

Voigt D, Döllinger M, Braunschweig T, Yang A, Eysholdt, Lohscheller J. (2010). Classification of functional voice disorders based on phonovibrograms. Artif Intell Med 49, 51–59.

Voigt D, Döllinger M, Yang A, Eysholdt, Lohscheller J. (2010). Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods. Comput Methods Programs Biomed 99, 275–288.

Wang W, Jones P, Partridge D. (1998). Ranking pattern recognition features for neural networks. In: Singh S. (Hrsg.): Advances in Pattern Recognition. 232–241, Berlin: Springer.

Weismer G. (1985). Speech breathing: Contemporary views and findings. In: Daniloff R (ed.): Speech science. 47–72, San Diego: College Hill Press.

Wendler J, Seidner W. (1997). Die Sängerstimme: phoniatrische Grundlagen der Gesangsausbildung. Berlin.

Wendler J. (2005). Stroboskopie. In: Wendler J, Seidner W, Eysholdt U. (Hrsg.): Lehrbuch der Phoniatrie und Pädaudiologie. 117, Stuttgart: Thieme.

Werth K, Voigt D, Döllinger M, Eysholdt U, Lohscheller J. (2010). Clinical value of acoustic voice measures: a retrospective study. Eur Arch Oto-Rhino-L 267, 1261–1271.

Whitehurst GJ. (1984). Interrater agreement for journal manuscript reviews. Am Psychol 39, 22–28.

Winkler R, Sendlmeier W. (2006). EGG open quotient in aging voices – changes with increasing chronological age and its perception. LPV 31, 51–56.

Winkworth AL, Davis PJ, Ellis E, Adams RD. (1994). Variability and consistency in speech breathing during reading: lung volumes, speech intensity, and linguistic factors. J Speech Hear Res 37, 535–556.

Wittenberg T, Tigges M, Mergell P, Eysholdt U. (2000). Functional imaging of vocal fold vibration: digital multislice high-speed kymography. J Voice 14, 422–442.

Wolf A, Swift J, Swinney H, Vastano J. (1985). Determining Lyapunov exponents from a time series. Physica 16D, 285–317.

Wolfe V, Cornell R, Fitch J. (1995a). Sentence/vowel correlation in the evaluation of dysphonia. J Voice 9, 297–303.

Wolfe V, Fitch J, Cornell R. (1995b). Acoustic prediction of severity in commonly occurring voice problems. J Speech Hear Res 38, 273–279.

Wolfe VI, Ratusnik DL. (1988). Acoustic and perceptual measurements of roughness influencing judgments of pitch. J Speech Hear Res 53, 15–22.

Wong D, Ito M, Cox N, Titze IR. (1991). Observation of perturbation in a lumped element model of the vocal folds with application to some pathological cases. J Acoust Soc Am 89, 383–394.

Wurzbacher T, Schwarz R, Döllinger M, Hoppe U, Eysholdt U, Lohscheller J. (2006). Model based classification of non-stationary vocal fold vibrations. J Acoust Soc Am 120, 1012–1027.

Wuyts FL, De Bodt MS, Molenberghs G, Remacle M, Heylen L, Millet B, Lierde KV, Raes J, Van de Heyning PH. (2000). The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approch. J Speech Hear Res 43, 796–809.

Yamaguchi H, Shrivastav R, Andrews ML, Niimi S. (2003). A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. Folia Phoniatr Logop 55, 147–157.

Yanagihara, N. (1967). Significance of harmonic changes and noise components in hoarseness. J Speech Hear Res 10, 531–541.

Yang A, Lohscheller J, Berry DA, Becker S, Eysholdt U, Voigt D, Döllinger M. (2010). Biomechanical modeling of the three-dimensional aspects of human vocal fold dynamics. J Acoust Soc Am 127, 1014–1031.

Yu P, Garrel R, Nicollas R, Oaknine M, Giovanni A. (2007). Objective voice analysis in dysphonic patients: New data including nonlinear measurements. Folia Phoniatr Logop 59, 20–30.

Yu P, Ouaknine M, Revis J, Giovanni A. (2001). Objective voice analysis for dysphonic patients: A multiparametric protocol including acoustic and aerodynamic measurements. J Voice 15, 529–542.

Yu P, Revis J, Wuyts FL, Zanaret M, Giovanni A. (2002). Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. Folia Phoniatr Logop 54, 271–281.

Yumoto E, Gould WJ, Baer T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. J Acoust Soc Am 71, 1544–1550.

Zellner B. (1994). Pauses and the temporal structure of speech. In: Keller E (ed.): Fundamentals of speech synthesis and speech recognition. 41–62, Chichester: John Wiley.

Zwirner P, Michaelis D, Kruse E. (1996). Acoustic voice analysis. On documentation of voice rehabilitation after laser surgery laryngeal carcinoma resection. HNO 44, 514–520.

Zyski B, Bull G, Mc Donald W, John M. (1984). Perturbation analysis of normal and pathologic larynges. Folia Phoniatr 36, 190–198.

Appendix A

Nordwind und Sonne

Einst stritten sich Nordwind und Sonne, wer von ihnen beiden wohl der Stärkere wäre, als ein Wanderer, der in einen warmen Mantel gehüllt war, des Weges kam. Sie wurden einig, daß derjenige für den Stärkeren gelten sollte, der den Wanderer zwingen würde, seinen Mantel abzunehmen. Der Nordwind blies mit aller Macht, aber je mehr er blies, desto fester hüllte sich der Wanderer in seinen Mantel ein. Endlich gab der Nordwind den Kampf auf. Nun erwärmte die Sonne die Luft mit ihren freundlichen Strahlen, und schon nach wenigen Augenblicken zog der Wanderer seinen Mantel aus. Da mußte der Nordwind zugeben, daß die Sonne von ihnen (beiden) der Stärkere war.

Appendix B

Correlations and mean difference (%) between the 1st and the 2nd measurement in vowels

Variable	*/a/*		*/e/*	
	Pearson's r	*Mean distance (%)*	*Pearson's r*	*Mean distance (%)*
Intensity	0.93	5	0.95	5
Jitter	0.62	13	0.85	14
Shimmer	0.63	43	0.51	48
EGG Jitter	0.88	216	0.76	724
EGG Shimmer	0.86	39	0.81	53
FMF	0.47	176	0.58	195
EGG FMF	0.53	96	0.45	253
IC	0.87	9	0.80	15
OQ	0.69	75	0.75	26
GNE	0.86	13	0.82	14
HNR	0.90	34	0.92	14
LTAS	0.93	5	0.95	5
LLE	0.36	205	0.17	388
AI	0.82	32	0.81	20
SHR	0.73	712	0.76	795

Appendix C

Group means and standard deviations (in parentheses) for objective voice measures as a function of R, B and H followed by corresponding Spearman's rank correlation coefficients and the number of significant contrasts according to Mann-Whitney U-test statistics. Data is stratified by vowel, signal or sex.

Variable	*Voice Quality*	*Vowel*	*Mean (SD) Grade 0*	*Mean (SD) Grade 1*	*Mean (SD) Grade 2*	*Mean (SD) Grade 3*	r_s	*Number of significant contrasts*
Int (dB)	R	/a/	69.9 (4.6)	71.8 (5.2)	71.3 (5.3)	71.3 (5.1)	ns	1
		/e/	70.5 (5.8)	72.1 (5.0)	71.7 (5.9)	70.6 (4.6)	ns	0
	B	/a/	72.9 (4.7)	71.8 (4.7)	70.2 (5.6)	70.5 (5.9)	-0.13	0
		/e/	73.1 (4.6)	72.1 (4.8)	70.3 (5.6)	70.5 (7.5)	-0.15	1
	H	/a/	71.9 (3.9)	72.1 (4.6)	71.2 (5.3)	70.7 (5.7)	ns	0
		/e/	73.4 (4.4)	72.1 (4.9)	71.7 (5.1)	70.3 (6.3)	-0.11	0
Jitter (%)	R	/a/	0.11 (0.07)	0.39 (.7)	0.65 (0.87)	1.35 (1.24)	0.54	3
		/e/	0.10 (0.07)	0.21 (0.19)	0.52 (1.12)	0.91 (1.06)	0.44	3
	B	/a/	0.32 (0.74)	0.46 (0.73)	0.75 (1,07)	0.87 (0.86)	0.31	3
		/e/	0.12 (0.07)	0.28 (0.38)	0.60 (1.30)	0.57 (0.65)	0.35	2
	H	/a/	0.10 (0.06)	0.32 (0.68)	0.51 (0.74)	1.10 (1.15)	0.51	3
		/e/	0.10 (0.07)	0.16 (0.11)	0.41 (0.98)	0.71 (0.87)	0.45	2
EGG Jitter (%)	R	/a/	10. 8 (38.5)	14.5 (36.2)	18.1 (34.9)	21.7 (32.4)	0.35	2
		/e/	10.3 (33.4)	13.3 (32.9)	15.1 (31.3)	21.6 (36.4)	0.30	1
	B	/a/	13.4 (37.8)	11.4 (33.2)	17.5 (27.8)	41.9 (52.9)	0.30	2
		/e/	11.8 (32.6)	10.7 (27.4)	14.2 (30.1)	40.1 (53.1)	0.22	2
	H	/a/	26.1 (61.2)	7.4 (22.4)	16.2 (36.1)	27.4 (41.4)	0.39	2
		/e/	22.3 (52.7)	8.7 (27.2)	12.0 (24.7)	26.5 (45.0)	0.30	2
Shimmer (%)	R	/a/	5.4 (3.4)	7.7 (4.5)	9.5 (4.6)	12.8 (6.0)	0.39	3
		/e/	3.6 (2.0)	4.9 (3.7)	6.5 (4.5)	10.7 (5.6)	0.38	2
	B	/a/	6.5 (3.7)	6.9 (3.8)	10.7 (4.8)	14.7 (6.1)	0.45	2
		/e/	3.5 (2.1)	4.7 (3.1)	7.0 (3.9)	13.1 (7.2)	0.46	2
	H	/a/	6.6 (5.8)	6.1 (3.1)	8.6 (4.1)	12.9 (6.0)	0.47	2
		/e/	3.5 (2.6)	3.9 (2.0)	5.4 (3.3)	10.7 (6.2)	0.47	2
EGG Shimmer (%)	R	/a/	15.0 (23.2)	10.9 (18.6)	25.6 (18.7)	26.3 (16.4)	0.26	2
		/e/	11.9 (14.5)	19.1 (16.1)	25.0 (18.8)	29.5 (17.1)	0.30	2
	B	/a/	22.8 (24.4)	18.5 (17.1)	25.0 (18.1)	35.8 (19.1)	0.25	2
		/e/	17.6 (15.5)	18.4 (17.1)	25.3 (17.7)	32.3 (17.3)	0.26	2
	H	/a/	21.8 (36.1)	15.8 (14.5)	24.2 (19.8)	29.3 (17.2)	0.33	2
		/e/	14.0 (19.5)	15.1 (13.8)	23.7 (18.1)	28.3 (18.2)	0.32	1
FMF (%)	R	/a/	1.6 (5.0)	5.0 (9.5)	11.5 (14.5)	29.3 (31.1)	0.58	3
		/e/	1.7 (4.7)	3.5 (7.9)	10.7 (21.4)	23.0 (26.9)	0.46	3
	B	/a/	3.2 (8.1)	7.6 (13.6)	13.1 (22.6)	18.7 (17.1)	0.33	3
		/e/	0.8 (0.3)	5.8 (11.7)	12.7 (23.9)	14.6 (25.3)	0.34	2
	H	/a/	0.7 (0.2)	3.7 (8.2)	8.4 (13.3)	22.5 (25.5)	0.53	3
		/e/	0.8 (0.2)	2.6 (6.1)	7.7 (17.9)	17.9 (24.3)	0.46	2
IC	R	/a/	3.94 (0.85)	4.95 (1.23)	5.81 (1.34)	7.23 (1.46)	0.56	3
		/e/	3.35 (0.82)	4.27 (1.28)	4.99 (1.66)	6.45 (1.59)	0.47	3
	B	/a/	4.41 (0.91)	4.97 (1.36)	6.03 (1.40)	7.05 (1.51)	0.49	3
		/e/	3.68 (0.69)	4.25 (1.41)	5.25 (1.55)	6.44 (1.93)	0.46	3
	H	/a/	3.8 (0.8)	4.5 (0.9)	5.4 (1.2)	7.0 (1.4)	0.61	3
		/e/	3.4 (0.6)	3.7 (0.8)	4.7 (1.5)	6.2 (1.7)	0.55	2
OQ	R	/a/	0.64 (0.15)	0.62 (0.16)	0.62 (0.17)	0.63 (0.14)	ns	0
		/e/	0.65 (0.15)	0.63 (0.16)	0.61 (0.15)	0.63 (0.17)	ns	0
	B	/a/	0.66 (0.14)	0.60 (0.15)	0.63 (0.17)	0.70 (0.19)	ns	0
		/e/	0.64 (0.14)	0.61 (0.14)	0.62 (0.16)	0.70 (0.23)	ns	1
	H	/a/	0.73 (0.13)	0.61 (0.14)	0.61 (0.17)	0.65 (0.16)	ns	1
		/e/	0.70 (0.15)	0.62 (0.14)	0.61 (0.17)	0.65 (0.18)	ns	0
GNE	R	/a/	0.78 (0.14)	0.66 (0.19)	0.59 (0.19)	0.53 (0.20)	-0.34	2
		/e/	0.82 (0.14)	0.7 (0.19)	0.62 (0.20)	0.57 (0.21)	-0.34	2
	B	/a/	0.80 (0.12)	0.72 (0.15)	0.51 (0.17)	0.39 (0.17)	-0.61	3
		/e/	0.84 (0.12)	0.75 (0.16)	0.57 (0.18)	0.41 (0.19)	-0.58	3
	H	/a/	0.85 (0.08)	0.73 (0.15)	0.61 (0.19)	0.50 (0.19)	-0.47	3
		/e/	0.85 (0.12)	0.78 (0.16)	0.64 (0.18)	0.56 (0.22)	-0.45	2
HNR (dB)	R	/a/	20.2(4.2)	15.3 (5.3)	12.1 (5.0)	7.8 (6.3)	-0.52	3
		/e/	22.6 (3.8)	18.2 (5.2)	16.6 (6.3)	11.6 (7.1)	-0.40	3
	B	/a/	17.5 (4.1)	15.8 (5.4)	11.3 (5.2)	6.5 (6.1)	-0.50	2
		/e/	20.4 (4.3)	19.3 (4.8)	15.2 (5.6)	9.4 (8.6)	-0.43	2
	H	/a/	20.2 (4.5)	17.5 (4.0)	13.4 (5.0)	8.1 (6.1)	-0.59	3
		/e/	22.0 (4.4)	20.2 (3.6)	17.8 (5.5)	11.5 (7.1)	-0.49	3

Variable	*Voice Quality*	*Vowel*	*Mean (SD) Grade 0*	*Mean (SD) Grade 1*	*Mean (SD) Grade 2*	*Mean (SD) Grade 3*	r_s	*Number of significant contrasts*
LTAS	R	/a/	36.8 (4.6)	38.7 (5.3)	38.2 (5.3)	38.3 (5.1)	ns	1
		/e/	37.4 (5.8)	38.8 (5.1)	38.4 (6.1)	37.1 (4.5)	ns	0
(dB/Hz)	B	/a/	39.9.(4.7)	38.7 (4.7)	37.2 (5.6)	37.2 (6.2)	-0.14	0
		/e/	39.7 (4.7)	38.9 (4.8)	37.1 (5.6)	36.5 (7.8)	-0.16	1
	H	/a/	38.9 (3.9)	39.0 (4.6)	38.1 (5.3)	37.5 (5.9)	ns	0
		/e/	40.1 (4.6)	38.8 (4.9)	38.5 (5.2)	36.7 (6.5)	-0.12	0
LLE (bit/s)	R	/a/	276 (265)	437 (565)	391 (376)	633 (488)	0.19	1
		/e/	416 (474)	514 (495)	516 (545)	920 (1171)	0.12	1
	B	/a/	417 (460)	362 (397)	456 (408)	721 (881)	0.14	1
		/e/	588 (610)	457 (383)	609 (683)	860 (1276)	ns	0
	H	/a/	298 (293)	363 (414)	387 (397)	626 (666)	0.21	1
		/e/	416 (600)	490 (467)	458 (422)	849 (1020)	0.14	1
AI	R	/a/	193 (79)	326 (130)	434 (153)	533 (129)	0.57	3
		/e/	84 (49)	143 (89)	166 (103)	228 (142)	0.33	2
	B	/a/	323 (166)	354 (157)	390 (158)	466 (166)	0.2	1
		/e/	119 (78)	146 (88)	153 (90)	250 (180)	0.2	2
	H	/a/	251 (141)	289 (130)	384 (150)	493 (150)	0.48	2
		/e/	117 (94)	118 (68)	154 (93)	217 (137)	0.35	2
SHR	R	/a/	0.06 (0.17)	0.15 (0.23)	0.18 (0.23)	0.24 (0.17)	0.27	2
		/e/	0.18 (0.26)	0.15 (0.25)	0.16 (0.23)	0.15 (0.22)	0.11	1
	B	/a/	0.19 (0.26)	0.17 (0.24)	0.14 (0.19)	0.15 (0.16)	ns	0
		/e/	0.25 (0.28)	0.17 (0.26)	0.12 (0.21)	0.08 (0.12)	-0.14	0
	H	/a/	0.07 (0.14)	0.14 (0.24)	0.16 (0.24)	0.22 (0.17)	0.24	1
		/e/	0.29 (0.27)	0.16 (0.27)	0.14 (0.23)	0.15 (0.20)	ns	1
F0 SD	R	Sp	3.2 (2.4)	4.2 (2.4)	7.3 (5.6)	12.7 (10.1)	0.52	3
		Lx	3.4 (1.2)	4.4 (2.7)	5.9 (5.2)	9.2 (6.9)	0.43	2
(st)	B	Sp	4.9 (3.1)	5.1 (4.1)	7.3 (7.6)	9.9 (7.9)	0.17	1
		Lx	3.7 (1.3)	4.9 (4.2)	5.5 (3.1)	9.6 (8.6)	0.24	2
	H	Sp	4.7 (3.6)	3.7 (1.9)	5.9 (4.9)	10.2 (8.6)	0.40	2
		Lx	4.5 (1.2)	3.9 (1.9)	4.7 (2.8)	8.9 (7.5)	0.36	3
FMF (%)	R	Sp	16.2 (9.6)	20.8 (9.9)	31.6 (17.1)	46.7 (25.7)	0.52	3
		Lx	18.1 (5.5)	21.5 (10.6)	26.8 (14.3)	40.2 (23.1)	0.43	2
	B	Sp	23.9 (11.6)	23.7 (13.5)	30.2 (21.8)	38.7 (22.1)	0.17	1
		Lx	19.2 (5.8)	23.1 (11.9)	26.3 (11.9)	39.7 (29.6)	0.24	2
	H	Sp	22.7 (13.9)	19.2 (8.2)	26.6 (16.1)	39.9 (23.0)	0.40	2
		Lx	22.8 (5.4)	19.8 (8.0)	22.9 (10.8)	37.4 (22.2)	0.36	3
PLF (%)	R	Sp	1.4 (1.3)	3.8 (6.3)	15.4 (24.7)	21.1 (23.1)	0.46	1
		Lx	2.5 (2.4)	7.2 (10.8)	14.6 (25.2)	23.8 (25.9)	0.37	2
	B	Sp	3.5 (5.2)	10.7 (20.6)	11.6 (18.6)	3.9 (6.6)	ns	2
		Lx	3.4 (3.9)	12.1 (21.4)	13.2 (19.4)	9.4 (17.6)	ns	1
	H	Sp	1.1 (0.3)	4.3 (6.4)	9.6 (19.3)	18.3 (24.6)	0.24	1
		Lx	3.2 (3.1)	6.2 (10.0)	11.4 (20.3)	19.4 (26.0)	0.25	2
80R (st)	R		5.6 (1.7)	6.6 (3.0)	9.3 (4.5)	16.6 (8.9)	0.48	2
	B		7.2 (2.7)	7.8 (4.0)	10.1 (7.6)	9.9 (7.1)	ns	0
	H		7.3 (2.4)	6.3 (2.3)	7.8 (3.9)	13.5 (8.3)	0.31	1
IFx (%)	R	Sp	13.2 (6.1)	14.88 (4.8)	18.29 (5.4)	27.2 (6.3)	0.53	2
		Lx	22.5 (14.4)	23.6 (11.5)	27.4 (13.3)	38.3 (11.7)	0.39	2
	B	Sp	16.7 (5.6)	15.7 (5.6)	18.8 (7.2)	21.7 (9.1)	0.19	1
		Lx	25.5 (12.1)	22.1 (7.5)	29.6 (12.9)	42.7 (22.6)	0.28	2
	H	Sp	19.0 (8.7)	13.8 (3.5)	16.6 (5.2)	23.3 (8.2)	0.39	3
		Lx	31.4 (17.4)	20.9 (7.4)	24.5 (9.4)	37.6 (18.1)	0.36	3
Jitter (%)	R	Sp	4.7 (3.8)	6.5 (4.1)	8.9 (4.1)	19.5 (13.3)	0.50	3
		Lx	35.6 (35.9)	35.5 (35.1)	46.1 (38.8)	62.7 (44.9)	0.23	2
	B	Sp	7.1 (2.9)	7.4 (5.2)	9.4 (9.0)	15.3 (10.7)	0.19	1
		Lx	32.0 (31.8)	32.1 (27.6)	49.7 (39.4)	88.4 (54.6)	0.35	2
	H	Sp	6.7 (4.3)	5.6 (3.1)	7.8 (4.1)	15.2 (11.5)	0.42	2
		Lx	32.9 (47.1)	28.7 (21.7)	40.6 (35.0)	67.1 (50.1)	0.34	2
Shimmer (%)	R	Sp	11.7 (3.9)	15.1 (5.7)	19.2 (5.8)	31.2 (9.6)	0.62	3
		Lx	28.1 (11.3)	33.4 (14.2)	38.4 (15.6)	43.4 (13.5)	0.28	1
	B	Sp	14.6 (4.1)	16.2 (5.6)	19.8 (9.6)	27.3 (11.7)	0.30	2
		Lx	31.0 (14.9)	32.6 (13.3)	40.3 (15.7)	44.6 (13.6)	0.30	1
	H	Sp	13.7 (5.1)	14.1 (3.8)	16.9 (5.4)	26.9 (10.8)	0.49	2
		Lx	29.9 (11.4)	29.9 (12.6)	36.7 (15.3)	43.3 (14.0)	0.33	2
MPT (s)	R		16.8 (5.6)	16.8 (6.6)	13.2 (5.7)	12.3 (6.4)	-0.30	1
	B		18.4 (5.9)	16.2 (6.7)	13.12 (5.6)	11.0 (4.7)	-0.35	2
	H		21.6 (5.1)	17.3 (6.6)	14.9 (5.8)	11.0 (5.5)	-0.42	3

Variable	*Voice Quality*	*Vowel*	*Mean (SD) Grade 0*	*Mean (SD) Grade 1*	*Mean (SD) Grade 2*	*Mean (SD) Grade 3*	r_s	*Number of significant contrasts*
PQ (ml/s)	R		185 (43)	210 (110)	232 (99)	265 (150)	ns	1
	B		189 (43)	199 (91)	236 (103)	338 (181)	0.29	2
	H		173 (39)	181 (48)	230 (127)	272 (122)	0.29	2
VC (ml)	R	m	3200(1151)	3551 (851)	2934 (828)	3127 (962)	ns	1
		f	2881 (680)	2567 (759)	2350 (851)	2071 (811)	-0.29	
	B	m	4100 (690)	3326 (819)	2942 (784)	3264(1150)	-0.24	0
		f	2790 (863)	2490 (780)	2328 (806)	-	ns	
	H	m	5000 (0)	3442 (800)	3331 (814)	2830 (939)	-0.30	0
		f	3125 (899)	2610 (690)	2454 (846)	2060 (796)	-0.28	
Int (dB)	R		67.9 (3.6)	68.8 (4.4)	68.3 (4.3)	68.7 (3.8)	ns	0
	B		68.6 (4.4)	67.9 (4.3)	69.4 (4.3)	69.6 (2.9)	ns	1
	H		66.4 (2.8)	68.3 (4.3)	68.4 (4.6)	69.4 (3.5)	ns	0
OQ	R		0.75 (0.17)	0.77 (0.18)	0.79 (0.18)	0.85 (0.15)	ns	0
	B		0.72 (0.18)	0.75 (0.18)	0.83 (0.16)	0.92 (0.1)	0.34	2
	H		0.73 (0.18)	0.75 (0.18)	0.77 (0.18)	0.87 (0.15)	0.24	1
Rtime (s)	R		46.6 (11.4)	46.3 (7.6)	48.7 (6.4)	51.8 (9.5)	0.24	1
	B		42.8 (6.4)	46.9 (7.8)	50.6 (7.1)	51.2 (10.3)	0.33	2
	H		40.8 (7.9)	45.4 (6.3)	47.7 (7.5)	52.9 (9.1)	0.36	1
Ptime (s)	R		9.9 (7.1)	8.5 (3.9)	9.8 (3.6)	11.5 (4.4)	0.25	1
	B		6.6 (2.0)	8.5 (3.5)	10.5 (3.5)	14.8 (7.4)	0.46	3
	H		6.8 (2.4)	7.8 (3.1)	9.2 (3.6)	12.7 (5.6)	0.42	2
Npauses	R		16.6 (8.7)	15.5 (5.8)	16.6 (5.3)	19.5 (6.3)	0.20	0
	B		13.6 (3.9)	14.8 (5.4)	18.2 (4.9)	24.1 (8.4)	0.44	2
	H		14.6 (5.6)	14.5 (5.0)	15.9 (5.5)	20.8 (7.0)	0.35	1
P/100 syl	R		9.1 (4.7)	8.5 (3.3)	9.3 (3.4)	10.7 (3.5	0.20	2
	B		6.6 (2.1)	8.5 (3.5)	10.6 (3.6)	14.8 (7.5)	0.43	2
	H		6.8 (2.9)	7.8 (3.1)	9.2 (3.6)	12.7 (5.6)	0.35	2
P (%)	R		0.19 (0.05)	0.17 (0.05)	0.19 (0.05)	0.21 (0.06)	0.21	1
	B		0.15 (0.03)	0.18 (0.05)	0.21 (0.05)	0.28 (0.09)	0.46	3
	H		0.16 (0.04)	0.17 (0.04)	0.19 (0.05)	0.23 (0.07)	0.39	2
Plength	R		0.57 (0.13)	0.54 (0.08)	0.59 (0.11)	0.58 (0.09)	0.17	1
	B		0.48 (0.09)	0.58 (0.11)	0.58 (0.08)	0.59 (0.11)	0.22	1
	H		0.47 (0.09)	0.53 (0.1)	0.58 (0.1)	0.60 (0.1)	0.30	2
S/P	R		4.5 (1.7)	5.1 (1.6)	4.4 (1.7)	3.9 (1.5)	-0.21	2
	B		5.9 (1.5)	5.0 (1.6)	4.2 (1.3)	3.1 (1.8)	-0.46	3
	H		5.4 (1.9)	5.3 (1.5)	4.7 (1.6)	3.7 (1.6)	-0.39	2
Sylbp	R		13.1 (5.2)	13.1 (4.3)	12.2 (4.7)	10.3 (3.4)	-0.20	2
	B		14.4 (4.1)	13.7 (4.6)	10.7 (2.9)	8.7 (4.5)	-0.43	2
	H		13.9 (5.1)	13.8 (4.1)	12.7 (4.5)	9.8 (3.8)	-0.35	2
Srate	R		4.1 (0.9)	4.0 (0.6)	3.8 (0.5)	3.6 (0.6)	-0.25	1
(syl/s)	B		4.3 (0.6)	3.9 (0.7)	3.7 (0.5)	3.7 (0.8)	-0.33	2
	H		4.6 (0.7)	4.1 (0.6)	3.8 (0.6)	3.5 (0.6)	-0.37	3
Arate	R		5.1 (0.9)	4.8 (0.6)	4.7 (0.5)	4.6 (0.8)	-0.18	1
(syl/s)	B		5.1 (0.7)	4.8 (0.7)	4.6 (0.5)	5.0 (0.6)	ns	3
	H		5.5 (0.7)	4.9 (0.6)	4.8 (0.6)	4.6 (0.6)	-0.23	1

Appendix D

Results for the analysis of variance for the group median data as a function of R, B and H.

Measure	*Mann-Whitney-U*			*Mann-Whitney-U*			*Mann-Whitney-U*		
	z			*z*			*z*		
	Roughness			*Breathiness*			*Hoarsenes*		
	0/1	1/2	2/3	0/1	1/2	2/3	0/1	1/2	2/3
Ji /a/	-5.1 **	-4.3 **	-4.1 **	-2.5 *	-2.4 *	-2.2 *	-2.5 *	-4.6 **	-4.9 **
Ji /e/	-4.5 **	-2.9 **	-3.4 **	-3.3 **	-3.0 **	-1.5	-1.8	-4.1 **	-4.1 **
Shi /a/	-2.8 **	-3.6 **	-2.8 **	-0.3	-6.2 **	-3.1 **	-0.6	-5.0 **	-4.8 **
Shi /e/	-1.8	-2.9 **	-4.0 **	-2.5 *	-5.1 **	-3.6 **	-1.2	-3.0 **	-6.1 **
EGG Ji /a/	-3.5 **	-2.8 **	-1.4	-0.5	-3.6 **	-2.3 *	-1.0	-3.8 **	-3.4 **
EGG Ji /e/	-3.6 **	-1.7	-1.8	-0.6	-2.0 *	-2.3 *	-2.0 *	-2.8 **	-2.1 *
EGG Shi /a/	-2.5 **	-2.3 *	-0.4	0.5	-3.1 **	-2.6 **	-0.8	-3.2 **	-2.3 *
EGG Shi /e/	-2.6 **	-2.4 *	-1.5	0.2	-3.5 **	-2.0 *	-0.7	-3.7 **	-1.8
LTAS /a/	-1.9 *	0.9	-0.4	1.1	1.7	-0.1	0.1	1.0	-0.2
LTAS /e/	-1.2	0.1	1.4	0.8	2.2 *	0.3	0.7	0.3	1.7
HNR /a/	4.4 **	5.1 **	3.6 **	1.4	6.1 **	4.1 **	2.0 *	6.5 **	5.8 **
HNR /e/	4.6 **	2.0 *	4.1 **	1.4	5.3 **	3.2 **	2.0 *	3.9 **	5.9 **
GNE /a/	3.3 **	2.9 **	1.6	3.0 **	7.9 **	3.7 **	2.7 **	4.5 **	3.8 **
GNE /e/	3.1 **	3.1 **	1.1	3.4 **	7.0 **	4.1 **	1.4	5.8 **	2.4 *
FMF /a/	-4.6 **	-5.5 **	-4.2 **	-3.3 **	-2.1 *	-2.4 *	-2.2 *	-4.5 **	-5.3 **
FMF /e/	-3.0 **	-3.7 **	-4.1 **	-4.8 **	-2.4 *	-1.0	-1.8	-3.7 **	-4.7 **
Int /a/	-1.9 *	0.9	-0.3	1.0	1.7	-0.2	0.1	1.0	0.2
Int /e/	-1.3	0.2	0.9	1.1	2.2 *	-0.3	0.9	0.4	1.2
IC /a/	-4.3 **	-5.1 **	-4.5 **	-2.1 *	-5.8 **	-3.2 **	-2.1 *	-5.7 **	-6.8 **
IC /e/	-4.3 **	-3.5 **	-4.5 **	-2.1 *	-5.4 **	-3.0 **	-1.5	-5.2 **	-5.9 **
LLE /a/	-1.4	-0.2	-3.3 **	0.8	-2.1 *	-1.3	-0.4	-0.6	-3.5 **
LLE /e/	-1.5	0.3	-2.3 *	0.6	-0.5	-0.9	-1.4	0.3	-2.8 **
AI /a/	-5.2 **	-5.2 **	-3.5 **	-0.8	-1.6	-2.1 *	-1.1	-4.6 **	-4.7 **
AI /e/	-4.2 **	-1.5	-2.7 **	-2.1 *	-0.5	-2.8 **	-0.6	-2.9 **	-3.5 **
OQ /a/	-0.8	-0.3	0.5	-1.8	1.3	1.5	-2.6 **	-0.1	1.3
OQ /e/	-1.5	-0.3	0.2	-1.1	0.8	2.6 **	-1.7	-0.2	1.7
NC /a/	-3.1 **	-2.9 **	-1.7	-2.8 **	-8.0 **	-3.6 **	-2.6 **	-4.6 **	-3.9 **
NC /e/	-2.8 **	-3.4 **	-1.0	-3.5 **	-7.2 **	-3.7 **	-1.3	-5.7 **	-2.5 *
SHR /a/	-1.7	-2.0 *	-2.3 *	0.7	0.3	-0.9	-0.4	-0.8	-3.6 **
SHR /e/	1.1	-2.1 *	-0.4	1.3	1.1	0.7	1.9 *	0.4	-1.9
FMF	-3.6 **	-5.2 **	-3.4 **	0.5	-1.9	-2.0 *	0.3	-3.2 **	-4.6 **
EGG FMF	-1.3	-3.7 **	-4.8 **	-1.6	-2.4 *	-1.9 *	2.3 *	-2.2 *	-5.7 **
F0 SD	-3.6 **	-5.2 **	-3.4 **	0.5	-1.9	-2.0 *	0.3	-3.2 **	-4.6 **
EGG F0 SD	-1.2	-3.7 **	-4.8 **	-1.6	-2.4 *	-1.9 *	2.3 *	-2.2 *	-5.7 **
IFx	-1.6	-5.3 **	-6.4 **	1.4	-3.3 **	-1.4	2.2 *	-3.7 **	-5.4 **
EGG IFx	-1.2	-3.1 **	-5.2 **	1.4	-4.5 **	-2.5 *	2.3 *	-2.9 **	-5.1 **
80R	-1.7	-4.7 **	-4.6 **	0.6	-1.8	0.5	1.6	-1.8	-4.7 **
PLF	-0.8	-6.4 **	-1.3	-2.4 *	0.8	2.5 *	-1.5	-1.7	-2.4 *
EGG PLF	-1.9	-3.1 **	-2.9 **	-3.3 **	0.1	1.1	-0.6	-2.3 *	-2.2 *
Ji	-2.6 **	-4.8 **	-4.7 **	0.8	-1.4	-3.6 **	1.2	-4.0 **	-4.9 **
EGG Ji	-0.6	-2.1 *	-2.3 *	-0.7	-3.4 **	-4.6 **	-0.9	-2.4 *	-3.9 **
Shi	-3.7 **	-6.2 **	-6.1 **	-1.3	-2.7 **	-2.9 **	-0.9	-3.7 **	-6.4 **
EGG Shi	-1.9 *	-2.1 *	-1.6	-0.9	-3.7 **	-1.1	0.1	-3.1 **	-2.8 **
OQ	0.9	0.8	1.6	0.5	3.7 **	2.6 **	0.6	0.9	3.5 **
VC	-0.7	2.5 *	-0.1	1.9	0.6	-1.6	1.6	-0.2	1.9
MPT	0.1	4.1 **	1.4	2.0 *	3.8 **	1.7	2.6 **	2.4 *	4.9 **
PQ	-0.5	-1.9 *	-0.4	0.6	-3.1 **	-3.2 **	-0.4	-2.6 **	-2.6 **
Npauses	0.1	-1.5	-1.5	-0.6	-3.7 **	-2.6 **	-0.1	-1.4	-3.5 **
Plength	1.2	-3.8 **	0.6	-4.9 **	-0.3	-0.7	-2.1 *	-3.7 **	-1.0
P/100 syl	0.1	-2.4 *	-2.1 *	-0.9	-5.2 **	-3.7 **	-0.2	-2.2 *	-5.1 **
S/P	-2.1 *	3.6 **	1.5	2.9 **	4.3 **	4.3**	-0.3	3.4 **	4.5 **
Srate	0.3	3.4 **	1.7	2.9 **	4.1 **	0.3	2.6 **	2.4 *	4.2 **
Arate	0.9	2.4 *	0.4	2.3 *	2.4 *	-2.9 **	2.7 **	1.2	1.9
Sylbp	-0.1	2.3 *	2.1 *	0.9	5.3 **	3.7 **	0.1	2.2 *	5.1 **
Int	-1.3	1.1	0.1	0.9	-2.3 *	-0.3	-1.8	-0.1	-1.3
Rtime	-0.2	-2.4*	-1.2	-2.1*	-2.9**	-0.1	-1.8	-1.6	-3.0**
Ptime	0.91	-2.6**	-1.23	-2.2*	-3.2**	-2.2*	-0.5	-2.3*	-3.5**
P(%)	1.4	-2.5*	-1.1	-2.1*	-3.1**	-3.0**	0.2	-2.4*	-3.1**

** significant at .99 confidence level

*significant at .95 confidence level

Appendix E

Percentage of variance explained by single predictor models. Data are given for roughness (R), breathiness (B) and hoarseness (H)[24].

Speech Task	*Variable*	*R*	*B*	*H*
Vowels	*IC /a/*	0.33	0.23	0.37
	AI /a/	0.33	-	0.22
	Jitter /a/	0.29	-	0.25
	FMF /a/	0.28	0.11	0.25
	IC /e/	0.24	0.2	0.29
	Jitter /e/	0.23	-	0.21
	FMF /e/	0.19	-	0.17
	Shimmer /a/	0.17	0.2	0.22
	Shimmer /e/	0.16	0.23	0.24
	HNR /a/	0.16	0.16	0.21
	AI /e/	0.13	0.1	0.12
	NC /e/	0.12	0.3	0.19
	NC /a/	0.11	0.32	0.21
	EGG Jitter /a/	0.11	-	0.14
	GNE /a/	0.1	0.4	0.22
	GNE /e/	0.1	0.38	0.19
	HNR /e/	0.1	0.12	0.16
Speech	*F0 SD*	0.28	-	0.16
	IFx	0.31	-	0.15
	EGG IFx	0.14	0.1	0.12
	EGG F0 SD	0.19	-	0.14
	80R	0.28	-	0.15
	Jitter	0.27	-	0.18
	Shimmer	0.41	0.12	0.27
	PLF	0.2	-	-
	EGG PLF	0.13	-	-
	EGG Jitter	-	0.13	0.12
	OQ	-	0.11	-
Aerodynamic measures	*MPT*	-	0.11	0.18
	PQ	-	0.1	-
Speech	*P/100 syl*	-	0.18	0.11
	S/P	-	0.24	0.14
	Sylbp	-	0.18	0.11
	P(%)	-	0.23	0.14
	Srate	-	-	0.18

[24] Only variables that fulfilled the 10 % explained variance criterion are shown. If the distribution of an independent variable was skewed, values were logtransformed or squared.

Abstract

The objective of this study was to predict three perceptual voice qualities from objective voice parameters by means of quadratic discriminant analysis (QDA) and artificial neural networks (ANN). Classification experiments were performed using vowels and connected speech material obtained from 145 dysphonic and 5 normal subjects. 8 speech and language therapists rated the voices according to the RBH perceptual scale. The study design provided for four types of instrumental measures including measures from mid-vowel segments and connected speech as well as aerodynamic and prosodic measures. More than 50 objective measures were examined for their association with perceived voice quality and predictive power. Measurements were made on electroglottographic and acoustic signals. All measures could be obtained automatically without high personal and equipment cost. No restrictions were imposed on the spectrographic signal type and the choice of diagnoses. Both classification methods showed nearly equal prediction accuracy and solved the four-class classification problem on average with at least 70 % of correctly classified voices in each examined voice quality, implicating almost a 3-fold chance. In both classification methods, neither intermediate-grade dysphonia nor extreme dysphonia grades were systematically misclassified. Although predictions of the two methods did not significantly differ, they showed great difference in the number of variables needed to achieve the reported prediction accuracy. The ANN method used consistently more variables to classify voice quality. Measures from connected speech and prosodic measures significantly improved the overall prediction accuracy. The main application of the results of this study is expected to be screening for voice disorders and monitoring voice therapy progress.

Acknowledgements

This work would not have been possible without the help and assistance of many people. First and foremost, I would like to thank my project supervisor, Prof. Dr. med. R. Schönweiler - who inspired me for the topic on the basis of his previous work - and Prof. Dr.. med. R. Linder, for their patience, kindness and guidance over the past few years. I am grateful to Prof. Dr. med. R. Schönweiler for the opportunity to work with patients having pathologic voices and for giving me freedom to explore many issues on my own and at the same time supporting me with constructive criticism at different stages of my research. I am grateful to all the support I have received from Prof. Dr. med. R. Linder, especially for helping me to sort out technical details of my work, for simulating the artificial neural networks and reading through my draft copies. In challenging me to hold to a high research standard, my project supervisors have truly made the completion of this project an excellent learning experience.

I would like to extend a special thank you to the whole staff of the Phoniatrics and Pediatric Audiology division of the University Clinic in Lübeck, especially to Ms G. Rehm, Mr. R. Fenske, Ms Chr. Mahlstedt and Mr. N. Wong for helping me with data collection. Special thanks also to Dipl.-Ing. H. Jäger and Dipl.-Phys. F. Landwehr for assistance in maintaining the experimental equipment and support in technical questions.

I am also greately indebted to Dipl.-Ing. & Dipl. Techn. Red. R. Meyer-Henke, Dipl.-Inf. I. Komlev and BS Inf. N. Prokoschenko for helping me to develop and trouble-shoot the website with evaluation protocols, to D. Genin, Ph. D., for statistical advice and assistance in interpretation of LLE data and to Dipl.-Inf. L. Kramer who helped me with LLE calculation. I thank Mr. Walton for his kind assistance in revising the English text of the manuscript.

Finally, I would like to express my special gratitude to 8 expert raters. Any mistakes and misinterpretations of the experimental data are the autor’s full responsibility.

Lebenslauf

Persönliche Angaben

Name:	Elena Kramer
Geburtsdaten:	02.05.1974 Nowosibirsk, Rußland
Staatsangehörigkeit:	deutsch

Schulbildung

1981-1991	Besuch der Mittelschule in Kriwodanowka, Rußland
1991-1992	Elektrotechnische Universität, Nowosibirsk, Vorstudium mit Schwerpunkten Mathematik und Physik
1992-1995	Pädagogische Universität, Studium im Fach Germanische Philologie
1996	Anerkennung bisheriger Schul- und Studienleistungen als fachgebundenes Abitur

Studium

1996-2005	Studium an der Universität Hamburg
09/2003	MA Betriebswirtschaftslehre mit Schwerpunkten Industriebetriebslehre und Wirtschaftsenglisch
2/2005	MA in Phonetik und Englische Sprache, Literatur und Kultur

Beruf

2005	Angestellte bei Lingcom GmbH, Forchheim
2007-2009	Wissenschaftliche Mitarbeiterin, Universitätsklinikum Schleswig-Holstein, Campus Lübeck

Dissertation

03/2007-05/2007	Vorbereitung und Prüfung von Aufnahmetechnik
06/2007-11/2007	Sammlung von Stimmproben
12/2007-04/2008	Erstellung einer Webseite mit Bewertungsprotokollen
05/2008-12/2008	Werbung von Experten und Sammlung von Bewertungsprotokollen
01/2009-03/2010	Datenauswertung
Seit 04/2010	Mutterschutz und Elternzeit

Die aus der Dissertation resultierenden Veröffentlichungen:

Kramer E, Linder R, Schönweiler R. (2007). Psychoakustische Skalierung electroakustischer Heiserkeitsparameter. Poster. 24. Wissenschftliche Jahrestagung der Deutschen Gesellschaft für Phoniatrie und Pädaudiologie, 28. – 30. September, Innsbruck, Österreich.

Kramer E, Schönweiler R, Linder R. (2009). Subharmonische Anteile im Spektrum gehaltener Vokale bei pathologischen Stimmen. Vortrag. 76. Kongress der Deutschen Gesellschaft für Sprach- und Stimmheilkunde e.V., 26. – 29. März, Bochum, Deutschland.

Kramer E, Schönweiler R, Linder R. (2009). F0-statistics in dysphonic voices. Vortrag. 8th Pan European Voice Conference, August 26th – August 29th, Dresden, Germany.

Kramer E, Schönweiler R, Linder R. (2009). Akustische Stimmanalyse: Vergleich der Klassifikation durch künstliche neuronale Netze (KNN) und nichtlineare Diskriminanzanalyse (DA). Vortrag. 26. Wissenschftliche Jahrestagung der Deutschen Gesellschaft für Phoniatrie und Pädaudiologie, 11. – 13. September, Leipzig, Deutschland. In: Gross, M; am Zehnhoff-Dinnesen, A (Hrsg.): Aktuelle phoniatrisch-pädaudiologische Aspekte, Bd. 17, S. 105-108.

Kramer E, Schönweiler R, Linder R. (2010). Vorhersage der Behauchtheit in stimmgestörten Patienten. Vortrag. 27. Wissenschftliche Jahrestagung der Deutschen Gesellschaft für Phoniatrie und Pädaudiologie, 17. – 19. September, Aachen, Deutschland. In: Gross, M; am Zehnhoff-Dinnesen, A (Hrsg.): Aktuelle phoniatrisch-pädaudiologische Aspekte, Bd. 18, S. 154-157.

Kramer E, Schönweiler R, Linder R. (2011). Bequeme Tonlage und Intensität in stimmgesunden und stimmkranken Personen. Poster. 77. Kongress der Deutschen Gesellschaft für Sprach- und Stimmheilkunde e.V., 24. – 26. März, Göttingen, Deutschland.