University of Amsterdam
Amsterdam Center for Language and Communication

Toward an Objective Multidimensional Evaluation of Voice Quality in Head And Neck Cancer

Xinyu Zhang

Under the Supervision of

Dr. Rob van Son

Second Reader:

Dr. David Weenink

A thesis submitted in partial fulfillment of the requirements for the degree of Research Master of Arts

in

Linguistics

University of Amsterdam

July, 2021

Abstract

The evaluation of voice quality is an essential part of the treatment and rehabilitation of head and neck cancer. The aim of the current study was to develop multi-parametric acoustic models for the assessment of breathiness and roughness, respectively.

96 concatenated voice samples of connected speech and sustained vowels were perceptually judged for breathiness and roughness by two speech-language pathologists from the Netherlands Cancer Institute. The same voice samples were acoustically analyzed. Stepwise linear regression yielded a three-variable acoustic model for breathiness, and a two-variable acoustic model for roughness. Strong correlations were found between the models and the auditory-perceptual ratings of both breathiness and roughness.

Acknowledgments

Writing a thesis during a global pandemic is decidedly predetermined to be filled with difficulties. However, my supervisor Dr. Rob van Son, in his magical ways, not only rid all COVID-related hurdles that he could possibly remove, but also made my experience of writing the thesis enjoyable. In addition, Rob has also been a great teacher who excels at explaining complex concepts. I am grateful to have been a mentee of Rob's.

I would like to thank Klaske van Sluis and Anne Kornman for lending their expertise as speech-language pathologists and serving as expert listeners for the perceptual experiment.

Special thanks also go to Dr. David Weenink (who is kind enough to agree to be a second reader), Dr. Paul Boersma (who supervised my previous thesis), Dr. Silke Hamann (who has been teaching me since day one), and other phoneticians and phonologists at ACLC.

Contents

==========

Ethical approval:

Permission from the Institutional Review Board (IRB) of the Netherlands Cancer Institute was obtained for this study (NKI-AVL IRBd20-024).

==========

1 Introduction

1.1 Head and Neck Cancer and its Effect on Voice Quality

Voicing in speech is generated from the vibration of the vocal folds caused by the interaction of the high velocity and low pressure when air passes through, commonly referred to as the Bernoulli effect. Cancers of the head and neck, which amount to 65000 new cases on average per year globally (Chow, 2020), can have an impact on the quality of voice.

When the cancer is located on the vocal folds, the neoplasm will increase the mass of the vocal folds, and the tumor bulk and tissue invasion will result in a pair of vocal folds that are mismatched structurally and bio-mechanically. The leading edge of the affected vocal fold typically becomes irregular and sometimes displaced. The lesions can also cause stiffness which results in a local or more generalized loss of mobility and inconsistent vibratory patterns. Additionally, the lesion can promote diplophonia, either by direct extension, reactive inflammation, and/or fibrosis. The lack of glottal closure due to vocal fold apposition would cause unwanted air leakage during phonation (Orlikoff & Kraus, 1996; Ford & Connor, 2000). Another possible effect is abnormal activation of supraglottic muscles (Ford & Connor, 2000). In Agarwal et al. (2009), grade 3[1] breathiness was observed in 86% of the patients before treatment, and was persistent in 54% of the patients after radiotherapy; grade 3 hoarseness was 24% before and 16% after radiotherapy. Interviews with one of

[1]The subjective assessment carried out in Agarwal et al. (2009) uses a three-point scale, with Grade 1 being the closest to unaffected voice quality.

the practicing Speech-Language Pathologists at the Department of Head and Neck Oncology and Surgery of The Netherlands Cancer Institute also indicate that roughness and breathiness are the two most common voice changes among individuals with head and neck cancer both before and after the treatment.

Common options for the treatment of pharyngeal and hypopharyngeal carcinoma are CO_2 laser surgery, radiotherapy, and chemotherapy (Balm, 2014). A combination of more than one modalities might be used at more advanced stages (Tan, Stoker, & Smeele, 2014). Treatment for head and neck cancer can change the voice quality both for the better and for the worse (Orlikoff & Kraus, 1996), depending on the severity of dysphonia prior to treatment (Starmer, Tippett, & Webster, 2008; Verdonck-De Leeuw et al., 1999).

Table 1 is Orlikoff and Kraus (1996)'s summary of the effect of laryngeal radiation used in cancer treatment. Orlikoff and Kraus (1996) also points out that radiation is known to be damaging to both salivary and mucous glands. They predict that the dehydration of vocal folds caused by radiation could consequently lead to epithelial changes that make them vibrate irregularly. This was indeed confirmed by Roh, Kim, and Kim (2006) where the researchers induced xerostomia as well as reductions in resting and stimulated whole salivary flow rates in healthy individuals by administering glycopyrrolate, and compared the vocal function of the xerostomic group to a control group by acoustic, aerodynamic, and laryngostroboscopic measurements, and concluded that vocal function can be affected by xerostomia. However, another study by Roh, Kim, and Cho (2005) found that the mean salivary flow rate in limited radiotherapy patients did not differ significantly from the control group,

and that the decrease was most severe in the wide-field radiotherapy group. Certain chemotherapy agents are also said to have similar dehydration effects (Orlikoff & Kraus, 1996; Collins, McDonald, Stanley, Donovan, & Bonebrake, 1993).

Muscle atrophy	Loss of muscle bulk
Fibrosis	Formation of a diffuse and disorganized matrix of fibrous connective tissue within the muscle and lamina propria of the mucosa
Telangiectasis	Dilation of blood vessels, associated with engorgement (hyperemia) and inflammatory redness (erythema)
Keratosis	Formation of a scaly or plaque-like layer of keratin not ordinarily found in vocal-fold epithelium
Reineke's edema	Effusion and buildup of fluid in the superficial lamina propria, just beneath the surface epithelium

Table 1: Possible effects of laryngeal radiation used in cancer treatment (Orlikoff & Kraus, 1996)

Voice quality after laser excision is generally believed to be heavily dependent on the amount of tissue resected (Vilaseca et al., 2008), and the depth of the resection (Starmer et al., 2008).

A randomized trial conducted in Finland (Aaltonen et al., 2014) reported similar overall dysphonic severity after radiotherapy and laser surgery in male patients with early laryngeal carcinoma (T1aN0M0), but with laser surgery resulting in higher breathiness and a wider glottal gap. In Rydell, Schalén, Fex, and Elner (1995), the radio-

therapy group also had better results both in acoustic and perceptual assessments, as compared to the laser cordectomy group. Similarly, Krengli et al. (2004) reports better results in acoustic measurements in the irradiated patient group than those undergone laser surgery. Loughran, Calder, MacGregor, Carding, and MacKenzie (2005) found no significant difference between the laser excision and the radiotherapy group in any perceptual measurements. A meta-analysis conducted by Lee, Hong, Kim, and Hong (2019) found no significant difference between the perceptual evaluation of voice quality after laser surgery compared to radiotherapy for early-stage glottic cancer, although the assessments by acoustic measurements were better in patients that underwent radiotherapy.

In short, head and neck cancer, as well as its treatment, could affect phonation in several ways: the shape and texture of the vocal folds, the proper closure of vocal folds, hydration of epithelial tissues, and muscle control.

Acoustically, the incomplete closure of vocal folds, which consequently leads to the leakage of unphonated air, increases the component of noise and decreases the harmonics-to-noise ratio (HNR) (Agarwal et al., 2009). It is also reported that such increased transglottic airflow mainly results in turbulent noises in high frequencies (Carrara-de Angelis, Feher, Barros, Nishimoto, & Kowalski, 2003). Additionally, irregular vibrations of the vocal folds may also lead to perturbations in voicing.

It is also worth noting that the variation of voice quality between individuals is high even given the same tumor type and treatment, as there are other factors such as muscle usage, habits of vocalization, gastroesophageal reflux, and other, psychological and emotional, fac-

tors (Morrison, 1997).

2 Voice Quality Measurement

Voice quality has been studied since at least as early as Roman times. Cicero considered the practice of one's voice as an essential part for the training of an orator (Cicero, 55 B.C./1976, Lib. i. §156). Terms such as "harshness", "cracked", and "clearness" have been used to describe voice quality by Austin in Chironomia (Austin, 1806, p.33).

The European Laryngological Society (ELS) recommends five key components of clinical voice evaluation: perception, videostroboscopy, acoustics, aerodynamics, and subjective rating by the patient (Dejonckere et al., 2001).

The evaluation methods for voice quality can be roughly divided into two general categories: the subjective and the objective. Objective methods further include instrumental and acoustic analysis.

2.1 Subjective Measurements

Some of the most commonly used subjective measurement scales are: the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) (Kempster, Gerratt, Abbott, Barkmeier-Kraemer, & Hillman, 2009), the Grade-Roughness-Breathiness-Asthenia-Strain Scale (GRBAS) (Hirano, 1981), and the Roughness-Breathiness-Hoarseness (RBH) scheme (Nawka, Anders, & Wendler, 1994).

The GRBAS evaluates voice quality in five dimensions, namely, Grade, Roughness, Breathiness, Asthenia, and Strain, each on a four-point scale (Hirano, 1981). Carding, Wilson, MacKenzie, and Deary

(2009)'s overview article concluded that the GRBAS has the best reliability when compared with the CAPE-V, the Buffalo Voice Profile, and Vocal Profile Analysis. The GRBAS does not have requirements for the types of stimulus needed for evaluation, but previous research suggests that, at normal loudness level, inter-rater reliability is higher when the SLPs were provided with read text than when the stimuli came from sustained vowels (Bele, 2005). The GRBAS is the only perceptual scale used in the Netherlands, where the current study is conducted.

The CAPE-V is developed by Kempster et al. (2009) as a clinical and research protocol to promote standardized evaluation and documentation of auditory-perceptual assessment of abnormal vocal quality. The CAPE-V specifies the tasks and stimuli used for the evaluation process. Three tasks are involved for the speaker: vowel prolongation, text reading, and conversational speech. Three corner vowels are elicited at a steady pitch for the prolonged vowel samples. A set of specifically designed sentences targeting different features are used for stimuli for the read speech task. Conversational speech is prompted by asking the patient to describe their voice problems to the evaluator. This part is deemed the most crucial of the three tasks in the CAPE-V and is expected to be assessed throughout the evaluation session. Six aspects are evaluated in the CAPE-V, each on a 100 mm visual analog scale: Overall Severity, Roughness, Breathiness, Strain, Pitch, and Loudness. Besides the six aspects, two unlabeled scales are also included for the clinician to document additional features such as nasality, spasm, tremor, intermittent aphonia, falsetto, glottal fry, or weakness (Kempster et al., 2009). Solomon, Helou, and Stojadinovic (2011) shows that speech-

language therapists rated severity, roughness, breathiness, as well as strain all at a lower rating when in a laboratory setting compared to a clinical setting, when using the CAPE-V scale in both settings.

The RBH index, suggested by Nawka et al. (1994), is a shortened version of the GRBAS scale, where the overall severity is represented by Hoarseness (H, similar to the G dimension in GRBAS), and are comprised of Roughness (R) and Breathiness (B). The RBH is especially used in Germany, and has been criticized to be more reliable interrater-wise for male voices than for female voices (Koreman, Pützer, & Just, 2004).

In clinical settings, speech-language pathologists also take into account the patient's self-reported perceptual impression of their own voice by using e.g. the voice-related quality-of-life questionnaire (VRQoL) (Hogikyan & Sethuraman, 1999), and/or the Voice Handicap Index (Jacobson et al., 1997).

Overall, such perceptual rating scales are useful but not flawless.

Given that the perceptual ratings are essentially the listener's subjective response to the objective acoustic signal, the perceptual rating can be said to be a function of both the listener and the signal. As such, perceptual scales also need to accurately model the behavior of the listeners. According to Kreiman and Gerratt (1998), multi-dimensional perceptual scales imply a set of assumptions for the measurements to be meaningfully compared between voices and listeners.

First, such scales assume that listeners' general impression of voice quality can be separated into distinct aspects such as breathiness and roughness. Additionally, paradigms such as the GRBAS also require listeners to be able to selectively attend to individual

dimensions of voice quality. Finally, all perceptual scales are built on the assumption that the scales are constant across voices and listeners, as if voice quality can be treated as a characteristic of the voices themselves instead of the subjective reaction they evoke in listeners[2] (Kreiman & Gerratt, 2000a).

There have also been criticisms of the terms in perceptual scales not being well-defined. In the appendix of Kreiman and Gerratt (2000a), the authors listed nine different definitions for the perceived quality of "breathiness" by scholars from 1986 to 1995 and eight different definitions for the perceived quality of "roughness" in studies from 1981 to 1995.

Empirical evidence of between-listener variability exists. Studies have also shown that listener's expectations may affect their judgment of voice quality (Ghio, Révis, Merienne, & Giovanni, 2013). Apart from the effect of listener-blindedness, other factors could also affect the validity of perceptual rating scales for voice quality. Kreiman and Gerratt (1998) analyzed data from seven studies, and results show that for the "moderately pathologic" range, more than 60%, and up to 78%, of the variances in subjective ratings can be caused by factors other than the *de facto* difference in the quality of the voice being rated. In Rammage, Peppard, and Bless (1992), expert listeners (each with a minimum of 7 years experience as a voice clinician) were trained before the perceptual rating, and yet still, the listeners agreed well on the overall dysphonic severity, but not on separate dimensions. Interviews with the speech-language pathologists at the NKI confirm that in diagnoses, speech-language pathologists focus more on the overall severity of dysphonia than sin-

[2] More on inter-rater reliability in §4.3.1.

gle aspects such as whether a voice is more rough or more breathy.

Kreiman, Gerratt, and Ito (2007) identified four factors that account for the variability in perceptual measurements: instability of listeners' memory standards for levels of a certain quality, ability to isolate single dimensions in a complex context (most important), measurement scale resolution, and the magnitude of the attribute being measured.

2.2 Instrumental Measurements

Instrumental measurements only rely on the input signal and are not observer-dependent, hence provide high test-retest reliability. While the test-retest reliability of instrumental measurements is high, the validity of instrumental measurement scales (i.e. the extent to which the parameters measure what they are intended to measure) is not guaranteed, which leads to the motivation of the current study: to search for reliable instrumental measurements for particular voice traits.

Instrumental measurements include aerodynamics, stroboscopy, and acoustic measurements.

2.2.1 Aerodynamic measurements

Pulmonary phonation is powered by the air filled in the lungs above the resting expiratory tidal range (Hixon, Goldman, & Mead, 1973). Therefore, the airflow, air pressure, as well as volumes of air all play a role in the variation of the quality of phonation. The study of aerodynamics of voice production, therefore, can provide insights into an individual's voice quality.

Direct measurement of air volume requires the use of mouthpieces or face masks, which usually interferes with speech production. Due to the relatively little restriction of the vocal tract during vowel production, direct measures of glottal airflow rate are commonly estimated from oral airflow rate during vowel phonation. However, this restricts the samples collected to be sustained vowels, which does not fully represent daily conversational speech, besides involving the use of mouthpieces and face masks. Direct measurement of subglottal air pressure usually requires invasive procedures such as the insertion of a hypodermic needle or a thin catheter (Hillmen & Kobler, 2000).

Such assessment is also dependent on fundamental frequency and type of speech task. There is currently no normative data for aerodynamic measures that is universally accepted or applied in clinical work (Ziethe, Patel, Kunduk, Eysholdt, & Graf, 2011).

2.2.2 Laryngeal Stroboscopy

Laryngeal stroboscopy involves a strobe light controlled to flash at the frequency at which the patient's vocal folds vibrate. Multiple snapshots per glottal cycle are then captured and combined into a simulated slow-motion view of the vocal fold vibration.

The technique provides a direct look at the articulatory factors that contribute to the timbre of a voice, such as the exact way in which the vocal fold mucosa moves and the way glottal closure takes place. This is helpful when investigating why a specific voice does not have desired characteristics, as it provides morphological and biomechanical information about the speaker (Cranen & de Jong, 2000).

In reality, jitter is always present in speech, making it difficult to determine the rate at which to capture the image, which may result in an inaccurate reconstruction of the vibration cycle. Besides the probable inaccuracy, glottal leakage is difficult to measure by laryngostroboscopy. In addition, endostroboscopy is an invasive procedure that requires anesthesia, and may lead to unnecessary bleeding or infection and further worsen the voice quality of the individual undergoing the procedure.

More importantly, laryngeal stroboscopy alone is not enough to make diagnostic decisions. More sources of information are needed even before the stroboscopy to determine which details deserve special attention as well as to interpret the images.

2.3 Acoustic Measurements

The methods of acoustic measurements are appealing because they are not intrusive in nature. So far, there is not a consensus about a standard set of parameters to use for the acoustic measurement of voice quality, nor is there sufficient psychoacoustic evidence. Not even the type of voice samples to collect and process has been agreed upon.

Moreover, the mechanisms in articulation that lead to the perception of roughness and breathiness are complex, making it difficult to precisely prescribe a set of acoustic measurements for such voice characteristics.

2.3.1 The Acoustics of Breathiness and Roughness

Breathiness:

The majority of research on the acoustics of breathiness is based on languages in which breathiness is phonemic, instead of the type of breathiness that is caused by anatomical changes.

According to Stevens (2000, p.87), the incomplete glottal closure in breathy phonation allows a path for airflow, which then lessens the abruptness of the cessation of the airflow during the closing phase, causing a smoother transition into a steady airflow between the approximation and parting of the vocal folds. This leads to a smaller negative peak if one were to plot out the derivative of the airflow volume velocity (i.e. the prototypical amplitude time waveform). Consequently, a source spectrum of the incomplete closure condition would be weaker in the higher ($>$ 2000 Hz) frequencies.

The source spectra of breathy phonation are also more sinusoidal than that of non-breathy phonation in general (Klatt & Klatt, 1990; Hillenbrand & Houde, 1996)[3]. Hillenbrand, Cleveland, and Erickson (1994) state that "rounding" a signal to a sinusoid results in a higher relative amplitude in the first harmonic. Evidence from several languages confirms that breathier speech indeed have higher amplitude in the first harmonic and weaker energy in upper harmonics, e.g. in Jalapa Mazatec (Garellek & Keating, 2011; Blankenship, 2002), Tagalog (Blankenship, 2002), !Xóõ (Ladefoged, 1983), Newar[4](Ladefoged, 2003), and Zapotec (Ladefoged, 2003). It should also be mentioned that higher amplitude in the first harmonic is not always the most salient perceptual cue for breathiness. For instance,

[3]Ladefoged (2003, p.172) points out that this is not always true supraglotally, and that the airflow might be so turbulent in very breathy voices that the waveform might end up more like random noise with no regular vocal fold movements. More details in the following paragraphs.

[4]In breathy nasals.

Klatt and Klatt (1990) showed that perceptual judgment of breathiness is more correlated with aspiration noise than with first harmonic amplitude, and that the amount of change in the amplitude of the first harmonic that is needed to effect a change in the perceptual classification between "breathy" and "clear" greatly exceeded the actual H1 difference measured from speech data.

The combination of the two characteristics above (i.e., the increase of the amplitude in H1 and the decrease of amplitude in the higher harmonics caused by the less sharp glottal pulses) would in theory indicate a steeper spectral slope subglotally. However, note that breathy voices also have more turbulence and noise excitation especially above 2000 Hz, which raises the energy in the higher frequencies and flattens the slope. Therefore, the eventual spectral slope is not necessarily steeper in breathy voices compared to modal voices. In fact, observations from Ladefoged (1983) show that breathy vowels have "a less falling spectrum" in !Xóõ. Findings from the literature about spectral slope and spectral tilt in breathy voices are also mixed. Another reason for the mixed results in spectral slopes, as pointed out by Ladefoged (2003), is that there are other factors that also contribute to the relative amplitude of harmonics and the spectral slope. Such factors include vowel quality (e.g., distance between formants), pitch, and stress. The relationship between spectral slope and breathiness is also shown to differ depending on which part of the utterance is being measured as well as the gender of the speaker (Garellek & Keating, 2011; Klatt & Klatt, 1990).

In terms of additive noise and periodicity, the periodicity measurements in Hillenbrand et al. (1994) explained approximately 80%

of the variance in perceived breathiness[5]. Klich (1982, p.574) concluded in the literature review that "the excessive fricative noise produced in the larynx and superimposed on the laryngeal tone" is the primary distinguishing characteristic of breathy vowels. Periodicity in breathy voice is difficult to quantify especially when the vowels are not especially prolonged or produced at a steady fundamental frequency (Ladefoged, 1983). Periodicity measures also differ depending on the type of stimuli. For example, the stressed vowels with a high fundamental frequency tend to have perfectly periodic spectra in Klatt and Klatt (1990), even for the most breathy speakers. Nonetheless, the aspiration noise in F3 region is one of the only two statistically significant acoustic correlates for breathiness among a total of ten being tested by Klatt and Klatt (1990).

Roughness:

Roughness in pathological voices is a more complicated mechanism compared to breathiness. The term "roughness" is used to describe multiple types of disordered phonation. For instance, Imaizumi (1986, p.458) listed seven types of voices that have been defined as "rough" in the literature. Among the seven, the definitions of roughness can be roughly divided into two subcategories: the existence of additive noise, and irregularities in periodicities. From experiments conducted in Imaizumi (1986), three types of roughness were found: voices that are partially modulated, voices with very low fundamental frequencies and rich harmonics hence have a deeper modulation in the amplitude envelope, and voices that "consist of an alternating repetition of two segments".

[5]Though not all measured by perturbation.

Zwicker and Fastl (1990) explained the psychoacoustics of roughness by manipulating the amplitude and frequency modulations in a 1000 Hz pure tone. They found a number of acoustic dependencies of roughness. Namely, for amplitude modulation, the degree of modulation[6] and the modulation frequency determine the roughness level; for frequency modulation, the level of roughness in sensation is determined by the width of critical band-rate deviation and modulation frequency. More specifically, in amplitude modulation, the sensation of roughness increases as the degree of modulation increases up to 1. In terms of the modulation frequency in amplitude modulation, the sensation of roughness increases as the modulation frequency increases, reaches a maximum roughness (depending on the center frequency of the signal[7]), then decreases as the modulation frequency increases. When the modulation is frequency modulation, the resulted roughness is much larger than that from amplitude modulation in general. The relationship between modulation frequency and sensation of roughness in frequency modulation is similar to that in amplitude modulation (for a 1k Hz tone, the maximum perceived roughness is reached at 70 Hz in both amplitude modulation and frequency modulation). Relating this to human hearing, Zwicker and Fastl (1990, p.234) concluded that since the perception of roughness is influenced by the frequency resolution and temporal resolution of the human hearing system, in the sense that human ears are only able to process "changes in excitation level or in specific loudness at all places along the critical-band rate scale", the model for rough-

[6] See Gaudernack (1934, p.820-822) for a detailed definition of the degree of modulation.

[7] The frequency at which maximum roughness is reached increases as the center frequency increases.

ness should be based on the differences made in excitation levels. In such a model, roughness is a function of modulation frequency and temporal masking depth. Zwicker and Fastl also note that when the sensation of roughness decreases after reaching a peak as modulation frequency increases, the perception becomes "hearing three audible tones". Vassilakis (2005, p.121) explains this by likening the hearing process to a frequency analysis performed by the ear: when the rate of the amplitude fluctuation in the perceived signal is smaller than the critical bandwidth (i.e. the bandwidth of the hypothetical analysis filters), the ear perceives the signal as a single tone with either fluctuation in loudness or with roughness; when the fluctuation rate is larger than the critical bandwidth, then the ear will be able to analyze the signal as a combination of multiple pitches. To put it another way, the perception of roughness can be psycho-physiologically explained as the auditory system being unable to resolve the frequencies whose differences are smaller than the frequency resolution of the basilar membrane.

This is probably the reason why diplophonia is considered by clinicians as a subtype of roughness as well as a symptom of laryngeal pathology (Ward, Sanders, Goldman, & Moore, 1969). Diplophonia is often found in patients with asymmetry in their vocal folds (Wong, Ito, Cox, & Titze, 1991; Moore, 1976). The asymmetry can be that of texture, shape, length, tension, stiffness, mass, and so forth.

Acoustically, diplohonic voices would have subharmonic series like those present in creaky voice[8] (Klatt & Klatt, 1990; Cavalli & Hirson, 1999). Some researchers even consider the presence of sub-

[8] Although, Dejonckere and Lebacq (1983) warns that diplophonia "must be distinguished from vocal fry".

harmonics to be the prerequisite for the perception of diplophonia[9] (Dejonckere & Lebacq, 1983; Wong et al., 1991), although there is no consensus on the exact structure of said subharmonics. Cavalli and Hirson (1999) studied the acoustic correlates to the perception of diplophonia, and found that diplophonia is related to "all MVDP measures[10] except *number of subharmonics*" even when only considering "the presence of two distinctive pitches" as diplophonia. However, this lack of correlation between perceived diplophonia and *number of subharmonics as measured by MVDP* could also be due to the lack of accuracy in MVDP's measurement of subharmonics, since the same study also found a low correlation between the number of subharmonics observed spectrographically and the number of subharmonics as measured by MVDP. Cavalli and Hirson (1999) eventually concluded that subharmonics are "neither necessary nor sufficient to perceive double voice".

Theoretically, diplophonia would also correlate with higher jitter and shimmer measurements. But the acoustic measurement of jitter and shimmer measurements in diplophonia is likely to be unreliable depending on different pitch tracking algorithms.

2.3.2 Types of Stimuli

Voice quality evaluations made from sustained vowels and that made from running speech do not correspond well (Barsties v. Latoszek,

[9]Caveat: the definition of diplophonia seems to differ between researchers. Some only consider the presence of more than one distinct pitches as diplophonia (e.g., Dejonckere & Lebacq, 1983), while others include other subtypes of pathological roughness into diplophonia.

[10]Namely, number of subharmonics, jitter, shimmer, relative average perturbation, and amplitude perturbation quotient.

Ulozaitė-Stanienė, Petrauskas, Uloza, & Maryn, 2019, p.184), and there has been a protracted controversy in whether to use sustained vowels or connected speech as voice samples to measure voice quality.

On the one hand, sustained vowels can be extracted in a more "controlled" environment for measuring purposes, and are relatively free from speaker idiosyncrasies. Sustained vowels are also more appropriate to be used for perturbation measures, as perturbation measures are essentially cycle-to-cycle variations, and such between-cycle variations can only be theoretically meaningful if they are measured from sustained vowels.

On the other hand, not all patients can sustain a vowel for the length required for perturbation measures. More importantly, what voice quality judgments are based on in real life is connected speech rather than sustained vowels. Moreover, connected speech contains cues such as rapid voice onsets and termination, F0 and amplitude variation, speaking rate, as well as voice breaks (Parsa & Jamieson, 2001), all of which may contribute to the perception of vocal quality. For instance, while most sustained-vowel-based measurements are made on the steady part of a vowel, which is usually the middle part, De Krom (1995) found that the acoustic information contained in the onset of a vowel correlates best with perceived roughness. In other words, said "idiosyncrasies" that can be easily avoided when only soliciting sustained vowels may well serve a role in how an individual's voice is being perceived in day-to-day life. Therefore, connected speech samples as stimuli are more ecologically valid for voice quality judgments. However, with the greater amount of useful information carried in connected speech, also comes greater difficulty in measurements (more on this in §2.3.4).

In general, sustained vowels are used for perturbation measures, and connected speech is used for spectral/cepstral measures in acoustic evaluations[11]. This is so because perturbation measures can only be theoretically meaningful when measured from sustained vowels, and spectral/cepstral measures don't necessarily require vowels hence connected speech can be used.

2.3.3 Perturbation Measures

Perturbation measurement refers to the comparison of cycle-to-cycle differences in a voice signal. The most common parameters used are jitter and shimmer. Jitter measures the difference of period length in neighboring vibration cycles. Shimmer measures the between-cycle difference of amplitude.

Given the nature of the perturbation measurements, it would only make sense to apply them on single sustained vowels, and this is also the case in reality that perturbation measures are rarely applied to connected speech.

Even when only applied to sustained vowels, some drawbacks still remain with perturbation measures. Besides sustained vowels not being sufficiently representative of an individual's voice quality, perturbation measures rely heavily on the detection of exact vibration cycles. In the case of disordered speech, however, the irregularity in phonation makes it difficult, sometimes hardly possible, to locate the boundaries of cycles (Awan & Roy, 2005). The pre-treatment recordings of 32 out of 145 patients (22.06%) in Carding et al. (2004) could not be assessed with perturbation measures due to this partic-

[11] Exceptions exist, such as the AVQI.

ular reason. The substantial portion of patients whose voice cannot be evaluated using perturbation metrics creates a diagnostic gap in such methods.

The fact that perturbation measurements require the detection of periodicity, and that disordered voices typically have substantial aperiodicity, may reduce the reliability of such measurements as parameters for disordered voices (Halberstam, 2004). Moers et al. (2012, p.420) also pointed out that the fact that periodic analyses require stable phonation and the lack of stability in phonation in dysphonia "may be the most probable reason for the pertinent differences across studies".

2.3.4 Spectral and Cepstral Measures

Spectral and cepstral measurements[12] are not time-based, therefore using such measures eliminates the issue of the difficulty in determining cycle boundaries in some disordered voices (Moers et al., 2012).

Commonly used spectral and cepstral measurements include the slope of the long-term-average-spectrum (LTAS), the tilt of the LTAS, as well as cepstral peak prominence (CPP).

The LTAS is the spectra taken at multiple time points (at a particular sampling rate) averaged across the time axis of the signal. The slope of the LTAS refers to the difference of energy in lower frequencies (usually 0 - 1000 Hz) and the energy in the higher frequencies (usually 1000 - 10000 Hz). The tilt of the LTAS is calculated by first fitting a trendline through the LTAS and then comparing the

[12] See Figure 8 for an illustration of the relationship between a waveform, a spectrum, and a cepstrum.

energy difference between the lower and higher frequency ranges. Radiotherapy and chemotherapy for head and neck cancer both leave individuals with turbulent noises in higher frequencies (Carrara-de Angelis et al., 2003). But this does not mean that the LTAS slope is necessarily flatter in this patient group. As was mentioned in §2.3.1, there are two opposing factors affecting the slope here, namely the decreased source harmonics and the increased turbulent noises, and neither is necessarily always stronger than the other.

Cepstral peak prominence, as the name suggests, is a measure of the prominence (peakiness) of the cepstral peak. A cepstral peak is the highest peak in the cepstrum, which is a log power spectrum of a log power spectrum (Hillenbrand & Houde, 1996). The log spectrum of a sound signal is the energy at different frequency bins. The log spectrum of the above log spectrum then shows the amplitude by the periods of different components in the signal. The amplitude at the period of the fundamental frequency would show up as the most prominent peak in a cepstrum. Therefore, a more prominent cepstral peak would indicate a more distinguishable harmonic structure. The prominence of the cepstral peak, defined by Hillenbrand et al. (1994), is essentially the relative amplitude of the most prominent peak in the cepstrum. The cepstral peak prominence is calculated by first fitting a regression line onto the cepstrum, and then calculating the difference between the actual amplitude of the peak and the amplitude of the peak as predicted by the regression.

Hillenbrand et al. (1994) also proposed a modification to the CPP that noticeably improves the prediction accuracy: the smoothed cepstral peak prominence (CPPS). The smoothing in CPPS refers to the averaging between adjacent time frames and the averaging between

quefrency frames. More specifically, each cepstrum is first replaced with the mean of several cepstral frames to its left and several cepstral frames to its right on the time axis. Then each cepstral magnitude is replaced by a running average of the magnitude of several quefrency bins to its left and several quefrency bins to its right. CPPS explained 96% and 92% of the variance in breathiness in sustained vowels and connected speech respectively in Hillenbrand and Houde (1996) even without having to filter for voicing in connected speech or to correct errors in pitch tracking. Halberstam (2004) provided evidence that CPPS is a more valid parameter for overall voice quality than perturbation measures when used on speech samples that are concatenations of a sustained vowel and some connected speech.

2.3.5 Objective Multi-parametric Scales

Since, as can be seen from the previous subsections, voice quality does not have a single perfect acoustic correlate, an accurate objective acoustic measurement for voice quality are likely to be multi-parametric. The two most often used objective multi-parametric scales that currently exist are the Acoustic Voice Quality Index (AVQI) (Maryn, Corthals, Van Cauwenberge, Roy, & De Bodt, 2010)[13] and the Cepstral Spectral Index of Dysphonia (Awan, Roy, & Dromey, 2009). Both scales have been confirmed to have high accuracy and reliability. The AVQI has even been validated in multiple languages. However, one could argue that even though AVQI "works in practice", it does not make perfect theoretical sense in some details. For

[13] See §4.2 for a more elaborate description of the AVQI.

instance, perturbation measurements in the AVQI are made on a concatenation that includes samples of connected speech.

Further, multi-parametric objective scales for specific dimensions of voice quality (such as breathiness and roughness) barely exist. To the best knowledge of the current author, the only existing dimension-specific multi-parametric scale is the Acoustic Breathiness Index (Barsties v. Latoszek, Maryn, Gerrits, & De Bodt, 2017) for the B (Breathiness) dimension in the GRBAS scale. Similar to the AVQI, the ABI uses a concatenation of a sustained vowel and a phonetically balanced sentence as input. Acoustic measurements in the ABI include CPPS, jitter, shimmer, glottal-to-noise excitation ratio, high-frequency noise, period standard deviation, and H1-H2 difference.

The present study aims to find acoustic models for breathiness[14] and roughness respectively, and base the input on more ecologically valid measurements.

3 Perceptual Experiment

3.1 Voice Samples

Forty-five participants with relatively high Acoustic Voice Quality Index (AVQI) scores[15] were selected from a group of laryngeal can-

[14] The breathiness model in the current study differs from the ABI in the sense that we aim to find multi-parametric models similar to the AVQI for different dimensions so that both a general score as well as dimension-specific scores can be produced from the same set of measurements.

[15] Higher AVQI scores indicate worse overall voice quality. See §4.2 for a description of the AVQI.

cer patients consisting of 78 men and 15 women[16], all of whom had laryngeal carcinoma (stage CIS-T2N0M0) and were treated for glottic, supraglottic, or subglottic tumors with laser surgery or radiation therapy. The AVQI scores ranged from 1.32 to 8.63, with a mean of 4.40, representing a clinical population with comparatively severe dysphonia.

Samples of connected speech were selected from the existing recordings of the aforementioned patients reading a content-neutral text in Dutch (*80 dappere fietsers* "80 brave cyclists"). A total of 100 samples were selected, each with a length of 5 seconds. The speech samples were randomized, to avoid the effect of the listener's expectation on the ratings, as previous research (Ghio et al., 2013) has shown that the SLPs' knowledge of whether a speaker is in the pre-treatment or post-treatment period affected their judgment of the speaker's voice quality.

3.2 Perceptual Ratings of Breathiness and Roughness

Two expert listeners (practicing speech-language pathologists at the Department of Head and Neck Oncology and Surgery of The Netherlands Cancer Institute) participated in the listening experiment. The listening experiment was conducted in an interactive interface written with and hosted on *Akoúste* (Van Son, 2020), where the listener is presented with two visual analog scales, one for breathiness and another for roughness, together with a button to play the voice sam-

[16]Due to the confidentiality of patient data, the exact gender ratio of the selected patients (i.e., whose recordings were used in the perceptual experiment) was not available to the current author.

ple, and the instruction "The speakers might have voice problems. Please evaluate the roughness and breathiness of the voice". Figure 1 is a screenshot of the user interface of the listening experiment. The listeners can listen to each recording multiple times by repeatedly clicking on the "Speech" button, if they so choose.

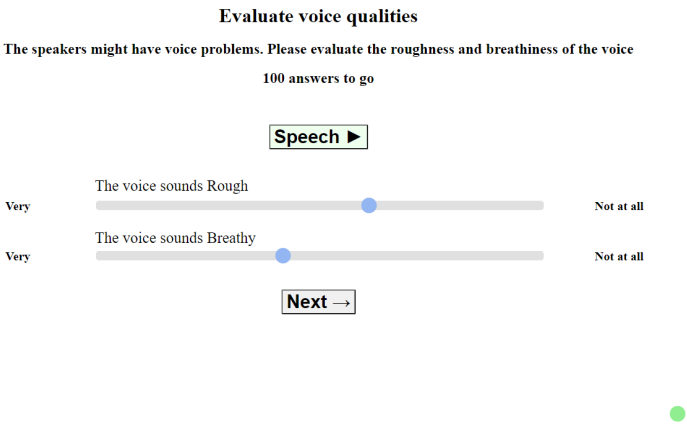


Figure 1: The user interface of the listening experiment, courtesy of Van Son (2020)

Four stimuli[17] served as practice items and the ratings of these were not included in the final results. The visual analog scale, instead of other equal-appearing interval scales such as the Likert scale, was

[17] which were the last four stimuli in the list before randomization in reverse order.

adopted in order to allow for a more differentiated evaluation[18].

The visual analog scale in the current study is labeled as "Very" on the left end, and "Not at all" on the right end. The perceptual ratings from the visual analog scale were converted into numbers between 0 and 1000 (inclusive) for statistical analysis, with the lower numbers reflecting more severe voice problems on each dimension.

The speech-language pathologists graded the roughness and breathiness of each voice sample according to the GRBAS scale (Hirano, 1981), where roughness is defined as representing "a psycho-acoustic impression of the irregularity of vocal fold vibrations" which "corresponds to the irregular fluctuations in the fundamental frequency and/or the amplitude of the glottal source sound" (ibid. p.83), and breathiness is defined as "a psycho-acoustic impression of the extent of air leakage through the glottis" that is "related to turbulence"(ibid. p.83).

3.3 Statistical Analyses

Statistical analyses were done for two purposes, using R (R Core Team, 2020): one being to find out the correlation between breathiness and roughness ratings, in order to process the data for further analyses; another being to determine the between-evaluator effect of each of the two targeted voice quality dimensions.

[18] Wuyts, De Bodt, and Van de Heyning (1999) claims that visual analog scales would result in higher disagreement between raters compared to equal-appearing interval rating scales. However, it could be argued that the high agreement from equal-appearing interval scales could be a result of binning scores into fewer categories, which causes a loss of accuracy. The decision was made to use Visual Analog Scales in the present study due to the assumption that listeners perceive voice qualities on a continuum rather than in intervals (Bele, 2005) and for the purpose of finer judgment and more accurate statistical analysis.

3.3.1 Between-Dimension Correlations

To check the correlation between the roughness and breathiness ratings, the breathiness and roughness scores were standardized to z-scores. The z-scores of the roughness rating per speech sample between the listeners were then averaged. The same is done for the breathiness ratings. Linear regression was then run on the mean breathiness z-scores against the mean roughness z-scores.

Interviews with the speech-language pathologists at The Netherlands Cancer Institute, as well as the literature (e.g., Shrivastav, Eddins, & Anand, 2012), indicate that roughness and breathiness usually co-present and co-vary. This is borne out by our data.

Figure 2 shows the correlation between the roughness and breathiness ratings, after standardization. The solid blue line represents the linear regression slope, and the shaded area indicates the 95% confidence interval.

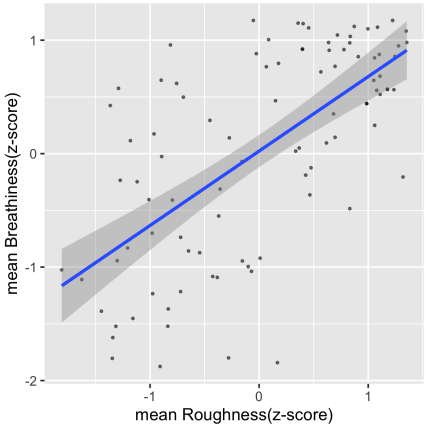


Figure 2: Correlation of the standardized roughness rating and breathiness rating of each speech sample. Blue solid line: linear regression slope; shaded area: 95% confidence interval.

To de-correlated the breathiness and roughness ratings so that the two are statistically orthogonal to each other, a linear regression model is first fitted on the z-scores of the two variables. Then the fitted value of each breathiness z-score as predicted by the corresponding roughness z-score was extracted and subsequently subtracted from the original z-scores of each breathiness rating, generating a new decorrelated z-score for breathiness. Figure 3 shows the result of decorrelation between the z-scores of breathiness and roughness. The resulting *Decorrelated Breathiness* will be used in the analyses in subsequent analyses (§3.3.2 and §4.2.2).

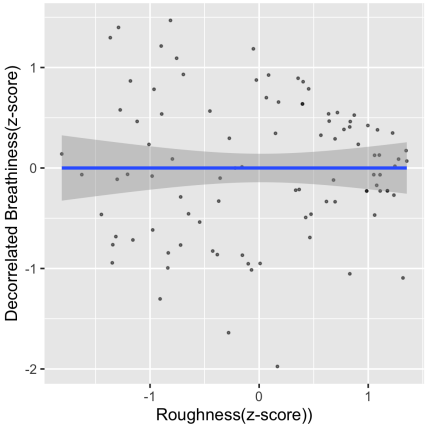


Figure 3: Correlation of the average z-scores of each speaker's roughness ratings and decorrelated-breathiness ratings. Solid blue line: linear regression slope; shaded area: 95% confidence interval.

3.3.2 Between-Evaluator Effect

Figures 4 shows the distribution of scores between two dimensions within each listener, and Figure 5 shows the comparison of scores between evaluators within each dimension. The plots show a slight bias against the mid-point (where score = 500), especially in SLP1. This might be due to the default position of the selection point on the scale in the GUI of the perceptual experiment.

Two models were built to assess the between-evaluator effects, one for roughness, and another for breathiness. One linear regression model was run on the perceptual ratings, with the perceptual scores for roughness as the dependent variable, and evaluator as the predictor, while assigning a random slope for each speaker. Corre-

spondingly, the linear regression model for the between-evaluator effect in breathiness uses breathiness ratings as the dependent variable, and the evaluator as the predictor, also assigning a random slope for each speaker.

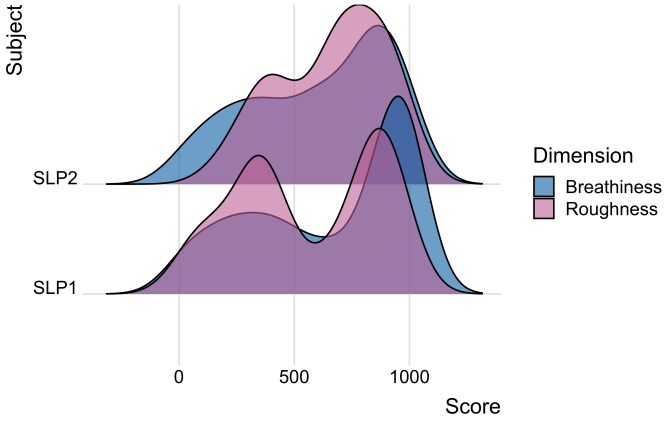


Figure 4: Within-listener comparison of the distribution of scores across dimensions

Figure 6a shows the breathiness scores that the two listeners gave to each speech sample. Scores from the two listeners are plotted against each other, with the red dotted line (slope = 1) indicating an ideal one-to-one correspondence between the two listeners where they rated each token exactly the same score. The blue solid line shows the regression slope through the actual data. Figure 6b shows the correlation between the roughness ratings given by the two listeners, with the dotted line indicating the ideal correlation and the solid line as the actual regression.

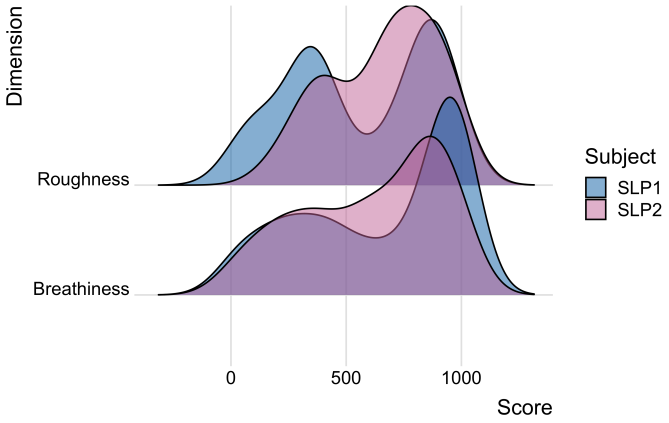
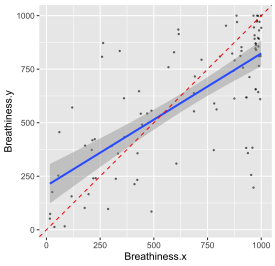


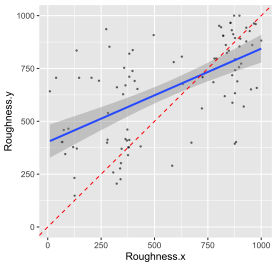
Figure 5: Within-dimension comparison of the distribution of scores across SLPs

The breathiness ratings between two speech-language pathologists were not significantly different (assigning a slope to each *Patient*, t = 1.11).

The agreement on roughness ratings between the two speech-language pathologists who participated was less ideal compared to that of the breathiness ratings. This is in accordance with the trend in the literature. The roughness ratings between evaluators were significantly different (assigning each *Patient* a slope, t = - 3.02).



(a) Inter-rater agreement of breathiness ratings



(b) Inter-rater agreement of roughness ratings

Figure 6: Inter-rater agreement of breathiness and roughness

4 Acoustic Analyses

4.1 Patient Sample and Recordings

The same patient population was selected for the acoustic experiment, which consisted of 45 laryngeal carcinoma patients (stage CIS-T2N0M0) who underwent laser surgery or radiation therapy for glottic, supraglottic, or subglottic tumors. The participants are all severely dysphonic in at least one of the three samples collected from each of them.

Existing recordings of the patients were used for the acoustic analyses. The recordings were made at three time points, namely, pre-treatment, six months after treatment, and 12 months after treatment. The speech samples were recorded in the speech-language pathologists' office during the patients' visits. Both sustained vowels and connected speech were recorded during each session. For sustained vowels, the patients were asked to produce three corner

vowels (/a:/, /i:/, and /u:/) and sustain them for at least 3 seconds. For connected speech, each patient was recorded reading a standard text in Dutch (*80 dappere fietsers* "80 brave cyclists") of about 150 words, during each visit. A total of 104 recordings were available.

4.2 Methods

Acoustic measurements were implemented through Praat (Boersma & Weenink, 2020), using the Acoustic Voice Quality Index (AVQI) model (Maryn et al., 2010), adding in the measurement for jitter (local) measured on the sustained vowels[19].

The AVQI is a multi-parametric objective model that measures overall voice quality, with both high sensitivity and high specificity for multiple languages (Barsties v. Latoszek et al., 2019). The specific weight assigned to each variable in the regression formula for the AVQI score is as follows: AVQI = [(3.295 − (0.111 ∗ CPPS) − (0.073 ∗ HNR) − (0.213 ∗ Shimmer) + (2.789 ∗ ShdB) − (0.032 ∗ Slope) + (0.077 ∗ Tilt)) ∗ 2.208] + 1.797, according to Maryn et al. (2010)[20]. Figure 7 shows an example of a report generated by the AVQI (with jitter value added in the upper right corner) from one of the voice samples that was included in the current study. The voice sample in Figure 7 has an AVQI score of 4.38, which is representative of the sub-population sampled in the current study (where mean AVQI =

[19]The measurement of jitter is one of the established acoustic parameters commonly used to measure perturbation, and was initially included in Maryn et al. (2010) for model selection, but did not end up in the final model of the AVQI. Jitter is included in the current study because it is one of the acoustic parameters routinely used in clinical practice to assess vocal function and monitor patient progress over the course of therapy

[20]The version of AVQI used in the current study is v.2.03.

4.40).

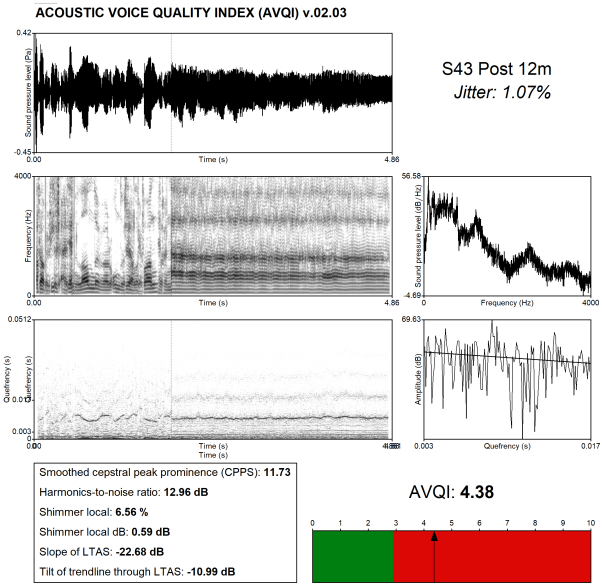


Figure 7: A sample report generated from the AVQI, with jitter (local) value added

4.2.1 Data Sampling and Processing

The sampling and processing of the acoustic data is described below.

For sustained vowels, three seconds of recording per patient was selected from one of the sustained vowels they produced during each of their visits to the speech-language pathologists. The vowel from

which the recording was sampled was usually /a:/, and the sampled part is usually the mid-section of the vowel. In cases where no vowels were successfully sustained for longer than three seconds, the steady sections of two vowels were selected and concatenated.

For connected speech, four consecutive seconds of recording was selected from each of the text recordings of each patient made at each of their visits to the speech-language pathologists. The selected segments were usually from the beginning sentences of the recordings.

The two types of recordings from each speaker at each time point were then concatenated, generating one audio file for each patient at each time point. Table 2 illustrates the selection of recordings[21].

[21] In practice, not all samples in all stages for each patient were present because of omissions or technical errors. Also, some samples had to be excluded due to low recording quality.

Subject ID	Time point	Files
S1	T1	continuous speech (4 sec) + sustained vowel (3 sec)
	T2	continuous speech (4 sec) + sustained vowel (3 sec)
	T3	continuous speech (4 sec) + sustained vowel (3 sec)
S2	T1	continuous speech (4 sec) + sustained vowel (3 sec)
	T2	continuous speech (4 sec) + sustained vowel (3 sec)
	T3	continuous speech (4 sec) + sustained vowel (3 sec)
⋮		⋮
S45	T1	continuous speech (4 sec) + sustained vowel (3 sec)
	T2	continuous speech (4 sec) + sustained vowel (3 sec)
	T3	continuous speech (4 sec) + sustained vowel (3 sec)

Table 2: Selection of Recordings (T1= pre-treatment; T2 = 6 months post-treatment; T3 = 12 months post-treatment)

Given that continuous speech contains pauses and voiceless segments, the AVQI model first employs a voicing detection algorithm to extract the voiced segments from the continuous speech section of each of the concatenated files. The voicing detection criteria used in the AVQI model follows Parsa and Jamieson (2001), according to Maryn et al. (2010), where each 30 ms frame of speech is evaluated at a time, and it is determined as voiced if it fulfills all three of the following requirements (Parsa & Jamieson, 2001):

- energy $>$ 30% overall signal energy;
- zero-crossing rate $<$ 1500 Hz;
- normalized autocorrelation peak $>$ 0.3.

4.2.2 Acoustic Measurements

A total of seven acoustic measurements were made on the concatenation of the sustained vowel and the voiced segments in the connected speech produced by the same speaker at the same time point in the treatment trajectory. Out of the seven acoustic measures, three were time-based (namely, Jitter, Shimmer, and ShdB), two were frequency-based (Slope and Tilt), one was quefrency-based (Smoothed Cepstral Peak Prominence), and one was a glottal noise measure (Harmonics-to-noise ratio).

Jitter refers to the cycle-to-cycle variation in frequency in the same signal, and shimmer refers to the amplitude variance between cycles in a signal[22]. Harmonics-to-noise ratio (HNR) is a measurement for harmonicity by quantifying the ratio of additive noise in a signal (Ferrand, 2002). The Smoothed Cepstral Peak Prominence (CPPS) is the average prominence of peaks in the cepstra of a signal. To get a cepstrum (first defined in Borgert, Healy, & Tukey, 1963, p.213) from a waveform, a Fourier Transform is first performed on the waveform. This results in a spectrum, in which the signal is transformed to the frequency domain from the time domain as it was in the waveform. The product of the magnitude of each frequency component and the complex conjugate of the magnitude is then divided by the total magnitude, resulting in a power spectrum. The natural logarithm of the power spectrum is then taken of the power spectrum. The final step is an inverse Fourier Transform on the spectrum, the output of which is a cepstrum. The dependent variable of the cepstrum is now *quefrency* (first defined by Borgert

[22] ShdB is shimmer measured in decibels.

et al., 1963), which can be understood as "the frequency of frequencies", with the unit being "cycles per cps", reverting it back into the time domain ($\frac{c}{c/s} = s$). In other words, a cepstrum is, in essence, "a spectrum of the spectrum". Figure 8 shows a flowchart of the process of getting a cepstrum from a waveform.

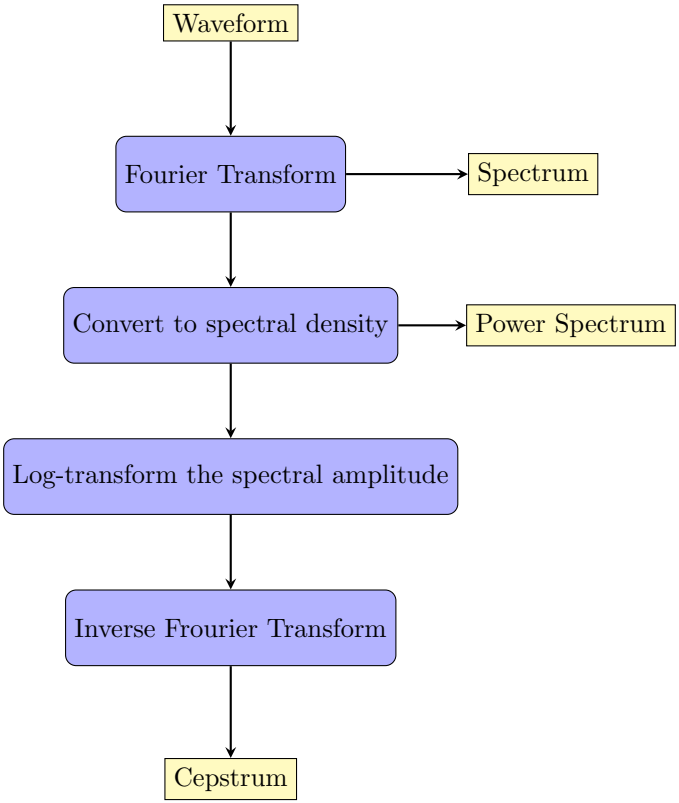


Figure 8: The process of getting a cepstrum from a waveform

The peak prominence of a cepstrum is calculated by fitting a trend line over the cepstrum and then getting the difference between each peak to the trend line. Smoothing is done in two steps according to Hillenbrand and Houde (1996): in each window, the cepstra are first smoothed across time, meaning that the frame in the middle of the window is averaged with its adjacent frames; the second step is smoothing across quefrency bins. Namely, in each window, the magnitude in the middle quefrency bin is averaged with the magnitudes in the quefrency bins that are immediately lower and higher to itself.

The benefit of transforming the voice signal into a cepstrum lies in the fact that the computation of a cepstrum is essentially the deconvolution (Childers, Skinner, & Kemerait, 1977, p.1429) of the source spectrum and the filter spectrum. Computing the logarithm of the spectrum deconvolutes the spectral signal from the *product* of the source and the filter to the *sum* of the two, so that the fast variation from the source and the slow variation from the filter would end up landing on the different ends. This in turn makes it easier to observe the source signal without the influence of the filter. In the case of voice quality measurement, converting a voice signal to a cepstrum makes it possible to examine the vibration happening at the vocal folds without the filtering effect of the vocal tract. This is especially useful when the stimuli are chunks of connected speech instead of sustained vowels.

Table 3 summarizes the acoustic metrics used in the present study.

Measurement	Description	Type
CPPS	Smoothed Cepstral Peak Prominence	Quefrency-based
HNR	Harmonics-to-Noise Ratio	Glottal noise measure
Jitter	Cycle-to-cycle frequency variation	Time-based
Shimmer	Cycle-to-cycle amplitude variation	Time-based
ShdB	Shimmer in dB	Time-based
Slope	Slope of LTAS	Frequency-based
Tilt	Tilt of trend line fitted on LTAS	Frequency-based

Table 3: List of acoustic measurements used

4.3 Statistical Analyses

The present study follows the convention of considering the mean of the clinicians' perceptual evaluations of voice quality as the ground truth upon which other methods are validated (Fritzen, Hammarberg, Gauffin, Karlsson, & Sundberg, 1986; Maryn, Roy, De Bodt, Van Cauwenberge, & Corthals, 2009; Kramer, 2011; Moers et al., 2012, etc.)

It is worth noting that voice quality is the listener's perception of an acoustic signal. The acoustic signal itself does not possess the qualities. Rather, the perception of certain qualities is evoked in individual listeners as they hear the signal (Kreiman & Gerratt, 2000a). That is, the quality of a voice is to some extent "in the ears of the perceiver". Thus, the perceptual judgment of voice quality is subjective, and the variance between (and *within*, although irrelevant here) listeners is unavoidable.

Therefore, by definition, there is a certain percentage of variance in the perceptual ratings in the data collected in the present

study that is caused by between-listener variation in the listening experiment, hence not explainable by acoustic measures.

4.3.1 The Explainable

Kreiman, Gerratt, Kempster, Erman, and Berke (1993) concluded that the mapping between physical signals and psychological qualities is not a constant or linear process. Instead, several other factors, such as the listeners' individual perceptual habits and biases (Solomon et al., 2011) , listeners' overall sensitivity to the quality being judged, as well as how well the quality is being defined, also contribute to inter-rater variability of observed voice ratings, even among highly experienced expert listeners.

With regard to whether listeners can attend to different dimensions of voice quality separately, the answer is still somewhat unknown. Kreiman and Gerratt (2000b) use a binary classification task (primarily breathy vs. primarily rough) to investigate whether the disagreement stems from the validity of scales or the manner in which the qualities are being measured. Their results show that the 15 expert listeners that participated in the study agreed better when a voice sample is neither breathy nor rough. Otherwise, the agreement between listeners seem rather random (unanimous agreement for "primarily rough" = 3/160 voices (1.9%), unanimous agreement on "primarily breathy" = 5/160 voices (3.1%))[23].

[23]However, it can be argued that the design of Kreiman and Gerratt (2000b) requires a very high acuity from the listeners, given the high co-occurrence of the two features in the pathological voices. Listener agreement would be higher if the stimuli came from e.g. synthesized voice samples (C.f. Hillenbrand, 1988; Alwan, Bangayan, Gerratt, Kreiman, & Long, 2000) where the breathiness and roughness were simulated and had more distinction between their respective severities.

Comparing the above with the results in §3.3.2, the between-listener variation in the present perceptual experiment is quite low.

As was shown in §3.3.2, where the *Adjusted* R^2 is 0.023 for Roughness, and *Adjusted* R^2 is 0.040 for Decorrelated Breathiness, it can be said that 2.3% of the variance in Roughness, and 4% of the variance in Decorrelated Breathiness, is caused by between-listener variations in the perceptual experiment.

Thus, we can calculate the highest percentage in the variance of Roughness and the variance of Decorrelated Breathiness explainable by acoustic data (i.e. that is not caused by between-listener variation): For Roughness, the highest possible percentage explainable by acoustic measurements $= 1 - 2.3\% = 97.7\%$. For Decorrelated Breathiness, the highest possible percentage explainable by acoustic measurements $= 1 - 4\% = 96\%$.

4.3.2 The Explained

To delineate which factors are the most explanatorily valid for the two traits (i.e., breathiness and roughness), stepwise model selections were then performed with the seven acoustic parameters mentioned above as potential predictors. Breathiness and decorrelated roughness were used as the dependent variables, respectively. For the stepwise model selection, Bayesian Information Criterion (BIC, Schwarz et al., 1978) was adopted instead of the default Akaike Information Criterion (AIC). The reason for choosing BIC over AIC is that BIC assigns more importance to specificity over sensitivity when selecting parameters, as compared to AIC (Dziak, Coffman, Lanza, Li, & Jermiin, 2020). Consequently, BIC penalizes the com-

plexity of models more heavily and tends not to include unnecessary predictors, hence reducing the likelihood of model overfitting, especially in cases like the current study where the data set is relatively small.

Roughness

Results show that the most relevant acoustic correlates for roughness are HNR and Tilt. In the best model selected using BIC, HNR is positively correlated to perceptual roughness ratings[24] ($\beta = 0.08474$) , while the tilt of the LTAS is negatively correlated ($\beta = -0.15297$) with perceptual roughness ratings ($Adjusted\ R^2 = 0.3292$). Both correlations are significant ($p < 0.0001$ for HNR, and $p < 0.05$ for Tilt). The regression formula generated from the model is as follows:

$$\widehat{meanRoughnessZScore} = -2.81821 + 0.08474 * HNR - 0.15297 * Tilt$$

Figure 9 is a visualization of the roughness model fitted on the perceptual and acoustic data collected in the current study.

[24] Higher Roughness and Breathiness ratings indicate less severe vocal problems. See §3.2.

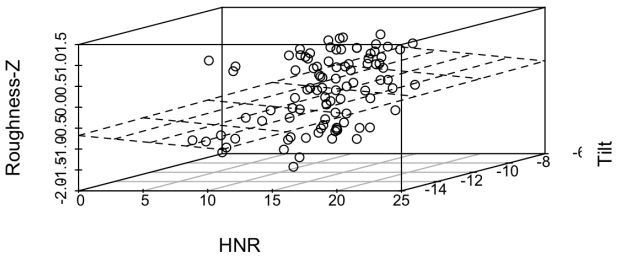


Figure 9: Correlation of the perceptual ratings of roughness to the best acoustic predictors for roughness

Figures 10 and 11 show the effect of each predictor separately.

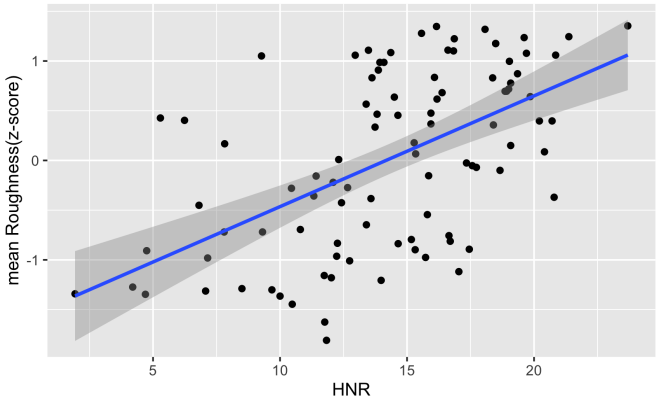


Figure 10: Correlation of HNR to the perceptual ratings of roughness

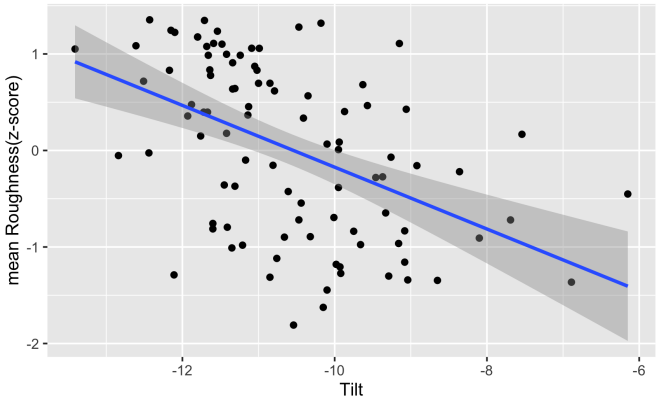


Figure 11: Correlation of spectral tilt to the perceptual ratings of roughness

Decorrelated Breathiness

According to the stepwise linear regression, the most relevant predictors for the decorrelated perceptual ratings for breathiness are CPPS and HNR. CPPS is positively correlated ($\beta = 0.22166$) with the perceptual ratings of breathiness that have been decorrelated from roughness, while HNR is negatively correlated ($\beta = -0.08569$) with the decorrelated perceptual breathiness ratings. Both correlations are significant ($p < 1^{-9}$ for CPPS, and $p < 0.001$ for HNR). The *Adjusted* R^2 of the model is 0.364. The regression formula generated from the model is as follows:

$$\widehat{meanDecorr.BreathinessZScore} = -1.19941 - 0.08569 * \text{HNR} + 0.22166 * \text{CPPS}$$

Figure 12 is a visualization of the correlation between CPPS, HNR, and perceptual ratings for the decorrelated perceptual ratings of breathiness according to the above model, fitted on the data collected.

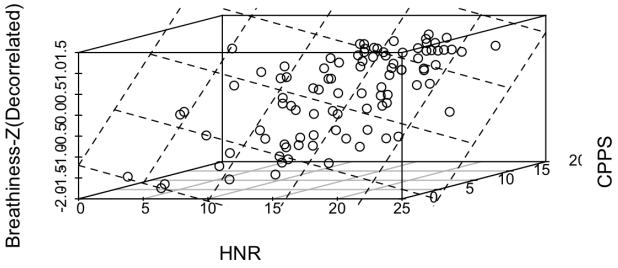


Figure 12: Correlation of the best acoustic predictors for the decorrelated Breathiness to the decorrelated perceptual ratings of Breathiness

Figures 13 and 14 show the influence of each predictor separately.

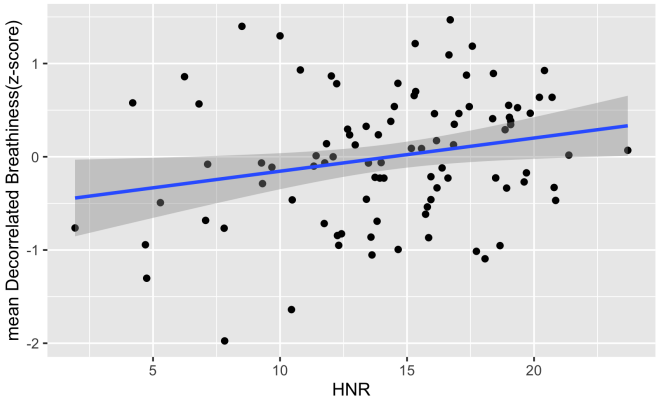


Figure 13: Correlation of HNR to the perceptual ratings of the decorrelated Breathiness

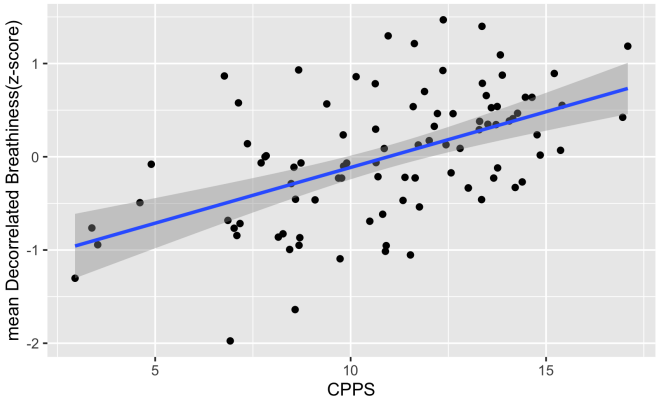


Figure 14: Correlation of Cepstral Peak Prominence (Smoothed) to the perceptual ratings of the decorrelated Breathiness

Reconstructed Breathiness

To model the breathiness that is present in reality (albeit correlated with roughness) as a dependent variable, I now re-construct the predictors from the two models above. The mean of the z-scores of the breathiness ratings made by the two expert listeners (before decorrelation) was used as the dependent variable. Variables chosen by the stepwise model selections above (i.e., HNR, tilt, and CPPS) were included as predictors. The linear regression result shows that CPPS ($p < 1^{-9}$, $\beta = 0.25450$) is positively correlated with perceptual breathiness ratings, while HNR ($p < 0.05$, $\beta = -0.04923$) and Tilt ($p < 0.1$, $\beta = -0.10715$) correlate negatively with perceptual breathiness ratings (adjusted r-squared = 0.5831). The regression formula generated from the model is as follows:

$$\begin{aligned} \widehat{meanBreathinessZScore} = & -3.18284 \\ & -0.04923 * \text{HNR} + 0.25450 * \text{CPPS} \\ & -0.10715 * \text{Tilt} \end{aligned}$$

Figures 15, 16 and 17 show the effect of each predictor separately.

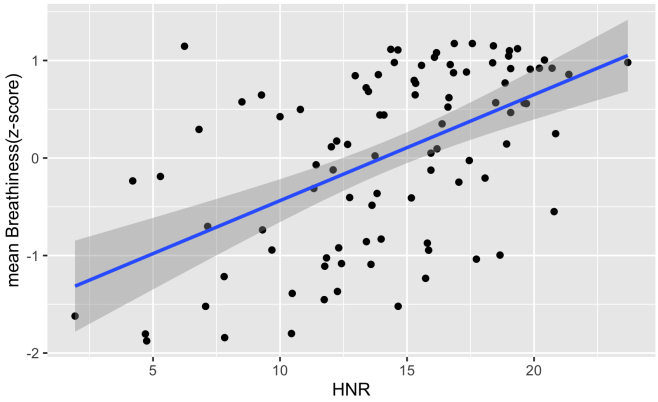


Figure 15: Correlation of HNR to the perceptual ratings of Breathiness

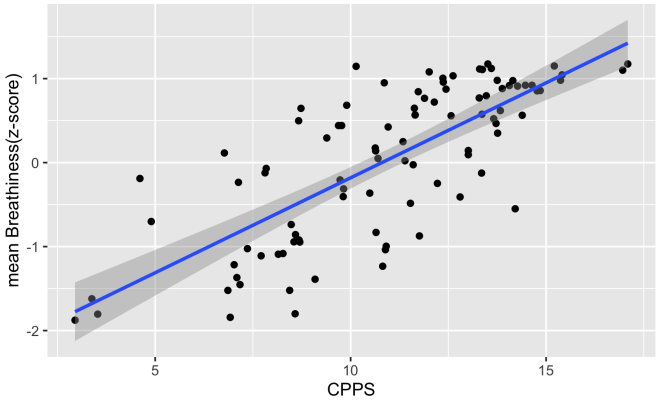


Figure 16: Correlation of Cepstral Peak Prominence to the perceptual ratings of Breathiness

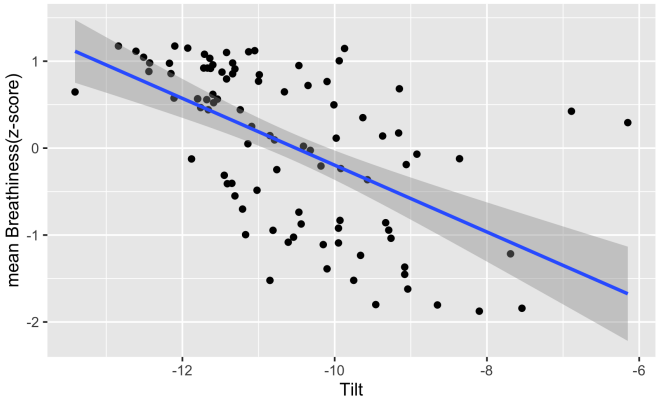


Figure 17: Correlation of Spectral Tilt to the perceptual ratings of Breathiness

Consider that 2.67% of the variance in the perceptual ratings for roughness and 4.0% for breathiness is caused by inter-evaluator variability (see §4.3.1), I calculate below the percentage explained by the models used.

For roughness, the linear regression has an *Adjusted* R^2 of 0.364. Combining this with the proportion of the total variance in roughness ratings explainable by acoustics, the variables selected by the model explains $\frac{0.364}{1-0.023} = 37.26\%$ of the explainable variance. Similarly, for breathiness, the variables selected by the model explains $\frac{0.583}{1-0.04} = 60.73\%$ of the variance explainable by acoustics.

Linearity, Normality, and Other Model Assumptions

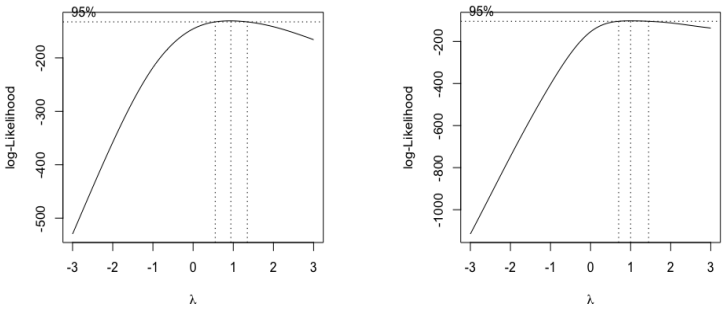
Given that the statistical analyses were done by using linear re-

gressions, there is reason to check the assumptions of linear regression in order to ensure the validity of the results[25]. To do so, the data for $PerceivedRoughness \sim \text{HNR} + \text{Tilt}$ was passed through a Box-Cox transformation (Box & Cox, 1964). The same was done for the data for $DecorrelatedBreathiness \sim \text{HNR} + \text{CPPS}$.

The Box-Cox transformation assumes the relationship between Y and X is in the form of $y^\lambda \sim x$ as opposed to $y \sim x$, and calculates the best λ value in such a way that the distribution is normal and that the model bias induced by heteroscedasticity is minimized. This prevents the model from becoming overly complex or filled with spurious interaction effects.

The best λ for the Roughness model is 0.9393 (95% confidence interval 0.5757..1.3030). For the Decorrelated Breathiness model, the best λ is 1 (95% confidence interval 0.0758 .. 1.4242). Figures 18a and 18b are visualizations of the best λ values for both models and their 95% confidence intervals.

[25]It is true that human hearing perception is more logarithmic than linear for frequency and intensity. This is taken care of by measuring relevant values such as HNR, Slope, and Tilt in dB. The purpose of the current step is to check whether the relationship between the variables in the data is indeed linear, as presupposed by all linear regression models.



(a) Roughness (b) Decorrelated-Breathiness

Figure 18: The best λ and their 95% confidence intervals for the two models

As the best λs for both models are close to 1, and their respective 95% confidence intervals both contain 1, it is not necessary to transform the data, and the relationship is highly likely to be linear. Further, this serves as additional confirmation that the models meet the assumptions for linear regression and are unlikely to have included unnecessary variables or interactions.

5 Discussion

Given that the existing objective measures for voice quality only give a general score of the overall quality, the current study aimed to model voice quality multidimensionally, focusing on the aspects of roughness and breathiness, by using acoustic measurements on both sustained vowels and continuous speech. Step-wise regressions show

that the best predictors for roughness are HNR and Tilt, and for breathinss, the best acoustic predictors are HNR, CPPS, and Tilt.

The currently existing voice quality assessments (e.g. the AVQI) have some limitations. For one, they only give a general score of the overall voice quality. The current study aimed to model different aspects of voice quality separately, specifically, for roughness and breathiness, as these aspects are most relevant for the target population of the study.

Another drawback of the currently existing objective voice quality measurements is that they rely heavily on sustained vowels and vowel-specific measurements. The current study took an attempt to develop models that are more appropriate for connected speech samples.

The motivations for building such models are as follows. First, in terms of practicality, not all post-treatment speakers can produce a sustained vowel of the required length of three seconds for reliable perturbation measurements. More importantly, sustained vowels are not what a listener usually bases her judgment of voice quality on in everyday communication situations. Sustained vowels that are the same lengths as the samples required by the measurements rarely, if ever, occur in unscripted speech. Moreover, even if such sustained vowels do occur in day-to-day speech, sustained vowels alone do not reflect the vocal characteristics of a speaker. For instance, sustained vowels do not show the rapid onset or termination in temporal modulation that would occur in connected speech. Additionally, for the perturbation measurements to be reliable, a stable part of the sustained vowel needs to be selected to be used as input. However, the sustained vowels produced by a substantial portion of patients sim-

ply do not contain a stable part at all. This hence leaves a diagnostic gap in the population that can be examined by such methods.

The final models found by the current study were able to explain 38% of the roughness variation, and 60% of the breathiness variation in the data. One out of the two predictors that best explain roughness, and two out of three predictors that best explain breathiness are suited to be used on continuous speech alone. Comparatively, among the six variables used in the AVQI, three out of six are vowel-specific measurements[26]. However, the fact that the final models generated by the step-wise regression in the current study did not select any perturbation measures might have to do with the input used for the perturbation measurements being not ideal. Namely, the perturbation measurements used in the current study were obtained from the AVQI scheme. As was mentioned in §2.3.2, the input used for perturbation measurements in AVQI is the concatenation of a sustained vowel and the voiced parts of the four-second connected speech sample[27]. The nature of the input thus makes the perturbation readings difficult to interpret since perturbation measures are theoretically meant to be used on sustained vowels only. Improvements can be made in future research by making more accurate and reliable perturbation measurements, should the purpose be to compare the validity of different types of measurements.

As subjective judgment is considered the ground truth for the acoustic analyses, and although inter-rater variation is taken into account in the results, inter-rater disagreement is still worth men-

[26] Even so, those measurements were still made on vowel-speech concatenations in the AVQI.

[27] Exception: the measurement for jitter was added to the script by linguists working at the NKI, and only used the sustained vowel as input. See §4.2.

tioning. The agreement between the two expert listeners who participated in the listening experiment turned out to be higher than what is reported in the literature. This could be interpreted in either direction. On the one hand, it could mean that the subjective scores of the current study are more reliable than average. On the other hand, it is also possible that the listeners are not diverse enough since both of the expert listeners work within the same clinic and might have had similar influences during training. Future research should recruit more, ideally multi-institutional, expert listeners. Similarly, the acoustic analyses conducted in the current study were based on 96 speech samples of 45 speakers, all of whom underwent treatment for laryngeal carcinoma of the same stage. Further research needs to be done to validate the results on a wider population by including more speakers and more severities in dysphonia.

Nonetheless, the fact that the between-evaluator agreement is lower for roughness compared to breathiness is also indicative of the difference in the complexity of the two perceptual voice qualities. The aerodynamics involved in breathiness is mostly only the unintended excessive aspiration during phonation. Whereas any type of atypical vibration can cause, and be perceived as, roughness. Kreiman and Gerratt (2000b) also confirm that the unanimously agreed "primarily rough" voices in their experiments are acoustically heterogeneous, while breathy ones are acoustically similar. Given that laryngeal carcinoma can happen at any location in the larynx and the treatment can leave a myriad of types of irregularities in the vocal folds, roughness is bound to present in wider varieties and be more difficult to model than breathiness.

For the same reason, irregular shaping of the vocal folds can lead

to both excessive aspiration (due to lack of proper closure) and irregular vibration at the same time. This explains the high correlation between breathiness and roughness in this particular patient population. However, the high co-occurrence does not mean that the two qualities are too similar to be distinguished. Breathiness and roughness are different both articulatorily and auditorily. A case in point would be that whisper phonation and creaky phonation are phonemic in many languages (Gujerati, Hausa, Jalapa Mazatec, !Xóõ, to name a few. See, e.g., Ladefoged, 1983 and Ladefoged & Johnson, 2014 for more examples). Moreover, research has shown that human listeners can tell apart breathiness and roughness even when the two features are not phonemic in their native languages (e.g., Kreiman & Gerratt, 2000b). The high co-occurrence of the breathiness and roughness is most likely due to their shared origin i.e., the irregular shaping and atypical texture of the vocal folds after treatment, rather than any acoustic or auditory resemblance.

6 Conclusion

Stepwise linear regression analyses yielded two acoustic models for the multiparametric measurement of breathiness and roughness, respectively. The correlation between the estimates made by the breathiness model and the auditory-perceptual rating is strong (*Adjusted* $R^2 = 0.583$), so is the correlation between the estimates made by the roughness model and the auditory-perceptual rating (*Adjusted* $R^2 = 0.364$).

6.1 Strengths

Existing objective measurements for voice quality usually use a concatenation of a sustained vowel and some continuous speech as input, and only provide a score for the overall "goodness" of the voice quality. The current study aimed to give a multidimensional objective assessment, as opposed to an overall score, for voice quality in the treatment of laryngeal carcinoma. The end results made improvements on the existing objective judgment tools in two aspects: dimension specificity, and types of measurements. The dimension-specific ratings generated in the current study give a more comprehensive evaluation of the voice, and the improvement made in the type of measurements lays a foundation for potentially further improvement in the ecological validity of objective evaluations of voice quality.

6.1.1 Dimension Specificity

In terms of dimension specificity, the models generated in the current study provide scores for two dimensions of voice quality: roughness and breathiness, instead of one general goodness score. The models were able to explain 38% of the variation in roughness, and 60% of the variation in breathiness, with the inter-rater variability taken into account.

6.1.2 Types of Measurements

In regard to measurement types, none out of the five parameters that are deemed most explanatorily useful are perturbation measures. The use of spectral and cepstral type of variables make the

models more suitable to be used with connected-speech-only inputs, without having to rely on sustained vowels which do not represent the vocal characteristics of the speaker, to begin with. Additionally, using non-perturbation type measurements makes the results more interpretable when the input includes connected speech, comparing to what is used in e.g., the AVQI.

This change in the type of measurements hence makes it possible for future research to explore more on multiparametric-multidimensional objective measurements that only use connected speech as input, and hence potentially help widen the population that can be assessed by multidimensional objective models, as well as increase the ecological validity of such type of objective evaluations.

6.2 Limitations

6.2.1 Sample Size

The acoustic analysis in the present study was carried out on a sample size of 45 speakers and 96 voice samples, of which all speakers have the same type of cancer with similar severity, and all connected speech samples were in the Dutch language. On the one hand, this limits the noise in the data and contributes to the reliability of the results for the target population of this research project. On the other hand, one obvious improvement that can be made is to validate the results on a larger-sized sample, on a wider patient population with more diverse types of voice problems, and in other languages[28].

Besides the sample size of the speakers in the acoustic analysis,

[28] Ladefoged (1983) P. 351 "One person's voice disorder is another person's phoneme."

improvements can also be made in the sample size of the listening experiment. The present study follows the convention of treating the subjective judgment of speech-language pathologists as the ground truth in validating acoustic measures. However, such ground truth was only collected from two speech-language pathologists who work at the same institute. Although the agreement between the two expert listeners is high, a more representative sample of perceptual judgment can be obtained by e.g., a multi-center study with a wider range of speech-language pathologists.

6.2.2 Stimuli

Another limitation of the present study concerns the type and quality of recordings. Aside from the fact that all the recordings were collected during the patients' visit to the speech-language pathologists instead of being recorded in a more ideal setting such as a sound-proof speech lab, the types of stimuli could also be improved.

The aim of the study is to develop models that can give judgments on specific dimensions while hoping to also end up with measurements that are suitable to be used on connected speech. However, the data used in the analyses for stepwise regression were measured on the concatenation of a sustained vowel and several seconds of read speech. Although it could be argued that having included sustained vowels in the spectral- and cepstral-based measurements are unlikely to have had a negative impact on the results, further research can be done to compare and validate the predictive power of the non-perturbation type measurements by collecting acoustic data measured from connected speech alone.

In the same vein, as was mentioned in §5, some of the perturbation-type measurements (Shimmer, ShdB) were measured on the concatenated voice samples. Although this is based on the extensively validated and established scheme of the AVQI, the dismissal of perturbation measurements in the final models might not have done justice to the significance of perturbation measurements in measuring voice quality in general, given that the values were not measured on what they were intended to be measured from, i.e. sustained vowels only. Nevertheless, regarding the purpose of the current study, concatenated samples consisting of a sustained vowel and voiced parts of connected speech are closer to the ideal input of connected speech alone, in terms of the reliability in model selection.

References

Aaltonen, L.-M., Rautiainen, N., Sellman, J., Saarilahti, K., Mäkitie, A., Rihkanen, H., ... others (2014). Voice quality after treatment of early vocal cord cancer: a randomized trial comparing laser surgery with radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, *90*(2), 255–260.

Agarwal, J. P., Baccher, G. K., Waghmare, C. M., Mallick, I., Ghosh-Laskar, S., Budrukkar, A., ... others (2009). Factors affecting the quality of voice in the early glottic cancer treated with radiotherapy. *Radiotherapy and Oncology*, *90*(2), 177–182.

Alwan, A., Bangayan, P., Gerratt, B. R., Kreiman, J., & Long, C. (2000). Analysis by synthesis of pathological voices using the klatt synthesizer. In R. D. Kent & M. J. Ball (Eds.), *Voice quality measurement* (p. 307-335). San Diego, CA: Singular Publishing.

Austin, G. (1806). *Chironomia; or, a treatise on rhetorical delivery.* T. Cadell and W. Davies.

Awan, S. N., & Roy, N. (2005). Acoustic prediction of voice type in women with functional dysphonia. *Journal of voice*, *19*(2), 268–282.

Awan, S. N., Roy, N., & Dromey, C. (2009). Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model. *Clinical linguistics & phonetics*, *23*(11), 825–841.

Balm, A. J. M. (2014). Laryngeal and hypolaryngeal cancer: intervention approaches. In C. J. v. As-Brooks & E. C. Ward (Eds.), *Head and neck cancer : Treatment, rehabilitation, and outcomes* (2nd ed., p. 151-172). Oxford, UK: Plural Publishing, Inc.

Barsties v. Latoszek, B., Maryn, Y., Gerrits, E., & De Bodt, M. (2017). The acoustic breathiness index (abi): A multivariate acoustic model for breathiness. *Journal of Voice*, *31*(4), 511–e11.

Barsties v. Latoszek, B., Ulozaitė-Stanienė, N., Petrauskas, T.,

Uloza, V., & Maryn, Y. (2019). Diagnostic accuracy of dysphonia classification of dsi and avqi. *The Laryngoscope*, *129*(3), 692–698.

Bele, I. V. (2005). Reliability in perceptual analysis of voice quality. *Journal of Voice*, *19*(4), 555–573.

Blankenship, B. (2002). The timing of nonmodal phonation in vowels. *Journal of Phonetics*, *30*(2), 163–191.

Boersma, P., & Weenink, D. (2020). *Praat: doing phonetics by computer*. Retrieved from http://www.praat.org

Borgert, B., Healy, M., & Tukey, J. (1963). The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe craking. In *Proc. symp. on time series analysis* (pp. 209–243).

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243.

Carding, P., Steen, I., Webb, A., Mackenzie, K., Deary, I., & Wilson, J. (2004). The reliability and sensitivity to change of acoustic measures of voice quality. *Clinical Otolaryngology & Allied Sciences*, *29*(5), 538–544.

Carding, P., Wilson, J. A., MacKenzie, K., & Deary, I. J. (2009). Measuring voice outcomes: state of the science review. *The journal of laryngology and otology*, *123*(8), 823.

Carrara-de Angelis, E., Feher, O., Barros, A. P. B., Nishimoto, I. N., & Kowalski, L. P. (2003). Voice and swallowing in patients enrolled in a larynx preservation trial. *Archives of Otolaryngology–Head & Neck Surgery*, *129*(7), 733–738.

Cavalli, L., & Hirson, A. (1999). Diplophonia reappraised. *Journal of Voice*, *13*(4), 542–556.

Childers, D. G., Skinner, D. P., & Kemerait, R. C. (1977). The cepstrum: A guide to processing. *Proceedings of the IEEE*, *65*(10), 1428–1443.

Chow, L. Q. (2020). Head and neck cancer. *New England Journal of Medicine*, *382*(1), 60–72.

Cicero, M. T. (55 B.C./1976). *De oratore, Books I & II* (E. W. Sut-

ton, H. Rackham, et al., Eds. & Trans.). Harvard University Press.

Collins, M., McDonald, R., Stanley, R., Donovan, T., & Bonebrake, C. F. (1993). Severe paradoxical dysphonia in two young women. *American Journal of Speech-Language Pathology*, *2*(3), 52–55.

Cranen, B., & de Jong, F. (2000). Laryngostroboscopy. In R. D. Kent & M. J. Ball (Eds.), *Voice quality measurement* (p. 257-267). San Diego, CA: Singular Publishing.

Dejonckere, P. H., Bradley, P., Clemente, P., Cornut, G., Crevier-Buchman, L., Friedrich, G., ... Woisard, V. (2001). A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *European Archives of Oto-rhino-laryngology*, *258*(2), 77–82.

Dejonckere, P. H., & Lebacq, J. (1983). An analysis of the diplophonia phenomenon. *Speech Communication*, *2*(1), 47–56.

De Krom, G. (1995). Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments. *Journal of Speech, Language, and Hearing Research*, *38*(4), 794–811.

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in bioinformatics*, *21*(2), 553–565.

Ferrand, C. T. (2002). Harmonics-to-noise ratio: an index of vocal aging. *Journal of voice*, *16*(4), 480–487.

Ford, C. N., & Connor, N. P. (2000). Phonatory effects of mass lesions. In R. D. Kent & M. J. Ball (Eds.), *Voice quality measurement* (pp. 377–384). San Diego, CA: Singular Publishing.

Fritzen, B., Hammarberg, B., Gauffin, J., Karlsson, I., & Sundberg, J. (1986). Breathiness and insufficient vocal fold closure. *Journal of Phonetics*, *14*(3-4), 549–553.

Garellek, M., & Keating, P. (2011). The acoustic consequences of phonation and tone interactions in jalapa mazatec. *Journal of the International Phonetic Association*, *41*(2), 185–205.

Gaudernack, L. (1934). Some notes on the practical measurement of the degree of amplitude modulation. *Proceedings of the Institute of Radio Engineers*, *22*(7), 819–846.

Ghio, A., Révis, J., Merienne, S., & Giovanni, A. (2013). Top-down mechanisms in dysphonia perception: the need for blind tests. *Journal of Voice*, *27*(4), 481–485.

Halberstam, B. (2004). Acoustic and perceptual parameters relating to connected speech are more reliable measures of hoarseness than parameters relating to sustained vowels. *ORL*, *66*(2), 70–73.

Hillenbrand, J. (1988). Perception of aperiodicities in synthetically generated voices. *The Journal of the Acoustical Society of America*, *83*(6), 2361–2371.

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, *37*(4), 769–778.

Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research*, *39*(2), 311–321.

Hillmen, R. E., & Kobler, J. B. (2000). Aerodynamic measures of voice production. In R. D. Kent & M. J. Ball (Eds.), *Voice quality measurement* (pp. 245–255). San Diego, CA: Singular Publishing.

Hirano, M. (1981). *Clinical examination of voice*. Wien: Springer.

Hixon, T. J., Goldman, M. D., & Mead, J. (1973). Kinematics of the chest wall during speech production: Volume displacements of the rib cage, abdomen, and lung. *Journal of speech and hearing research*, *16*(1), 78–115.

Hogikyan, N. D., & Sethuraman, G. (1999). Validation of an instrument to measure voice-related quality of life (v-rqol). *Journal of voice*, *13*(4), 557–569.

Imaizumi, S. (1986). Acoustic measures of roughness in pathological voice. *Journal of Phonetics*, *14*(3-4), 457–462.

Jacobson, B. H., Johnson, A., Grywalski, C., Silbergleit, A., Jacob-

son, G., Benninger, M. S., & Newman, C. W. (1997). The voice handicap index (vhi) development and validation. *American Journal of Speech-Language Pathology*, *6*(3), 66–70.

Kempster, G. B., Gerratt, B. R., Abbott, K. V., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, *87*(2), 820–857.

Klich, R. J. (1982). Relationships of vowel characteristics to listener ratings of breathiness. *Journal of Speech, Language, and Hearing Research*, *25*(4), 574–580.

Koreman, J., Pützer, M., & Just, M. (2004). Correlates of varying vocal fold adduction deficiencies in perception and production: methodological and practical considerations. *Folia phoniatrica et logopaedica*, *56*(5), 305–320.

Kramer, E. (2011). *Predicting perceptual voice quality from objective voice parameters in dysphonic patients* (PhD dissertation). Universität zu Lübeck.

Kreiman, J., & Gerratt, B. (2000a). Measuring vocal quality. In R. D. Kent & M. J. Ball (Eds.), *Voice quality measurement* (pp. 73–101). San Diego, CA: Singular Publishing.

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America*, *104*(3), 1598–1608.

Kreiman, J., & Gerratt, B. R. (2000b). Sources of listener disagreement in voice quality assessment. *The Journal of the Acoustical Society of America*, *108*(4), 1867–1876.

Kreiman, J., Gerratt, B. R., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America*, *122*(4), 2354–2364.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: review, tu-

torial, and a framework for future research. *Journal of Speech, Language, and Hearing Research*, *36*(1), 21–40.

Krengli, M., Policarpo, M., Manfredda, I., Aluffi, P., Gambaro, G., Panella, M., & Pia, F. (2004). Voice quality after treatment for t1a glottic carcinoma radiotherapy versus laser cordectomy. *Acta oncologica*, *43*(3), 284–289.

Ladefoged, P. (1983). The linguistic use of different phonation types. In B. D & J. Abbs (Eds.), *Vocal fold physiology: Contemporary research and clinical issues* (p. 351-360). San Diego, CA: College-Hill Press.

Ladefoged, P. (2003). *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Wiley-Blackwell.

Ladefoged, P., & Johnson, K. (2014). *A course in phonetics*. Cengage learning.

Lee, S. H., Hong, K. H., Kim, J. S., & Hong, Y. T. (2019). Perceptual and acoustic outcomes of early-stage glottic cancer after laser surgery or radiotherapy: a meta-analysis. *Clinical and Experimental Otorhinolaryngology*, *12*(3), 241.

Loughran, S., Calder, N., MacGregor, F., Carding, P., & MacKenzie, K. (2005). Quality of life and voice following endoscopic resection or radiotherapy for early glottic cancer. *Clinical Otolaryngology*, *30*(1), 42–47.

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. *Journal of voice*, *24*(5), 540–555.

Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., & Corthals, P. (2009). Acoustic measurement of overall voice quality: a meta-analysis. *The Journal of the Acoustical Society of America*, *126*(5), 2619–2634.

Moers, C., Möbius, B., Rosanowski, F., Nöth, E., Eysholdt, U., & Haderlein, T. (2012). Vowel-and text-based cepstral analysis of chronic hoarseness. *Journal of Voice*, *26*(4), 416–424.

Moore, G. P. (1976). Observations on laryngeal disease, laryngeal

behavior and voice. *Annals of Otology, Rhinology & Laryngology*, *85*(5), 553–564.

Morrison, M. (1997). Pattern recognition in muscle misuse voice disorders: How i do it. *Journal of Voice*, *11*(1), 108–114.

Nawka, T., Anders, L. C., & Wendler, J. (1994). Die auditive beurteilung heiserer stimmen nach dem rbh-system. *Sprache Stimme Gehör*, *18*(3), 130–133.

Orlikoff, R. F., & Kraus, D. H. (1996). Dysphonia following nonsurgical management of advanced laryngeal carcinoma. *American Journal of Speech-Language Pathology*, *5*(3), 47–52.

Parsa, V., & Jamieson, D. G. (2001). Acoustic discrimination of pathological voice. *Journal of Speech, Language, and Hearing Research*.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rammage, L. A., Peppard, R. C., & Bless, D. M. (1992). Aerodynamic, laryngoscopic, and perceptual-acoustic characteristics in dysphonic females with posterior glottal chinks: a retrospective study. *Journal of Voice*, *6*(1), 64–78.

Roh, J.-L., Kim, A.-Y., & Cho, M. J. (2005). Xerostomia following radiotherapy of the head and neck affects vocal function. *Journal of clinical oncology*, *23*(13), 3016–3023.

Roh, J.-L., Kim, H. S., & Kim, A.-Y. (2006). The effect of acute xerostomia on vocal function. *Archives of Otolaryngology–Head & Neck Surgery*, *132*(5), 542–546.

Rydell, R., Schalén, L., Fex, S., & Elner, Å. (1995). Voice evaluation before and after laser excision vs. radiotherapy of t1a glottic carcinoma. *Acta oto-laryngologica*, *115*(4), 560–565.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Shrivastav, R., Eddins, D. A., & Anand, S. (2012). Pitch strength of normal and dysphonic voices. *The Journal of the Acoustical Society of America*, *131*(3), 2261–2269.

Solomon, N. P., Helou, L. B., & Stojadinovic, A. (2011). Clinical

versus laboratory ratings of voice using the cape-v. *Journal of Voice*, *25*(1), e7–e14.

Starmer, H. M., Tippett, D. C., & Webster, K. T. (2008). Effects of laryngeal cancer on voice and swallowing. *Otolaryngologic Clinics of North America*, *41*(4), 793–818.

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). MIT press.

Tan, I. B., Stoker, S. D., & Smeele, L. E. (2014). Oral, oralpharyngeal, and nasopharyngeal cancer: intervention approaches. In C. J. v. As-Brooks & E. C. Ward (Eds.), *Head and neck cancer : Treatment, rehabilitation, and outcomes* (2nd ed., p. 103-120). Oxford, UK: Plural Publishing, Inc.

Van Son, R. (2020). *Akouste*. Retrieved from http://doi.org/10.5281/zenodo.3712142

Vassilakis, P. N. (2005). Auditory roughness as a means of musical expression. *Selected Reports in Ethnomusicology*, *12*(119-144), 122.

Verdonck-De Leeuw, I. M., Hilgers, F. J., Keus, R. B., Koopmans-Van Beinum, F. J., Greven, A. J., De Jong, J. M., ... Bartelink, H. (1999). Multidimensional assessment of voice characteristics after radiotherapy for early glottic cancer. *The Laryngoscope*, *109*(2), 241–248.

Vilaseca, I., Huerta, P., Blanch, J. L., Fernández-Planas, A. M., Jiménez, C., & Bernal-Sprekelsen, M. (2008). Voice quality after CO2 laser cordectomy—what can we really expect? *Head & Neck: Journal for the Sciences and Specialties of the Head and Neck*, *30*(1), 43–49.

Ward, P. H., Sanders, J. W., Goldman, R., & Moore, G. P. (1969). Lxvii diplophonia. *Annals of Otology, Rhinology & Laryngology*, *78*(4), 771–777.

Wong, D., Ito, M. R., Cox, N. B., & Titze, I. R. (1991). Observation of perturbations in a lumped-element model of the vocal folds with application to some pathological cases. *The Journal of the Acoustical Society of America*, *89*(1), 383–394.

Wuyts, F. L., De Bodt, M. S., & Van de Heyning, P. H. (1999). Is the reliability of a visual analog scale higher than an ordinal

scale? an experiment with the grbas scale for the perceptual evaluation of dysphonia. *Journal of Voice*, *13*(4), 508–517.

Ziethe, A., Patel, R., Kunduk, M., Eysholdt, U., & Graf, S. (2011). Clinical analysis methods of voice disorders. *Current Bioinformatics*, *6*(3), 270–285.

Zwicker, E., & Fastl, H. (1990). *Psychoacoustics: Facts and models* (Vol. 22). Springer-Verlag.

Appendices

A R Script for Perception Experiment

```
AVQI_ <- read.csv("AVQI_results_new.csv", header =
    TRUE)
nrow(AVQI_)

listener1 <- read.csv("Results_SLP1.csv", header =
    TRUE, sep = ";")
nrow(listener1)

#Cmd-F "Stimuli/" replaced all with "" before
    loading:

listener2 <- read.csv("Results_SLP2.csv", header =
    TRUE, sep = ";")
nrow(listener2)

#Removing Speaker1 in AVQI:

library(dplyr)

AVQI <- AVQI_ %>%
  filter(!Speaker == "S1")

#Renaming columns (and deleting "Num" to reduce
    total row numbers in later steps):
SLP1 <- listener1 %>%
  rename(Roughness = Answer1, Breathiness =
    Answer2)%>%
  select(-Num)

SLP2 <- listener2 %>%
  rename(Roughness = Answer1, Breathiness =
    Answer2)%>%
```

```
  select(-Num)

SLP <- rbind(SLP1,SLP2)

#Merging the two tables by matching "Speaker"
    *and* "T":
library(dplyr)
#Inserting an identifier for each token in the
    AVQI measurement results by coombining
    ``Speaker'' and ``T'':
AVQI$dummy <- paste(AVQI$Speaker, "-", AVQI$T)
glimpse(AVQI)

#Creating an identifier for each token in the
    results from listening experiment by combining
    ``Speaker'' and ``T'':
SLP$token <- paste(SLP$Speaker, "-", SLP$T)
glimpse(SLP)

#merge AVQI results and Listening Experiment
    results by matching token identifiers
    ("left-join" only appends AVQI entries that has
    SLP rating correspondents):
library(dplyr)
merged <-left_join(SLP,AVQI, by = c("token" =
    "dummy"))%>%
  rename(Speaker = Speaker.x)#%>%
head(merged)

#deleting redundant rows:
table <- merged %>%
  select(-A, -T.y, -Z, -Version, -Speaker.y)
table

#writing new table:
write.csv(table, file = "merged.csv")

##linear regression:
```

```
library(lme4)
#setting contrast:
contrast <- cbind(c(0.5, -0.5)) #SLP1,SLP2
colnames(contrast) <- c("SubjectSLP1-SubjectSLP2")
table$Subject <- as.factor(table$Subject)
contrasts(table$Subject) <- contrast

#decorrelating breathiness from roughness:
brfromtable <- lm(Breathiness ~ Roughness, data =
    table)

brCorr <- fitted(brfromtable, table$Roughness)
table$BDecorr <- (table$Breathiness - brCorr)

lmerBreathy <- lmer(Breathiness ~ Subject +
    (Subject | Speaker), data = table, REML = TRUE,
    na.action = "na.omit")
summary(lmerBreathy)

lmerRough <- lmer(Roughness ~ Subject + (Subject |
    Speaker), data = table, REML = TRUE, na.action
    = "na.omit")
summary(lmerRough)
```

B R Script for Acoustic Analysis

```
## Generating the Files

library(dplyr)
#reading in listening test results:
listener1 <- read.csv("Results_SLP1.csv", header =
    TRUE, sep = ";")
listener2 <- read.csv("Results_SLP2.csv", header =
    TRUE, sep = ";")
```

```
#renaming the two answers, removing the column
    "Num",
#and adding a dummy variable (= speaker + T) as an
    identifier for when joining the datasets:
SLP1 <- listener1 %>%
  dplyr::select(-Num) %>%
  rename(Roughness = Answer1, Breathiness =
    Answer2)
SLP1$dummy <- paste(SLP1$Speaker, "-", SLP1$T)
nrow(SLP1) #96
SLP2 <- listener2 %>%
  rename(Roughness = Answer1, Breathiness =
    Answer2)%>%
  dplyr::select(-Num)
SLP2$dummy <- paste(SLP2$Speaker, "-", SLP2$T)
nrow(SLP2) #96

AVQI <- read.csv("AVQI_results_new.csv")
#also adding an dummy variable (= speaker + T) as
    an identifier when joining the dataset to the
    SLP table:
AVQI$dummy <- paste(AVQI$Speaker, "-", AVQI$T)

#converting "Jitter" as a numeric value from a
    factor (this introduces new NAs):
AVQI$Jitter <- as.numeric(AVQI$Jitter)

#defining the mean and sd of the Brethiness and
    Roughness scores:
library(dplyr)
mB1 <- mean(SLP1$Breathiness)#SLP1
sdB1 <- sd(SLP1$Breathiness)#SLP1
mB2 <- mean(SLP2$Breathiness)#SLP2
sdB2 <- sd(SLP2$Breathiness)#SLP2
mR1 <- mean(SLP1$Roughness)#SLP1
sdR1 <- sd(SLP1$Roughness)#SLP1
mR2 <- mean(SLP2$Roughness)#SLP2
sdR2 <- sd(SLP2$Roughness)#SLP2
```

```
#adding a column containing z-values for each SLP
    score table:
zSLP1 <- SLP1 %>%
  mutate(zBreathy = (Breathiness-mB1)/sdB1, zRough
    = (Roughness-mR1)/sdR1) %>%
  select(-A, -Z, -T)
zSLP2 <- SLP2 %>%
  mutate(zBreathy = (Breathiness-mB2)/sdB2, zRough
    = (Roughness-mR2)/sdR2) %>%
  select(-A, -Z, -T)

#joining the scores from two SLPs, by matching the
    dummy identifier, and removing redundant rows
zSLP <- full_join(zSLP1, zSLP2, by = "dummy") %>%
  select(-Speaker.x, -Speaker.y)

#calculating the mean of z scores for each token
    of recording:
zSLP$mzB = rowMeans(zSLP[,c('zBreathy.x',
    'zBreathy.y')], na.rm=TRUE)
zSLP$mzR = rowMeans(zSLP[,c('zRough.x',
    'zRough.y')], na.rm=TRUE)
zSLP$mR = rowMeans(zSLP[,c('Roughness.x',
    'Roughness.y')], na.rm=TRUE)
zSLP$mB = rowMeans(zSLP[,c('Breathiness.x',
    'Breathiness.y')], na.rm=TRUE)

#join the two SLP results tables with AVQI by
    matching both Speaker and T:
zAVQI_ <-inner_join(AVQI,zSLP, by = "dummy")
zAVQI <- na.omit(zAVQI_) #removing rows containing
    NAs

#write the file for later convenience:
write.csv(zAVQI, "zAVQI.csv")
#zAVQI.csv now has the AVQI data and the SLP
    ratings combined and matched,
```

```
#and with the variables in the right classes

## Stepwise model comparison

#decorrelating breathiness from roughness:
breathyRough <- lm(mzB ~ mzR, data = zAVQI)
mzBCorr <- fitted(breathyRough, zAVQI$mzR)
zAVQI$mzBDecorr <- (zAVQI$mzB - mzBCorr)

#using step to determine the best linear model:
#roughness
zRoughModel.tmp <- lm(mzR ~ CPPS + HNR + Jitter +
    Shimmer + ShdB + Slope + Tilt, zAVQI)
#using BIC instead of AIC
zRoughModel <- step(zRoughModel.tmp, k =
    log(nrow(zAVQI)))

#breathiness
zBrDecorrModel.tmp <- lm(mzBDecorr ~ CPPS + HNR +
    Jitter + Shimmer + ShdB + Slope + Tilt, zAVQI)
#using BIC instead of AIC (by changing k):
zBrDecorrModel <- step(zBrDecorrModel.tmp, k =
    log(nrow(zAVQI)))

#summarizing the two models:
summary(zRoughModel)
summary(zBrDecorrModel)

## Generating the output

#running the models:
rough.model <- lm(mzR ~ HNR + Tilt, data = zAVQI)
summary(rough.model)

breathyDecorr.model <- lm(mzBDecorr ~ HNR + CPPS,
    data = zAVQI)
summary(breathyDecorr.model)
```

```
#reconsrtucting de facto breathiness:
#(using "mzB" instead of "mzBDecorr" as the
    independent variable here
breathy.model <- lm(mzB ~ HNR + CPPS + Tilt, data
    = zAVQI)
summary(breathy.model)

##checking model assumptions

#transforming negative values for Box-Cox:
library(dplyr)
pzAVQI<- zAVQI %>% mutate(
  slope = - Slope,
  tilt = - Tilt,
  pmzR = 2 + mzR,
  pmzBDecorr = 2 + mzBDecorr
  )

#writing out the model in transformed values
    (positive):
p.rough.model <- lm(pmzR ~ HNR + tilt, data =
    pzAVQI)
p.breathyDecorr.model <- lm(pmzBDecorr ~ HNR +
    CPPS, data = pzAVQI)

library(MASS)
#boxcox for roughness:
bcR = boxcox (p.rough.model, lambda = seq(-3, 3))

#extract best lambda:
best.lamR = bcR$x[which(bcR$y == max(bcR$y))]
best.lamR #=0.9393
#95% conofidence interval
ciR = bcR$x[bcR$y > max(bcR$y) - 1/2 *
    qchisq(.95,1)]
ciR #=0.5757..1.3030

#boxcox for decorrelated breathiness
```

```
bcB = boxcox (p.breathyDecorr.model, lambda =
    seq(-3, 3))

#extract best lambda
best.lamB = bcB$x[which(bcB$y == max(bcB$y))]
best.lamB #=1

ciB = bcB$x[bcB$y > max(bcB$y) - 1/2 *
    qchisq(.95,1)]
ciB
```