ACOUSTIC MEASUREMENT OF OVERALL VOICE QUALITY IN SUSTAINED VOWELS AND CONTINUOUS SPEECH

Youri Maryn

Promotors
Prof. Dr. P. Van Cauwenberge
Prof. Dr. M. De Bodt

Ghent University
Faculty of Medicine and Health Sciences
Department of Otorhinolaryngology

Acoustic measurement of overall voice quality in sustained vowels and continuous speech

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor in Health Sciences

Youri Maryn

Promotors
Prof. Dr. Paul Van Cauwenberge
Prof. Dr. Marc De Bodt

Gent, 2010

Acoustic measurement of overall voice quality in sustained vowels and continuous speech.
(Doctoral thesis, Ghent University, Belgium)

(Promotors: Prof. Dr. Paul Van Cauwenberge and Prof. Dr. Marc De Bodt)

Cover design & lay-out: Youri Maryn

Printing: Nevelland Graphics, cvba-so.

ISBN 978-90-9024980-3
NUR 896

TABLE OF CONTENTS

DANKWOORD (ACKNOWLEDGMENTS)

Ik geloof dat wetenschappelijk onderzoek (en doctoreren in het bijzonder) een roeping is, op voorwaarde dat er geroepen wordt. In de roes van gevoelens die opborrelen bij het finaliseren van dit proefschrift, besef ik het fortuin van de vele roepende mensen rond mij. U was talrijk en constructief, u bevoorraadde aan troost en relativering, u zorgde voor het positieve in mijn leven. Erzonder was dit doctoraat nooit gelukt.

In allereerste instantie wil ik mijn promotor Prof. Dr. Marc De Bodt (Universiteit Antwerpen, Universiteit Gent) bedanken. Ik heb het bijzondere privilege gehad om u in de zomer van 1997 te leren kennen in het kader van mijn graduaatsthesis en later mijn eerste klinische stage in het Universitair Ziekenhuis van Antwerpen. U heeft mij indertijd geïnspireerd om universitaire studies aan te vatten en u was als promotor ook nauw betrokken in de realisatie van mijn licentiaatsthesis ... en nu is er mijn doctoraatsproefschrift. Als een echte mentor, en bovenal ook als een dierbare vriend, heeft u mij op verkenning doorheen stem en spraak meegenomen en gestimuleerd om logopedie op een serieuze en wetenschappelijke basis te beoefenen. Ik durf niet te denken welke logopedist ik zou zijn indien ik u nooit had ontmoet. Uw bijdrage, in veel meer dan alleen dit doctoraat, is fenomenaal en er is zoveel waarvoor ik u wil bedanken. Geconfronteerd echter met het besef dat ik daarvoor de gesofisticeerde taal tekortschiet, verkies ik mijn erkentelijkheid met een citaat uit uw eigen doctoraatsproefschrift uit te drukken: *"Ik hoop dat dit werk een aanmoediging zal zijn voor collega's om de weg van het wetenschappelijk onderzoek in te slaan"*. Marc, ik hoop hieraan tegemoet te zijn gekomen. Bedankt voor de inspiratie, het was de stuwkracht doorheen al mijn wetenschappelijke activiteiten.

Ook mijn promotor Prof. Dr. Paul Van Cauwenberge (Universiteit Gent) wens ik oprecht te bedanken voor zijn onontbeerlijke rol in dit project. Ondanks uw eigen drukke agenda als Diensthoofd Neus-, Keel- en Oorheelkunde, als Decaan van de Faculteit Geneeskunde en Gezondheidswetenschappen of zelfs als Rector van de Universiteit Gent heeft u altijd de kostbare tijd gevonden om met veel aandacht en interesse naar mij te luisteren. Als een echte promotor heeft u mij veilig doorheen de procedurele aspecten van het zelfstandig wetenschappelijk onderzoek geloodst en heeft u mij telkenmale geadviseerd en gestimuleerd in het zoeken naar antwoorden. Uw toegankelijkheid, ervaring en betrokkenheid hebben een onvergetelijke indruk nagelaten. Bedankt!

Prof. Dr. Paul Corthals (Hogeschool Gent, Universiteit Gent), ook uw rol in de totstandkoming van dit proefschrift kan nooit genoeg belicht worden. Al heel vroeg in mijn opleiding tot logopedist heeft u mij weten te boeien met epische verhalen over spraak, fonetiek en akoestiek en met meting in de rol van de held. U heeft mij doen realiseren dat deze elementen een zeer krachtige combinatie kunnen

vormen en ik hoop dat dit proefschrift hiervan getuigt. Heel vaak heb ik u bestookt met vragen over moeilijke akoestische en soms statistische kwesties, en u heeft mij als een trouwe kompaan altijd en met minimale latentie van erg betrouwbare antwoorden voorzien. Bedankt Paul, het was mij een waar genoegen om samen het pad van de akoestische fonetiek te bewandelen. Dat we onderweg vriendschap zijn tegengekomen, is onvervangbaar.

Een bijzonder woord van dank wens ik ook te richten aan Prof. Dr. Nelson Roy (University of Utah). U ontmoeten tijdens het congres van de Austrian Voice Institute in 2006 heeft een imposante impact gehad, zowel op mijn logopedisch-vocologische activiteiten als op het finale resultaat van mijn wetenschappelijke onderzoeken in dit doctoraat. Ik heb u met mijn Engelstalige teksten frequent geconsulteerd voor advies inzake fraseologie, spelling en inhoud. Het resultaat van uw revisies was altijd verbluffend en getuigend van toewijding en ongekende precisie. U heeft mij geleerd om efficiënt bepaalde argumenten te weerleggen terwijl andere te verdedigen, met een aantal fraaie manuscripten tot gevolg. Nelson, ik dank u voor het delen van uw vakkundige en wetenschappelijke expertise zonder enige drempel.

Ook Prof. Dr. Dimitar Deliyski (University of South Carolina) wil ik afzonderlijk bedanken. Ik ervaar ons eerste gesprek in 2007 tijdens de Pan-European Voice Conference in Groningen nog steeds als een doorslaggevende factor in mijn huidige visie op de bestaande akoestische methodes in het klinisch stemonderzoek. Ondanks uw talrijke wetenschappelijke activiteiten heeft u nooit getalmd om samen te sleutelen aan een van de teksten in dit doctoraat. Ik geloof steevast dat uw expertise inzake digitale signaalverwerking in cruciale mate heeft bijgedragen tot de publicatie ervan. Bovendien heeft u mij gedemonstreerd wat diplomatie kan betekenen in het wetenschappelijk proces. Dimitar, hartelijke dank om zo bereikbaar te zijn.

Prof. Dr. Kristiane Van Lierde (Universiteit Gent), ook u heeft bij mij al van in het begin een bijzondere indruk achtergelaten. We hebben elkaar in 1998 leren kennen in het kader van mijn stage in het Universitair Ziekenhuis Gent en mijn graduaatsthesis. U heeft zich vervolgens geëngageerd om mij te begeleiden in mijn licentiaatsproefschrift en tenslotte ook in dit doctoraat. Me door u gesteund voelen in al deze projecten heeft een ontegensprekelijk gunstige invloed uitgeoefend op het eindresultaat. Kristiane, oprecht dank voor het vertrouwen en de vele voorbeelden van professionele integriteit.

Ik wens de andere leden van de examencommissie – met name Prof. Dr. Paul Boon, Prof. Dr. Felix de Jong (Katholieke Universiteit Leuven), Prof. Dr. Ingeborg Dhooge (Universiteit Gent), Prof. Dr. P. Santens (Universiteit Gent), Prof. Dr. John Van Borsel (Universiteit Gent), Prof. Dr. Van de Heyning

(Universiteit Antwerpen) – te bedanken voor de constructieve opmerkingen en de positieve waardering van dit proefschrift.

Er zijn ook enkele mensen in het AZ Sint-Jan Brugge-Oostende AV die ik met een speciaal woord van dank wens te vermelden. Dr. Hans Rigauts (Directeur-Generaal en Hoofdgeneesheer-Directeur), u heeft steeds positief gereageerd op mijn vraag om aan (buitenlandse) congressen deel te nemen. Het opbouwen van internationale contacten en netwerken, het grensoverschrijdend discussiëren over ideeën en resultaten, het opvolgen van de meest recente ontwikkelingen, ... zijn allemaal essentiële onderdelen van modern wetenschappelijk onderzoek. U heeft mij nooit geremd in mijn sciëntifieke activiteiten, integendeel. Rekening houdende met de gunstige invloed van enkele internationale contacten op dit proefschrift, wens ik u graag te danken voor deze indirecte sponsoring en vooral ook voor de vrijheid die u mij biedt. Eén van de missies van ons ziekenhuis is 'innovatieve referentiezorg voor iedereen'. Met dit doctoraat hoop ik er een steentje toe bij te dragen. Daarnaast wens ik de stafleden van de Dienst voor Neus-, Keel- en Oorziekten & Gelaats- en Halschirurgie – met name Dr. Rudolf Kuhweide, Dr. Stephan Vlaminck, Dr. Tom Vauterin, en in het bijzonder Dr. Catherine Dick – expliciet te erkennen voor de boeiende samenwerking, het consistent doorverwijzen van personen met een aan stem gerelateerde klacht en het ontvankelijk zijn voor overleg en discussie. U heeft mij vanaf mijn aanwerving in het ziekenhuis geschraagd door uw vertrouwen in het logopedisch luik van het klinisch stemonderzoek en we hebben sedertdien veel patiënten samen bejegend. Door u kan ik bijna dagelijks het beeldschone design en de verbluffende krachten van de larynx bewonderen. Bedankt, want zonder u had ik geen subjecten en proefgroepen en bijgevolg ook geen doctoraat. En dan zijn er natuurlijk nog de directe collegae van de Dienst voor Logopedie en Audiologie – met name de logopedisten Christelle Vanmaele, Caroline Vandenbruaene en Ines Verté, en de audiologen William Damman en Janne Dedeyne. Ik dank jullie voor de noodzakelijke steun, het begrip en de interesse. Door jullie heb ik het gevoel dat het geheel wel degelijk veel meer is dan de som der delen.

Verder dank ik nog Fons Mertens en Dr. Bernadette Timmermans, beiden bron van inspiratie, voor hun vrijwillige interventies en de uren die we samen hebben doorgebracht voor het beoordelen van de stemsamples. Onderschat uw bijdrage in dit doctoraat niet, want zonder u was het er nu niet. Ik dank vervolgens ook mijn collegae bestuursleden van de Vlaamse Vereniging voor Logopedisten – met name Pol De Meyere, Dr. Ronny Boey, Chris De Bal, Marleen D'Hondt, Dr. Louis Heylen en Stefaan Lefevere – voor hun geduld, onophoudelijke interesse en toewijding. Vaak heb ik troost, afleiding en vooral enthousiasme gezocht en bij u gevonden. Bovendien heb ik van jullie geleerd wat het beroep van logopedist waard is. Bedankt! Stefaan wens ik nog bijzonder te erkennen voor zijn editoriale expertise tijdens het nalezen van het finale manuscript.

Dr. Gwen Van Nuffelen verdient eveneens een aparte vermelding. Doctoreren gaat namelijk gepaard zowel hoogtepunten als laagtepunten, en het is zeker bemoedigend om hierover te kunnen converseren met een verwante geest op een evenwijdig parcour. Gwen, bedankt voor uw niet aflatende sympathie.

Peter en Liesbet, Sven en Kathy, Klaas en Veronique, Kim en Heidi, Johan, ware vrienden en vriendinnen, bedankt om telkens fijne ervaringen te garanderen. Erzonder had ik dit doctoraat nooit tot de ware proporties kunnen herleiden. Ik wil tevens mijn schoonouders, Agnes en Wilfried, erkennen voor de aanhoudende steun en openhartigheid. U bent vaak de bron van stabiliteit in een jong en soms woelig leven. Tevens wil ik Freddy bedanken voor de aanhoudende interesse en de onvoorwaardelijke betrokkenheid.

Annie Timperman en Martial Maryn, mama en papa, het schrijven van een doctoraat is niet gemakkelijk, maar het betekent niets in vergelijking tot de omzeggens onmogelijke taak om u te bedanken voor uw aandeel in wie ik vandaag ben en wat ik vandaag doe. Achter de definitie van de beste ouders beitel ik rotsvast uw namen, gevolgd door ontzaglijk veel oude en jonge verhalen om dat te illustreren. Bedankt voor uw geduld in mij, voor uw trots in mij, voor uw toewijding in mij en voor uw hulp in de vele keuzes. Meer had ik niet nodig, maar ik besef nu wel dat het erg veel was. Ik draag dit doctoraat deels aan u op.

Gedeelde smart is halve smart, en gedeeld geluk is dubbel geluk! Liesbeth, mijn zo dierbare Liesje, in de prille fase van dit doctoraat ben je dansend en met overtuiging in mijn leven gekomen, en sedertdien hebben we samen enkele imposante hoofdstukken van het ware leven ontsluierd. Er is zoveel dat je mij geleerd hebt, aangeprezen en afgeraden. Er is zoveel koers dat je ons gegeven hebt, en ik beken gelukkig te zijn met de bestemming. Door jou sta ik nergens alleen. Door jou mis ik nooit rust. Dit doctoraat is voor pakweg negentig percent gerealiseerd in vrije tijd. Je hebt mij herhaaldelijk moeten missen en hoewel present was ik vaak afwezig. Bedankt voor je kracht, je engagement in ons en de kleine en grote verrassingen. Met dit doctoraat hoop ik je fier te maken. Ik draag het ook aan jou op, en natuurlijk ook aan Emiel, onze bron van ontelbaar veel glimlachen en puur geluk.

Youri Maryn,
14 januari 2010.

1

GENERAL INTRODUCTION

PHONATION, VOICE QUALITY AND DYSPHONIA

Phonation is the term used to describe the physical and physiological processess of vocal fold vibration (Titze, 1994). The term *voice* refers to the result of phonation, i.e., the acoustic signal generated by the larynx and vocal tract (Mathieson, 2001). The perception of voiced sound can be described in terms of four dimensions: pitch, loudness, phonetic identification and voice quality (Titze, 1994). *Pitch* is the perceived "height" of the voice signal. It correlates highly with fundamental frequency (i.e., F_0 or the rate vocal fold oscillation) but is also influenced by loudness and voice quality (Titze, 1994; Debruyne & Buekers, 1998). Differences in pitch are illustrated in Figure 1.1. *Loudness* refers to the perception of the "magnitude" or "strength" of a sound. It is especially related to the intensity or sound pressure level of a sound but varies also with the pitch and spectral properties of the sound (Debruyne & Buekers, 1998; Kent & Read, 2002). Differences in loudness are represented in Figure 1.2. Physiologically, pitch is determined by vocal fold length, tension, and resistance to subglottic pressure. Loudness chiefly depends on the quantity of respiratory airflow and the subglottic pressure (Van den Berg, 1958; Iwata, 1988; Nishizawa et al., 1988; Titze, 1994; Jiang et al., 2000). *Phonetic identification* is related to the perception of the characteristic features of the phonemes. For example, vowels are perceived and classified on the basis of their two lowest formants (i.e., resonance frequencies of the vocal tract). The frequency of the first formant (F_1) is inversely related to tongue height. The frequency of the second formant (F_2) is associated with the anterior-posterior position of the tongue (Kent, 1997; Kent & Read, 2002). This is illustrated by a F_1-F_2-plot, such as Figure 1.3. However, when it comes to the voiced-voiceless distinction, phonetic identification depends on the timing of articulatory manoeuvres and laryngeal phonatory activity. None of the other features that typify phonemes is of phonatory nature. The last dimension, *voice quality* (also known as vocal quality or timbre), has been defined by the American National Standards Institute (ANSI, 1960, p.45) as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar". Titze (1994) described voice quality as a poorly defined term "that includes all the leftover perceptions after pitch, loudness and phonetic category have been identified". Consequently, since voice quality is everything except pitch, loudness and phonetic category, it includes all perceptual dimensions of the spectral envelope and its changes in time (Kreiman & Gerratt, 1998). Voice quality thus is a multidimensional perceptual construct, in contrast to pitch, loudness and the voiced phonemes that are monodimensional (i.e. for which there is a single acoustic correlate: fundamental frequency, intensity and formant frequency, respectively). In contrast to pitch, for example, voice quality can not be quantified by a single measure or rating. At the level of the vocal folds, voice quality mainly varies with (a) the sufficiency of vocal fold adduction and (b) the regularity of the vocal fold vibration pattern (Mathieson, 2001).

Figure 1.1 Three acoustic waveforms (x-axis: time in seconds; y-axis: sound pressure in Pascal) illustrate the terms "pitch" and "fundamental frequency": (top) relatively low voice with a fundamental frequency of 119.35 Hz; (middle) intermediate height of voice with a fundamental frequency of 251.14 Hz; (bottom) relatively high voice with a fundamental frequency of 392.66 Hz. The black solid lines originate from natural voice recordings. The grey dotted lines represent sinusoidal patterns with the same number of cycles per second.

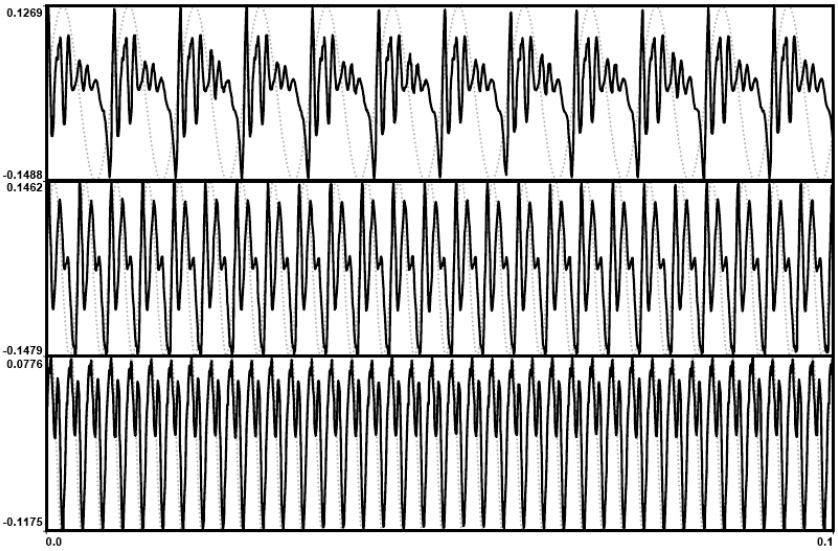


Figure 1.2 Three acoustic waveforms (x-axis: time in seconds; y-axis: sound pressure in Pascal) illustrate the terms "loudness" and "intensity": (top) relatively soft voice with an intensity of 57.99 dB; (middle) intermediate strength of voice with an intensity of 69.69 dB; (bottom) relatively loud voice with an intensity of 79.25 dB.

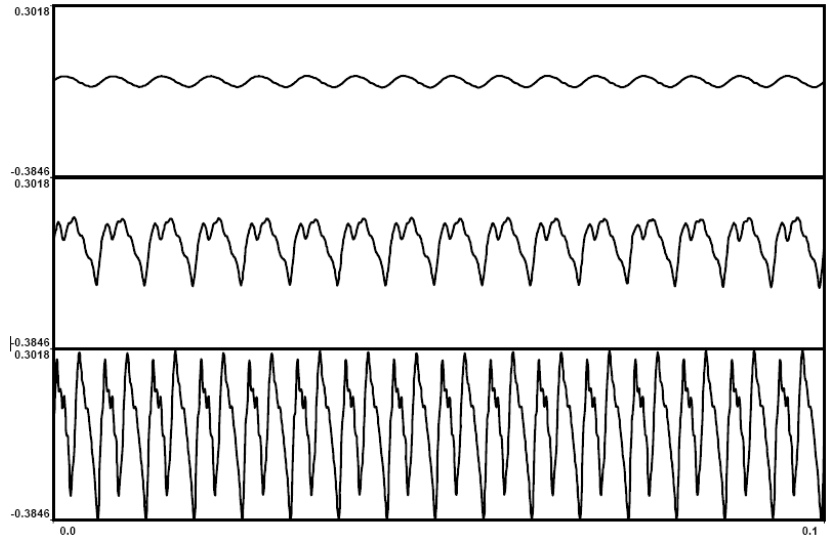
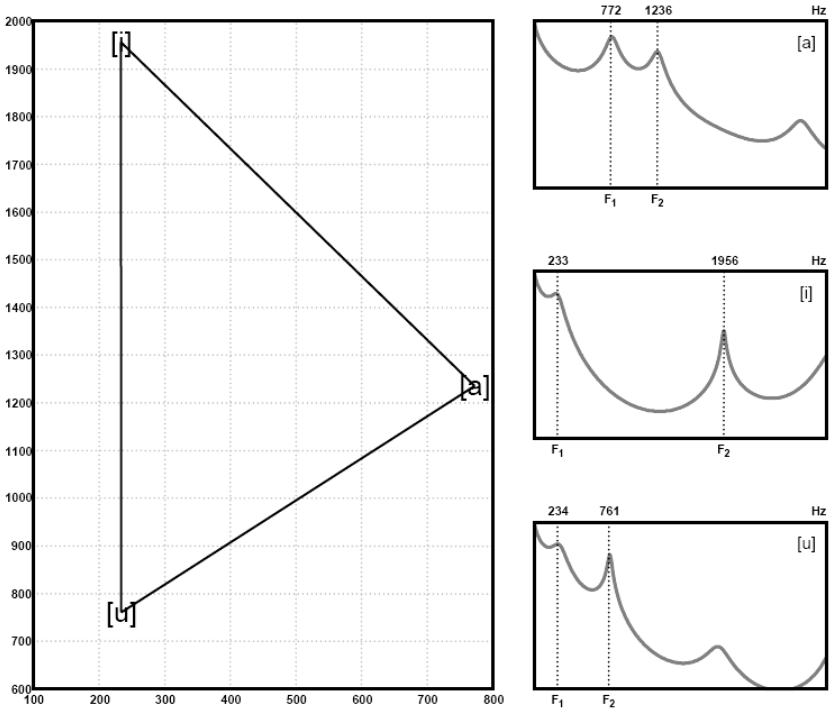


Figure 1.3 A F_1-F_2-chart (x-axis: F_1 in Hertz; y-axis: F_2 in Hertz) to illustrate the phonetic identification cues for three Dutch vowels: [a] as in "zaak", [i] as in "ziek" and [u] as in "zoek" (left). The smaller graphs on the right are LPC-smoothed Fourier spectra of these three vowels.



The etiologic classification of voice disorders, as proposed by Mathieson (2001), is provided in Table 1.1. Irrespective of their etiology, voice disorders can provoke various degrees of phonatory disturbance. The condition in which the vocal folds do not vibrate at all, is called *aphonia* (i.e., a complete absence of the voice). When the vocal folds vibrate abnormally, there is *dysphonia* (Titze, 1994; Mathieson, 2001). Dysphonia can affect one or more of the voice-related dimensions. It can pertain to pitch. For example, in an adolescent with puberphonia (prepubertal voice due to failure in laryngeal mutation), the typical feature of the dysphonia will be an excessively high vocal pitch. Acoustic (as illustrated in Figure 1.1) as well as electroglottographic methods can easily be used for clinical measurement of fundamental frequency (i.e., the physical attribute of pitch) (Baken & Orlikoff, 2000). Sometimes, dysphonia can be characterized by abnormal loudness. In some cases of unilateral vocal fold paralysis, for instance, inadequate vocal fold adduction hampers vocal intensity. As a result, the person speaks too quiet (Mathieson, 2001). Intensity is typically measured via acoustic methods (see Figure 1.2) (Baken & Orlikoff, 2000). Finally, dysphonia can manifest itself in voice quality. As a matter of fact, dysphonia involves abnormal voice quality much more frequently than abnormal pitch or loudness (Dejonckere, 1995). "Overall

voice quality" pertains to the general degree of voice quality abnormality. However, instead of studying voice quality as a whole, many authors have focused on single aspects of voice quality, of which "breathiness" and "roughness" are the best known and most commonly used (Fairbanks, 1940; Murry et al., 1977; Hirano, 1981; De Bodt, 1997; Kreiman & Gerratt, 1998). A disordered voice quality may be the result of an insufficient vocal fold adduction during phonation. This will result in audible air leakage through the glottis. The perceptual attribute of this air turbulence is breathiness (Hirano, 1981; Kreiman & Gerratt, 1998; Kreiman & Gerratt, 2000). Voice quality may also be disordered as a consequence of irregularity in the vibration of the vocal folds. In this case, irregular fluctuations in the frequency, the amplitude and/or the oscillation pattern of the vocal fold vibrations give rise to the perception of roughness (Hirano, 1981; Kreiman & Gerratt, 1998; Kreiman & Gerratt, 2000).

The clinical measurement of disordered overall voice quality (and roughness and breathiness) will be addressed in the next paragraph. From this point on, the term dysphonia will be used as a synonym for disordered voice quality (either generally or specifically in terms of roughness or breathiness).

Table 1.1 Etiological taxonomy of voice disorders (Mathieson, 2001).

Behavioral voice disorders	
Hyperfunctional	*Muscle tension dysphonia, without changes in the vocal fold mucosa*
	Muscle tension dysphonia, with changes in the vocal fold mucosa: vocal fold nodules, polypoid mucosa (edema), granuloma, polyp, haemorrhage, contact ulcer, chronic laryngitis
Psychogenic	*Anxiety state, conversion symptom, delayed pubertal voice change (puberphonia/mutational falsetto), trans-sexual conflict*
Organic voice disorders	
Structural	*Congenital: laryngeal web, tracheal and/or vocal tract stenosis, laryngomalacia, tracheoesophageal fistula, laryngeal cleft, laryngeal atresia, sulcus vocalis, vergeture*
	Acquired: trauma, vocal tract stenosis, presbylarynx
Neurogenic	*Peripheral: recurrent laryngeal nerve paralysis/paresis, superior laryngeal nerve paralysis/paresis*
	Central: (pseudo-)bulbar palsy, cerebellar ataxia, benign essential tremor, extrapyramidal condition, dyspraxia/apraxia, focal dystonia (spasmodic dysphonia), post-CVA syndromes, multiple lesions (motor neuron disease, multiple sclerosis, etc.)
Endocrinological	*Thyrotoxicosis, myxoedema, male sexual retardation, female virilisation, adverse drug therapy*
Disease and inflammation	*Neoplasm (benign/malign), papillomatosis, cyst, laryngitis (acute/chronic), autoimmune disease, cricoarythenoid rheumatoid arthritis, gastric reflux, allergy, syphilis, fungal infection, tuberculosis*

CLINICAL MEASUREMENT OF VOICE QUALITY AND DYSPHONIA

Clinical measurement

Measurement is the process of determining the existence, characteristics, size, and/or quantity of some variable through systematic recording and organization of observations. Effective measurement enables researchers to proceed systematically from collecting data to the data analysis phase, where conclusions are drawn (Frey et al., 1991). Measurement, as the process of assigning numbers to variables, is methodologically and clinically very interesting for several puposes. First, it can be used to describe the quality or quantity of an existing variable. Second, it can be implemented to make absolute decisions based on a criterion or standard of performance. Third, measurement can be used for choosing between two courses of action. Fourth, it can be applied to evaluate change or progress as a response to treatment. Fifth, measurement can lead to comparison and discrimination between individuals and groups. Sixth, measurement enables evaluation of the predictive or concurrent relationship between variables (Portney & Watkins, 2000; Mathieson, 2001). The clinical importance of measurement in the realm of voice disorders lies in the accumulation of diagnostic information about the patient, the disorder and related features, as a basis for diagnosis and treatment planning. Clinical measurement thereby conforms to the current trend of evidence-based practice that promotes de-emphasis of intuition and unsystematic clinical experience as solitary grounds for decision making (Sackett et al., 1996). It is therefore important to note that clinical voice-related measurement complements the eyes and ears of the examiner and furnishes quantitative data for diagnostic as well as therapeutic purposes (Jiang et al., 1999).

The clinical measurement of voice and its disorders traditionally relies on a multidimensional protocol in which facultative outcomes of electromyographic, aerodynamic, laryngoscopic (vibratory), acoustic, perceptual and psychosocial investigations are combined (Hirano, 1981; De Bodt, 1997; Dejonckere et al., 2001). Since voice quality is auditory-perceptual by nature, its primary measurement technique is the auditory-perceptual rating scale. Since voice quality is an attribute of the acoustic wave emanating from the vocal tract of the speaker, numerous acoustic measuring techniques have been proposed as well. The relevance of both perceptual rating and acoustic analysis for measuring voice quality will be outlined in the next paragraph.

Perceptual measurement

Voice quality is the result of auditory processing of the acoustic voice signal by the listener. Certain acoustic cues elicit a percept of voice quality or dysphonia in the listener. Consequently, the degree of dysphonia is measured best by means of a standardized auditory-perceptually based method and many applications of this type of approach have been described (Wilson, 1979; Askenfelt

& Hammarberg, 1986; Laver et al., 1986; De Bodt et al., 1996). Two of these auditory-perceptual voice quality rating protocols have been described particularly often. They deserve special mentioning because of their wide-spread use and clinical feasibility: the GRBAS equal-appearing interval scales (the acronym for Grade, Roughness, Breathiness, Asthenicity and Strain) specified by Hirano (1981), and the CAPE-V hybrid visual analog scales (the acronym for Consensus Auditory-Perceptual Evaluation of Voice) described by Hillman et al. (2003) and Kempster et al. (2008). The perceptual nature of judgments/ratings have prompted several lines of research concerning the reliability of auditory-perceptual evaluations of voice quality and dysphonia.

Methodological issues with perceptual measures

The first issue pertains to the listeners and their *test-retest reliability in rating*, i.e., the *intra-listener agreement*: how consistent is a listener in his rating when confronted more than once with the same voice sample? De Bodt et al. (1997) asked listeners to rate the voice quality of twelve dysphonic subjects two times with an interval of two weeks. The overall test-retest reliability of all GRBAS scales, when assessing ratings at two consecutive times, resulted in a moderate agreement of 43 %. Furthermore, Kreiman & Gerratt (2000) pooled the relevant data and study results of Kreiman et al. (1993), Kreiman et al. (1994), Rabinov et al. (1995), Kreiman & Gerratt (1996) and Chhetri (1997). They found that in a group of expert listeners, only 38.6 % produced identical second ratings for the same sample. However, the level of agreement increased to 76.8 % when ratings were allowed to differ by one scale value, which led Kreiman & Gerratt (2000) to the conclusion that individual listeners are able to make reasonably consistent judgments of voice qualities.

The second listener-related issue deals with *inter-listener agreement*: how well do listeners agree in their ratings when confronted with the same voice samples? Kreiman & Gerratt (2000) pooled data and results related to inter-listener concordance of the abovementioned reports. Across these reports, pairs of expert listeners agreed exactly in 26.7 % of all cases rated. When ratings between two expert listeners were permitted to differ by one scale value, the inter-listener agreement increased to 63.7 %. Clearly, different listeners use different perceptual strategies to evaluate voice quality and dysphonia. Whether or not this is due to different degrees of experience of the listeners involved, has been investigated in several studies, but contradictory results have emerged. Through exposure (professional experience and/or explicit training) to various types and degrees of dysphonia, listeners develop individual internal standards along the severity continuum. Consequently, listeners may differ in the amount of detail present in their internal representation of voice qualities (Eadie & Baylor, 2005). De Bodt et al. (1997), on the contrary, found that experienced listeners generally rate more consistently than inexperienced listeners. However, they could not find a statistically significant difference between both groups. Kreiman et al. (1990) and

Kreiman et al. (1992), on the other hand, suggested that inexperienced listeners as a group use more similar perceptual strategies than expert listeners as a group. Because experienced listeners vary in their experience with dysphonia, their internal standards vary more than those of naive listeners. Naive listeners generally lack internal standards for dysphonia and therefore use the same strategies as for normal voices which results in more similar ratings (Kreiman et al., 1992). Furthermore, according to Eadie & Baylor (2006), training of inexperienced listeners seems to refine perceptual judgments and to significantly improve the inter-listener reliability for breathiness ratings in connected speech and vowels as well as for roughness ratings for vowel stimuli. Training also resulted in a significant betterment of intra-listener reliability for overall dysphonia (or grade) ratings in connected speech. Much earlier, Bassich & Ludlow (1986) had already reported that eight hours of training were required for inexperienced listeners to attain 80 % inter-listener reliability. The importance of training and providing anchor stimuli in perceptual evaluations of dysphonia, obtaining more similarity in the internal standards between listeners and thereby augmenting the inter-listener reliability, was also demonstrated in the results of Chan & Yiu (2002) and corroborated by the findings of Shrivastav et al. (2005).

The third research issue deals with the influence of the *level of dysphonia* on the reliability of the perceptual ratings: are listeners as consistent in their ratings for slight dysphonia as they are in their ratings of intermediate and severe levels of dysphonia? Kreiman & Gerratt (2000) reported that inter-listener agreement varies with the level of dysphonia. For midrange levels of dysphonia, listeners showed less consistency than for scale end-point levels of dysphonia (normal and very slight dysphonia and severe dysphonia). It thus seems easier for listeners to rate (nearly) normal and extremely dysphonic voice samples than to rate intermediate levels of dysphonia.

The fourth question is also related to the nature of voice stimuli: the variation of perceptual reliability with *type of dysphonia*. Does reliability differ between different types of dysphonia such as overall severity, breathiness and roughness? The results of Dejonckere et al. (1993) indicated that overall severity of dysphonia (or grade) is rated least ambiguously, followed by roughness and than breathiness. De Bodt et al. (1997), De Bodt (1997), Yamaguchi et al. (2003), Webb et al. (2004) and Eadie & Baylor (2006) found very similar results. This implies that overall severity of dysphonia is the most salient of all perceptual voice quality dimensions, making it easier for a listener to rate the voice signal as a whole than producing ratings of particular dimensions, such as breathiness (audible air turbulence) or roughness (audible irregularities in vocal fold vibration).

The fifth question again concerns the voice stimuli, more specifically the *speaking task (or sample type)*: are continuous speech samples rated differently when compared to sustained vowel samples? de Krom (1994) conducted a perceptual study in which six listeners rated GRBAS in four sample types (connected speech, and three segments of a sustained vowel) recorded from 78 subjects. Results indicated that sample type did not significantly affect the

reliability of the rating. These findings were corroborated by a very similar study of Revis et al. (1999). De Bodt (1997) also investigated the difference in dysphonia ratings between sustained vowels and connected speech. Three listeners judged samples from 451 subjects. The results indicated that listeners gave significantly higher overall severity and breathiness rates for sustained vowel samples. Zraick et al. (2005) asked three expert listeners to rate the overall dysphonia severity in sustained vowels and connected speech (picture descriptions and reading samples) from 29 dysphonic patients. They found that all listeners rated overall dysphonia more severely in a sustained vowel than in running speech fragments. These differences were statistically significant between sustained vowel and continuous speech elicited via reading. Furthermore, the results of both studies of De Bodt (1997) and Zraick et al. (2005) are in accordance with Wolfe et al. (1995), who reported slightly but significantly more severe scores for vowels than for continuous speech. Complementary explanations are (a) that dysphonia might be more prominent and therefore estimated more severe in sustained vowels and (b) that, during sustained vowels, listeners can draw all attention to dysphonia-related phenomena, whereas in connected speech the attention for dysphonia is somehow diverted by other speech-related phenomena. Furthermore, the inter-listener reliability was found to be slightly higher for continuous speech samples (De Bodt, 1997). These results were partially confirmed by the findings of Bele (2006). For overall voice quality and breathiness (and nine other voice-related items), ratings of text readings at normal loudness turned out to be more reliable than ratings of sustained vowels. For roughness, the opposite was found. Collectively, it can be concluded that not all speaking tasks yield similar auditory-perceptual dysphonia ratings with the same level of reliability. It is thus reasonable to take both sustained vowels and continuous speech into account in clinical dysphonia assessment protocols.

The sixth question concerns the rating task itself, more precisely the use of different *scale types*: does the type of rating scale (ordinal equal-appearing interval scale versus visual-analog scale) affect the auditory-perceptual evaluation of dysphonia? The ordinal equal-appearing interval provides (typically) a 4-point or 7-point scale on which the listener has to indicate one point according to the severity of the dysphonia, e.g. 0 indicating normophonia and 3 indicating severe dysphonia. The GRBAS protocol (Hirano, 1981) operates with this type of scale. On a visual-analog scale, perceived severity of dysphonia is appraised on a horizontal 100-millimeter line with its left endpoint demarcating normophonia and its right endpoint demarcating most severe dysphonia. Rating is done by placing a check mark at a distance corresponding to the perceived degree of dysphonia. The CAPE-V protocol (Hillman, 2003; Kempster et al., 2008) utilizes a hybrid visual-analog with annotated intervals scale for several voice qualities. The main advantage of the ordinal equal-appearing interval scale is to limit inter- and intra-listener variability by suggesting a range for every degree of dysphonia (Yu et al., 2002). However, such scales do not allow the listener to entirely express his/her perceptive acumen (Gerratt et al., 1993). Visual-analog scales, on the other hand,

allow much more in-between points – between 0 and 100 points (and even more), compared to the ordinal 4 or 7 points – thereby providing more rating freedom and better discrimination (Yu et al., 2002). However, by doing so, it is assumed that variability between ratings increases, since there are no boundaries specified (Kreiman et al., 1993). Wuyts et al. (1999) asked 29 listeners to rate the same 13 pathologic voice samples twice (with an interlude of two weeks) with two versions of the GRBAS scales: the original equal-appearing 4-point interval scales and visual-analog scales. Although the inter-listener reliability of the ratings was found to be acceptable for both scale types, it was higher for the ordinal version than for the visual-analog version for all GRBAS subscales. Wuyts et al. (1999), as well as Kreiman et al. (1993) therefore concluded that reliability between listeners decreases with increasing freedom of judgment. Yiu & Ng (2004) compared the reliability of breathiness and roughness ratings on an 11-point equal-appearing interval scale and on a visual-analog scale. Their results showed that raters demonstrated a significantly higher intra-and inter-rater reliability for the equal-appearing interval scale. In a similar study, Karnell et al. (2007) examined the reliability of GRBAS and CAPE-V severity ratngs (both denoting the dimension overall severity of dysphonia) in the same group of 103 purposely selected subjects with various types of voice disorders and degrees of dysphonia. Intra- as well as inter-listener reliability was acceptable for both rating scales (however, with CAPE-V yielding slightly better reliability scores than GRBAS). Additionally, the correlation between the two rating scales was very strong (r_s=0.96). Karnell et al. (2007) considered the combined use of both scale types, as in the CAPE-V scales, to be indispensable in a complete assessment of dysphonia.

It can be concluded that auditory-perceptual rating is the most relevant measure of voice quality. However, the implementation of training and explicit anchor points for various types and degrees of dysphonia is necessary to maximize reliability of clinical as well as experimental auditory-perceptual evaluations of voice quality. Finally, given the fact that auditory-perceptual measurement of voice quality and dysphonia is subjective by definition and has inherent sources of variability, many researchers have tried to find objective measures that correlate well with the abovementioned percepts. Among these objective measures, the acoustic measures are particularly interesting, as will be described in the next paragraph.

Acoustic measurement

Voice quality judgment is the cognitive response to the auditory perception of a voice signal (Shrivastav, 2003). However, the only information the listener can depend on for the rating of someone's voice quality, is communicated via the acoustic waveform. Measurements that investigate the acoustic waveform via acoustic algorithms, can be regarded as objective surrogates for subjective perception. Luckily, from a clinical point of view, capturing and analyzing acoustic waveforms of the voice signal is noninvasive, commonly available, relatively

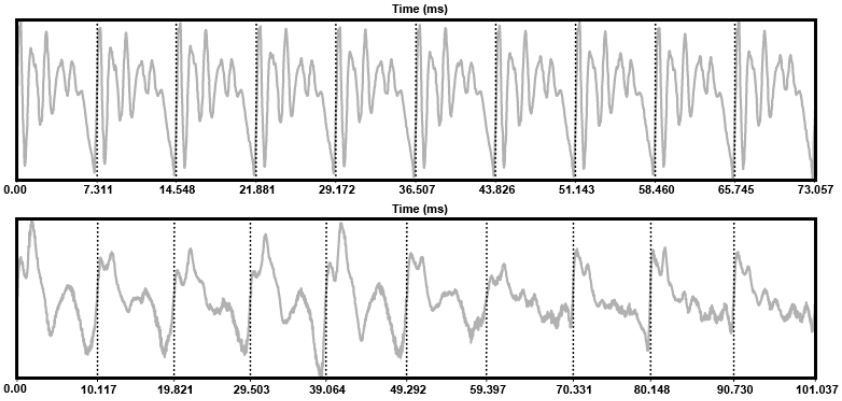
inexpensive, easily applicable, and, consequently, attractive for clinical voice assessment (Dejonckere et al., 1996; Parsa & Jamieson, 2000). All these advantages, combined with the reliability issues of perceptual methods, have called forth many new acoustic methods and investigations into their value. The papers of Lieberman in 1961 and 1963 certainly inaugurated acoustic analysis of voice quality via digital algorithms, resulting in a large amount of acoustic voice quality measures, as evidenced by Buder (2000). His tabulation of more than hundred algorithms derived from more than five-hundred scientific reports provides a quite complete coverage of acoustic measures that have been used in voice (quality) analysis. His goal was to organize the acoustic algorithms into a small set of optimally comprehensive and mutually exclusive categories. This resulted in a set of fifteen categories. Table 1.2 lists the categories together with a typical example. It is interesting to examine the methodological quality of these measures (where available), i.e. test-retest reliability, inter-program reliability, validity, standardization, and normative data. The only groups of measures for which some or all of these methodological outcomes have been investigated are fundamental frequency perturbation measures, amplitude perturbation measures, waveform perturbation measures and spectral noise measures.

Table 1.2 The fifteen categories in the tabulation of Buder (2001). One illustrative measure is given per category.

Category	Example
I F0 statistics	Median of F_0
II Short-term F_0 perturbations	Percent jitter
III Long-term F_0 perturbations	F_0 tremor frequency
IV Amplitude statistics	Median of intensity
V Short-term amplitude perturbations	Percent shimmer
VI Long-term amplitude perturbations	Amplitude tremor frequency
VII F_0/amplitude covariations	Voice range profile
VIII Waveform perturbation	Yumoto's harmonics-to-noise ratio
IX Spectral measures: spectrographic measures	Yanagihara's classification of standard narrowband spectrograms
X Spectral measures: Fourier and LPC spectra	Spectral tilt
XI Spectral measures: LTA spectra	Alfa-parameter
XII Spectral measures: cepstra	Cepstral peak prominence
XIII Inverse filtering measuring (radiated signal)	Pitch amplitude
XIV Inverse filtering measuring (flow-mask signal)	Closing quotient
XV Dynamics	Correlation dimension

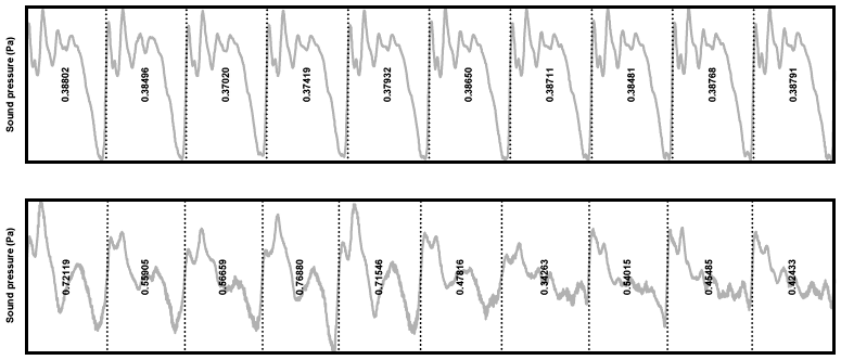
Fundamental frequency perturbation (or jitter) is the variability of the fundamental frequency (or, reciprocally, of the fundamental period) from one cycle (or period) to the next. It is a measurement of how much a given period differs from the period that immediately follows it (Baken & Orlikoff, 2000). The principle of jitter measurement is illustrated in Figure 1.4.

Figure 1.4 Irregularity in duration (T) (jitter) between ten adjacent voice cycles. The upper waveform shows data from a normophonic person. The lower waveform was produced by a person with rough vocal quality. Zero-crossings were handmarked and served as cycle boundaries. Jitter percent (also known as jitter factor) was calculated using the Baken & Orlikoff (2000) equation $[\Sigma|T_i-T_{i-1}|/(n-1)]/[\Sigma(T_i)/(n)]*100$. For the normophonic waveform, there was a jitter percent of 0.509 %. For the dysphonic waveform, there was a jitter percent of 4.763 %. This illustrates that in rough voices the vocal fold vibration is less regular than in voices with a normal voice quality.



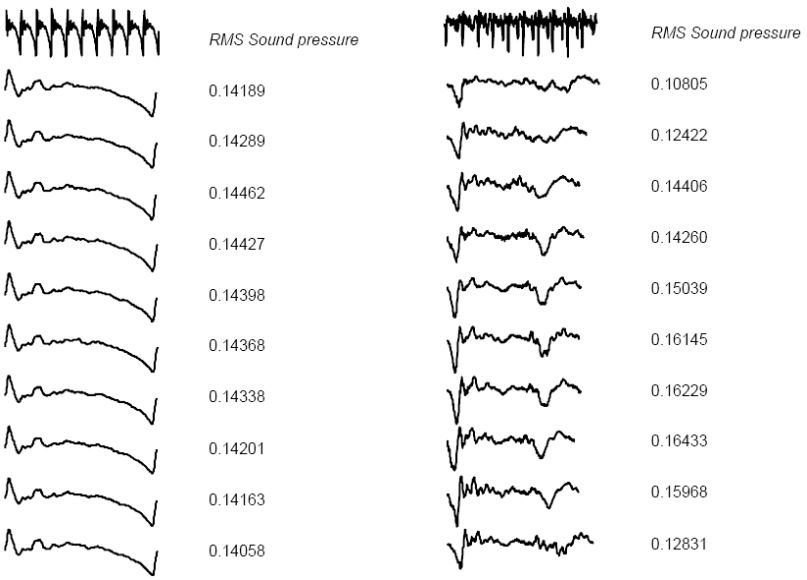
Amplitude perturbation (shimmer) refers to the instability in amplitude between consecutive pairs of two or more adjacent periods (Baken & Orlikoff, 2000). The principle of shimmer is demonstrated in Figure 1.5.

Figure 1.5 Amplitude irregularities (shimmer) in ten adjacent voice cycles. The upper waveform is normophonic. The lower waveform is produced by a person with rough vocal quality. Zero-crossings were handmarked and served as cycle boundaries. For every cycle, the difference between the maximal and minimal sound pressure levels was calculated to quantify the factor A (i.e., amplitude). The formula to calculate shimmer in dB was (Baken & Orlikoff, 2000): $\Sigma|20\log_{10}(A_i/A_{i-1})|/n-1$. For the normophonic waveform, there was a shimmer of 0.102 dB. For the dysphonic waveform, there was a shimmer of 2.005 dB. Again, this illustrates that in rough voices the vocal fold vibration is less regular than in voices with a normal voice quality.



Perturbation measures traditionally have been related to roughness (as the perceptual attribute of irregular vocal fold vibration). A typical waveform perturbation metric is the harmonics-to-noise ratio (HNR), proposed by Yumoto and his colleagues (Yumoto et al., 1982; Yumoto, 1983; Yumoto et al., 1984). The idea behind it is illustrated in Figure 1.6.

Figure 1.6 Waveform perturbations in ten adjacent voice cycles. The waveforms on the left are normophonic. The waveforms on the right are produced by a person with breathy voice quality. Zero-crossings were handmarked and served as cycle boundaries. The final sound pressure levels (A) are based on the root-mean-square (A_{RMS}). Waveform-based HNR was calculated according to Yumoto et al., 1982: $H=n*[\Sigma(A_{RMS})/n]^2$; $N=\Sigma[A_{RMS}-(\Sigma(A_{RMS})/n)]^2$; $HNR=10*\log10(H/N)$. For the normophonic waveform, this resulted in HNR=41.15 dB. For the breathy waveform there was HNR= 18.06 dB.



Instead of starting from the time-domain, spectral noise measures are frequency-based. This is demonstrated in Figure 1.7. Typically, these noise measures have been associated with breathiness (as the perceptual attribute of glottal air leakage).

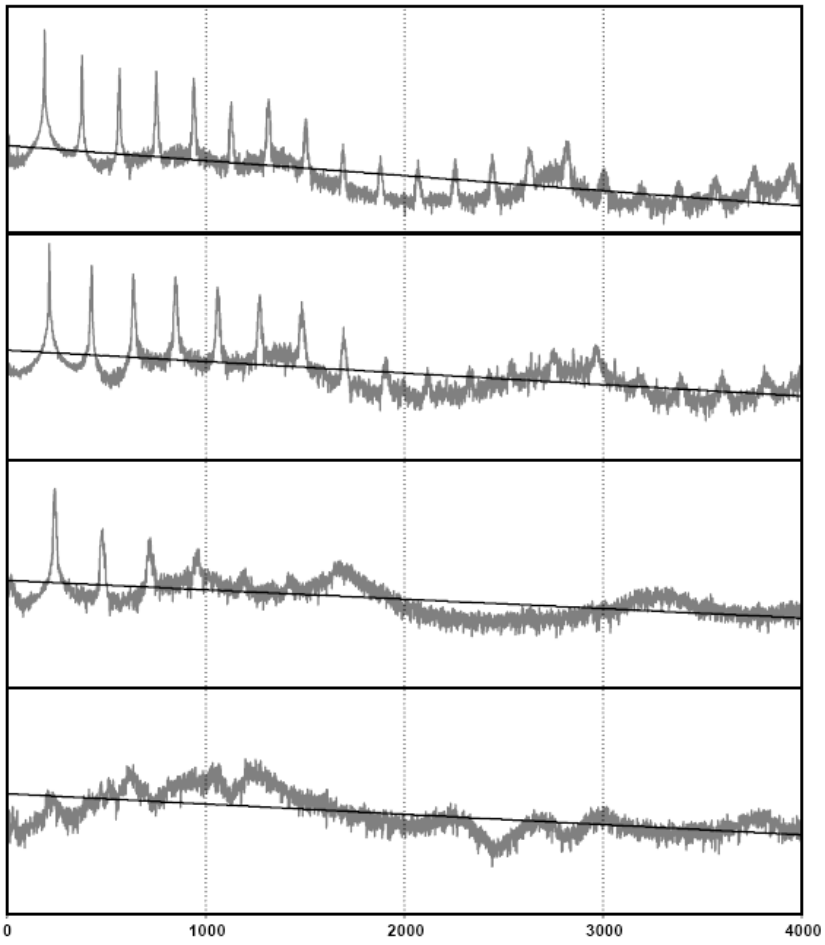
Methodological issues with acoustic measures

The first methodological issue is the *test-retest reliability* or the *intra-program agreement*: how consistent are the outcomes of acoustic measures on repeated recordings, or, how many consecutive recordings are required to do a representative acoustic analysis? This is an important issue when acoustic measurement is used for the monitoring of treatment progress and outcome. Scherer et al. (1995) investigated the test-retest reliability for jitter, shimmer and

HNR in 24 subjects. It was concluded that for voices with low perturbation, at least 6 recordings were needed to establish a representative average perturbation. Normal voice signals required 15 recordings. For voice signals with high perturbation, more than 15 recordings were necessary. The HNR measure of Yumoto et al. (1982) appeared to be less sensitive to test-retest variability. Still, an average of approximately 10 recordings was required to establish a representative HNR value, regardless of measure value (low, normal, high). Based on these results, there seems to be considerable test-retest variability in the perturbation measures. Bough et al. (1996) also investigated the test-retest reliability of jitter. They asked fourteen subjects to produce fifteen samples of a sustained vowel. The intraclass correlation between repeated recordings ranged from 0.956 to 0.984, which, in contrast to the results found by Scherer et al. (1995), gives evidence of good reliability in jitter measures. Based on these contrasting findings, further investigation of test-retest reliability of all measures is recommended.

The second methodological issue is the *inter-program reliability*: how different are the outcomes for acoustic measures when different computer programs are used? There are many commercially programs available for measuring voice perturbation, e.g. Multi–Dimensional Voice Program a.k.a. MDVP (Kay Elemetrics, Lincoln Park, USA), Computerized Speech Lab a.k.a. CSL (Kay Elemetrics, Lincoln Park, USA), Dr. Speech (Tiger Electronics DRS, Seattle, USA), SoundScope (GM Instruments, Cambridge, USA) and CSpeech a.k.a. TF32 (Paul Milenkovic, Madison, USA). There are also computer programs freely available on the world wide web, such as Praat (Paul Boersma, Institute of Phonetic Sciences, Amsterdam, The Netherlands) and Speech Filing System a.k.a. SFS (Mark Huckvale, University College London, London, UK). The inter-program reliability for these and/or other computer programs has been investigated by several authors. Karnell et al. (1995) did a comparison of jitter and shimmer results from three computer programs. They found correlation coefficients ranging from 0.29 to 0.64 for jitter and from 0.26 to 0.75 for shimmer. There were statiscally significant differences for various comparisons. The authors concluded that the perturbation programs clearly do not result in comparable outcomes. Bielamowicz et al. (1996) also compared perturbation measures of four computer programs. For jitter, correlation coefficients ranged from 0.33 to 0.80. For shimmer, correlation coefficients ranged from 0.81 to 0.89. For HNR, correlation coefficients varied from 0.23 to 0.81. Statistically significant results were found for several measures. It was concluded that there is reasonable inter-program reliability for shimmer measures across different severity levels. Jitter and HNR, on the other hand, are much less reliable across computer programs. Smits et al. (2005) compared jitter, shimmer and HNR data derived from two programs and found correlation coefficients of 0.26, 0.69 and 0.74, respectively. Collectively, jitter measures are the least reliable, seriously compromising their clinical application in the assessment of dysphonia. Shimmer measures were found to have an acceptable inter-program reliability.

Figure 1.7 Four Fourier spectra (x-axis: frequency in Hertz; y-axis: sound pressure level in decibel/Hertz) with linear trend lines to illustrate various degrees of breathiness. The slope of the trend lines is calculated by subtracting the mean energy (in dB) between 0-1 kHz from the mean energy (in dB) between 1-4 kHz. The first spectrum is obtained from a sample with normal voice quality. Harmonics are prominently present up to 4000 Hz (and beyond), HNR=28.170 dB, and slope-of-trend-line=-13.277 dB. The second spectrum illustrates slight breathiness. Harmonics start to give way to noise from 2000 Hz on, and there is a HNR of 22.193 dB and the slope-of-trend-line=-10.130 dB. The third spectrum illustrates moderate breathiness. From 1000 Hz on there is only noise, HNR=16.609 dB, and slope-of-trendline=-8.279 dB. The fourth spectrum illustrates the most severe degree of breathiness (aphonia). There are no more spectral harmonics. They have been replaced by noise, HNR=1.685 dB, and slope-of-trend-line=-9.036 dB.



The third methodological issue is *validity*: to which extent can an acoustic algorithm measure what it is actually intended to measure? The validity of perturbation and many other acoustic measures has been investigated frequently. In this introduction only a few examples of relevant research outcomes on criterion-

related concurrent validity (a.k.a. predictive validity)[1] of jitter and shimmer are given in the purposely formatted Table 1.3. Obviously, inconsistent and even contradictory results have been found. Whereas some reports underscore the ability of jitter and shimmer to measure different voice quality dimensions, other reports failed to demonstrate this same ability.

Table 1.3 Examples of correlation coefficients between perceptual evaluation (for overall voice quality, roughness and breathiness) and jitter and shimmer measures, with r=0.60 as a demarcation point between strong and weak predictors a correlation coefficient.

	Jitter percent		Absolute shimmer	
	r<0.60	r≥0.60	r<0.60	r≥0.60
Overall voice quality	0.07[a]	0.61[b]	0.31[b]	0.73[c]
Roughness	0.57[b]	0.68[c]	0.55[c]	0.66[d]
Breathiness	-0.17[b]	0.63[e]	0.07[b]	0.63[c]

[a]: Plant et al. (1997), [b]: Kreiman et al. (1990), [c]: Dejonckere et al. (1996), [d]: Martin et al. (1995), [e]: Wolfe & Martin (1997)

Kreiman & Gerratt (2000) give four options to explain this varying validity results. First, the acoustic measure may be insufficiently related to the validly measured perception. Second, the acoustic measure may be sufficiently associated with the validly measured perception, but at the same time there may be problems with the procedure used to make the acoustic measurement. Third, there can be a problem with the techniques used to estimate the relation between acoustic and perceptual measures. In this case, a true association may exist but its estimation is blurred due to sample size, or the particular selection of speakers, the speaking

[1] *Criterion-related validity* implies that the outcome of one instrument (in this case the acoustic measure) can be used as a substitute measure for an established gold standard criterion test (in this case the auditory-perceptual rating). It can be tested as concurrent or predictive validity. *Concurrent validity* is studied when the measurement to be validated and the criterion measure are taken at relatively the same time (concurrently). This approach to validation is useful in situations when a new or untested measure is potentially more efficient, easier to administer, more practical, or safer than another more established method, and is being proposed as an alternative instrument. *Predictive validity* attempts to establish that a measure will be a valid predictor of some future criterion score. To assess this validity, a target test is given at one session and is followed by a period of time after which a criterion score is obtained (Portney & Watkins, 2000, pp. 82-87). The names of both types of validity are sometimes used as synonyms. A typical statistical measure for investigating the degree of concurrent validity is the correlation coefficient. A measuring instrument can also be designed as a diagnostic tool and its *diagnostic validity* is then evaluated in terms of its ability to accurately assess the presence or absence of the target condition (in this case dysphonia). Diagnostic validity is typically investigated with statistical measures such as sensitivity, specificity, predictive value, receiver operating characteristic curve (Portney & Watkins, 2000, pp. 92-93) and likelihood ratio (Dollaghan, 2007).

task, etc. Fourth, there might be a problem with the perceptual measurement. Prompted by the inconsistencies illustrated in Table 1.3 and by the large amount of acoustic algorithms intended to measure voice quality, we conducted a meta-analysis in Chapter 3.

The fourth issue is standardization of methods. Many items are related to this issue: nomenclature of acoustic phenomena and acoustic measures, microphone-to-mouth distance, off-axis angle of the microphone, type of microphone, room acoustics and data acquisition environment, digital sampling frequency, analysis software and acoustic algorithms, speaking task (i.e., test utterance), etc. The need for standardization in acoustic voice analysis has been outlined by Titze (1994b) and the consensus-based report of Titze (1995) has provided a state-of-the-art on the purpose and methods of acoustic analysis of voice signals. However, relevant research has been focused on standardizing the methods for jitter and shimmer. To date, there are several other acoustic measures (e.g., cepstral measures, nonlinear dynamics measures, etc.), for which there is almost no standardization of acoustic methods. One could therefore ask whether the standards for perturbation measures can also be applied for the other acoustic measures. If not, new standards should be provided for the acoustic measures that yield reasonable reliability and validity outcomes.

The fifth issue is the availability of norm-referencing data. During the diagnostic decision-making phase, it is important to objectively reveal whether someone is normophonic or dysphonic. Concerning voice quality and dysphonia, voice clinicians should therefore rely on norm-references for their acoustic measures. Normative data are available for perturbation measures of several computer programs across different languages: American English (Deliyski, 1993), Flemish (Smits et al., 2005), Brazilian Portuguese (Naufel de Felipe et al., 2006), etc. However, there are no normative values for many of the other and possibly more valid acoustic measures such as the cepstral measures and the autocorrelation-based measures.

SCOPE AND GOALS

It is obvious from this introduction, that there is a need for further research in the field of clinical voice quality measurement. A special need is to make voice quality measurement more realistic and increase its ecological validity (i.e., its ability to represent daily voice use patterns). Therefore it is necessary to implement continuous speech in clinical measurement protocols. The final goal of the research presented in this thesis was to study the feasibility of the implementation of continuous speech (in combination with sustained vowels) in the standardized auditory-perceptual rating protocols and the objective acoustic measurement of overall voice quality. Therefore, the following subgoals were pursued.

Regarding the comparison of two computer systems and programs for voice perturbations measurement:

1 to study the agreement and difference between two computer programs commonly used for voice perturbation measures;

2 to study the agreement and difference between two computer systems commonly used for voice perturbation measures.

Regarding the acoustic measurement of overall voice quality in voice samples containing a concatenation of continuous speech and sustained vowel:

3 to study research results in literature concerning the validity of acoustic algorithms to measure overall voice quality in sustained vowels, in order to range and shorten the list of such algorithms;

4 to study research results in literature concerning the validity of acoustic algorithms to measure overall voice quality in continuous speech, in order to range and shorten the list of such algorithms;

5 to study the feasibility of concatenating samples of sustained vowels and continuous speech in the perceptual as well as the acoustic measurement of overall voice quality;

6 to study the criterion-related concurrent validity of various acoustic metrics for the measurement of overall voice quality in the concatenated voice samples;

7 to construct a statistical model for the acoustic measurement of overall voice quality and dysphonia severity in both sample types and based on a clinically representative sample of patients with various types and degrees of dysphonia;

8 to study the criterion-related concurrent validity and the diagnostic accuracy of this statistical model for acoustic measurement;

9 to internally and externally cross-validate the predictive power and the diagnostic accuracy of this statistical model for acoustic measurement;

10 to study the responsiveness to change of this statistical model for acoustic measurement;

11 to study the criterion-related concurrent validity of various acoustic metrics for the measurement of overall voice quality in the special population of patients after total laryngectomy.

Regarding the behavioral management of voice disorders and voice-related phenomena:

12 to systematically review the literature on the effects of acoustic biofeedback in the management of phonatory disorders and vocal performance.

REFERENCES

ANSI (1960). *USA standard: acoustical terminology (S1.1).* New York: American National Standards Institute.

Askenfelt, A.G., & Hammarberg, B. (1986). Speech waveform perturbation analysis: a perceptual-acoustical comparison of seven measures. *Journal of Speech and Hearing Research, 29,* 50-64.

Baken, R.J., & Orlikoff, R.F. (2000). *Clinical measurement of speech and voice (2nd edition).* San Diego: Singular Publishing Group.

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

Bielamowicz, S., Kreiman, J., Gerratt, B.R., Dauer, M.S., & Berke, G.S. (1993). Comparison of voice analysis systems for perturbation measurement. *Journal of Speech and Hearing Research, 39,*126-134.

Bough, I.D., Heuer, R.J., Sataloff, R.T., Hills, J.R., & Cater, J.R. (1996). Intrasubject variability of objective voice measures. *Journal of Voice, 10,* 166-174.

Buder, E.H. (2000). Acoustic analysis of voice quality: a tabulation of algorithms 1902-1990. In R.D. Kent, & M.J. Ball (Eds.), *Voice quality measurement* (pp. 119-244). San Diego: Singular Publishing Group.

Chan, K.M., & Yiu, E.M. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language and Hearing Research, 45,* 111-126.

Chhetri, D.K. (1997). *Treatment of voice disorders related to unilateral paralysis of the vocal cord.* Unpublished senior medical student thesis, University of California.

De Bodt, M. (1997). *A framework for voice assessment: the relation between subjective and objective parameters in the judgement of normal and pathological voice.* Unpublished doctoral dissertation, University of Antwerp.

De Bodt, M.S., Wuyts, F.L., Van de Heyning, P.H., & Croux, C. (1997). Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice, 11,* 74-80.

De Bodt, M., van de Heyning, P.H., Wuyts, F.L., & Lambrechts, L. (1996). The perceptual evaluation of voice disorders. *Acta Oto-Rhino-Laryngologica Belgica, 50,* 283-292.

Debruyne, F., & Buekers, R. (1998). Interdependency between intensity and pitch in the normal speaking voice. *Acta Oto-Rhino-Laryngologica Belgica, 52,* 201-205.

Dejonckere, P.H., Obbens, C., de Moor, G.M., & Wieneke, G.H. (1993). Perceptual evaluation of dysphonia: reliability and relevance. Folia Phoniatrica, 45, 76-83.

Dejonckere, P.H. (1995). Principal components in voice pathology. *Voice, 4,* 96-105.

Dejonckere, P.H., Remacle, M., Fresnel-Elbaz, E., Woisard, V., Crevier-Buchman, L., & Millet, B. (1996). Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Revue de Laryngologie, Otologie et Rhinologie, 117,* 219-224.

Dejonckere, P.H., Bradley, P., Clemente, P., Cornut, G., Crevier-Buchmann, L., Friedrich, G., Van de Heyning, P., Remacle, M., & 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. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society. *European Archives of Otorhinolaryngology, 258,* 77-82.

de Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech and Hearing Research, 37,* 985-1000.

Deliyski, D.D. (1993). *Acoustic model and evaluation of pathological voice production.* Proceedings of Eurospeech, Berlin, Germany.

Dollaghan, C.A. (2007). *The handbook for evidence-based practice in communication disorders.* Baltimore: MD Brookes.

Eadie, T.L., & Baylor, C.R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice, 20,* 527-544.

Fairbanks, G. (1940). *Voice and articulation drillbook.* New York: Harper & Brothers.

Frey, L.R., Botan, C.H., Friedman, P.G., & Kreps, G.L. (1991). *Investigating communication, an introduction to research methods.* Englewood Cliffs: Prentice-Hall.

Fritzell, B., Hammarberg, B., Gauffin, J., Karlsson, I., & Sundberg, J. (1986). Breathiness and insufficient vocal fold closure. *Journal of phonetics, 14,* 549-553.

Gerratt, B.R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. (1993). Comparing internal and external standards in voice quality judgements. *Journal of Speech and Hearing Research, 36,* 14-20.

Hillman, R. (2003). *Overview of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), instrument developed by ASHA Special Interest Division 3.* Paper presented at 32nd Symposium: Care of the Professional Voice, Philadelphia, USA.

Hirano, M. (1981). *Clinical examination of voice.* New York: Springer-Verlag.

Hirano, M., Hibi, S., Terasawa, R., & Fujiu, M. (1986). Relationship between aerodynamic, vibratory, acoustic and psychoacoustic correlates in dysphonia. *Journal of Phonetics, 14,* 445-456.

Iwata, S. (1988). Aerodynamic aspects for phonation in normal and pathologic larynges. In O. Fujimura (Ed.), *Vocal physiology: voice production, mechanisms and functions* (pp. 423-431). New York: Raven Press.

Jiang, J.J., Shah, A.G., & Hanson, D.G. (1999). Voice measurement: importance of voice analysis and measurement to patient management. *Current Opinion in Otolaryngology & Head and Neck Surgery, 7,* 119-124.

Jiang, J., Lin, E., & Hanson, D.G. (2000). Vocal fold physiology. *Otolaryngologic Clinics of North America, 33,* 699-718.

Karnell, M.P., Hall, K.D., & Landahl, K.L. (1995). Comparison of fundamental frequency and perturbation measurements among three analysis systems. *Journal of Voice, 9,* 383-393.

Karnell, M.P., Melton, S.D., Childes, J.M., Coleman, T.C., Dailey, S.A., & Hoffman, H.T. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (VRQOL and IPVI) documentation of voice disorders. *Journal of Voice, 21,* 576-590.

Kempster, G.B., Gerratt, B.R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R.E. (2008). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 17,* epub ahead of print.

Kent, R.D., & Read, C. (2002). *Acoustic analysis of speech (2nd edition).* Albany: Delmar.

Kent, R.D. (1997). *The speech sciences.* Clifton Park: Delmar.

Kreiman, J., Gerratt, B.R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research, 33,* 103-115.

Kreiman, J., Gerratt, B.R., Precoda, K., Berke, G. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35,* 512-520.

Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A., & Berke, G.S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36,* 21-40.

Kreiman, J., Gerratt, B. R., & Berke, G.S. (1994). The multidimensional nature of pathologic voice quality. *Journal of the Acoustical Society of America, 96,* 1291-1302.

Kreiman, J., & Gerratt, B.R. (1996). The perceptual structure of pathologic voice quality. *Journal of the Acoustical Society of America, 100,* 1787-1795.

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

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

Laver, J., Hiller, S., Mackenzie, J., & Rooney, E. (1986). An acoustic screening system for the detection of laryngeal pathology. *Journal of Phonetics, 14,* 517-524.

Lieberman, P. (1961). Perturbations in vocal pitch. *Journal of the Acoustical Society of America, 33,* 597-603.

Lieberman, P. (1963). Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. *Journal of the Acoustical Society of America, 35,* 344-353.

Martin, D., Fitch, J., & Wolfe, V. (1995). Pathologic voice type and the acoustic prediction of severity. *Journal of Speech and Hearing Research, 38,* 765-771.

Mathieson, L. (2001). *The voice and its disorders (6th edition).* London: Whurr Publishers.

Murry, T., Singh, S., & Sargent, M. (1977). Multidimensional classification of abnormal voice qualities. *Journal of the Acoustical Society of America, 61,* 1630-1635.

Naufel de Felipe, A.C.N., Grillo, M.H.M., & Grechi, T.H. (2006). Standardization of acoustic measures for normal voice patterns. *Brazilian Journal of Otorhinolaryngology, 72,* 659-664.

Nishizawa, N., Sawashima, M., & Yonemoto, K. (1988). Vocal fold length in vocal pitch change. In O. Fujimura (Ed.), *Vocal physiology: voice production, mechanisms and functions* (pp. 75-82).New York: Raven Press.

Plant, R.L., Hillel, A.D., & Waugh, P.F. (1997). Analysis of voice changes after thyroplasty using linear predictive coding. *Laryngoscope, 107,* 703-709.

Portney, L.G., & Watkins, M.P. (2000). *Foundations of clinical research: applications to practice (2nd Ed.).* Upper Saddle River: Prentice-Hall.

Rabinov, C.R., Kreiman, J., Gerratt, B.R., & Bielamowicz, S. (1995). Comparings reliability of perceptual ratings of roughness and acoustic measures of jitter. *Journal of Speech and Hearing Research, 38,* 26-32.

Revis, J., Giovanni, A., Wuyts, F., & Triglia, J.M. (1999). Comparison of different voice samples for perceptual analysis. *Folia Phoniatrica et Logopaedica, 51,* 108-116.

Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B., & Richardson, W.S. (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal, 312,* 71-72.

Scherer, R.C., Vail, V.J., & Guo, C.G. (1995). Required number of tokens to determine representative voice perturbation values. *Journal of Speech and Hearing Research, 38,* 1260-1269.

Shrivastav, R. (2003). The use of an auditory model in predicting perceptual ratings of breathy voice quality. *Journal of Voice, 17,* 502-512.

Shrivastav, R., Sapienza, C.M., Nandur, V. (2005). Application of psychometric theory to the measurement of voice quality using rating scales. *Journal of Speech, Language and Hearing Research, 48,* 323-335.

Smits, I., Ceuppens, P., & De Bodt, M.S. (2005). Comparative study of acoustic voice measurements by means of Dr. Speech and Computerized Speech Lab. *Journal of Voice, 19,* 187-196.

Titze, I.R. (1994). *Principles of voice production.* Englewood Cliffs: Prentice Hall.

Titze, I.R. (1994). The G. Paul Moore Lecture: toward standards in acoustic analysis of voice. *Journal of Voice, 8,*1-7.

Titze, I.R. (1995). *Workshop on acoustic voice analysis: summary statement.* Iowa City: National Center for Voice and Speech.

Van den Berg, J. (1958). Myoelastic-aerodynamic theory of voice production. *Journal of Speech and Hearing Research, 1,* 227-244.

Webb, A.L., Carding, P.N., Dreary, I.J., MacKenzie, K., Steen, N., & Wilson, J.A. (2004). The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Otorhinolaryngology, 261,* 429-434.

Wilson, D.K. (1979). *Voice problems of children.* Baltimore: Williams & Wilkins.

Wolfe, V.I., Cornell, R., & Fitch, J. (1995). Sentence/vowel correlation in the evaluation of dysphonia. *Journal of Voice, 9,* 297-303.

Wolfe, V., & Martin, D. (1997). Acoustic correlates of dysphonia: type and severity. *Journal of Communication Disorders, 30,* 403-416.

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,* 508-517.

Yamagushi, H., Shrivastav, R., Andrews, M.L., & Niimi, S. (2003). A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. *Folia Phoniatrica et Logopaedica, 55,* 147-157.

Yiu, E.M., & Ng, C. (2004). Equal appearing interval and visual analoque scaling of perceptual roughness and breathiness. *Clinical Linguistics & Phonetics, 18,* 211-229.

Yu, P., Revis, J., Wuyts, F.L., Zanaret, M., & Giovanni, A. (2002). Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. *Folia Phoniatrica et Logopaedica, 54,* 271-281.

Yumoto, E., Gould, W.J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America, 71,* 1544-1549.

Yumoto, E. (1983). The quantitative evaluation of hoarseness. *Archives of Otolaryngology, 109,* 48-52.

Yumoto, E., Sasaki, Y., & Okamura, H. (1984). Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. *Journal of Speech and Hearing Research, 27,* 2-6.

Zraick, R.I., Wendel, K., & Smith-Olinde, L. (2005). The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *Journal of Voice, 19,* 574-581.

2

PERTURBATION MEASURES OF VOICE: A COMPARATIVE STUDY BETWEEN MULTI–DIMENSIONAL VOICE PROGRAM AND PRAAT

Youri Maryn
Paul Corthals
Marc De Bodt
Paul Van Cauwenberge
Dimitar Deliyski

This chapter has been published in:
Folia Phoniatrica et Logopaedica, 2009;61:217-226.

ABSTRACT

Background/aims: frequency and amplitude perturbations are inherent to voice acoustic signals. The assessment of voice perturbation is influenced by several factors, including the type of recording equipment used and the measurement extraction algorithm applied. In the present study, perturbation measures provided by two computer systems (a purpose-built professional voice analysis apparatus and a personal computer-based system for acoustic voice assessment) and two computer programs (Multi–Dimensional Voice Program and Praat) were compared.

Methods: correlations and inferential statistics for seven perturbation measures (absolute jitter, percent jitter, relative average perturbation, pitch perturbation quotient, shimmer in dB, percent shimmer, and amplitude perturbation quotient) in 50 subjects with various voice disorders are presented.

Results: results indicate statistically significant differences between the two systems and programs, with Multi-dimensional voice program yielding consistently higher measures than Praat. Furthermore, correlation analyses show weak to moderate proportional relationships between the two systems and weak to strong proportional relationships between the two programs.

Conclusion: based on literature and the proportional relationships and differences between the two systems and programs under consideration in this study, one can state that one can hardly compare frequency perturbation outcomes across systems and programs and amplitude perturbation outcomes across systems.

INTRODUCTION

Minor disturbances in the frequency and the amplitude of the voice signal, called perturbations, are unavoidably present even when one tries to produce a perfectly steady sound (Titze et al., 1994). In patients with a voice problem, perturbation may become worse and result in a more severe deviation from the normal voicing pattern. Perceptually, this may be interpreted as dysphonia and described using labels like hoarse, breathy and rough. Popular acoustic metrics to assess dysphonia are jitter and shimmer, denoting short term (cycle-to-cycle) variability in fundamental frequency and amplitude respectively. A comprehensive review on this topic can be found in Baken and Orlikoff (2000). Since Lieberman (1963) introduced the concept of perturbation analysis in the area of voice and speech, the demand for reliable, valid and objective voice analyses has motivated acoustic voice research and perturbation measurements have undergone considerable refinement. The availability of user-friendly personal computer systems has made quantitative voice and speech analysis commonly-accessible (Read et al., 1992, Howard, 2001). A well known commercially-available computer system for voice analysis, the Computerized Speech Lab (CSL) by Kay Elemetrics (currently known as KayPentax) (2004), offers several perturbation measures in its Multi–Dimensional Voice Program (MDVP) (Kay Elemetrics,

2003). An example of freely-available personal computer-based analysis software is Praat (Boersma, 2001; Boersma & Weenink, 2005). It also provides perturbation measures in a voice report.

Acoustic voice analysis based on perturbation measures has long been subject to debate. A key issue is validity, in particular concurrent criterion-related validity with perceptual evaluation as the bench mark for voice quality assessment. Several authors have found significant relationships between perceptual evaluation and acoustic perturbation. For example, Eskenazi et al. (1990) point to jitter (percent) as a predictor for breathiness and hoarseness, in contrast to pitch perturbation quotient (PPQ) and amplitude perturbation quotient (APQ). Dejonckere et al. (1996) found significant correlations between jitter (percent) and roughness, between shimmer (percent) and breathiness, and between shimmer (percent) as well as noise-to-harmonics ratio (NHR) and Hirano's (1981) Grade index for perceptual voice assessment. Wolfe and Martin (1997) revealed significant correlations between jitter (percent) and breathiness and between shimmer and hoarseness, an inclusive term the authors use for indicating glottal noise and roughness. However, such perturbation-quality relationships do not always emerge. For example, Bhuta et al. (2004) reported significant multivariate correlations between MDVP noise parameters (voice turbulence index or VTI, noise-to-harmonics ratio or NHR and soft phonation index or SPI) and perceptual GRBAS (Hirano, 1981) voice evaluation, but individual perturbation measures were not observed to be significant correlates. De Bodt (1997) could not find any meaningful objective acoustic correlate for perceptual GRBAS ratings. Differences in judge experience, voice samples used, type and severity of pathology, and data acquisition hardware and software often lead to inconsistent research findings. A more profound discussion on the validity of acoustic metrics for voice quality is beyond the scope of this article, and interested readers are referred to Kreiman and Gerratt (2000).

Another issue concerns the differences in measuring outcome between computer systems and between computer programs. Since every computerized speech recording and analysis system has its own configuration for data acquisition such as microphone type and localization relative to the source (Titze & Winholtz, 1993; Winholtz & Titze, 1997), presence or absence of external amplifying hardware (as in the case of Kay Elemetrics' CSL), type of personal computer with its typical hardware and software settings for recording and the properties of its internal sound card (Deliyski et al., 2005a; Deliyski et al., 2005b), use of external digital recording apparatus such as digital audio tape or minidisc (Winholtz & Titze, 1998), analysis and processing program (Bielamowicz et al., 1993; Karnell et al., 1995; Smits et al., 2005), and measurement algorithms (Rabinov et al., 1995), etc., differences in any of these system related items can lead to more or less *intersystem differences* in perturbation measurements. Collectively, Deliyski et al. (2006) investigated the extent and the order in which gender, microphone, number of tokens, type of environmental noise, level of environmental noise, data acquisition system and software influence perturbation measures on 80,000 audio

recordings. Although all of the factors were considered to be influential, they concluded that the most prominent effect on perturbation measures was exercised by analysis software, followed by gender and type of microphone (Deliyski et al., 2006).

When the same recording is analyzed using different software, keeping all other system related factors invariant, the differences in results must be due to the programs (as for example between Dr. Speech [Tiger Electronics DRS, Seattle, WA] and CSL) and more specifically their settings such as sampling rate, method of fundamental period extraction (Titze & Liang, 1993; Howard, 2001; Boersma, 2001; Awan & Scarpino, 2004; Roark, 2006), perturbation algorithm (Rabinov et al., 1995), etc. Especially the F_0 extraction algorithm seems to be of crucial importance in voice perturbation measures. Titze & Liang (1993) investigated the performance of three event-detection F_0 extraction methods, cycle-to-cycle waveform-matching, zero-crossing, and peak-picking. They stated that peak-picking yields higher perturbation values than zero-crossing and that waveform-matching provides the lowest perturbation values. Furthermore, they concluded that waveform-matching performs best in signals with a frequency variation below 6 % per cycle (p. 1133). Possible reasons why this is so are profoundly explored and discussed by Roark (2006). Differences in any of the program related items can lead to *interprogram differences* in perturbation measurements. Such interprogram differences in perturbation outcomes have been investigated by Bielamowicz et al. (1993), Karnell et al. (1995) and Smits et al. (2005). Comparison of the fundamental frequency measures among these three studies revealed near-perfect correlations and non-significant differences, illustrating very strong agreement for mean fundamental frequency. For frequency perturbation and amplitude perturbation on the other hand, there was a very poor to moderately high agreement with statistically significant differences between several computer programs. These data were more recently confirmed in the study of Deliyski & Shaw (2006), who found moderate to very strong correlations between frequency and amplitude perturbation measures of three different programs. In general, these differences were attributed to the use of different fundamental frequency extraction methods in the perturbation measurements of the various systems. These studies confirm the earlier review of Read et al. (1992), who concluded that the systems generally perform quite well but differ greatly in how their operations are performed.

This study was undertaken to: (a) investigate the intersystem differences between two commonly used systems for computerized perturbation measurements (CSL with MDVP and a common desktop PC-system with Praat) with dissimilar microphone type, microphone placement, external hardware, computer, and installed software; (b) examine the interprogram differences between two frequently utilized acoustic analysis programs (MDVP and Praat) for voice samples recorded with CSL.

These issues are especially interesting when clinicians, for instance, aim to relate data obtained by different systems and/or programs or when clinicians want to compare data with normative statistics. To the knowledge of the authors, a

comparative study between data collected in dysphonic patients by means of these two systems or programs has not been done yet despite the fact that both are widely known and used in the clinical and scientific realm of voice disorders.

METHODS

Subjects

Fifty patients participated in this study. The participants were recruited on an informed consent basis from the ENT case load of the Sint-Jan General Hospital in Bruges in the course of a 1-year period. They all presented with various voice disorders and had been referred for multidimensional voice assessment by staff otolaryngologists. There were 23 males with a mean age of 51 years and an age range from 13 to 74 years. The 27 females had a mean age of 36 years, ranging from 14 to 71. All laryngological diagnoses were made with a flexible transnasal chip-on-tip laryngoscope. Table 2.1 summarizes laryngoscopic findings. The scores on the Voice Handicap Index (VHI; Jacobson et al., 1997), as a quantification of the amount of disability caused by a voice disorder, had an average of 51 and ranged from 19 to 106. The scores on the Dysphonia Severity Index (DSI; Wuyts et al., 2000), as an objective and multiparametric estimate of (disordered) voice quality, ranged from -15.55 to 4.58 with a mean of -1.54. This group of subjects can be considered to be clinically representative for the population of voice disordered patients, reflecting different age groups, different degrees of dysphonia and voice complaints, and non-organic as well as organic laryngeal pathologies.

Table 2.1 List of laryngeal pathologies with their relative occurrence in the group of this study.

	Number	Percentage
Non-organic	20	40
Nodules	7	14
Polyp	5	10
Cyst	2	4
Polypoid mucosa (edema in Reinke's space)	4	8
Granuloma	1	2
Leukoplakia	2	4
Unilateral vocal fold paralysis	9	18
Total	**50**	**100**

Recordings

From every subject, a voice sample was simultaneously recorded using the two systems. Recording settings are summarized in Table 2.2. The subjects were asked to produce sustained phonation of the vowel /a/ at a comfortable pitch and

loudness. The simultaneous recording of the sustained vowel resulted in identical 3 second samples of an oscillographically-steady portion of the vowel (excluding voice onset and offset). Concerning the oscillographic steadiness of the samples, decisions were made based on the presence or absence of gross signs of instability (e.g. unvoiced segments, voice breaks, etc.) while looking at the real-time waveform in MDVP (with screen width equal to 3 seconds). When the first trial was not sufficiently long or oscillographically too unsteady for further research, more trials were undertaken until an acceptable recording was obtained. After recording, all samples were saved in wave format on the hard disks of both computer systems. Acoustic analyses were done on these pairs of files. Recordings from the CSL-system (with MDVP) and the PC-system (with Praat) were utilized for investigating intersystem differences. For interprogram differences, recordings from the CSL-system were analyzed in both MDVP and Praat. The ambient noise level in the laboratory room, measured with a Larson & Davis 800B precision integrating sonometer (Larson & Davis Laboratories, Provo, UT), was 36 dB_A. The voice recordings had an intensity range from 70.08 dB_{SPL} to 85.14 dB_{SPL}, resulting in signal-to-noise ratios (SNR) ranging from 34.08 dB to 39.14 dB. Although the recommended SNR-level was 42 dB (Deliyski et al., 2005c), SNR-levels above 30 dB are still acceptable (Deliyski et al., 2006).

Table 2.2 Recording and acquisition settings of the two computer systems used in this study.

	System 1 COMPUTERIZED SPEECH LAB (CSL)	System 2 PERSONAL COMPUTER (PC)
Microphone		
Type	AKG C420 head-mount condenser microphone with balanced output (AKG Acoustics, 2000)	Shure Prologue 14H desktop dynamic microphone (Shure, 2003)
Mouth-to-microphone angle	± 45° (left)	± 45° (right)
Mouth-to-microphone distance	± 5 cm	± 15 cm
Computer		
Type	Fujitsu Siemens Scenic P300 desktop computer	Fujitsu Siemens Scenic T desktop computer with a built-in soundcard
External hardware		
	Computerized Speech Lab model 4500 (CSL) (Kay Elemetrics, 2004)	/
Program		
Name	Multi-Dimensional Voice Program (MDVP) (Kay Elemetrics, 2003)	Praat (Boersma, 2001; Boersma & Weenink, 2005)
Model/version	Model 5105, Version 2.6.2	Version 4.4.01
Sample rate	44100 Hz	44100 Hz
F0 extraction method	Signum-encoded autocorrelation followed by pitch-synchronous peak detection with linear interpolation (Deliyski, 1993)	Autocorrelation with sinc interpolation followed by waveform-matching (Boersma, 1993; Boersma, 2004)

Acoustic measures

The following seven perturbation measures were obtained in MDVP as well as in Praat. There were four frequency perturbation parameters: absolute jitter, percent jitter, relative average perturbation and pitch perturbation quotient. The appellations of the parameters with similar order of perturbation function in Praat are: jitter local absolute, jitter local, jitter rap and jitter ppq5 respectively. There were three amplitude perturbation parameters: shimmer in dB, percent shimmer and amplitude perturbation quotient. The appellations of similar parameters in Praat are: shimmer dB, shimmer local and shimmer apq11 respectively. Profound elaboration regarding the F_0 extraction algorithms and the perturbation extraction algorithms of MDVP and Praat is provided in Deliyski (1993) and Boersma (1993), respectively.

Statistics

All statistical analyses were done using SPSS for Windows version 12.0 (SPSS, Chicago, Illinois, USA). First, all data were explored for the presence of outlying and extreme data. Outliers are defined as data with values between 1.5 and 3 times the interquartile range. Extremes are defined as data with values more than 3 times the interquartile range. Because outliers and extremes can dramatically influence and thus grossly distort the absolute value of r, they were omitted from the dataset, excluding between 3 and 7 data points per measure of the two systems and programs. Second, for the comparison of the two systems as well as the two programs, two kinds of statistics were employed. Pearson product-moment correlation coefficients (r) were calculated in order to determine the degree of correspondence among the 7 perturbation measures produced by both systems or programs respectively. Furthermore, as an important proportional relationship between two measures does not necessarily imply equality of the actual values produced by these programs or systems, differences were evaluated by means of the t test for 2 dependent samples.

RESULTS

Comparison of the systems

Descriptive statistics for the perturbation measurements derived from the two systems are shown in Table 2.3. In Table 2.4, the Pearson bivariate correlation scores for the different pairs of simultaneously recorded vowel samples are summarized. For all frequency and amplitude perturbation measures, the correlation values showed a weak to moderate relationship. As an example of the results in Table 2.4, the regression line in the scatterplot in Figure 2.1 illustrates the moderate correlation between the values of percent jitter obtained with the two systems.

Table 2.3 Descriptive statistics for the data in the two systems.

Perturbation measure: name, unit (system)	n	Mean ± SE	SD	Min	Max	Range
Absolute jitter, μs (1)	46	119.79 ± 9.57	64.92	19.97	295.60	275.63
Absolute jitter, μs (2)	45	51.52 ± 4.58	30.39	12.31	149.89	137.58
Percent jitter, % (1)	47	1.93 ± 0.16	1.13	0.38	5.08	4.70
Percent jitter, % (2)	44	0.79 ± 0.07	0.46	0.17	1.93	1.76
Relative average Perturbation, % (1)	47	1.16 ± 0.10	0.70	0.22	3.02	2.80
Relative average Perturbation, % (2)	45	0.45 ± 0.04	0.28	0.07	1.12	1.05
Pitch perturbation Quotient, % (1)	47	1.13 ± 0.10	0.65	0.23	3.11	2.88
Pitch perturbation Quotient, % (2)	46	0.48 ± 0.04	0.29	0.11	1.25	1.14
Shimmer in dB, dB (1)	45	0.38 ± 0.03	0.19	0.01	0.87	0.86
Shimmer in dB, dB (2)	47	0.33 ± 0.02	0.11	0.12	0.64	0.52
Percent Shimmer, % (1)	45	4.50 ± 0.35	2.35	0.81	11.31	10.50
Percent Shimmer, % (2)	47	3.69 ± 0.17	1.16	1.41	6.82	5.41
Amplitude perturbation Quotient, % (1)	46	3.44 ± 0.25	1.72	1.09	8.43	7.34
Amplitude perturbation Quotient, % (2)	46	2.60 ± 0.11	0.74	1.11	4.37	3.26

Based on the results of the t test for 2 dependent samples (also in Table 2.3), there is a statistically significant difference between the two systems for all pairs of perturbation measures. Perturbation values of the CSL-system were consistently higher than those of the PC-system, especially for the frequency perturbations. For percent jitter, such a difference is illustrated in the box-and-whiskerplot (displaying the upper quartile, lower quartile, and interquartile ranges of a data set) in Figure 2.2.

Table 2.4 Pearson correlation and statistical difference values for variability between two commonly used computer systems for voice perturbation measurement.

	Intersystem correlation r	Intersystem difference t
Absolute jitter	0.360*	7.463***
Percent jitter	0.442**	7.325***
Relative average perturbation	0.470**	7.716***
Pitch perturbation quotient	0.481**	7.653***
Shimmer in dB	0.455**	2.569*
Percent shimmer	0.332*	2.455*
Amplitude perturbation quotient	0.325*	3.469***

r: Pearson product-moment correlation coefficients
t: value of the t-test for dependent samples
**: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$*

Figure 2.1 Scatterplot with linear regression line to illustrate the moderate correlation of percent jitter values in the intersystem analysis between measurements in the CSL-system with MDVP and the PC-system with Praat r=0.44).

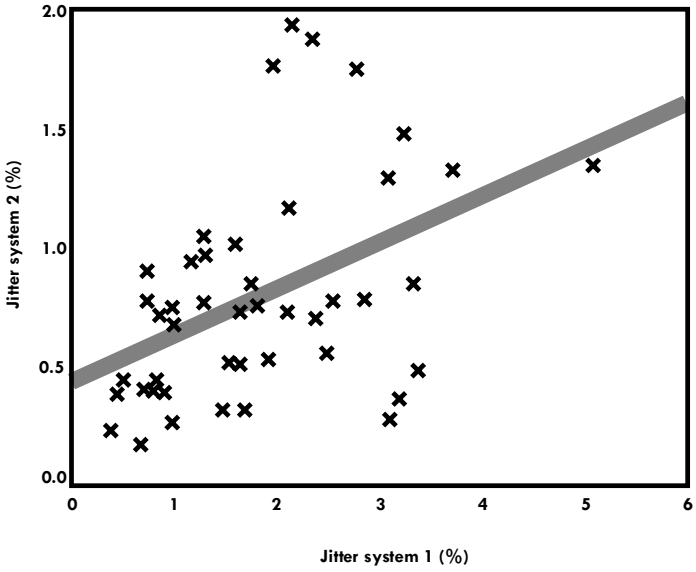
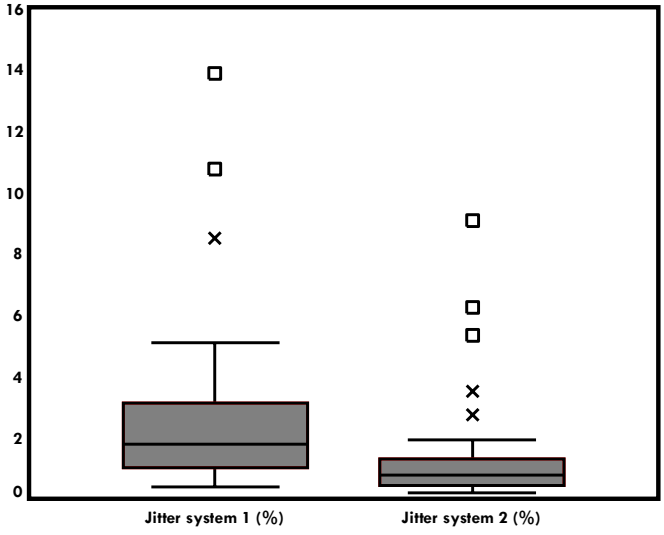


Figure 2.2 Box-and-whiskerdiagram to illustrate the statistically significant intersystem difference in percent jitter values between measurementsin the CSL-system with MDVP and the PC-system with Praat (×: outliers, □: extremes).



Comparison of the programs

Table 2.5 represents the descriptive statistics for the perturbation measurements derived from the two programs. Table 2.6 summarizes the Pearson correlation coefficients for the vowel samples recorded with the CSL-system and analyzed with MDVP and Praat. For all frequency perturbation measures, the results indicate a weak (for percent jitter and pitch perturbation quotient) to moderate (for absolute jitter and relative average perturbation) proportional relationship between MDVP and Praat. As an example, Figure 2.3 illustrates the weak correlation between the values of percent jitter obtained with the two programs. Regarding the amplitude perturbations, a moderate correlation was found for shimmer in db and there was a strong correlation for percent shimmer and amplitude perturbation quotient (as demonstrated in the scatterplot with regression line of Figure 2.4).

Table 2.5 Descriptive statistics for the data in the two programs.

Perturbation measure: name, unit (system)	n	Mean ± SE	SD	Min	Max	Range
Absolute jitter, μs (1)	46	119.79 ± 9.57	64.92	19.97	295.60	275.63
Absolute jitter, μs (2)	45	43.48 ± 5.25	34.45	10.38	213.61	203.23
Percent jitter, % (1)	47	1.93 ± 0.16	1.13	0.38	5.08	4.70
Percent jitter, % (2)	44	0.62 ± 0.05	0.33	0.17	1.91	1.74
Relative average Perturbation, % (1)	47	1.16 ± 0.10	0.70	0.22	3.02	2.80
Relative average Perturbation, % (2)	45	0.33 ± 0.03	0.16	0.07	0.69	0.62
Pitch perturbation Quotient, % (1)	47	1.13 ± 0.10	0.65	0.23	3.11	2.88
Pitch perturbation Quotient, % (2)	46	0.37 ± 0.03	0.21	0.11	1.23	1.12
Shimmer in dB, dB (1)	45	0.38 ± 0.03	0.19	0.01	0.87	0.86
Shimmer in dB, dB (2)	47	0.31 ± 0.03	0.21	0.07	0.95	0.88
Percent Shimmer, % (1)	45	4.50 ± 0.35	2.35	0.81	11.31	10.50
Percent Shimmer, % (2)	47	3.69 ± 0.37	2.43	0.85	10.79	9.94
Amplitude perturbation Quotient, % (1)	46	3.44 ± 0.25	1.72	1.09	8.43	7.34
Amplitude perturbation Quotient, % (2)	46	2.81 ± 0.27	1.76	0.78	6.84	6.06

For all pairs of perturbation measures, a statistically significant difference between the two programs was found. Looking at the box-and-whiskerdiagrams of Figure 2.5, where the results for percent jitter are serving as an example for all the other frequency perturbation measures, there is almost no overlap of the interquartile ranges between the two programs. The MDVP measures are consistently higher than the Praat measures. For the amplitude perturbations there

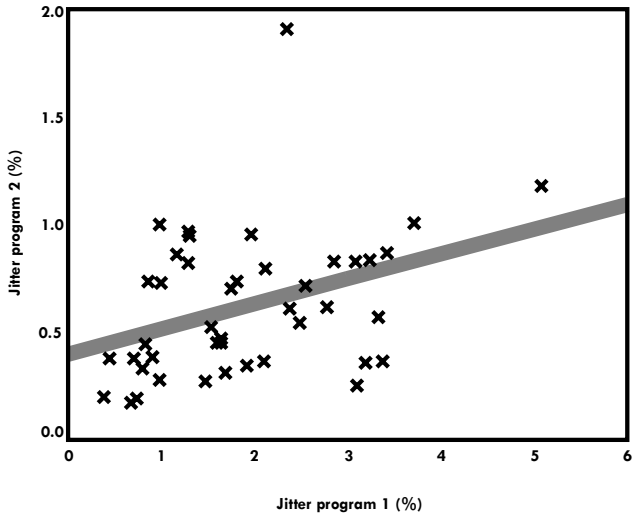
is more overlap, and thus there is less difference between similar measures, as evidenced by the lower t-values in Table 2.6.

Table 2.6 Pearson correlation and statistical difference values for variability between two frequently utilized acoustic analysis programs for voice perturbation measurement.

	Intersystem correlation r	Intersystem difference t
Absolute jitter	0.470**	8.669***
Percent jitter	0.366*	8.644***
Relative average perturbation	0.412**	9.059***
Pitch perturbation quotient	0.370*	8.527***
Shimmer in dB	0.542**	2.371*
Percent shimmer	0.780**	3.338**
Amplitude perturbation quotient	0.870**	4.577***

r: Pearson product-moment correlation coefficients
t: value of the t-test for dependent samples
**: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$*

Figure 2.3 Scatterplot with linear regression line to illustrate the weak correlation of percent jitter values in the interprogram analysis between measurements in MDVP and Praat, both acquired with the CSL-system (r=0.37).

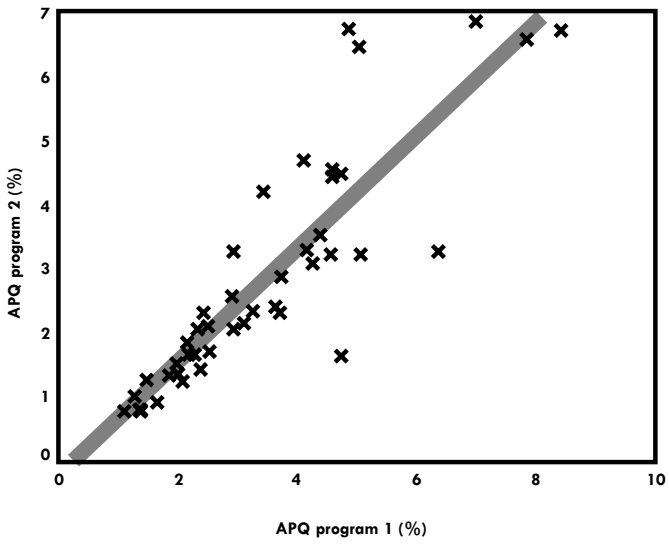


DISCUSSION

This study reports on the differences and similarities of perturbation measures obtained by two computer-based acoustic analysis programs (MDVP and

Praat) and systems (CSL with MDVP and a personal computer with Praat), when examining a corpus of 3 second segments of sustained vowel /a/ obtained from 50 patients with various voice disorders.

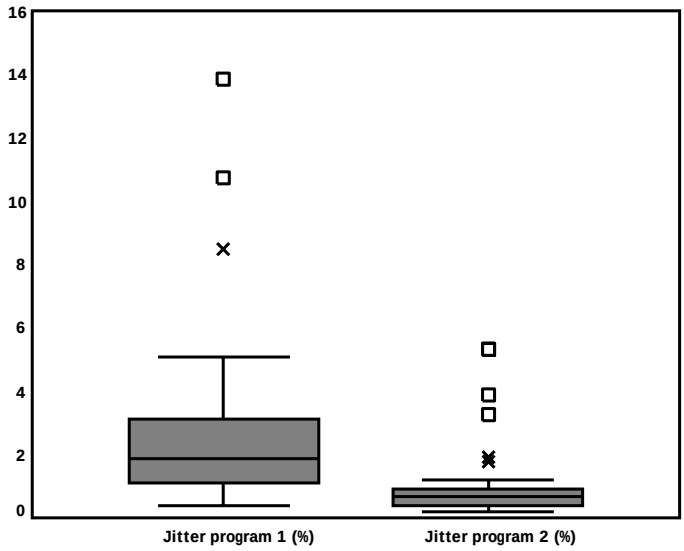
Figure 2.4 Scatterplot with linear regression line to illustrate the strong correlation of amplitude perturbation quotient values in the interprogram analysis between measurements in MDVP and Praat, both acquired with the CSL-system (r=0.87).



Before discussing the results of this investigation, attention is to be drawn to the data that were excluded from the dataset. In this study, statistical exploration was chosen to be the basis upon which data (outliers and extremes) were excluded. Another method for excluding perturbation data (expressed in percentage) from further analyses is the implementation of the threshold of 5%, since perturbation measures less than about 5% have been found to be reliable (Titze & Liang, 1993; Awan & Scarpino, 2004). Practised on the frequency perturbation data from the CSL-system, both methods exclude almost the same data. For percent jitter, three values (8.494%, 10.738%, 13.835%) were omitted based on statistical exploration. Only one value higher (5.075%) than 5% remained in the dataset. However, this is a very laminar value. For relative average perturbation, statistical exploration also excluded three values (4.748%, 6.366%, 7.849%) and there were no other values above 5%. For pitch perturbation quotient, also three values (5.582%, 7.170%, 8.639%) were excluded on the basis of statistical explorations and again there were no other values above 5%. The three values that were excluded across all frequency perturbation measures originate from the same three voice recordings: recording 38 (unilateral vocal fold paralysis), recording 40 (hyperfunctional dysphonia with ventricular hyperadduction) and recording 44 (unilateral vocal fold paralysis).

Visual investigation of the narrowband spectrograms revealed type 3 signals in all three recordings (with near-absence of harmonics), according to the classification of Titze (1995). There was 98%, 98% and 100% agreement in exclusion of data between these two methods for percent jitter, relative average perturbation and pitch perturbation quotient, respectively. The threshold of 5% can not be utilized for absolute jitter and shimmer in dB, since both are not expressed as a percentage.

Figure 2.5 Box-and-whiskerdiagram to illustrate the statistically significant interprogram difference in percent jitter values between measurements in the CSL-system with MDVP and the PC-system with Praat (×: outliers, □: extremes).



Several studies have already investigated the interprogram differences in acoustic vocal perturbation measurements (Bielamowicz et al., 1993; Karnell et al., 1995; Smits et al., 2005; Deliyski et al., 2006; Deliyski & Shaw, 2006). Although Bielamowicz et al. (1993), Karnell et al. (1995) and Smits et al. (2005) found a very strong interprogram agreement in the fundamental frequency measurements, the analysis of voice perturbation measures yielded much less significant correlations. Furthermore, the correlations between the programs were higher for amplitude perturbation measures than for frequency perturbation measures (Smits et al., 2005; Deliyski & Shaw, 2006). Bielamowicz et al. (1993) explained this difference in frequency and amplitude perturbation by the fact that jitter is far more dependent on the exact placement of cycle boundaries than shimmer. Whereas minimal errors in placing these boundaries (e.g. due to F_0 tracking dissimilarities) markedly adds noise to frequency perturbations measurements, the effect of such errors is less detrimental to amplitude perturbations because they generally lack sufficient magnitude to eliminate an entire peak from a cycle. Smits et al. (2005)

compared the measurement of absolute jitter, relative (percent) jitter and relative (percent) shimmer between CSL and Dr. Speech software. They found Pearson correlation coefficients ranging from 0.26 (absolute jitter) and 0.31 (relative jitter) to 0.69 (relative shimmer). Deliyski & Shaw (2006) compared frequency and amplitude perturbation between MDVP, TF32 (formerly known as CSpeech, by Paul Milenkovic, Madison, WI) and Praat. For the frequency perturbation, they found moderate to very strong correlations (0.40, 0.44 and 0.90) and for the amplitude perturbation there were strong to very strong correlations (0.75 and 0.98). Their interprogram comparison between MDVP and Praat yielded correlations of 0.44 and 0.98 for relative average perturbation and percent shimmer, respectively. We found similar correlations (0.41 and 0.78) in our interprogram comparison of the same measures. In general, these results in the literature corroborate with the findings of the interprogram comparison in the present study: weak to moderate correlations for frequency perturbation measures and moderate to strong correlations for the amplitude perturbation measures (Table 2.6). It should be noted that the different programs utilized different F_0 tracking methods. A profound tutorial on F_0 extraction methods and the effects of discrepancies in F_0 extraction is given by Roark (2006).

Next to comparing two programs for perturbation measurement, this study also investigated the differences and similarities between two commonly used data acquisition systems: CSL with MDVP and a personal computer with Praat. The intersystem comparison for frequency perturbation measures yielded weak to moderate correlations and was therefore similar to the interprogram comparison. For the amplitude perturbation measures, on the other hand, the moderate to strong correlations from the interprogram comparison dropped to weak to moderate correlations in the intersystem comparison. This suggests that the amplitude perturbation measures are more susceptible for differences in the data acquisition and harmonizes with the results of Deliyski et al. (Deliyski et al., 2005a), who found a statistically significant impact of data acquisition environment and microphone on amplitude perturbation but not on frequency perturbation.

The present study also revealed differences between the perturbation measures stemming from both analysis programs/systems. Whereas all differences where statistically significant for all perturbation measures (with MDVP-values being consistently higher than Praat-values), the interquartile ranges in the box-and-whiskerplots are clearly less overlapping for the frequency perturbations than for the amplitude perturbations. In the case of the comparison between the two programs (MDVP and Praat), the recording hardware and acquisition were identical. Furthermore, the perturbation measures across the two programs were rather similar regarding the order of the perturbation function. Statistical differences between the actual values, on the other hand, can be explained by the dissimilarities between the two systems/programs: the pitch extraction algorithm was different. Praat utilizes an autocorrelation method with sinc-interpolation followed by a cycle-to-cycle waveform-matching period detection (Boersma, 1993; Boersma, 2004), while MDVP uses a combination of a signum-encoded

autocorrelation method followed by pitch-synchronous peak detection with linear interpolation (Deliyski, 1993). This important difference causes Praat measuring smaller perturbation values than MDVP.

As for the intersystem comparison, very similar results arose and analogous explanations can be given. The strong correlations could be attributed to similarities in computer apparatus, noise conditions and computation algorithms used in both systems. Both systems differed in the presence/absence of external preamplifying hardware, microphone type, and mouth-to-microphone angle and distance. Although earlier research states that perturbation measures depend on microphone and hardware characteristics (Titze & Winholtz, 1993; Winholtz & Titze, 1997; Deliyski et al., 2005), these dissimilarities did not have a drastic impact on the correlation coefficients in the present study, according to interprogram correlation results. Statistical differences are slightly smaller in the intersystem than in the interprogram variability study.

An additional comment is warranted regarding the number of subjects included in this study. In order to be representative for the population of voice disordered patients, the inclusion of more than fifty subjects can empower the results of this study. However, Karnell et al. (1995) and Deliyski & Shaw (2006) included only twenty pathologic and normal subjects, respectively and Smits et al. (2005) included one-hundred twenty but normophonic subjects. Bielamowicz et al. (1993) included a selection of fifty pathologic subjects, a number similar to the number of subjects in this study.

CONCLUSION

Based on the available literature and on the proportional relationships and differences between the two systems and programs under consideration in this study, one can state that one can hardly compare frequency perturbation outcomes across systems and programs and amplitude perturbation outcomes across systems. It is therefore important to have system-specific or program-specific normative data. The normative data for MDVP are present in its manual (Kay Elemetrics, 2003). For Praat, however, there are no such data available, inducing a direction for future research.

REFERENCES

AKG Acoustics (2000). C420: User instruction. MicroMic series II. München: AKG Acoustics Harman Pro.

Awan, S.N., & Scarpino, S.E. (2004). Measures of vocal F_0 from continuous speech samples: an interprogram comparison. *Journal of Speech-Language Pathology and Audiology, 28*, 122-131.

Baken, R.J., & Orlikoff, R.F. (2000). *Clinical measurement of speech and voice (2nd edition).* San Diego: Singular Publishing Group.

Bhuta, T., Patrick, L., & Garnett, J.D. (2004). Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice, 18,* 299-304.

Bielamowicz, S., Kreiman, J., Gerratt, B.R., Dauer, M.S., & Berke, G.S. (1993). Comparison of voice analysis systems for perturbation measurement. *Journal of Speech and Hearing Research, 39,* 126-134.

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

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glottal International, 5,* 341-345.

Boersma, P. (2004). Stemmen meten met Praat. *Stem-Spraak-Taalpathologie, 12,* 237-251.

Boersma, P., & Weenink, D. (2005). *Praat: doing phonetics by computer (Version 4.3.14) [Computer program]*. Amsterdam: Institute of Phonetic Sciences [cited 2005 May 5]. Available from: http://www.praat.org.

De Bodt, M. (1997). *A framework for voice assessment: the relation between subjective and objective parameters in the judgement of normal and pathological voice.* University of Antwerp: Unpublished doctoral dissertation.

Dejonckere, P.H., Remacle, M., Fresnel-Elbaz, E., Woisard, V., Crevier-Buchman, L., & Millet, B. (1996). Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Revue de Laryngologie-Otologie-Rhinologie, 117,* 219-224.

Deliyski, D.D. (1993). *Acoustic model and evaluation of pathological voice production.* Proceedings of Eurospeech, Berlin, Germany.

Deliyski, D.D., Evans, M.K., & Shaw, H.S. (2005a). Influence of data acquisition environment on accuracy of acoustic voice quality measurements. *Journal of Voice, 19,* 176-186.

Deliyski, D.D., Shaw, H.S., Evans, M.K. (2005b). Influence of sampling rate on accuracy and reliability of acoustic voice analysis. *Logopedics Phoniatrics Vocology, 30,* 55-62.

Deliyski, D.D., Shaw, H.S., & Evans, M.K. (2005c). Adverse effects of environmental noise on acoustic voice quality measurements. *Journal of Voice, 19,* 15-28.

Deliyski, D.D., Shaw, H.S., Evans, M.K., & Vesselinov, R. (2006). Regression tree approach to studying factors influencing acoustic voice analysis. *Folia Phoniatrica et Logopaedica, 58,* 274-288.

Deliyski, D., & Shaw, E. (2006). *Acoustic measurement of jitter and shimmer: inter- and intrasystem relationships.* Proceedings of the ASHA Convention, Miami Beach, USA.

Eskenazi, L., Childers, D.G., & Hicks, D.M. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research, 33,* 298-306.

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

Howard, D.M. (2001). *The real and non-real in speech measurements.* Proceedings of the 2nd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications MAVEBA, Florence, Italy.

Jacobson, B.H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., & Benninger, M.S. (1997). The Voice Handicap Index (VHI): development and validation. *American Journal of Speech-Language Pathology, 6,* 66-70.

Karnell, M.P., Hall, K.D., & Landahl, K.L. (1995). Comparison of fundamental frequency and perturbation measurements among three analysis systems. *Journal of Voice, 9,* 383-393.

Kay Elemetrics (2004). *Multi-Speech and CSL software: software instruction manual.* Lincoln Park: Kay Elemetrics.

Kay Elemetrics (2003). *Multi-Dimensional Voice Program (MDVP) model 5105: software instruction manual.* Lincoln Park: Kay Elemetrics.

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

Lieberman, P. (1963). Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. *Journal of the Acoustical Society of America, 35,* 344-353.

Rabinov, C.R., Kreiman, J., Gerratt, B.R., & Bielamowicz, S. (1995). Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter. *Journal of Speech and Hearing Research, 38,* 26-32.

Read, C., Buder, E.H., & Kent, R.D. (1992). Speech analysis systems: an evaluation. *Journal of Speech and Hearing Research, 35,* 314-332.

Roark, R.M. (2006). Frequency and voice: perspectives in the time domain. *Journal of Voice, 20,* 325-354.

Shure (2003). Model Prologue 14H user guide. Evanston: Shure.

Smits, I., Ceuppens, P., & De Bodt, M.S. (2005). Comparative study of acoustic voice measurements by means of Dr. Speech and Computerized Speech Lab. *Journal of Voice, 19,* 187-196.

Titze, I.R., & Liang, H. (1993). Comparison of F_0 extraction methods for high-precision voice perturbation measurements. *Journal of Speech and Hearing Research, 36,* 1120-1133.

Titze, I.R., & Winholtz, W.S. (1993). Effect of microphone type and placement on voice perturbation measurements. *Journal of Speech and Hearing Research, 36,* 1177-1190.

Titze, I.R. (1994). *Principles of voice production.* Englewood Cliffs: Prentice Hall.

Titze, I.R. (1995). *Workshop on acoustic voice analysis: summary statement.* Iowa City: National Center for Voice and Speech.

Winholtz, W.S., & Titze, I.R. (1997). Miniature head-mounted microphone for voice perturbation analysis. *Journal of Speech, Language and Hearing Research, 40,* 894-899.

Wolfe, V., & Martin, D. (1997). Acoustic correlates of dysphonia: type and severity. *Journal of Communication Disorders, 30,* 403-416.

Winholtz, W.S., & Titze, I.R. (1998). Suitability of minidisc (MD) recordings for voice perturbation analysis. *Journal of Voice, 12,* 138-142.

Wuyts, F.L., De Bodt, M.S., Molenberghs, G., Remacle, M., Heylen, L., Millet, B., Van Lierde, K., Raes, J., & Van de Heyning, P.H. (2000). The Dysphonia Severity Index: an objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language and Hearing Research, 43,* 796-809.

3

ACOUSTIC MEASUREMENT OF OVERALL VOICE QUALITY: A META-ANALYSIS

Youri Maryn
Marc De Bodt
Paul Van Cauwenberge
Nelson Roy
Paul Corthals

This chapter is published in:
Journal of the Acoustical Society of America, 2009;126:2619-2634.

ABSTRACT

Over the past decades, many acoustic markers have been proposed to be sensitive to and measure overall voice quality. This meta-analysis presents a retrospective appraisal of scientific reports which evaluated the relation between perceived overall voice quality and several acoustic-phonetic correlates. Twenty-five studies met the inclusion criteria and were evaluated using meta-analytic techniques. Correlation coefficients between perceptual judgments and acoustic measures were computed. Where more than one correlation coefficient for a specific acoustic marker was available, a weighted average correlation coefficient was calculated. This was the case in thirty-six acoustic measures on sustained vowels and in three measures on continuous speech. Acoustic measures were ranked according to the strength of the correlation with perceptual voice quality ratings. Acoustic markers with more than one correlation value available in the literature and yielding a homogeneous weighted r of 0.60 or above were considered to be superior. The meta-analysis identified four measures that met these criteria in sustained vowels and three measures in continuous speech. Although acoustic measures are routinely utilized in clinical voice examinations, the results of this meta-analysis suggests that caution is warranted regarding the concurrent validity and thus the clinical utility of many of these measures.

INTRODUCTION

Evaluation of voice quality is considered an essential, but controversial part of the assessment process in the field of voice pathology. In clinical as well as in research settings, two main approaches exist to describe the perceived severity of a voice disorder (Kreiman & Gerratt, 2000a). First, generic and/or global ratings such as "overall voice quality", also known as "G" (for "grade"), "severity of voice disorder", "severity of dysphonia", "overall abnormality", and "overall severity" have been used to capture a composite perceptual judgment of the degree of the perceived dysphonia. In contrast, other voice quality ratings pertain to single and very specific perceptual dimensions, the best known of which are roughness and breathiness. Recent evidence has suggested that perceptual rating of overall voice quality and other more specific perceptual dimensions is difficult, as such judgments depend on the listener's internal standard or scale for voice quality dimensions, on his/her sensitivity for this particular dimension, on fatigue, attention, exposure to various disordered voices and training in perceptual evaluation of voice quality (Kreiman et al., 1993; Eadie & Baylor, 2006). Furthermore, other aspects of voice quality judgments, such as type and range of the scale (Bele, 2005; Eadie & Doyle, 2002), or the type of sample to be evaluated, such as sustained vowel versus continuous speech (Bele, 2005; Zraick et al., 2005; Eadie & Baylor, 2006), can significantly affect the perceptual evaluation of voice quality.

In spite of these listener-related and other potential biases, many researchers have tried to correlate the outcome of acoustic-phonetic measures to vocal quality ratings and dysphonia severity. The replacement of analog recording systems with digital recording systems, the availability of automated analysis algorithms, the non-invasiveness of acoustic measures, combined with the fact that acoustic parameters provide easy quantification of dysphonia improvement during the treatment process, have lead to considerable interest in clinical voice quality measurement using acoustic analysis techniques.

In this regard, the correlation coefficient has emerged as the preferred index to determine the extent of the relationship or effect size between acoustic measures and listener judgments of dysphonia severity. The correlation coefficient as a measure of effect size, measures the strength and direction of a linear relationship between two variables. In the voice quality literature, perceived overall voice quality is treated as the criterion variable with the objective acoustic measure treated as the predictor variable. A correlation coefficient of 1 or -1 indicates a perfectly linear association between the two variables (positive or negative, respectively). This means that any change in the predictor variable corresponds with a proportional change in the criterion variable. A correlation coefficient of 0 indicates the total absence of a linear relationship between the variables. The degree of the linear relationship between criterion and predictor (i.e. correlation) counts as an indication of validity, or the extent to which the score of a measurement (i.e. the acoustic parameter) can be regarded as a valid measure of the criterion (i.e. the perceptual rating). Consequently, the higher the absolute correlation coefficient, the more the acoustic measure is said to reflect the perception of overall voice quality, and vice versa. The correlation coefficient is thus an important and frequently used statistic in voice quality research, especially to validate acoustic measures.

Although the correlation coefficient is the preferred metric to assess the strength of the acoustic-perceptual relationship, at least sixty possible acoustic determinants of overall voice quality with varying predictive power have been identified in the literature over the last four decades. Buder (2000) proposed a taxonomy of fifteen digital signal processing-based categories to help manage the wide array of acoustic measures. The large numbers of studies reviewed by Buder (2000) clearly differ substantially in the number of participants and the magnitude of correlation with perceptual judgments of voice quality. Furthermore, the signal processing strategies vary from classic spectrography to sophisticated statistics on sound wave microstructure. Whereas some authors examined the predictive power of resonance-based aspects, the majority of investigators focused on glottal rather than on supraglottal phenomena, seeking correlates of overall voice quality in the distribution of fundamental frequency, in waveform perturbations, in various spectral parameters (including cepstral coefficients and noise content of the glottal sound source), in glottal air flow models obtained by inverse filtering, or in models based on non-linear dynamics theory.

Although an impressive body of research exists which ostensibly assesses the utility of acoustic measurement to quantify voice quality and dysphonia severity, procedural differences in type and number of acoustic predictors, type of recorded material, analysis equipment and measurement scales, have made it almost impossible to qualitatively appraise the merits of these studies, and precisely define a subset of the most robust and sensitive acoustic measures. One approach to this seemingly intractable problem is to apply meta-analytic techniques. Meta-analysis refers to "the analysis of analyses", and is a statistical technique for amalgamating, summarizing, and reviewing previous quantitative research. Unlike traditional research methods, meta-analysis uses the summary statistics from individual studies as the data points for the purpose of integrating the findings. A key assumption of this analysis approach is that each study provides a different estimate of the underlying relationship within the population. By accumulating results across studies, one can gain a more accurate representation of the population relationship than is provided by the individual study estimators. In this way, meta-analyses permit confidence that the reported results are based on more than one study that found the same result (Frey et al., 1991; Lipsey & Wilson, 2001).

Meta-analysis reports findings in terms of effect sizes. Defining an effect size statistic that adequately represents the quantitative findings of an assortment of research reports in a standardized profile is essential to meta-analysis, as it permits meaningful numerical comparison and analysis (Lipsey & Wilson, 2001). The effect size provides information about the magnitude of the relationships observed across all studies and for subsets of studies. By treating individual correlation coefficients as indicators of effect size, meta-analysis can regroup study outcomes into homogeneous subsets and establish population effect sizes. The population effect size, i.e. the real relationship between a predictor variable (a specific acoustic measure) and the criterion variable (a voice quality rating), is estimated by a "weighted" average of all correlations available for a particular acoustic predictor. In addition to defining a weighted average of all effect sizes (i.e. correlation coefficients) in a meta-analysis, it is also important to know whether or not the various effect sizes all estimate the same population effect size. This is a question of homogeneity (or heterogeneity) of the effect size distribution, and a population effect size can only be interpreted reliably if the underlying data set is sufficiently homogeneous (Hunter et al., 1982). When the variability of effect sizes around their weighted mean is no larger than the dispersion expected from sampling error alone, the effect size distribution is considered to be homogeneous. By comparison, in a heterogeneous distribution, individual effect sizes differ from the weighted mean by more than the sampling error (Lipsey & Wilson, 2001). Multiple correlation coefficients resulting in a homogeneous weighted mean correlation are considered to confirm each other, thereby increasing the generalizability of the findings.

Given the large body of research, which relates acoustic measures to voice quality ratings, meta-analysis techniques can potentially reduce information

overload, and distill this large literature into a manageable and/or tractable set of conclusions. Therefore, the aim of this meta-analysis is twofold: (1) to retrospectively appraise the acoustic-phonetic predictors for overall voice quality (i.e., dysphonia severity) and (2) to establish population relationship estimates for several acoustic measures.

METHODS

In most research on assessment of voice quality, measurements have been completed on sustained vowels as compared to continuous speech. This preference for sustained vowels over continuous speech in acoustic as well as perceptual measurements of voice quality has been motivated by several factors (Askenfelt & Hammarberg, 1986; Parsa & Jamieson, 2001), such as: (a) sustained vowels represent relatively time-invariant vocal phonation whereas continuous speech involves quick and continuous alterations of glottal and supraglottal mechanisms; (b) in contrast to continuous speech, sustained mid-vowel segments do not contain non-voiced phonemes, rapid voice onsets and offsets or prosodic fundamental frequency and amplitude fluctuations; (c) sustained vowels are not affected by speech rate, vocal pauses, phonetic context and stress. However, sustained vowels may lack representation of daily speech and voice (Parsa & Jamieson, 2001; Eadie & Baylor, 2006) and continuous speech potentially contains perceptual cues which are often considered to be decisive in vocal quality evaluations (Askenfelt & Hammarberg, 1986). Since both sample types offer valuable information in voice quality measurements, the present meta-analysis focused on studies of sustained vowel as well as connected speech.

Search strategy

Relevant scientific reports were identified by a systematic electronic search of the Medline database and the corpus of online publications by the American Speech-Language-Hearing Association. The combination of (a) keywords referring to composite perceptual voice evaluations and (b) keywords related to the concepts of prediction by means of acoustic measures was used as a guide. Using information derived from the titles and abstracts, an initial set of pertinent articles was generated. Subsequently, a manual search for references in relevant literature sources was launched using the same guide. This manual search started from the sources cited in the initial set of articles garnered from the electronic search and from periodicals, book chapters, and various bibliographies likely to contain relevant references and texts.

Inclusion and exclusion of literature sources

In order to be included, a study had to report sufficient mathematical detail on bivariate correlation coefficients establishing the relation between perceptual

overall voice quality ratings of sustained vowels or continuous speech (the criterion variable) and one or more acoustic parameters derived from the same samples (the predictors). Studies citing relevant correlation coefficients were included, whether or not significance levels were reported, and every study describing auditory-perceptual ratings of overall quality (i.e., dysphonia severity) was included, regardless of the type of rating scale used.

Investigations of acoustic correlates of specific perceptual dimensions such as breathiness and roughness were not included in the meta-analysis, as the present study concentrated on "composite" or "global" overall voice quality correlates. Furthermore, reports on non-acoustic or non-objective correlates, such as aerodynamic measures or electroglottographic parameters were also excluded, as well as studies dealing with the relationship between the auditory-perceptual evaluation of overall voice quality and its visual-perceptual representation in narrowband spectrograms. Furthermore, since the present study aimed to focus on acoustic-auditory determinants of dysphonia severity, studies investigating the correlation between objective acoustic measures and visual inspection of spectrograms or other diagrams were excluded. Also, reports on parameters derived from synthesized vowel samples, were not included in this study. Reports lacking sufficient quantitative and critical information, such as number of subjects or type of samples, were also excluded.

Methodological articles related exclusively to the use and development of perceptual rating scales or acoustic algorithms, which did not provide inferential statistics on the validity of the acoustic measure(s), were also excluded. Studies appraising the diagnostic value of acoustic parameters (i.e. the power of a diagnostic tool to discriminate between presence or absence of a voice disorder), expressed as sensitivity, specificity, positive predictive value, negative predictive and/or area under the ROC curve, or outcomes of studies based on comparative statistics between normal and pathologic voices, as expressed in chi-square tests, Mann-Whitney U tests, t tests, etc., were not included, because the present study concentrated on the correlation coefficient as population effect size.

In addition, reports on multivariate analyses were excluded, unless bivariate (zero-order) correlation coefficients were clearly identified (as in Wolfe & Martin, 1997; Wolfe et al., 1997; Yu et al., 2001; Eadie & Baylor, 2006; Ma & Yiu, 2006). The reason for excluding multiple regression studies is based upon the assumption that some predictor variables are dropped from the initial set of possible predictors as a result of co-linearity. A relevant independent variable, correlating well with the criterion variable, may be dropped when, in the presence of other predictors, it does not substantially increase the amount of variance explained. This phenomenon makes it difficult to assess the separate contribution of each independent variable to the prediction of the criterion. Moreover, the algorithm of a multiple regression not only looks for a parsimonious equation, it also gives each remaining predictor a coefficient that can only to be interpreted in combination with the particular set of remaining predictors in the rest of the

equation. As a consequence, meta-analysts have yet to develop effect size statistics for multivariate statistical analyses (Lipsey & Wilson, 2001).

Finally, reliability of the auditory-perceptual ratings of voice quality, as an index upon which an acoustic measure is validated, is also an important consideration (Kreiman & Gerratt, 2000a, 2000b; Kreiman et al., 2007). Reliability of auditory-perceptual ratings is traditionally described in terms of within and between listener reliability, consistency, agreement, or concordance. Such intra- and inter-rater reliability is considered an important prerequisite for validity. High reliability clearly and precisely defines the perceptual construct to be measured by an acoustic parameter. In contrast, listener unreliability increases "non-experimental" or "error" variance, thereby reducing the true variance in the perceptual construct that is to be accounted for by the acoustic measure. Thus, the increase in error variance due to listener unreliability should decrease the predictive validity of the acoustic measure, as evaluated by a correlation coefficient. In single experiments, acceptable rater reliability is often considered an essential prerequisite before attempting to assess an acoustic measure's worth in estimating overall voice quality. However, across studies many statistics have been used to measure rater reliability, for instance Pearson's product-moment correlation coefficient, Cohen's kappa correlation coefficient, Cronbach's alpha correlation coefficient, and intraclass correlation coefficient, to mention only a few. Given the large number of studies reviewed in this meta-analysis, each using a variety of listeners (with differing levels of experience and training), and different scales with various interpretation guidelines so as to determine "adequate" reliability, we elected to treat listener reliability as a nuisance variable, and to not exclude any studies solely on the basis of their estimates of listener reliability. This decision is predicated upon the assumption that listener unreliability essentially contributes to error variance, and necessarily attenuates any investigator's ability to identify significant correlations between listener ratings and specific acoustic measures. By treating listener reliability/unreliability as a nuisance variable, one that would necessarily vary between studies and differentially contribute to error variance, we assumed that across studies, the most compelling acoustic-perceptual relationships would eventually surface, having survived the potentially attenuating effects of listener unreliability. Analogous to the listener reliability/unreliability, we also elected to treat between-study differences in data acquisition and processing methodology as a nuisance variable. Variety in room acoustics, microphone type and placement, software, analysis algortitms etc., also creates "error" variance, and similarly decreases the variance in the perceptual construct that is to be explained by the acoustic measure. The large number of studies, each with its own acoustical configuration and hardware and software settings, clearly limits our ability to directly compare the outcomes of the studies. However, a guiding principle of meta-analysis is that the consistency of the significant results/conclusions across studies is paramount, and robust relationships should withstand such methodological "noise" (regardless of the source of the noise i.e, listener unreliability, recording instrumentation and surroundings, computer software, etc).

We therefore elected to consider methodological variations in recording conditions/settings, data acquisition and analysis algorithms etc., as additional sources of "error" variance, and an inherent limitation of the meta-analysis.

Originally, eighty-five reports were considered. Based upon the aforementioned inclusion and exclusion criteria however, many reports were excluded, producing a final corpus of twenty-five studies upon which the meta-analysis was performed. Twenty-one studies involved measurements on sustained vowel samples (methodological aspects of these studies are summarized in Table 3.1). Seven studies involved measurement on continuous speech samples (methodological aspects of these studies are similarly summarized in Table 3.2). However, three studies contained information on both continuous speech and sustained vowels (Heman-Ackah et al., 2002; Halberstam, 2004; Eadie & Baylor, 2006), thus leaving a total of twenty-five studies (i.e., 28-3=25).

From these studies, a list of acoustic measures was generated. Subsequently, the measures were organized based on their description in the method section of the original publication. The tabulation of Buder (2000) was chosen as a loose framework to group the measures. Buder's tabulation was the first compilation of acoustic voice measures, as it presented a complete overview of acoustic measures in a comprehensive and consistently structured manner. It therefore served as a basis upon which the measures of this meta-analysis were considered to be similar or different. In studies which analyzed sustained vowels, there were sixty-nine acoustic predictors identified, whereas in the connected speech studies, twenty-six acoustic measures were reported. Eighty-seven acoustic markers have been identified as measures of overall voice quality in the included studies. Table 3.3 lists these acoustic measures alphabetically and provides (for every measure) references expanding upon the rationale and the digital signal processing underlying the measure.

Methodological aspects of the included studies

Relevant methodological features of the 25 included studies are displayed in Tables 3.1 and 3.2. The number of normal and pathologic/dysphonic participants were summed and reported in the meta-analysis as *n*, the number of subjects. Second, the type of sample (sustained vowel: type and duration, continuous speech: type) was listed for each study. Third, four aspects of the perceptual evaluation of the voice recordings were itemized, including: (1) the number of judges or listeners, (2) the type of rating scale, (3) the name given to the perceptual construct of overall voice quality, and (4) the estimates of intra- and inter-rater reliability.

Table 3.1 Methodological items (number of subjects, type of voice recording, organization and reliability of the perceptual ratings, and the acoustic measures for which a correlation coefficient was available) of the 21 studies included in this meta-analysis on sustained vowels.

Source	Subjects[a]			Voice Sample[b]		Perceptual evaluation[b]				
	N	P	T	Vowel	Dur.	Number of judges	Rating scale[c]	Perceptual construct	Intrarater reliability[d]	Interrater reliability[d]
Kojima et al. (1980)	28	30	58	/a/	NA	5	EAI (4)	Hoarseness	NA	NA
Yumoto et al. (1984)	0	87	87	/a/	3 s	8	EAI (4)	Hoarseness	NA	0.51 – 0.79 Sp
Hirano et al. (1986)	0	68	68	/e/	NA	NA	EAI (4)	G, grade	NA	NA
Prosek et al. (1987)	0	90	90	/a/	2 s	9	EAI (7)	Severity of voice disorder	0.90 Pe	0.82 Cr
	16	44	60			14		Hoarseness	NA	NA
Wolfe & Steinfatt (1987)	0	51	51	/a/, /i/	1 s	8	EAI (7)	Severity of dysphonia	89% Ag	0.95 Cr
								Hoarseness		
Feijoo & Hernández (1990)	64	57	121	/e/	NA	4	EAI (4)	G, grade	77.48% Ag	98.35% Ag
Kreiman et al. (1990)	0	18	18	/a/	1.67 s	10	EAI (7)	Overall abnormality	NA	NA
Wolfe et al. (1995)	20	60	80	/a/	1 s	22	EAI (7)	Overall severity	0.99 Cr	0.98 Cr
Dejonckere et al. (1996)	0	943	943	/a/	2 s	2	EAI (4)	G, grade	0.51 Co	0.87 Sp
Dejonckere & Wieneke (1996)	0	28	28	/a/	0.1 s	2	EAI (5)	Overall severity of hoarseness	NA	NA
De Bodt (1997)	98	634	732	/a/	3 s	1	EAI (4)	G, grade	NA	NA
Plant et al. (1997)	0	26	26	/i/	2 s	3	EAI (5)	Overall voice	0.86	NA

Wolfe et al. (2000)	0	20	20	/a/	1 s	11	EAI (7)	Abnormality	0.81 Pe	0.98 Cr
Yu et al. (2001)	21	63	84	/a/	2 s	6	EAI (4)	G, grade	Cons	Cons
Heman-Ackah et al. (2002)	0	14	14	/a/	1 s	2	EAI (4)	G, grade	NA	0.83 Pe
Halberstam (2004)	0	60	60	/a/	1 s	2	EAI (7)	Hoarseness	0.89 NA	0.91 Cr
Eadie & Baylor (2006)	3	9	12	/a/	1 s	16	VAS	Overall severity	0.82 – 0.95 Pe	0.72 – 0.83 Pe
Gorham-Rowan & Laures-Gore (2006)	0	28^{ym}	28^{ym}	/a/	1 s	10	FMMEP	Hoarseness	-0.32 – 0.86 Pe	0.80 Cr
	0	28^{ew}	28^{ew}							
	0	28^{em}	28^{em}							
Yu et al. (2007)	38^{w} 20^{m}	270^{w} 121^{m}	308^{w} 141^{m}	/a/	2 s	4	VAS	G, grade	NA	NA

[a] *N = number of normal subjects, P = number of pathological or dysphonic subjects, T = total number of subjects, [ym] = young men, [ew] = elderly women, [em] = elderly men, [m] =men, [w] = women.*

[b] *NA = the information is not available in the original manuscript.*

[c] *EAI = equal-appearing interval scale with between brackets the number of points on the scale, VAS = visual analog scale, FMMEP = free modulus magnitude estimation paradigm.*

[d] *Sp = Spearman's rank-order correlation coefficient, Pe = Pearson's product-moment correlation coefficient, Co = Cohen's kappa correlation coefficient, Cr = Cronbach's alpha correlation coefficient, Ke = Kendall's coefficient of concordance, Ag = percentage of agreement/consistency between judgments, Cons = consensus between listeners without quantitative measure of reliability.*

Table 3.2 Methodological items (number of subjects, type of voice recording, organization and reliability of the perceptual ratings, and the acoustic measures for which a correlation coefficient was available) of the 21 studies included in this meta-analysis on continuous speech.

Source	Subjects[a]			Voice Sample	Perceptual evaluation[b]				
	N	P	T		Number of judges	Rating scale[c]	Perceptual construct	Intrarater reliability[d]	Interrater reliability[d]
Askenfelt & Hammarberg (1986)	0	41	41	Voiced segments, 40 s of reading a story	6	EAI (6)	Overall voice quality	0.86 – 0.98 Pe	NA
Qi et al. (1999)	0	87	87	1st and 2nd sentence from Rainbow passage	5	VAS	Overall voice quality	0.93 – 0.96 Pe	0.97 Cr
Heman-Ackah et al. (2002)	0	18	18	2nd sentence from Rainbow passage	2	EAI (4)	G, grade	NA	0.83 Pe
Halberstam (2004)	0	60	60	12 s from Rainbow passage	2	EAI (7)	Hoarseness	0.93 NA	0.97 Cr
Eadie & Doyle (2005)	6	24	30	2nd sentence from Rainbow passage	12	DME	overall severity	0.69 Pe	0.97 Cr
Eadie & Baylor (2006)	3	9	12	2nd sentence from Rainbow passage	16	VAS	overall severity	0.80 – 0.97 Pe	0.84 – 0.91 Pe
Ma & Yiu (2006)	41	112	153	/ba ba da bo/	4	EAI (11)	G, grade	≥ 0.90 Pe	0.86 – 0.91 Pe

[a] *N = number of normal subjects, P = number of pathological or dysphonic subjects, T = total number of subjects.*
[b] *NA = the information is not available in the original manuscript.*
[c] *EAI = equal-appearing interval scale with between brackets the number of points on the scale, VAS = visual analog scale, DME = direct magnitude estimation.*
[d] *Pe = Pearson's product-moment correlation coefficient, Cr = Cronbach's alpha correlation coefficient.*

Table 3.3 The eighty-seven acoustic measures in this study, with their full name and the sources referring to correlational studies in the present appraisal.

Acoustic measure	Sources included in study
Absolute jitter	Kreiman et al. (1990), De Bodt (1997), Wolfe et al. (1997), Halberstam (2004)
Amplitude perturbation quotient	De Bodt (1997), Wolfe et al. (1997), Heman-Ackah et al. (2002), Halberstam (2004), Gorham-Rowan & Laures-Gore (2006)
Amplitude perturbation quotient of residue signal	Prosek et al. (1987)
Area of voice range profile	Ma & Yiu (2006)
Breathiness index	Plant et al. (1997), Wolfe et al. (2000)
Cepstral peak magnitude (a.k.a. Magnitude of first rahmonic)	Dejonckere & Wieneke (1996)
Cepstral peak prominence	Wolfe & Martin (1997), Wolfe et al. (2000), Halberstam (2004), Eadie & Baylor (2006)
Cepstrum of excitation signal	Feijoo & Hernández (1990)
Coefficient of excess	Prosek et al. (1987)
Coefficient of variation of fundamental frequency	Wolfe et al. (1997)
Coefficient of variation of jitter	Kreiman et al. (1990)
Coefficient of variation of period	Wolfe & Steinfatt (1987)
Coefficient of variation of shimmer	Kreiman et al. (1990)
Compression of relative frequency differences	Askenfelt & Hammarberg (1986)
Cycle-of-cycle variation of waveform	Feijoo & Hernández (1990)
Difference between frequencies of second and first formant	Kreiman et al. (1990)
Directional perturbation factor	Askenfelt & Hammarberg (1986)
Fluctuation in amplitude	Hirano et al. (1986)
Fluctuation in fundamental frequency	Hirano et al. (1986)
Frequency-domain harmonics-to-noise ratio	Eadie & Doyle (2005)
Frequency of first formant	Kreiman et al. (1990)
Frequency of second formant	Kreiman et al. (1990)
Frequency of third formant	Kreiman et al. (1990)
Fundamental frequency	Yu et al. (2001), Yu et al. (2007), Ma & Yiu (2006)
Fundamental frequency range in voice range profile	Ma & Yiu (2006)

Jitter factor	Yu et al. (2001), Yu et al. (2007)
Jitter from Yumoto	Yumoto et al. (1984)
Jitter ratio	Wolfe & Steinfatt (1987), Dejonckere & Wieneke (1996)
Lowest fundamental frequency in voice range profile	Ma & Yiu (2006)
Lyapunov coefficient	Yu et al. (2001), Yu et al. (2007)
Maximum intensity in voice range profile	Ma & Yiu (2006)
Mean harmonic emergence between 500 and 1500 hz	Dejonckere & Wieneke (1996)
Minimum intensity in voice range profile	Ma & Yiu (2006)
Natural logarithm of standard deviation of period	Wolfe & Steinfatt (1987), Kreiman et al. (1990)
Noise-to-harmonics ratio	Wolfe et al. (1997)
Noise-to-harmonics ratio from mdvp	Dejonckere et al. (1996), De Bodt (1997), Heman-Ackah et al. (2002), Halberstam (2004), Gorham-Rowan & Laures-Gore (2006), Ma & Yiu (2006)
Normalized mean absolute period jitter	Feijoo & Hernández (1990)
Normalized mean absolute period shimmer	Feijoo & Hernández (1990)
Normalized noise energy	Feijoo & Hernández (1990)
Number of harmonics	Kreiman et al. (1990)
Partial period comparison	Kreiman et al. (1990)
Peakedness of relative frequency differences	Askenfelt & Hammarberg (1986)
Pearson r at autocorrelation peak	Wolfe et al. (2000)
Percent jitter	Kreiman et al. (1990), De Bodt (1997), Plant et al. (1997), Wolfe & Martin (1997), Wolfe et al. (1997), Halberstam (2004)
Percent shimmer	Kreiman et al. (1990), Dejonckere et al. (1996), De Bodt (1997), Wolfe & Martin (1997), Wolfe et al. (1997), Halberstam (2004), Ma & Yiu (2006)
Perturbation factor	Askenfelt & Hammarberg (1986)
Perturbation magnitude	Askenfelt & Hammarberg (1986)
Perturbation magnitude mean	Askenfelt & Hammarberg (1986)
Phonatory fundamental frequency range	De Bodt (1997), Yu et al. (2001), Halberstam (2004), Yu et al. (2007)
Pitch amplitude	Prosek et al. (1987), Plant et al. (1997), Eadie & Doyle (2005)
Pitch perturbation quotient	De Bodt (1997), Wolfe et al. (1997), Halberstam (2004)
Pitch perturbation quotient of residue signal	Prosek et al. (1987)
Power spectrum ratio	Wolfe et al. (2000)

Relative noise level	Hirano et al. (1986)
Residue signal power ratio	Plant et al. (1997)
Richness of high frequency harmonics	Hirano et al. (1986)
Shimmer in db	Wolfe et al. (1995), De Bodt (1997), Wolfe et al. (1997), Halberstam (2004)
Signal-to-noise ratio	Yu et al. (2001), Yu et al. (2007)
Signal-to-noise above 1000 hz	Yu et al. (2001), Yu et al. (2007)
Signal-to-noise ratio from Milenkovic	Wolfe & Martin (1997)
Signal-to-noise ratio from Qi	Qi et al. (1999), Eadie & Doyle (2005)
Smoothed amplitude perturbation quotient	De Bodt (1997), Wolfe et al. (1997), Halberstam (2004)
Smoothed cepstral peak prominence	Heman-Ackah et al. (2002), Halberstam (2004), Eadie & Baylor (2006)
Smoothed pitch perturbation quotient	De Bodt (1997), Wolfe et al. (1997), Heman-Ackah et al. (2002), Halberstam (2004)
Soft phonation index	De Bodt (1997)
Spectral distortion	Feijoo & Hernández (1990)
Spectral flatness of inverse filter	Prosek et al. (1987)
Spectral flatness of residue signal	Prosek et al. (1987), Eadie & Doyle (2005)
Spectral noise level above and under 6000 hz	Dejonckere & Wieneke (1996)
Spectral tilt	Eadie & Doyle (2005)
Spectral tilt of voiced segments	Eadie & Doyle (2005)
Standard deviation of cepstral peak prominence	Wolfe & Martin (1997)
Standard deviation of fundamental frequency	Wolfe et al. (1997)
Standard deviation of jitter	Kreiman et al. (1990), Wolfe & Martin (1997)
Standard deviation of partial period comparison	Kreiman et al. (1990)
Standard deviation of period	Wolfe et al. (2000)
Standard deviation of relative frequency differences	Askenfelt & Hammarberg (1986)
Standard deviation of shimmer	Kreiman et al. (1990), Wolfe & Martin (1997)
Standard deviation of signal-to-noise ratio from Milenkovic	Wolfe & Martin (1997)
Voice turbulence index	Halberstam (2004)

Statistics

Quantitative data from the selected scientific reports were analysed using statistical software packages for personal computers, including Microsoft Office Excel 2003 and Meta-Analysis Programs version 5.3 (Ralf Schwarzer, Department of Psychology, Freie Universität Berlin, Germany). Meta-analyses on correlation coefficients according to the Schmidt-Hunter method (Hunter et al., 1982) were performed on all predictive acoustic voice quality correlates for which more than one effect size was available. This method is based on four statistics. The first statistic is the *number of effect sizes (k)* or the number of available bivariate correlation coefficients for a given predictor. The second statistic is the total *number of subjects (N)*. The third statistic is the *population effect size* or the weighted mean correlation coefficient ($\overline{r_w}$). Correlation coefficients based on studies with large sample sizes digress less from the population effect size and therefore more weight is assigned to large N effect sizes (Hunter et al., 1982). If only one effect size is reported, a weighted effect size can not be calculated. In this case, there is no meta-analysis and the discussion is based on the initial and solitary r-value. While there is no firm criterion or universal consensus for evaluating the magnitude of correlation coefficients (Frey et al., 1991), we chose a correlation coefficient (r or $\overline{r_w}$) of 0.60 as the cutoff to distinguish between strong and weak acoustic predictors. Following the guidelines established by Franzblau (1958), this threshold intends to separate a "moderate" degree of correlation from a "marked" degree of correlation. It should be acknowledged however that other interpretations have been proposed, including Frey et al. (1991) for example, who recommended r of 0.70 to distinguish between moderate and marked correlations. We selected a less stringent correlation coefficient r=.60 in light of our decision to treat listener unreliability and methodological/procedural differences as sources of error variance (i.e, nuisance variables) which would potentially attenuate the strength of reported bivariate correlations across studies. The fourth statistic relates to the homogeneity or heterogeneity of the effect sizes. A population effect size can only be interpreted reliably if the underlying data set is sufficiently homogeneous (Hunter et al., 1982). Here one can rely on several indicators: (1) the residual standard deviation, (2) the percentage of observed variance accounted for by the sampling error and (3) the chi-square value. However, the prefered index for homogeneity is the population variance or its square root, called *residual standard deviation (SD_{res})*. This indicator, SD_{res}, is the variance left after the sampling error has been subtracted and thus it is the actual amount of variance (Hunter et al., 1982). Ideally, SD_{res} equals zero, meaning that all the observed variance is accounted for by sampling error and that the dataset of correlations is completely homogeneous. If the analysis however failed to identify a source of systematic variation in the data, SD_{res} is indicative of heterogeneity. As a rule of thumb, a set of effect sizes can be considered homogeneous when SD_{res} is less than ¼ of rw (Hunter et al., 1982; Lipsey & Wilson, 2001).

RESULTS

Sustained vowels

Twenty-one studies meeting the selection criteria were identified, the majority originated from the *Journal of Speech (Language) and Hearing Disorders* (6), *Journal of Voice* (3) and *Journal of Communication Disorders* (3). Other sources were *Acta Otorhinolaryngologica Belgica* (1), *Acta Otolaryngologica* (1), *Folia Phoniatrica et Logopaedica* (2), *Journal of Phonetics* (1), *Laryngoscope* (1), *ORL* (1), *Revue de Laryngologie-Otologie-Rhinologie* (1), and a chapter in volume VI of *Advances in Clinical Phonetics* (1). Relevant information concerning the methodology of the reports that were included in the meta-analysis can be found in Table 3.1. All 21 studies reported on pathologic or dysphonic voices, however only 8 studies also contained normal voices. The mean number of dysphonic voice samples was 115 (range 9 to 943). For the normal voices, the mean number was 34 (range 3 to 98). The total number of subjects was 116 on average and ranged from 12 to 943. In these studies, 146 distinct effect sizes (i.e. correlation coefficients) were reported, pertaining to 69 different acoustic predictors as displayed in Table 3.4.

All acoustic parameters and data on sustained vowels were extracted from the central portion of the recordings. The length of the mid-vowel segment varied from 0.1 to 3 seconds with a mean duration of 1.5 seconds. One second was the modal duration, occurring in 50% of the studies (the duration was not specified in 3 studies). The vowels [a:], [i:] and [e:] were analyzed in 86%, 19% and 10% of the studies, respectively. Substantial differences existed among the data acquisition systems that were used, which could potentially influence the outcome of acoustic measurements. For instance, recording equipment i.e. type of microphone and microphone localization relative to the sound source, type of hardware, processing algorithms, measurement algorithms, software settings such as sampling rate or method of fundamental period extraction varied among the studies and have been demonstrated to influence the outcome of acoustic measurements, particularly the outcomes of perturbation measures.

For the perceptual experiments, the number of judges ranged from 1 to 22, with a mean value of 8. The rating scale used was typically an equal-appearing interval scale, using 4, 5 or 7 points in 38%, 10% and 33% of studies, respectively. In two studies (Wolfe et al., 1997; Yu et al., 2007), a visual analog scale was used. In Gorham-Rowan & Laures-Gore (2006) a free modulus magnitude estimation paradigm was used. A variety of perceptual labels were used including: hoarseness, G (from grade), severity of voice disorder, severity of dysphonia, overall abnormality, overall severity, overall severity of hoarseness, abnormality and overall voice quality. A variety of estimates of inter- and intra-judge reliability estimates were used including: Spearman's rank-order correlation coefficient, Pearson's product-moment correlation coefficient, Cohen's kappa correlation coefficient, Cronbach's alpha correlation coefficient, Kendall's coefficient of

concordance and the percentage of agreement or consistency between judgments. As mentioned previously, the variety and range of methods to determine reliability hampers comparisons between studies. In general, intrajudge reliability fluctuated from rather low, as in Dejonckere & Wieneke (1996) and Gorham-Rowan & Laures-Gore (2006), to very high, as in Wolfe et al. (1995) and Prosek et al. (1987). Similar variability was observed for interjudge reliability.

Meta-analysis on correlation coefficients

The results of the meta-analysis on sustained vowels are summarized in Table 3.4 and Figure 3.1. For thirty-three of the sixty-nine acoustic predictors (48%), there was only 1 correlation coefficient available and consequently, no weighted mean correlation coefficient could be determined. For the remaining 36 acoustic determinants (52%), there was more than 1 correlation coefficient and the k-values ranged from 2 to 7. The most frequently investigated parameters were nhr from mdvp (k = 7), and the vocal perturbation measures amplitude perturbation quotient, percent jitter and percent shimmer (k = 6). For these thirty-six predictors, a $\overline{r_w}$ was calculated with Meta-Analysis Programs version 5.3. The organization of the meta-analysis on acoustic measures on sustained vowels is illustrated in Figure 3.1.

In the first subset there were fifty-two of the sixty-nine acoustic measures on sustained vowels with a (weighted) correlation coefficient below 0.60. Weighted correlation coefficients ranged from 0.11 for coefficient of excess and voice turbulence index to 0.56 for amplitude perturbation quotient of residuals and harmonic-to-noise ratio from Yumoto. In this subset, there were thirty-two markers with a k-value of 2 or more. The SD_{res} statistics indicated heterogeneity for eight measures. For the remaining twenty-four acoustic correlates with $\overline{r_w} < 0.60$ and $k \geq 2$, SD_{res} statistics showed homogeneity. The second subset consisted of seventeen acoustic measures with a (weighted) effect size equal to or above 0.60. In this subset of seventeen measures, there were four markers with a k-value of 2 or more. Weighted correlation coefficients ranged from 0.62 for smoothed cepstral peak prominence to 0.75 for pitch amplitude. Statistical homogeneity testing (SD_{res}) indicated that these four $\overline{r_w}$-values were based on a set of homogeneous effect sizes, indicating that these effect sizes are consistently equal to or above 0.60 (smoothed cepstral peak prominence: $\overline{r_w} = 0.62$, spectral flatness of residue signal: $\overline{r_w} = 0.69$, Pearson r at autocorrelation peak: $\overline{r_w} = 0.74$, pitch amplitude: $\overline{r_w} = 0.75$).

Table 3.4 Summary of the meta-analytic findings for the individual acoustic measures of overall voice quality in sustained vowels. The acoustic measures are ordered according to their effect size (r or $\overline{r_w}$).

Acoustic measure	k[a]	r or $\overline{r_w}$ [b]	SD_{res}[c]	Acoustic measure	k[a]	r or $\overline{r_w}$ [b]	SD_{res}[c]
Fluctuation in fundamental frequency	1	0.00	/	*Pitch perturbation quotient of residue signal*	2	0.47	Ho
Soft phonation index	1	0.01	/	*Coefficient of variation of percent jitter*	1	0.48	/
Standard deviation of signal-to-noise ratio from Milenkovic	1	0.06	/	*Coefficient of variation of fundamental frequency*	2	0.49	Ho
Frequency of second formant	1	0.07	/	*Percent jitter*	6	0.49	Ho
Standard deviation of cepstral peak prominence	1	0.11	/	*Cepstral peak prominence*	4	0.50	Ho
Voice turbulence index	2	0.11	He	*Natural logarithm of standard deviation of period*	2	0.51	He
Coefficient of excess	2	0.11	Ho	*Percent shimmer*	6	0.52	He
Frequency of third formant	1	0.14	/	*Spectral noise level above and under 6000 hz*	1	0.52	/
Ratio of amplitudes of first and second harmonic	1	0.15	/	*Relative average perturbation*	5	0.52	Ho
Number of harmonics	1	0.19	/	*Pitch perturbation quotient*	3	0.52	Ho
Fluctuation in amplitude	1	0.19	/	*Smoothed pitch perturbation quotient*	4	0.53	Ho
Richness of high frequency harmonics	1	0.19	/	*Jitter ratio*	2	0.53	Ho
Frequency of first formant	1	0.21	/	*Lyapunov coefficient*	3	0.54	He
Standard deviation of percent jitter	2	0.22	Ho	*Phonatory fundamental frequency range*	5	0.54	Ho
Breathiness index	3	0.22	Ho	*Harmonics-to-noise ratio from Yumoto*	3	0.56	He
Spectral flatness of inverse filter	2	0.25	Ho	*Amplitude perturbation quotient of residue signal*	2	0.56	Ho
Fundamental frequency	3	0.28	Ho	*Mean harmonic emergence between 500 and 1500 hz*	1	0.58	/
Coefficient of variation of percent shimmer	1	0.28	/	*Coefficient of variation of period*	1	0.62	/
Ratio of frequencies of second and first formant	1	0.32	/	*Smoothed cepstral peak prominence*	3	0.63	Ho
Difference between frequencies of second and first formant	1	0.33	/	*Standard deviation of partial period comparison*	1	0.67	/

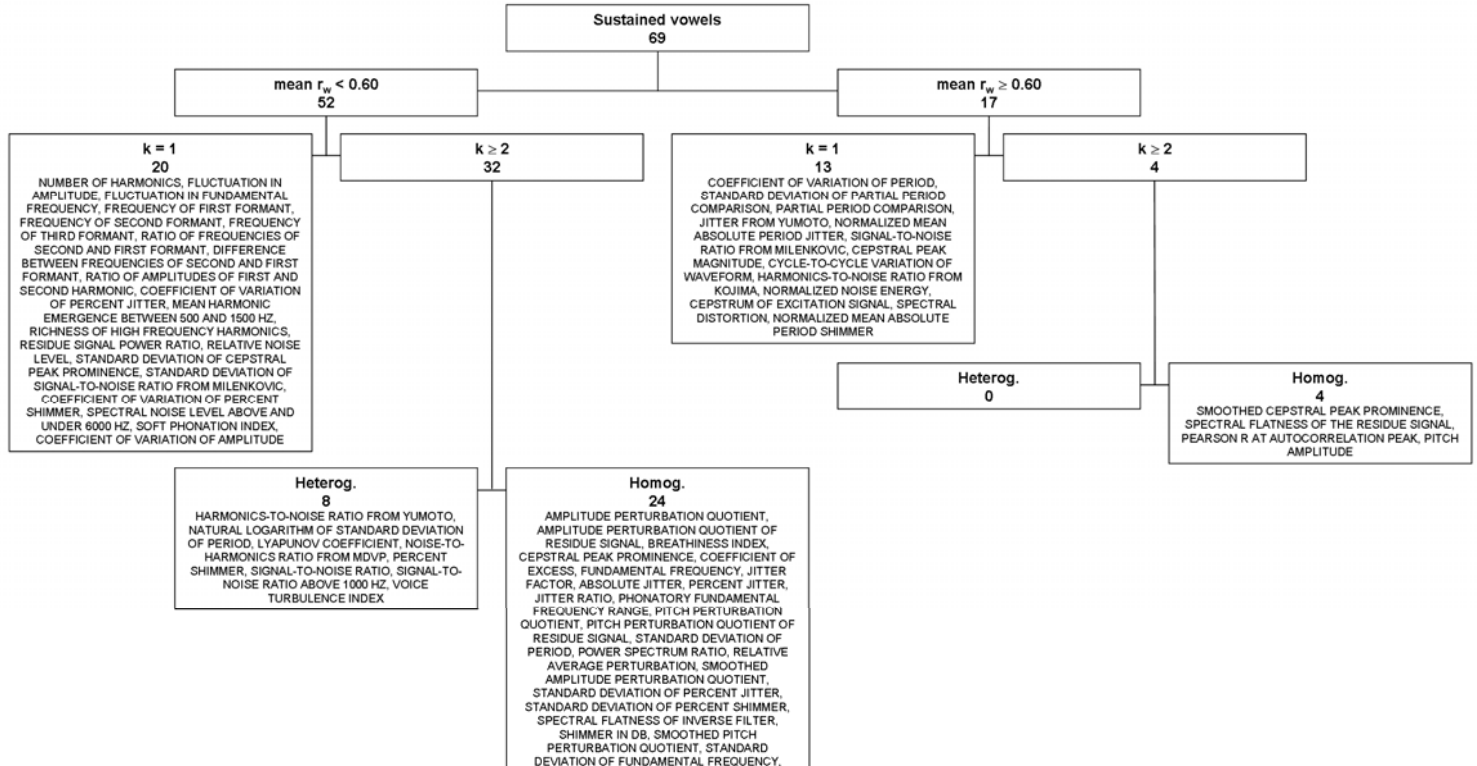
Residue signal power ratio	1	0.40	/	*Normalized mean absolute period jitter*	1	0.75	/
Relative noise level	1	0.40	/	*Pitch amplitude*	3	0.75	Ho
Standard deviation of percent shimmer	2	0.41	Ho	*Signal-to-noise ratio from Milenkovic*	1	0.76	/
Signal-to-noise ratio above 1000 hz	3	0.42	He	*Cepstral peak magnitude*	1	0.80	/
Jitter factor	3	0.42	Ho	*Cycle-to-cycle variation of waveform*	1	0.83	/
Power spectrum ratio	2	0.44	Ho	*Harmonics-to-noise ratio from Kojima*	1	0.87	/
Amplitude perturbation quotient	6	0.45	Ho	*Normalized noise energy*	1	0.88	/
Noise-to-harmonics ratio from mdvp	7	0.45	He	*Cepstrum of excitation signal*	1	0.90	/
Shimmer in db	4	0.45	Ho	*Spectral distortion*	1	0.93	/
Absolute jitter	4	0.47	Ho	*Normalized mean absolute period shimmer*	1	0.93	/
Standard deviation of fundamental frequency	2	0.47	Ho				

[a] *k = number of effect sizes available in the included literature.*

[b] *r = correlation coefficient (when k=1),* $\overline{r_w}$ *= mean weighted correlation coefficient (when k>1).*

[c] SD_{res} *= residual standard deviation, / = not applicable (when k=1), Ho/He = homogeneous/heterogeneous r or* $\overline{r_w}$ *(when k>1).*

Figure 3.1 Diagram illustrating the organisation of the meta-analysis for acoustic measures on sustained vowels. The second line in every box contains the number of acoustic measures.



Continuous speech

Seven studies using continuous speech samples met the inclusion criteria of this meta-analysis. These studies were published in *Journal of Voice* (4), *Journal of Speech (Language) and Hearing Research* (1), *Journal of the Acoustical Society of America* (1) and *ORL* (1). As shown in Table 3.5, there were 29 separate effect sizes pertaining to 26 distinct acoustic measures. Relevant information regarding the methodology of these seven reports is found in Table 3.2. Whereas all seven studies used pathologic or dysphonic voice samples, only three studies also investigated normal voices. The mean number of dysphonic voice samples was 50, (range 9 to 112). For the normal voices, the mean number was 7 (range 3 to 41). The mean number of subjects was 57 (range 12 to 153). All acoustic measures were extracted from recordings of continuous speech, most often from speakers reading from a text. With the exception of Askenfelt & Hammarberg (1986) and Ma & Yiu (2006), the so-called 'Rainbow passage' was read aloud and a portion (typically the second sentence) was extracted for further analysis.

As for auditory-perceptual evaluation, the mean number of judges employed across studies was 7, (range 2 to 16). In four studies the rating scale was an equal-appearing interval scale with 4, 6, 7 or 11 points. In two studies (Qi et al., 1999; Eadie & Baylor, 2006), a visual analog scale was used. In another study (Eadie & Doyle, 2005), direct magnitude estimation was used. The following labels were used to designate the perceptual construct that was to be evaluated: *hoarseness*, *G* (for grade), *overall severity,* and *overall voice quality*. Estimates of reliability of listener judgments included two types of statistics: Pearson's product-moment correlation coefficient and Cronbach's coefficient alpha. To evaluate intrajudge reliability, Pearson's r-values were uniformly reported. Where a range of r values was given (Askenfelt & Hammarberg, 1986; Qi et al., 1999; Eadie & Baylor, 2006), the lowest r-value was chosen to calculate a weighted average of intrajudge correlation across reports (Table 3.2). The intrajudge $\overline{r_w}$ was 0.81 which is indicative of homogeneous intrajudge reliability. It appears that listeners were generally consistent in their perceptual evaluations of continuous speech. Regarding interjudge reliability, only three studies provided a Pearson's r-value. Meta-analysis, again using the lowest r-value of the reported range, resulted in an interjudge $\overline{r_w}$ of 0.84, i.e. homogeneous interjudge reliability. This was corroborated by the three studies that used Cronbach's α, since they all mentioned an α-value of 0.97. Concerning the acoustic measures, Table 3.5 provides an overview of the determinants that were used to predict overall voice quality. As was the case for sustained vowel studies, there were considerable differences between studies' recording equipment and settings.

Meta-analysis on correlation coefficients

The results of the meta-analysis on continuous speech data are summarized in Table 3.5 and Figure 3.2. For 23 of the 26 (88%) acoustics measures cited, there

was only 1 effect size available. For the remaining 3 acoustic determinants (cepstral peak prominence, smoothed cepstral peak prominence and signal-to-noise ratio from Qi) there were 2 effect sizes (k = 2). For these 3 predictors, a $\overline{r_w}$ was calculated with Meta-Analysis Programs version 5.3. The organization of the meta-analysis on acoustic measures on continuous speech is illustrated in Figure 3.2.

As in our meta-analysis on sustained vowel data, a correlation coefficient of 0.60 was chosen as the threshold to distinguish between marked and weak predictors. In the first subset of sixteen acoustic measures with a (weighted) effect size below 0.60, k was always equal to 1, and therefore no meta-analysis was performed. In the second subset consisting of ten acoustic measures with a (weighted) effect size equal to or above 0.60, there were 3 markers with k = 2: signal-to-noise ratio from Qi, cepstral peak prominence and smoothed cepstral peak prominence. Meta-analysis for these 3 measures yielded $\overline{r_w}$-values of 0.69, 0.88 and 0.88, respectively. Furthermore, SD_{res} indicated that these 3 $\overline{r_w}$-values were based on a set of homogeneous effect sizes.

DISCUSSION

The present meta-analysis assessed the relationship between acoustic measures and perceptual judgments of overall voice quality. In Buder (2000) alone, more than one hundred acoustic algorithms were cited and numerous microcomputer-based software systems offering various acoustic voice quality parameters had been developed. The fact that correlations between perceptual ratings and acoustic measures vary substantially (Kreiman & Gerratt, 2000a), raises questions regarding the validity and usefulness of these acoustic determinants. This meta-analysis represented an attempt to synthesize the corpus of algorithms and measures, and to establish a hierarchy of predictors on a statistical basis. In total, twenty-five study reports were included. Twenty-one studies reported on one-hundred and fifty correlation coefficients for sixty-nine acoustic measures on sustained vowels. Seven studies identified twenty-nine correlation coefficients for twenty-six acoustic measures on continuous speech.

In the context of the present meta-analysis, a homogeneous $\overline{r_w}$ exceeding 0.60 was judged to be a critical index. For instance, the amplitude perturbation quotient measure on sustained vowels was cited in five studies with 0.41, 0.54, 0.63, 0.50, 0.41, and 0.71 as coefficients of correlation. The single r=0.71 value in particular (Halberstam, 2004), seems to identify amplitude perturbation quotient as a valid acoustic marker of for overall voice quality of sustained vowels. However, the r-values from other studies are less persuasive, thus the meta-analysis resulted in a smaller homogeneous $\overline{r_w}$ of 0.45. In contrast to the amplitude perturbation quotient example wherein the meta-analysis resulted in a relatively weak $\overline{r_w}$ of 0.45, the meta-analysis outcome of studies related to smoothed cepstral peak

Table 3.5 Summary of the meta-analytic findings for the individual acoustic measures of overall voice quality in continuous speech. The acoustic measures are ordered according to their effect size (r or $\overline{r_w}$).

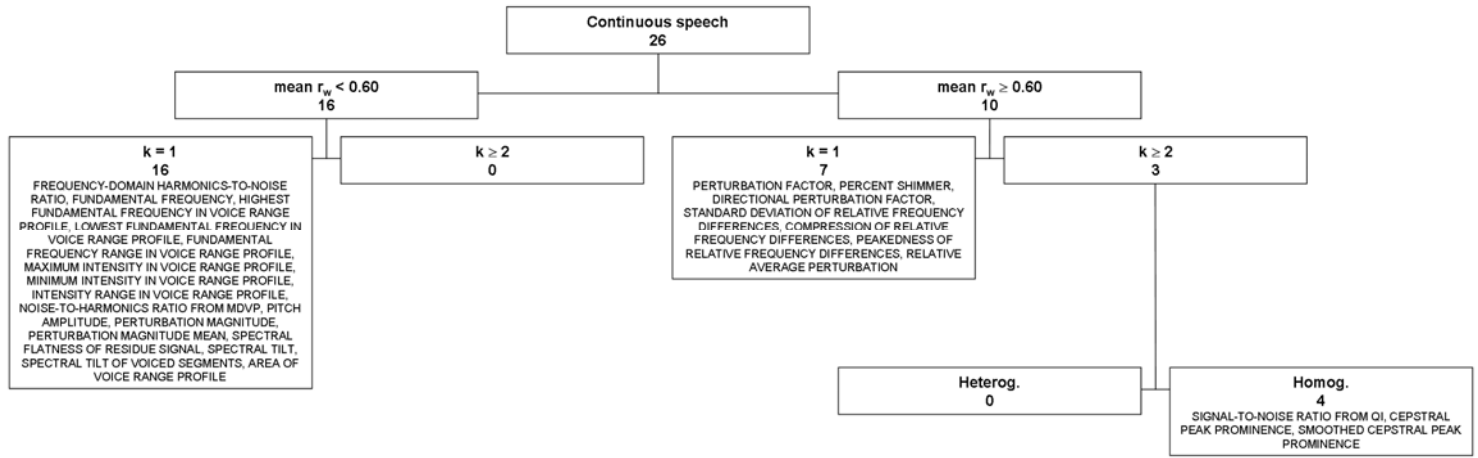
Acoustic measure	k[a]	r or $\overline{r_w}$ [b]	SD_{res}[c]	Acoustic measure	k[a]	r or $\overline{r_w}$ [b]	SD_{res}[c]
Perturbation magnitude	1	0.01	/	Spectral tilt	1	0.47	/
Maximum intensity in voice range profile	1	0.02	/	Pitch amplitude	1	0.58	/
Lowest fundamental frequency in voice range profile	1	0.09	/	Perturbation magnitude mean	1	0.59	/
Noise-to-harmonics ratio from mdvp	1	0.13	/	Perturbation factor	1	0.62	/
Fundamental frequency	1	0.18	/	Percent shimmer	1	0.62	/
Frequency-domain harmonics-to-noise ratio	1	0.26	/	Signal-to-noise ratio from Qi	2	0.69	He
Spectral flatness of residue signal	1	0.26	/	Directional perturbation factor	1	0.71	/
Spectral tilt of voiced segments	1	0.33	/	Standard deviation of relative frequency differences	1	0.71	/
Highest fundamental frequency in voice range profile	1	0.34	/	Compression of relative frequency differences	1	0.73	/
Intensity range in voice range profile	1	0.35	/	Peakedness of relative frequency differences	1	0.73	/
Fundamental frequency range in voice range profile	1	0.37	/	Relative average perturbation	1	0.75	/
Minimum intensity in voice range profile	1	0.38	/	Cepstral peak prominence	2	0.88	Ho
Area of voice range profile	1	0.43	/	Smoothed cepstral peak prominence	2	0.88	Ho

[a] *k = number of effect sizes available in the included literature.*

[b] *r = correlation coefficient (when k=1),* $\overline{r_w}$ *= mean weighted correlation coefficient (when k>1).*

[c] SD_{res} *= residual standard deviation, / = not applicable (when k=1), Ho/He = homogeneous/heterogeneous r or* $\overline{r_w}$ *(when k>1).*

Figure 3.2 Diagram illustrating the organisation of the meta-analysis for acoustic measures on continuous speech. The second line in every box contains the number of acoustic measures.



prominence seems to suggest a much stronger association. For instance, although Halberstam's (2004) r-value of 0.55 for smoothed cepstral peak prominence does not provide strong support for smoothed cepstral peak prominence as a valid measure of overall voice quality; combining this result with the Heman-Ackah et al. (2002) and the Eadie & Baylor (2006) results of r = 0.80 and r = 0.82 respectively, the final $\overline{r_w}$ is 0.63, which supports smoothed cepstral peak prominence as a promising acoustic marker of overall voice quality. Based on the meta-analysis of sustained vowel studies, four measures satisfied the requirement of a homogeneous $\overline{r_w} \geq 0.60$: (1) Pearson r at autocorrelation peak, (2) pitch amplitude, (3) spectral flatness of residue signal, (4) smoothed cepstral peak prominence. For continuous speech, three measures satisfied the criterion: signal-to-noise ratio from Qi, cepstral peak prominence and smoothed cepstral peak prominence. Consequently, these six measures are considered to be the most promising measures for the acoustic measurement of overall voice quality, as compared to the remaining eighty-one measures included in the original meta-analysis. The results of these six measures will be discussed in the next sections.

The first of these six measures is Pearson r at autocorrelation peak. To obtain this measure, correlations are calculated between the voice signal and delayed versions of the same signal (i.e., autocorrelation) at time lags between the minimally and maximally expected fundamental period. The Pearson moment-product correlation coefficient is computed at the highest peak of this autocorrelation function (i.e., the correlogram with "delay" or "time lag" on the abscissa and "correlation" on the ordinate). The rationale behind this measure is that more periodic voice signals display more prominent autocorrelation peaks, and vice versa. A perfectly periodic signal reveals a Pearson r at autocorrelation peak of 1.0, and the more the signal deviates from perfect periodicity, the more this correlation decreases (Hillenbrand & Houde, 1996). This measure of the autocorrelation function on the sound waveform, was treated by Wolfe et al. (2000) as an overall voice quality predictor in both male and female voices. Meta-analysis resulted in $\overline{r_w} = 0.74$ (k = 2). This result is not confirmed by other independent correlation studies. Although Hillenbrand & Houde (1996) indicated a correlation of 0.84 between breathiness ratings and Pearson r at autocorrelation peak for both sustained vowels and continuous speech, and concluded that Pearson r at autocorrelation peak is an accurate predictor of breathiness, further corroboration of its predictive validity is needed.

The second measure is pitch amplitude. To acquire this measure, the radiated voice signal is first inverse filtered via a linear predictive coding algorithm. The result of this inverse filtering is a residue signal, i.e., a series of impulses theoretically showing the moment of vocal tract acoustic excitation provided by glottal closure (enabling investigation of the signal provided by the laryngeal source instead of the entire vocal tract). Second, the autocorrelation function of this residue signal is calculated. Pitch amplitude is the amplitude of the maximum correlation (i.e., traditionally corresponding with the pitch) in the

correlogram and consequently is considered to be a measure of the strength of voicing periodicity (Prosek et al., 1987). Plant et al. (1997), as well as Prosek et al. (1987) used pitch amplitude in predictions of both disorder severity and hoarseness in sustained vowels. Meta-analysis of these two independent studies resulted in a $\overline{r_w}$ = 0.75 (k = 3). Although Eadie & Doyle (2005) reported an r value for pitch amplitude of only 0.58 when applied on continuous speech, further support for the value of measures based on inverse filtering is provided by Parsa & Jamieson (2001), who concluded that such measures are superior to perturbation measures for both continuous speech and sustained vowels. In part, Parsa and Jamieson arrived at their conclusion based upon measures of diagnostic accuracy, which included the area under the receiver operating characteristic curve (ROC). In the case for pitch amplitude, the area under the ROC curve for sustained vowels was 0.977 (perfect diagnostic accuracy = 1.00), and the rate of correct classification between normal and pathologic voice was 93.0 %. For continuous speech there was an area under ROC curve of 0.953 and a correct classification rate of 88.9 %. Parsa & Jamieson (2001) state that pitch amplitude provided the best classification among all measures extracted from continuous speech samples.

Based upon the results of the meta-analysis, the third acoustic measure which produced respectable raw correlation results was spectral flatness of residue signal. This measure also generates the residue signal as an output of the inverse filter. The spectrum is then derived from the residue signal, and finally the distribution of the frequencies in the spectrum is computed. The flatter the spectral distribution of the residue signal, the more the harmonics are considered to be masked by noise (Prosek et al., 1987). This measure was investigated as a predictor for both severity of voice disorder and hoarseness in sustained vowels by Prosek et al. (1987). Our meta-analysis on spectral flatness of residue signal results leads to $\overline{r_w}$ = 0.69 (k = 2). Although there is no confirmation from other independent correlations, Parsa & Jamieson (2001) who used discriminant analyses, supported the implementation of spectral flatness of residue signal as a valid discriminator between normal and pathological voices. For sustained vowels, spectral flatness of residue signal showed the largest area under ROC curve (0.996) and had the highest classification accuracy (96.5% correct). For continuous speech, the area under the ROC was 0.928 and 85.8 %, respectively. Furthermore, Parsa & Jamieson (2001) concluded that more commonly used measures (as jitter, shimmer and noise-to-harmonics ratio) did not perform as well as measures based on linear prediction modeling and inverse filtering.

The predictive power of the fourth measure, signal-to-noise ratio from Qi, was investigated in continuous speech only. This measure uses linear predictive coding and inverse filtering for the decomposition of speech samples into signal (i.e., waveform of the original signal) and noise (i.e., waveform of the signal with a typically random Gaussian distribution after removal of resonance-based and voice-based patterns). The ratio between the average root-mean-square amplitudes of the signal and the noise components can then be computed to quantify the acoustic properties of disordered voices (Qi et al., 1999). Combining the

independent results of Qi et al. (1999) and Eadie & Doyle (2005) leads to a homogeneous $\overline{r_w}$ = 0.69 (k = 2), which is promising. This measure was also examined in the studies of Parsa & Jamieson (2001). Although not as robust as pitch amplitude and spectral flatness of residue signal, signal-to-noise ratio from Qi demonstrated acceptable diagnostic precision (distinguishing normophonic from dysphonic individuals) in both sustained vowels (area under ROC curve: 0.945; classification rate: 81.6 %) and continuous speech (area under ROC curve: 0.903; classification rate: 79.6 %). In summary, measures and algorithms based on inverse filtering and linear prediction modeling appear to be very promising and useful in clinical settings, where patients present with heterogeneous voice qualities and severities.

The meta-analysis outcome for the fifth and sixth measure, the cepstral markers of cepstral peak prominence and smoothed cepstral peak prominence, can be summarized as follows. To obtain these two measures, one constructs a cepstrum (i.e., a log power spectrum of a log power spectrum, resulting in a graph with "quefrency" on the abscissa and "cepstral magnitude" on the ordinate). The highest cepstral peak is identified between the minimally and maximally expected fundamental period, and a linear regression line is drawn which relates quefrency to cepstral magnitude. The difference in amplitude between this cepstral peak and the corresponding value on the linear regression line exactly below the peak determines the cepstral peak prominence. Averaging (i.e., smoothing) of the cepstrum across time and across quefrency results in a smoothed cepstrum, and the difference between the highest peak and the corresponding value on the regression line in the smoothed cepstrum is called the smoothed cepstral peak prominence. The rationale behind these measures is that the more periodic a voice signal is, the more it displays a well-defined harmonic configuration in the spectrum and, subsequently, the more prominent the cepstral peak intends to be (Hillenbrand & Houde, 1996). For cepstral peak prominence on sustained vowels, meta-analysis of the Wolfe & Martin (1997), the Wolfe et al. (2000) and the Halberstam (2004) results yields a homogeneous $\overline{r_w}$ of 0.50 (k = 3). On continuous speech, however, meta-analysis on the findings of Halberstam (2004) and Eadie & Baylor (2006) results in $\overline{r_w}$ = 0.88 (k = 2). Heman-Ackah et al. (2002), Halberstam (2004) and Eadie & Doyle (2005) investigated smoothed cepstral peak prominence applied on sustained vowels. Meta-analysis of these three independent studies results in $\overline{r_w}$ = 0.63 (k = 3). Furthermore, Heman-Ackah et al. (2002) and Halberstam (2004) also provide correlation coefficients for continuous speech, which results in a homogeneous $\overline{r_w}$ = 0.88 (k = 2) after meta-analysis. In summary, on the basis of this meta-analysis the two cepstral measures, and smoothed cepstral peak prominence in particular, can be viewed as potentially the most accurate predictive acoustic algorithms or single correlates of overall voice quality. Additional evidence for the validity of cepstral measures can be found in Hillenbrand & Houde (1996) who found that cepstral peak prominence was among the most robust correlates of breathiness in sustained vowels as well as in continuous

speech. Other studies confirming this conclusion were conducted by de Krom (1993), who stated that cepstrum-based harmonics-to-noise ratio is a strong predictor of both roughness and breathiness in sustained vowels. Dejonckere & Wieneke (1996) found a correlation of 0.80 between overall severity of hoarseness and the amplitude of highest rahmonic (a.k.a. cepstral peak magnitude). The magnitude of this correlation far exceeded the correlation of the other acoustic measures in their study (jitter ratio, relative noise level above 6 khz, mean harmonic emergence between 0.5 and 1.5 khz). In a later study using factor analysis, Dejonckere (1998) reported that cepstral peak magnitude is negatively affected by irregularity in vocal fold vibration as well as by excessive glottal air leakage, bolstering the assertion that cepstral peak magnitude is sensitive to aspects that potentially contribute to overall dysphonia severity. Heman-Ackah et al. (2003) investigated the diagnostic validity of smoothed cepstral peak prominence on sustained vowels and continuous speech and of amplitude perturbation quotient, percent jitter, noise-to-harmonics ratio from mdvp, relative average perturbation and smoothed pitch perturbation quotient on sustained vowels only. They concluded that the smoothed cepstral peak prominence measures are good correlates of dysphonia and that, on average, smoothed cepstral peak prominence on continuous speech, performed better on measures of diagnostic precision such as sensitivity, specificity, positive predictive value and negative predictive value, as compared to traditional time-based measures of perturbation. They concluded that smoothed cepstral peak prominence "are reliable measures that should become routine in objective voice analysis (p. 332)". Finally, Awan & Roy (2006) conducted a study in which they used a cepstral measure they called expected cepstral peak prominence in a multiple regression procedure. This measure actually is the ratio of the cepstral peak prominence to the expected amplitude of the cepstral peak based on linear regression. It is very similar to the cepstral peak prominence measure described by Hillenbrand & Houde (1996). Awan & Roy (2006) indicate that expected cepstral peak prominence "may be the most significant component (pp. 44)" contributing to their four-factor model for predicting dysphonia severity. Collectively, measures derived from the cepstrum (such as cepstral peak prominence and smoothed cepstral peak prominence) can be used in sustained vowel as well as continuous speech samples because they do not rely on accurate fundamental period detection (Hillenbrand & Houde, 1996; Heman-Ackah et al., 2003), and they can be easily implemented in clinical settings.

In addition to these six k>1 measures with $\overline{r_w} \geq 0.60$, there were many k=1 measures. However, because a high correlation in one study can be offset by a low correlation in another study (and vice versa), caution is warranted when interpreting the outcome of a solitary r, and this applies to all acoustic measures with k=1 (e.g. r=0.93 for normalized mean absolute period shimmer on sustained vowels or r=0.26 for frequency-domain harmonics-to-noise ratio on continuous speech). Without further confirmation or rejection of the presented r-values, it is impossible to draw firm conclusions regarding these k=1 measures at this point in time.

Interestingly, the measures that were investigated most often (with $k \geq 5$) were the perturbation measures (amplitude perturbation quotient, percent jitter, percent shimmer, relative average perturbation), the noise measure noise-to-harmonics ratio from mdvp and voice range profile measure phonatory fundamental frequency range (see Table 3.5). Percent jitter is a measure of fundamental frequency or period perturbation. It measures the mean difference in fundamental frequency of adjacent periods relative to the mean fundamental frequency of all periods in the voice recording (Buder, 2000). Relative average perturbation is also a measure of fundamental frequency perturbation. This measure is similar to percent jitter, but uses a moving three-point smoothing and normalization to average the period before computing the mean deviation in period relative to the mean period of all periods (Buder, 2000). Percent shimmer, another measure similar to percent jitter, measures amplitude perturbations by computing the mean deviation in amplitude between adjacent cycles relative to the mean amplitude of all cycles (Buder, 2000). Amplitude perturbation quotient is another amplitude perturbation measure, but instead of working with adjacent cycles as percent shimmer, it first averages the amplitude of a moving number (i.e., an odd integer greater than one) of successive cycles before calculating the mean deviation in amplitude between cycle groups relative to the mean amplitude of all cycles (Buder, 2000). These perturbation measures are traditionally linked to the measurement of irregular voice fold vibrations. The noise-to-harmonics ratio from mdvp is a spectral measure that computes the ratio of the between-harmonic spectral magnitudes in the range from 1500 to 4500 Hz to the harmonic spectral magnitudes in the range from 70 to 4500 Hz (Buder, 2000). This measure is classically associated with measurements of additive noise at the level of the glottis. The phonatory fundamental frequency range in the voice range profile is one of the measures of the dispersion of the fundamental frequency and consists of subtracting the lowest from the highest possible fundamental frequency (Buder, 2000). According to De Bodt (1997), who reviewed the literature between 1991 and 1995, these are the most frequently mentioned measures in voice literature (except for F_0 and amplitude measures). Yet, on sustained vowels, these measures did not yield a $\overline{r_w} \geq 0.60$. Regarding jitter, meta-analysis yielded homogeneous $\overline{r_w}$ of 0.47, 0.49, 0.52 and 0.52 for absolute jitter, percent jitter, relative average perturbation and pitch perturbation quotient, respectively. Absolute jitter is the mean of the differences between the period or the fundamental frequency of adjacent cycles (Buder, 2000). Pitch perturbation quotient is the same as relative average perturbation, but with a smoothing factor of five cycles (Buder, 2000). Similarly, the meta-analysis for shimmer resulted in a homogeneous $\overline{r_w}$ of 0.45 for shimmer in db and amplitude perturbation quotient, and a heterogeneous $\overline{r_w}$ of 0.52 for percent shimmer. Regarding noise-to-harmonics ratio from mdvp, the measure most frequently encountered, a heterogeneous $\overline{r_w}$ of 0.45 was found. On continuous speech, there was a solitary correlation of 0.62 and 0.75 for percent shimmer and relative average perturbation, and 0.37 and 0.13 for phonatory

fundamental frequency range and noise-to-harmonics ratio from mdvp, respectively. Measures related to the voice range profile yielded a $\overline{r_w}$ of maximally 0.43. In general, the results of this meta-analysis confirm the apparent inferiority of perturbation measures as compared to other measures that do not depend upon accurate identification of cycle boundaries. This conclusion supports the findings of Parsa & Jamieson (2001), and is confirmed by Kreiman & Gerratt (2005), who concluded that "the associations between jitter, shimmer, and perceived voice quality are not sufficiently explanatory to justify continued reliance on jitter and shimmer as indices of voice quality (p. 2209)". As mentioned previously, F_0 and amplitude perturbation measures are especially susceptible for the influence of type of microphone and microphone localization relative to the sound source, type of hardware, processing algorithms, measurement algorithms, and software settings such as sampling rate and fundamental period extraction. Furthermore, F_0 and amplitude perturbation perturbation measures are not sensitive to differences in glottal waveform shape and additive glottal noise, and appear only reliable in nearly periodic voice signals (Titze, 1995; Parsa & Jamieson, 2001). This meta-analysis, combined with previous studies, seems to confirm that measures that do not rely on the extraction of the fundamental period in their calculation such as cepstral peak prominence, Pearson r at autocorrelation peak and pitch amplitude produce stronger relationships with perceptual judgements of overall severity of dysphonia in sustained vowels as well as continuous speech, and deserve further attention in clinical circles (Hillenbrand & Houde, 1996; Parsa & Jamieson, 2001).

Caveats and limitations

There are limitations regarding the present meta-analysis that restrict the generalizability of the findings, but identify areas for future research. It is important to acknowledge that current acoustic measures might not be sensitive predictors of perceived voice quality because of limitations of their algorithms and the theoretical models upon which they are based. First, this meta-analysis concentrated on the relationship between acoustic markers and overall voice quality. Additional meta-analytic research is needed to address the relationship between acoustic measures and specific vocal qualitiy attributes, such as breathiness and roughness. Meta-analytic techniques may improve the resolution of which acoustic measures best predict these specific voice qualities. Second, the presented meta-analysis is restricted to reports and findings based on correlation coefficients. In addition to the 69 measures on sustained vowels and the 26 measures on continuous speech, other measures have been discussed in the literature. But because no correlation coefficients were available, the value of these markers in voice quality measurement and their relative validity in comparison with the presented markers remains unclear. Future meta-analysis should also explore other effect size measures, aside from the correlation coefficient, to investigate the validity of acoustic markers. Third, overall voice quality can be investigated with measures other then acoustic measures. For example, certain

aerodynamic measures could also be worth exploring within this context, and thus meta-analysis investigating the association between aerodynamic measures and perceptual voice quality measurement is recommended. Fourth, the interpretation of the findings of the present meta-analysis is seriously restricted by the variability related to data acquisition. Whereas the significant influence of many items of the data acquisition system (e.g. microphone type and placement, environmental noise, software, etc.) on the outcome of perturbation measures already has been investigated, the impact of these items on other measures such as cepstral peak prominence and pitch amplitude remains unclear. Additional scrutiny of the impact of the data acquisition on the outcome of these measures is warranted. Fifth, the relationship between the auditory-perceptual rating and the acoustic measurement of overall voice quality relies greatly on the rationale and algorithm underlying the acoustic measure. However, as previously discussed, unreliability of listener ratings introduces perceptual "noise" and consequently tends to handicap the acoustic (or other) measurement of voice quality. While suggestions to improve rater reliability exist (e.g. Kreiman & Gerratt, 2000b; Eadie & Doyle, 2002; Bele, 2005; Eadie & Baylor, 2006; Yiu et al., 2007; Kreiman et al., 2007) few studies have estimated the true (absolute) impact of listener unreliabilty on the correlation between perception and acoustic measures. Furthermore, there is no universal standard distinguishing an acceptable from an unacceptable reliability estimate. Future research should address the criteria used to determine what precisely constitutes an acceptable level of listener reliability, and the impact of such criteria on the validation of acoustic voice quality measures.

CONCLUSIONS

The above-stated limitations notwithstanding, measures for which the meta-analysis resulted in a homogeneous $\overline{r_w}$ of at least 0.60 are Pearson r at autocorrelation peak, pitch amplitude, spectral flatness of residue signal and smoothed cepstral peak prominence on sustained vowels, and signal-to-noise ratio from Qi, cepstral peak prominence and smoothed cepstral peak prominence on continuous speech. Tables 3.4 and 3.5 present a hierarchy of the numerous predictive outcomes of acoustic markers measuring overall voice quality, but the reader is referred to the height of r or $\overline{r_w}$ as a quantity-based overview of the domain of acoustic voice quality measurement. Furthermore, the tables show the relative position of a given acoustic measure according to its concurrent validity as a measure of overall voice quality. In this regard, the present meta-analysis was able to effectively distill an extremely large number of potential acoustic measures to a subset of strong predictor variables. This should be particularly informative for voice practitioners in clinical settings who are faced with software packages that automatically generate a daunting number of acoustic measures ostensibly aimed to quantify dysphonia severity and track voice change following intervention. The

present meta-analysis confirmed that not all acoustic measures are created equal with respect to these clinical goals.

ACKNOWLEDGMENTS

The assistance by Jan Deman (Medical library, Sint-Jan General Hospital, Bruges, Belgium) for library work and article retrieval is greatly appreciated. The authors also would like to credit the associate editor and the three anonymous reviewers for the numerous valuable comments on earlier versions of this manuscript.

REFERENCES

Askenfelt, A.G., & Hammarberg, B. (1986). Speech waveform perturbation analysis: a perceptual-acoustical comparison of seven measures. *Journal of Speech and Hearing Research, 29,* 50-64.

Awan, S.N., & Roy, N. (2006). Toward the development of an objective index of dysphonia severity: a four-factor acoustic model. *Clinical Linguistics & Phonetics, 20,* 35-49.

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

Buder, E.H. (2000). Acoustic analysis of voice quality: a tabulation of algorithms 1902-1990. In R.D. Kent and M.J. Ball (Eds.), *Voice quality measurement (pp. 119-244).* San Diego: Singular Publishing Group.

De Bodt, M. (1997). *A framework of voice assessment: the relation between subjective and objective parameters in the judgement of normal and pathological voice.* Unpublished doctoral dissertation. Antwerp: University of Antwerp.

Dejonckere, P.H., Remacle, M., Fresnel-Elbaz, E., Woisard, V., Crevier-Buchman, L., & Millet, B. (1996). Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Revue de Laryngologie, Otologie et Rhinology, 117,* 219-224.

Dejonckere, P.H., & Wieneke, G.H. (1996). Cepstra of normal and pathological voices: correlation with acoustic, aerodynamic and perceptual data. In M.J. Ball and M. Duckworth (Eds.), *Advances in clinical phonetics, Volume 6, Studies in speech pathology and clinical linguistics (pp. 217-227).* Amsterdam: John Benjamins Publishing Company.

Dejonckere, P.H. (1998). Cepstral voice analysis: link with perception and stroboscopy. *Revue de Laryngologie, Otologie et Rhinology, 119*, 245-246.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research, 36,* 254-266.

Eadie, T.L., & Doyle, P.C. (2002). Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *Journal of the Acoustical Society of America, 11,* 3014-3021.

Eadie, T.L., & Doyle, P.C. (2005). Classification of dysphonic voice: acoustic and auditory-perceptual measures. *Journal of Voice, 19,* 1-14.

Eadie, T.L., & Baylor, C.R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice, 20,* 527-544.

Feijoo, S., & Hernández, C. (1990). Short-term stability measures for the evaluation of vocal quality. *Journal of Speech and Hearing Research, 33,* 324-334.

Franzblau, A.N. (1958). *A primer of statistics for non-statisticians.* New York: Harcourt, Brace & Company.

Frey, L.R., Botan, C.H., Friedman, P.G., & Kreps, G.L. (1991). *Investigating communication: an introduction to research methods.* Englewood Cliffs: Prentice-Hall.

Gorham-Rowan, M.M., & Laures-Gore, J. (2006). Acoustic-perceptual correlates of voice quality in elderly men and women. *Journal of Communication Disorders, 39,* 171-184.

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,* 70-73.

Heman-Ackah, Y.D., Michael, D.D., & Goding, G.S. (2002). The relationship between cepstral peak prominence and selected parameters of dysphonia. *Journal of Voice, 16,* 20-27.

Heman-Ackah, Y.D., Heuer, R.J., Michael, D.D., Ostrowski, R., Horman, M., Baroody, M.M., Hillenbrand, J., & Sataloff, R.T. (2003). Cepstral peak prominence: a more reliable measure of dysphonia. *Annals of Otology, Rhinology and Laryngology, 112,* 324-333.

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

Hirano, M., Hibi, S., Terasawa, R., & Masako, F. (1986). Relationship between aerodynamic, vibratory, acoustic and psychoacoustic correlates in dysphonia. *Journal of Phonetics, 14,* 445-456.

Hunter, J.E., Schmidt, F.L., & Jackson, G.B. (1982). *Meta-analysis, cumulating research findings across studies.* Beverly Hills: Sage Publications.

Kojima, H., Gould, W.J., Lambiase, A., & Isshiki, N. (1980). Computer analysis of hoarseness. *Acta Otolaryngologica, 89,* 547-554.

Kreiman, J., Gerratt, B.R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research, 33,* 103-115.

Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A., & Berke, G.S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for research. *Journal of Speech and Hearing Research, 36,* 21-40.

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

Kreiman, J., and Gerratt, B.R. (2000b). "Sources of listener disagreement in voice quality assessment," J. Acoust. Soc. Am. 108, 1867-1876.

Kreiman, J., & Gerratt, B. (2005). Perception of aperiodicity in pathological voice. *Journal of the Acoustical Society of America, 117,* 2201-2211.

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

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis.* Thousand Oaks: Sage Publications.

Ma, E., & Yiu, E. (2006). Multiparametric evaluation of dysphonic severity. *Journal of Voice, 20,* 380-390.

Parsa, V., & Jamieson, D.G. (2001). Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. *Journal of Speech, Language, and Hearing Research, 44,* 327-339.

Plant, R.L., Hillel, A.D., & Waugh, P.F. (1997). Analysis of voice changes after thyroplasty using linear predictive coding. *Laryngoscope, 107,* 703-709.

Prosek, R.A., Montgomery, A.A., Walden, E., & Hawkins, D.B. (1987). An evaluation of residue features as correlates of voice disorders. *Journal of Communication Disorders, 20,* 105-117.

Qi, Y., Hillman, R.E., & Milstein, C. (1999). The estimation of signal-to-noise ratio in continuous speech for disordered voices. *Journal of the Acoustical Society of America, 105,* 2532-2535.

Titze, I.R. (1995). *Workshop on acoustic voice analysis: summary statement.* Iowa City: National Center for Voice and Speech.

Wolfe, V.I., & Steinfatt, T.M. (1987). Prediction of vocal severity within and across voice types. *Journal of Speech and Hearing Research, 30,* 230-240.

Wolfe, V., Fitch, J., & Cornell, R. (1995). Acoustic prediction of severity in commonly occurring voice problems. *Journal of Speech and Hearing Research, 38,* 273-279.

Wolfe, V., & Martin, D. (1997). Acoustic correlates of dysphonia: type and severity. *Journal of Communication Disorders, 30,* 403-416.

Wolfe, V., Fitch, J., & Martin, D. (1997). Acoustic measures of dysphonic severity across and within voice types. *Folia Phoniatrica et Logopaedica, 49,* 292-299.

Wolfe, V.I., Martin, D.P., & Palmer, C.I. (2000). Perception of dysphonic voice quality by naïve listeners. *Journal of Speech, Language, and Hearing Research, 43,* 697-705.

Yu, P., Ouaknine, M., Revis, J., & Giovanni, A. (2001). Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements. *Journal of Voice, 15,* 529-542.

Yu, P., Garrel, R., Nicollas, R., Ouaknine, M., & Giovanni, A. (2007). Objective voice analysis in dysphonic patients: new data including nonlinear measurements. *Folia Phoniatrica Logopaedica, 59,* 20-30.

Yumoto, E., Sasaki, Y., & Okamura, H. (1984). Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. *Journal of Speech and Hearing Research, 27,* 2-6.

Zraick, R.I., Wendel, K., & Smith-Olinde, L. (2005). The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *Journal of Voice, 19,* 574-581.

4

TOWARD IMPROVED ECOLOGICAL VALIDITY IN THE ACOUSTIC MEASUREMENT OF OVERALL VOICE QUALITY: COMBINING CONTINUOUS SPEECH AND SUSTAINED VOWELS

Youri Maryn
Paul Corthals
Paul Van Cauwenberge
Nelson Roy
Marc De Bodt

This chapter has been published in:
Journal of Voice, 2009, Epub ahead of print.

ABSTRACT

To improve ecological validity, perceptual and instrumental assessment of disordered voice, including overall voice quality, should ideally sample both sustained vowels and continuous speech. This investigation assessed the utility of combining both voice contexts for the purpose of auditory-perceptual ratings, as well as acoustic measurement of overall voice quality. Sustained vowel and continuous speech samples from 251 subjects with (n=229) or without (n= 22) various voice disorders were concatenated and perceptually rated on overall voice quality by 5 experienced voice clinicians. After removing the non-voiced segments within the continuous speech samples, the concatenated samples were analyzed using 13 acoustic measures based upon fundamental frequency perturbation, amplitude perturbation, spectral and cepstral analyses. Stepwise multiple regression analysis yielded a six-variable acoustic model for the multiparametric measurement of overall voice quality of the concatenated samples (with a cepstral measure as the main contributor to the prediction of overall voice quality). The correlation of this model with mean ratings of overall voice quality resulted in r_s=0.78. A cross validation approach involving the iterated internal cross-correlations with 30 subgroups of 100, 50 and 10 samples confirmed a comparable degree of association. Furthermore, the ability of the model to distinguish voice disordered from vocally normal participants was assessed using estimates of diagnostic precision including receiver operating characteristic curve analysis, sensitivity and specificity, as well as likelihood ratios which adjust for base-rate differences between the groups. Depending upon the cutoff criteria employed, the analyses revealed an impressive area under ROC=0.895, and as well as respectable sensitivity, specificity, and likelihood ratios. The results support the diagnostic utility of combining voice samples from both continuous speech and sustained vowels in acoustic and perceptual analysis of disordered voice. The findings are discussed in relation to the extant literature, and the need for further refinement of the acoustic algorithm.

INTRODUCTION

Clinical assessment of dysphonia often relies on a combination of perceptual and acoustic measurement techniques. In general, there is the clinician's perceptual evaluation of voice quality, which is considered to be the gold standard upon which other methods are validated. Different kinds of rating scales and various rating systems, such as GRBAS (Hirano, 1981) or CAPE-V (Hillman, 2003; Kempster et al., 2008), have been proposed to standardize and quantify this perceptual assessment and to enhance its reliability (De Bodt et al., 1996; De Bodt, 1997; Bele, 2005). Because of the subjective nature of perceptual methods however, there are potentially several internal and external sources of bias involved[7] including, but not limited to: (1) experience of the listener and his/her

exposure to voice disorders (Kreiman et al., 1990; Kreiman et al., 1992; Kreiman et al., 1993; De Bodt et al., 1997; Wolfe et al., 2000; Eadie & Baylor, 2006), (2) degree of the patients' dysphonia (Rabinov et al., 1995), (3) type of auditory-perceptual rating scale (Kreiman et al., 1992; Dejonckere et al., 1993; Wuyts et al., 1999; Eadie & Doyle, 2002; Yu et al., 2002; Karnell et al., 2007) and, (4) speaking task or stimulus type (Zraick et al., 2005; Eadie & Baylor, 2006). Despite these problems, perceptual judgment of voice quality provides a measure that is readily accessible to all voice clinicians (Orlikoff et al., 1999) and it therefore remains an essential part of voice assessment (Hammarberg et al., 1980; Awan & Roy, 2006). The aforementioned problems however, have lead clinicians and researchers to develop various kinds of instrumental methods to "objectively" quantify the degree of overall voice quality disruption. Among these methods, acoustic measurements have become especially attractive due to their non-invasiveness, relatively low cost, and ease of application (Parsa & Jamieson, 2001). Acoustic analyses often provide a numerical output, which potentially captures the degree of dysphonia severity, permits tracking of treatment outcomes, and provides a means to communicate this information relatively easily to all stakeholders, for example voice clinicians, patients, third party payers, and physicians (Portney & Watkins, 2000). But, perhaps one of the most compelling arguments for the use of acoustic measures is consistency (Awan & Roy, 2006), or the fact that, for a given voice sample, the outcome remains unaltered as long as the algorithm behind the measurement remains unchanged. Given these advantages, it is no surprise that there is a vast body of research that addresses acoustic analysis algorithms and methods [see Buder (2000) for a comprehensive and complete overview] and investigates the relationship between acoustic measurement and the perceptual evaluation of voice quality [see Maryn et al. (submitted) for a meta-analysis].

In most voice clinics, acoustic measures are derived from sustained vowel samples and not from continuous speech samples. Several factors have contributed to this preference (Murry & Doherty, 1980; Askenfelt & Hammarberg, 1986; Parsa & Jamieson, 2001). First, a sustained vowel represents relatively time-invariant phonation whereas continuous speech involves rapid and frequent changes caused by glottal and supraglottal mechanisms. Second, in contrast to continuous speech, sustained mid-vowel segments do not contain non-voiced phonemes, fast voice onsets and terminations and prosodic fundamental frequency and amplitude fluctuations. Third, sustained vowels are not affected by speech rate, vocal pauses, phonetic context and stress. Fourth, classic fundamental frequency or period perturbation and amplitude perturbation measures strongly rely on pitch detection and extraction algorithms. As a consequence, they lose precision in continuous speech analyses, in which perturbation is significantly affected by intonational patterns, voice onsets and offsets, and unvoiced fragments (Parsa & Jamieson, 2001). Fifth, sustained vowels can be elicited and produced with less effort and in a more standardized manner than continuous speech. Sixth, there is no linguistic loading in a sustained vowel, resulting in relative immunity from influences related to dialect and region, language, cognition, etc. (Zraick et al., 2005).

While sustained vowel productions are certainly attractive for a variety of reasons, relying exclusively on this voice context seems not to provide the most ecologically valid voice assessment, one that is truly representative of daily speech and voice use patterns (Parsa & Jamieson, 2001; Eadie & Baylor; 2006). Vocal fluctuations related to voice onset, voice termination, voice breaks, etc., which are considered to be crucial in voice quality evaluation (Hammarberg et al., 1980), can have a relatively large impact on short signals. Furthermore, dysphonia symptoms usually emerge in conversational voice production instead of sustained vowels (with the exception of singing voice) and they are most often signalled by the patients themselves in continuous speech (Yiu et al., 2000). Additionally, certain voice disorders, like adductor spasmodic dysphonia, can be characterized by relatively normal voice during sustained vowel productions, whereas voice produced in connected speech is often more severely compromised (Roy et al., 2005). Stimulus type (sustained vowel versus continuous speech) is also an important issue in the perceptual evaluation of voice quality and has been investigated by several authors. Although de Krom (1994) and Revis et al. (1999) reported no significant difference between the ratings of a sustained vowel and running speech, Wolfe et al. (1995) found a significant difference between the ratings of both sample types. The latter finding was supported in part by Zraick et al. (2005), who reported a statistically significant difference between the judgements of sustained vowels and recordings of a picture description. Collectively, these findings highlight the need to base clinical voice assessment on more than just sustained vowel analyses, and it seems essential for perceptual and instrumental analyses to be based upon both sample types if it is to be considered ecologically valid (Hammarberg et al., 1980; Yiu et al., 2000).

The relationship between acoustic measures and perceptual analysis of voice has received considerable attention in the literature. Researchers have traditionally reported bivariate correlations between specific acoustic measures and auditory-perceptual judgements of overall voice quality. For instance, in their meta-analysis examining the predictive power of specific acoustic correlates, Maryn et al. (submitted) found evidence of moderate to strong correlations with overall voice quality for only a few acoustics markers (out of possible 69 acoustic measures). Aside from smoothed cepstral peak prominence (Hillenbrand & Houde, 1996) and pitch amplitude (Prosek et al., 1987), most of the other acoustic measures showed only moderate to very weak correlations with perceptual ratings of overall voice quality. In order to overcome the limited predictive power of single acoustic markers and also motivated by the multidimensionality of voice, several researchers have advocated and explored a multivariate approach for the prediction of voice quality and/or to discriminate among different perceptual categories/levels of dysphonia severity (Eskenazi et al., 1990; Wolfe et al., 1995; Giovanni et al., 1996; Wolfe et al., 1997; Piccirillo et al., 1998; Wuyts et al., 2000; Yu et al., 2001; Bhuta et al., 2004; Awan & Roy, 2006; Ma & Yiu, 2006). Table 4.1 summarizes relevant methodological items and the most salient outcomes of these multivariate

Table 4.1 Methodology and outcome of studies that used a multiparametric approach in the objective measurement of overall voice quality.

Source	Number of subjects	Multivariate statistical method	Objective measures included in multivariate model	Perceptual evaluation of overall voice quality		Outcome	
				Dimension	Scale	Absolute correlation	Classification accuracy (%)
Eskenazi et al. (1990)	16	Multiple linear regression analysis	• Pitch amplitude • Harmonics-to-noise ratio	Overall severity	EAI 7 points	0.75	/
Wolfe et al. (1995b)	80	Stepwise multiple regression analysis	• Relative average perturbation • Fundamental frequency	Quality of phonation	EAI 7 points	0.56	/
Giovanni et al. (1996)	245	Direct-entry discriminant function analysis	• Percent jitter • Corrected spectrum • Ratio of oral airflow to intensity (glottal leakage) • Duration of the attack period	G, grade	EAI 5 points	/	66.1
Wolfe et al. (1997)	51	Multiple regression analysis	• Noise-to-harmonics ratio • Standard deviation of fundamental frequency, Percent jitter, Relative average perturbation or Pitch perturbation quotient	Severity of dysphonia	EAI 7 points	0.61	/
Wolfe et al. (1997)	51	Multiple regression analysis	• Noise-to-harmonics ratio • Percent shimmer, Shimmer in dB or Amplitude perturbation quotient	Severity of dysphonia	EAI 7 points	0.63	/
Piccirillo et al. (1998)	33	Logistic regression analysis	• Subglottic pressure • Airflow at lips • Fundamental frequency range	G, grade	EAI 4 points	0.58	/

Wuyts et al. (2000)	387	Stepwise logistic regression analysis	• Maximum phonation time • Highest fundamental frequency • Softest intensity • Percent jitter	G, grade	EAI 4 points	/	49.9
Yu et al. (2001)	84	Stepwise discriminant function analysis	• Fundamental frequency range • Fundamental frequency • Lyapunov coefficient • Maximum phonation time • Estimated subglottic pressure • Total signal-to-noise ratio	G, grade	EAI 4 points	/	86.0
Bhuta et al. (2004)	37	Stepwise multiple regression analysis	• Voice turbulence index • Noise-to-harmonics ratio • Soft phonation index	G, grade	EAI 4 points	0.66	/
Awan & Roy (2006)	134	Stepwise multiple regression analysis	• Ratio of the amplitude of the cepstral peak prominence to the expected amplitude of the cepstral peak • Discrete Fourier transform ratio ($energy_{<4000Hz}/energy_{>4000\ Hz}$) • Logarithm of shimmer • Inverse square root of the pitch sigma	Severity of dysphonia	EAI 7 points	0.88	/
Ma & Yiu (2006)	153	Direct-entry discriminant function analysis	• Maximum phonation time • Peak intraoral pressure • Voice range profile area • Relative amplitude perturbation	G, grade	EAI 11 points	/	67.3

studies. All studies used an equal-appearing interval scale (with a varying number of points, however) to measure the perceptual severity of dysphonia or G. With the exception of Yu et al. (2001), the majority suggested a multivariate algorithm consisting of four (acoustic and/or aerodynamic) instrumental measures. The outcomes of these studies were expressed either as a correlation coefficient or in classification accuracy. The classification accuracy of four multivariate models ranged from 49.9% to 86.0%. The association between perception and instrumental measurement was investigated in two other studies, revealing absolute correlation coefficients of 0.58 and 0.88. Both statistics illustrate that the predictive validity of the multivariate approaches can vary from rather low to rather high. We reasoned that improved acoustic prediction of overall voice quality may be derived from combining both sustained vowels and connected speech contexts. There are very few studies in which concatenation of both stimulus types has been used for the clinical examination of overall voice quality, and in which correlation coefficients (as a statistic for concurrent validity) as well as conventional measures of diagnostic test performance/precision such as the receiver operating characteristic (ROC) analysis, sensitivity, specificity and likelihood ratios (LR), have been presented.

Therefore, this study was undertaken to investigate the feasibility and utility of including both stimulus types in overall voice quality (i.e. dysphonia severity) assessment consisting of perceptual and acoustic methods. The voiced segments of two sentences read aloud were concatenated with three seconds of the vowel /a/ into a single sound file. In a first experiment, the inter- and intrajudge reliability of perceptual overall voice quality ratings of the concatenated sound files were examined. In a second experiment, the criterion-related concurrent validity of several acoustic markers for the measurement of overall voice quality was studied. The individual correlations of acoustic markers with perceptual ratings were calculated and the concurrent validity, as well as the internal consistency of a multivariate model based on linear stepwise regression was investigated. Finally, the diagnostic precision of the model was assessed, using ROC analysis and estimates of sensitivity, specificity and likelihood ratios.

METHODS

Participants

Voice samples were provided by 22 vocally normal and 229 voice-disordered subjects on an informed consent basis. The voice-disordered subjects were recruited from the ENT caseload of the Sint-Jan General Hospital in Bruges, Belgium. All voice disordered participants presented with a variety of etiologies and were referred for voice assessment by staff otolaryngologists. Participants were selected consecutively over the course of a 2-year period. There were 149 females and 79 males and ages ranged from 8 to 85 years with a mean of 38.9 years (SD=19.5 years). The scores on the Dysphonia Severity Index (Wuyts et al., 2000)

ranged from -16.50 to 9.67 with a mean of -0.46. The scores on the Voice Handicap Index (Jacobson et al., 1997) ranged from 0 to 106 with a mean of 39.8. Laryngological diagnoses were made with a flexible transnasal chip-on-tip laryngostroboscope (Olympus ENF-V). Table 4.2 summarizes the variety of voice disorders included in the sample. This group of subjects is considered to be representative of a clinical population of voice disordered patients. It reflects different age and gender groups, different types and degrees of voice quality disruption and vocally induced disability, including non-organic as well as organic laryngeal pathologies. This study also included 19 females and 3 males without any voice disorder, aged from 19 to 48 years with a mean of 24.6 years. These subjects did not seek help and since they had no actual voice complaint or history of voice, speech or hearing problems, the assessment of these vocally normal subjects was limited to the recording of voice samples.

Table 4.2 List of laryngeal pathologies, with their absolute and relative occurrence in the voice-disordered group of this study.

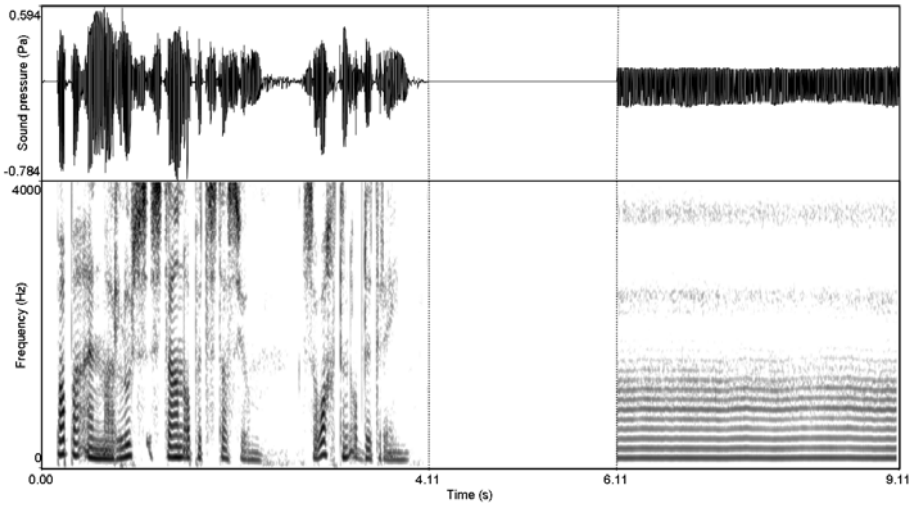
Voice disorder	Absolute number	Relative number
Functional dysphonia	81	35.5
Nodules	42	18.4
Polipoid mucosa (edema)	29	12.7
Paralysis/paresis	18	7.9
Polyp	11	4.8
Cyst	8	3.5
Acute laryngitis	5	2.2
Hemorrhage	4	1.8
Granuloma	4	1.8
Leukoplakia	4	1.8
Mutational falsetto	3	1.3
Tumor	3	1.3
Presbylarynx	3	1.3
Ventricular hypertrophy	2	0.9
Sulcus glottidis	2	0.9
Post-radiotherapy	2	0.9
Web	2	0.9
Post-phonosurgery	1	0.4
Larynxtrauma	1	0.4
Interarythenoidal pachyderm	1	0.4
Spasmodic dysphonia	1	0.4
Hyperkeratosis	1	0.4
Total	**228**	**100**

Voice samples

Every participant was asked to sustain the vowel /a/ for at least 5 seconds and to read aloud a phonetically balanced text (Van de Weijer & Slis, 1991; Van Lierde, 2001) using a comfortable pitch and loudness. Both voice samples were

recorded using an AKG C420 head-mounted condenser microphone (AKG Acoustic, 2000) and digitized at 44100 samples per second (Roark, 2006), i.e. a sampling rate of 44.1 kHz, and 16 bits of resolution using the Computerized Speech Lab (CSL model 4500; KayPentax, 2004). For the voice-disordered subjects, this was done at the beginning of a standard voice assessment. The samples were saved in .wav format. The vowel samples used in this study, were edited to include only the middle 3 seconds. The read text/connected speech samples were trimmed to include only the first two sentences. Finally, the voice samples were concatenated in the following order using Praat (Boersma, 2001; Boersma & Weenink, 2006): text segment, a pause of two seconds, followed by the 3 second sustained vowel segment. An example of the resulting concatenated waveform is given in Figure 4.1.

Figure 4.1 Oscillogram and narrowband-spectrogram (window length = 0.03 s) of a concatenated voice sample (derived from subject 2), as used in the perceptual evaluations of this study. There are three areas. The left portion reflects the first two sentences of the 'Papa en Marloes' text. The right area reflects the middle three seconds of a sustained /a/. Both samples were separated by two seconds of silence (area in the middle).



Overall dysphonia ratings

Five speech-language pathologists (two females and three males, with ages ranging from 27 to 59 years) were asked to rate each of the 251 concatenated voice samples. All listeners had previously participated more than once in post-academic courses on voice disorders and they all had at least five years of clinical experience judging voice quality and overall dysphonia severity. The listening experiment was performed in a quiet setting. The listeners were seated in a circle, equidistantly around two loudspeakers that emitted the voice samples in opposite directions. All

concatenated voice samples were presented in random order. All samples were judged within a 5 hour period of time. A 15 minute break was provided after each set of five rating sessions, i.e. after 45 minutes. Before the beginning of the listening experiment, all judges had confirmed that one particular concatenated voice sample represented normal voice quality. In order to establish an external standard of normal voice quality, all five rating sessions started with listening to this 'normal' voice sample as a referent to compare the 251 voice samples. By doing so, the authors intended to augment the reliability of the auditory-perceptual voice ratings (Chan & Yiu, 2002).

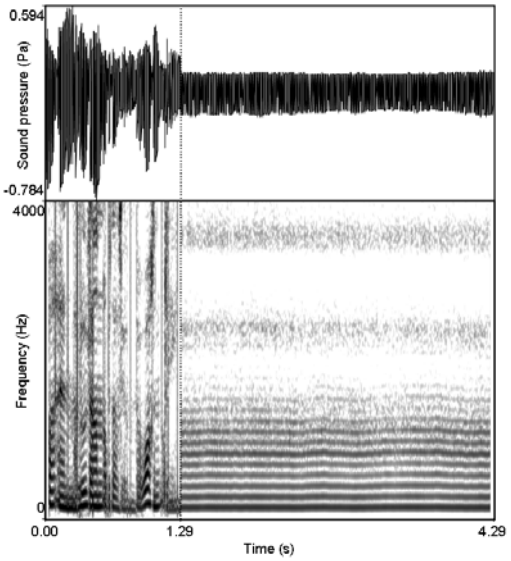
The five listeners were instructed to evaluate the severity of the perceptual dimension overall dysphonia (or Grade, "G"). Before judging the samples, this perceptual dimension was operationally defined following the description of Kreiman & Gerratt (2000). The "G" dimension was rated on a 4-point equal-appearing interval scale, as suggested by the Japan Society of Logopedics and Phoniatrics (Hirano, 1981): with a score of 0 representing the absence of a dimension and scores 1, 2 and 3 respectively corresponding with a slight, moderate and severe presence of G. Samples were repeated whenever one or more listeners were not confident in their judgement. At the end of the perceptual experiment, twenty-five randomized voice samples (i.e. 10 % of all samples) were repeated a second time in order to determine intra-rater reliability.

Acoustic measures

Digital copies of the recordings that were used for the perceptual evaluations, were selected for acoustic measurement. Since the majority of measures in this study pertain to voiced segments, a custom voicing detection algorithm was used to extract the voiced segments from the continuous speech files. The algorithm for detection and extraction of voiced segments was based on the three criteria proposed by Parsa & Jamieson (2001, pp. 332) and implemented in Praat. The programming script is provided in Appendix 4.1. Frames of 30 milliseconds were designated as voiced if [(a)] sound energy exceeded 30% of the overall signal energy, [(b)] zero crossing rate was below 1500 Hz, and [(c)] the normalized autocorrelation peak was above 0.3. Afterwards, the voiced continuous speech samples were concatenated with the sustained vowel sample of the same patient. An example of the resulting waveform is shown in Figure 4.2. Thirteen acoustic measures were derived from this material. The following eleven acoustic measures were derived using Praat: slope of the long-term average spectrum (Slope), tilt of the trend line through the long-term average spectrum (Tilt), jitter local (a.k.a. percent jitter), jitter rap (a.k.a. relative average perturbation), jitter ppq5 (a.k.a. pitch perturbation quotient), shimmer local (a.k.a. percent shimmer), shimmer local dB (a.k.a. shimmer in dB), shimmer apq11 (a.k.a. amplitude perturbation quotient), mean autocorrelation (mACF), noise-to-harmonics ratio (NHR) and harmonics-to-noise-ratio (HNR). The programming scripts that were used to obtain these measures in Praat are provided in Appendix 4.2. In addition,

the concatenated voice samples were analyzed using the computer program SpeechTool (Hillenbrand, 2006), obtained from Hillenbrand et al. (1994) and Hillenbrand & Houde (1996), which provided two cepstral measures: the cepstral peak prominence (CPP) and smoothed cepstral peak prominence (CPPs). In short, the measures on the concatenated samples (with only voiced fragments) included two spectral measures (slope and tilt), six perturbation measures (jitter local, jitter rap, jitter ppq5, shimmer local, shimmer local dB and shimmer apq11), three glottal noise measures (mACF, NHR and HNR), and two cepstral measures (CPP and CPPs).

Figure 4.2 Oscillogram and narrowband-spectrogram (window length = 0.03 s) of a concatenated voice sample (derived from subject 2), as used for the acoustic measures of this study. There are two areas. The left area reflects the concatenated voiced segments of the first two sentences of the 'Papa en Marloes' text. The right area reflects the middle three seconds of a sustained /a/.



Statistics

All statistical analyses were completed using SPSS for Windows version 12.0 (SPSS Inc., Chicago, Illinois, USA). In the first experiment, the intra-rater and inter-rater realibility of perceptual evaluation of overall voice quality (G) in concatenations of continuous speech and sustained vowel fragments (Figure 4.1) was explored. Two coefficients were used to determine listener agreement or reliability. Both statistics are nonparametric because G-ratings are on an ordinal scale. First, the Cohen kappa coefficient (κ) was calculated. This statistic, yielding values of $\kappa=1$ for perfect agreement and $\kappa=0$ when agreement is no better than by

chance, can be defined as a measure of the unanimity in the evaluations by multiple pairs of raters when they are rating the same object (Cohen, 1960). Guidelines for the interpretation of the κ statistic are provided by De Bodt et al. (De Bodt et al., 1997). Second, the Spearman rank-order correlation coefficient (r_s) was determined. This statistic reflects the degree to which a monotonic relationship exists between variables (Sheskin, 1997). Interpretation guidelines for r_s are provided by Frey et al. (1991).

For the second experiment, the predictive validity of the acoustic measurement of overall dysphonia severity (G) in the concatenated voiced samples was assessed (see Figure 4.2), and the following statistics were utilized. First, r_s and the coefficient of determination (r_s^2) between G and the thirteen acoustic measures were calculated as measures of concurrent validity. Second, stepwise multiple linear regression was executed to construct a statistical model representing the best combination of acoustic predictors for the overall degree of disordered voice. A multiple regression equation was constructed based on the unstandardized coefficients of the statistical model. In order to simplify clinical interpretation, the model was linearly rescaled in such a way that the outcomes of the equation resulted in a score between 0 and 10. This final model was called Acoustic Voice Quality Index or AVQI. Third, in order to investigate the criterion-related concurrent validity of AVQI, the correlation between G and AVQI was calculated with the Spearman rank-order correlation coefficient. Fourth, in order to examine the diagnostic utility of AVQI, several estimates of diagnostic precision were calculated (Portney & Watkins, 2000). For instance, the accuracy of a diagnostic test is commonly evaluated by the sensitivity and specificity of the test. Sensitivity is defined as the proportion of subjects with the disease (i.e., cases) who have a positive test, whereas the specificity is the proportion of subjects without the disease (i.e., non-cases) who have a negative test. In tests that yield continuous data like the AVQI employed in this study, several values of sensitivity and specificity are possible, depending on the cutoff point chosen to define a positive test. This trade-off between sensitivity and specificity can be displayed graphically using a technique known as the receiver operating characteristic (ROC) curve. To generate an ROC curve, the investigator selects several cutoff points and determines the sensitivity and specificity at each point. Sensitivity (or the true positive rate) is plotted on the Y-axis as a function of 1-specificity (the false positive rate) on the X-axis. An optimal diagnostic test is one that reaches the upper left corner of the graph. A test of no value follows the diagonal from the lower left to the upper right corners, suggesting that at any cutoff the true-positive rate is the same as the false-positive rate.

For the ROC-curve of AVQI, a voice was considered to be normal only when all five judges agreed on its normalcy (i.e. mean G = 0.0). On the other hand, a voice was considered dysphonic if one judge evaluated it at least as slightly dysphonic or G1 ($0.2 \leq$ mean G ≤ 3). The ability of AVQI to discriminate between normal and dysphonic voices was represented by the "area under ROC" i.e., A_{ROC}-statistic. The outcome of A_{ROC} is interpreted as a score between 1.0 (for perfect

discrimination between normal and dysphonic voices) and 0.5 (for chance-level diagnostic accuracy) (Portney & Watkins, 2000). ROC-statistics have been used previously to discriminate vocally normal from voice disordered subjects in several studies (Parsa & Jamieson, 2000; Parsa & Jamieson, 2001; Heman-Ackah et al., 2003; Umapathy et al., 2005). In order to facilitate clinical interpretation of AVQI-scores, a threshold-score to distinguish normal from disordered voice quality was derived from the ROC-curve, and positive and negative likelihood ratios were also calculated.

Likelihood ratios provide additional information about the value of a diagnostic test and help diminish problems with sensitivity, specificity related to the uneven number of normophonic and dysphonic subjects in the sample. The likelihood ratio incorporates both the sensitivity and specificity of the test and provides a direct estimate of how much a test result will change the odds of having a disease. The likelihood ratio for a positive result (LR^+) yields information regarding how the odds of the disease increase when the test is positive. Specifically, LR^+ is calculated by determining the ratio of true positive cases (sensitivity) to false positive cases (1-specificity) [i.e., LR^+=(sensitivity)/1-specificity] and gives information regarding the likelihood that an individual has a voice disorder. When LR^+ yields a number greater than 10, the value of the diagnostic test is high. If the LR^+ yields a value of 3, there is a moderate likelihood that the test suggests the person has the disorder, but is not conclusive and therefore should be interpreted with caution. If the test yields a LR^+ of 1, the diagnostic test does not help to diagnose a specific disorder. LR^- produces an estimate that helps determine whether an individual does not have a particular disorder when the diagnostic test does not identify them as such. LR^- gives information regarding how much the odds of the disease decrease when a test is negative. It is calculated by determining the ratio of false negative cases (1-sensitivity) to true negative cases (specificity) [LR^-=(1-sensitivity)/specificity]. Because the LR statistics consider sensitivity and specificity simultaneously, they are less vulnerable to sample size characteristics and base-rate differences between vocally normal and voice disordered participants (Dollaghan, 2007). Both LR^+ and LR^- were calculated for specific AVQI cut-off points (based on the ROC-curve).

Finally, a cross-validation procedure was undertaken. It is well known that when applied to a new set of data, different from the one upon which it was initially modelled, any predictive model may lose accounted variance (r_s^2) and concurrent validity. Therefore, correlation coefficients between G-scores and AVQI-scores were calculated for thirty randomly selected subgroups of one-hundred, fifty and ten voice samples. This method of cross-validation is similar to a method described by Awan & Roy (2006).

RESULTS

Reliability of auditory-perceptual ratings of concatenated samples

Figure 4.3 shows the frequency distribution of the mean G-ratings. The results for intra-rater reliability, based on 25 of the 251 voice samples, are represented in Table 4.3. The κ-statistic shows an average of 0.60 and ranges from 0.49 to 0.71. The r_s-statistic indicates a mean of 0.85 and ranges between 0.77 and 0.90. These results confirm moderate to high intra-rater reliability. The inter-rater agreement outcomes are shown in Table 4.4. The κ-statistic shows an average of 0.39 and ranges from 0.21 to 0.52. The r_s-statistic has a mean of 0.61 and ranges between 0.51 and 0.73. These outcomes indicate fair to moderate inter-rater agreement.

Figure 4.3 Frequency distribution of the mean auditory-perceptual overall voice quality ratings (average of G-scores of 5 experienced listeners) of the 251 concatenated voice samples.

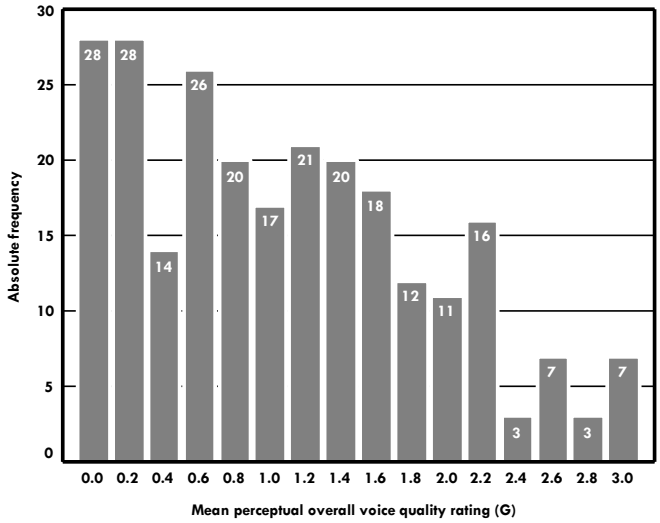


Table 4.3 Intra-rater reliability of the five listeners who rated overall voice quality on the concatenated voice samples: K and r_s.

	K	r_s
Rater 1	0.66	0.90
Rater 2	0.63	0.82
Rater 3	0.71	0.90
Rater 4	0.49	0.87
Rater 5	0.49	0.77

Table 4.4 *Matrix of inter-rater reliability between the five listeners who rated overall voice quality on the concatenated voice samples: κ and r_s.*

		Rater 2	Rater 3	Rater 4	Rater 5
Rater 1	κ	0.49	0.37	0.52	0.21
	r_s	0.64	0.54	0.68	0.51
Rater 2	κ		0.51	0.50	0.31
	r_s		0.63	0.73	0.56
Rater 3	κ			0.42	0.37
	r_s			0.62	0.61
Rater 4	κ				0.23
	r_s				0.60

Predictive validity of acoustic measures on concatenated samples

Table 4.5 lists the descriptive data for the thirteen acoustic variables in the group of 23 vocally normal cases and the 228 dysphonic subjects. The correlations (r_S) and coefficients of determination (r_S^2) between overall voice quality ratings and these thirteen acoustic measures are shown in Table 4.6. The highest absolute r_S-value was found for CPPs (r_S=0.71); followed by HNR (r_S=0.68), shimmer local dB (r_S=0.66) and CPP (r_S=0.65). The lowest absolute r_S-values were found for the frequency perturbation measures, the glottal noise measures NHR and mACF, and the spectral measures: slope (r_S=0.01), tilt (r_S=0.48), NHR (r_S=0.51) and jitter local (r_S=0.54). The strongest correlation identified is for CPPs (r_S=-0.71), explaining approximately 50 percent of the variation of G. With the exception of Slope, for which no significant correlation was found (r_S=0.01), all correlations were significant at the alpha=0.01 level. The stepwise multiple regression analysis revealed that a combination of six acoustic variables best predicted the overall dysphonia severity of voice recordings containing a concatenation of continuous speech as well as sustained vowels. The equation, based on the unstandardized coefficients of the regression, is:

$$\begin{aligned} \text{AVQI} = {} & 2.905 - 0.111 \times \text{CPPs} - 0.073 \times \text{HNR} - 0.213 \times \\ & \text{shimmer local} + 2.789 \times \text{shimmer local dB} - 0.032 \times \text{slope} + \\ & 0.077 \times \text{tilt} \end{aligned} \qquad \text{(Eq. 1)}$$

The outcomes of this equation range from -0.39 to 3.50. For practical clinical application, however, the equation is linearly rescaled in order to fall on a scale with values between 0 and 10. The resulting equation is:

$$\begin{aligned} \text{AVQI} = {} & (3.295 - 0.111 \times \text{CPPs} - 0.073 \times \text{HNR} - 0.213 \times \\ & \text{shimmer local} + 2.789 \times \text{shimmer local dB} - 0.032 \times \text{slope} + \\ & 0.077 \times \text{tilt}) \times 2.571 \end{aligned} \qquad \text{(Eq. 2)}$$

Table 4.5 Average (M), standard deviation (SD) and range (Min – Max) of the outcomes of the thirteen acoustic measures.

Measures	Normal (N=23)				Dysphonic (N=228)			
	M	SD	Min	Max	M	SD	Min	Max
Slope (dB)	-23.31	5.03	-33.15	-14.17	-24.86	4.99	-37.84	-8.15
Tilt (dB)	-10.51	0.73	-12.14	-8.62	-9.45	1.38	-13.81	-5.00
Jitter local (%)	0.98	0.18	0.69	1.31	1.60	1.08	0.71	7.50
Jitter rap (%)	0.46	0.10	0.29	0.63	0.80	0.60	0.30	3.99
Jitter ppq5 (%)	0.50	0.10	0.34	0.71	0.84	0.64	0.34	5.07
Shimmer local (%)	3.18	0.91	1.53	5.09	5.47	3.66	1.51	22.05
Shimmer local dB (dB)	0.31	0.05	0.21	0.42	0.52	0.31	0.21	1.89
Shimmer apq11 (%)	2.31	0.74	1.11	4.19	3.87	2.77	1.22	19.59
mACF	0.97	0.00	0.97	0.98	0.95	0.06	0.55	0.98
NHR	0.04	0.01	0.02	0.06	0.08	0.11	0.03	0.88
HNR (dB)	22.92	2.09	18.89	26.41	18.66	4.71	0.99	28.92
CPP	16.77	2.08	13.57	21.78	13.80	2.45	8.65	21.74
CPPs (dB)	8.05	0.94	5.97	10.16	6.41	1.81	0.89	10.94

Table 4.6 Correlation coefficients (r_s) and coefficients of determination (r_s^2) between the auditory-perceptual overall voice quality ratings (G) and the thirteen acoustic measures.

	Slope	Tilt	Jitter local	Jitter rap	Jitter ppq5	Shimmer local	Shimmer local dB
r_s	0.01	0.48	0.54	0.56	0.55	0.64	0.66
r_s^2	0.00	0.23	0.29	0.31	0.31	0.40	0.44

	Shimmer apq11	mACF	NHR	HNR	CPP	CPPs
r_s	0.61	-0.56	0.51	0.68	-0.65	-0.71
r_s^2	0.37	0.31	0.26	0.46	0.43	0.50

Inspecting the results, it is clear that there is a positive relationship between AVQI and G and thus the higher an AVQI score, the more disrupted the overall voice quality, and vice versa. The correlation between the outcome of AVQI and the G-scores was 0.78, revealing high concurrent (or predictive) validity. This proportional relationship between G and AVQI is illustrated in Figure 4.4. The

coefficient of determination was 0.61. Figure 4.4 also shows that the AVQI scores for subjects with severe dysphonia are higher than expected and consequently raises the possibility of a nonlinear polynomial trend between AVQI and G. However, closer investigation of second- and third-order polynomial relationships for the present data revealed no statistically significant difference between the linear and the nonlinear models.

Cross-validation of AVQI

The thirty iterated cross-validations yielded mean correlations of 0.77, 0.75 and 0.80 for randomized subgroups of one-hundred, fifty and ten voice samples, respectively. These results are almost identical to the original correlation for all 251 voice samples. Figure 4.5 represents the distribution of these cross-validation correlations. For example, the thirty correlations for one-hundred randomly chosen voice samples show that the validity of AVQI can range from 0.670 to 0.857 (mean = 0.769; standard error = 0.009; standard deviation = 0.047). The correlations for fifty randomly chosen voice samples lie between 0.633 and 0.852 (mean = 0.751; standard error = 0.011; standard deviation = 0.058), and between 0.462 and 0.963 for thirty times ten randomly selected voice samples (mean = 0.805; standard error = 0.021; standard deviation = 0.118). These results confirm the stability of the AVQI across subsets of voices.

Figure 4.4 Scatterplot to illustrate the concurrent validity of AVQI (the two dotted lines above and under the regression fit line delineate the upper and lower boundaries of the 95 % prediction interval).

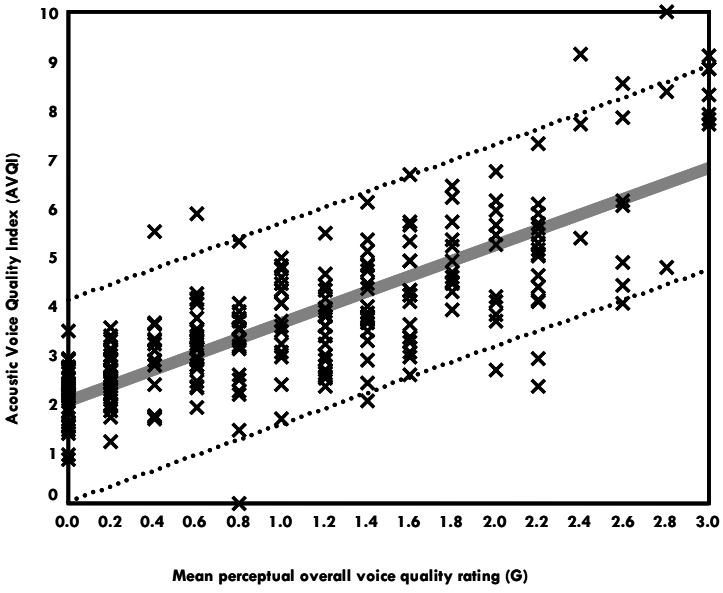
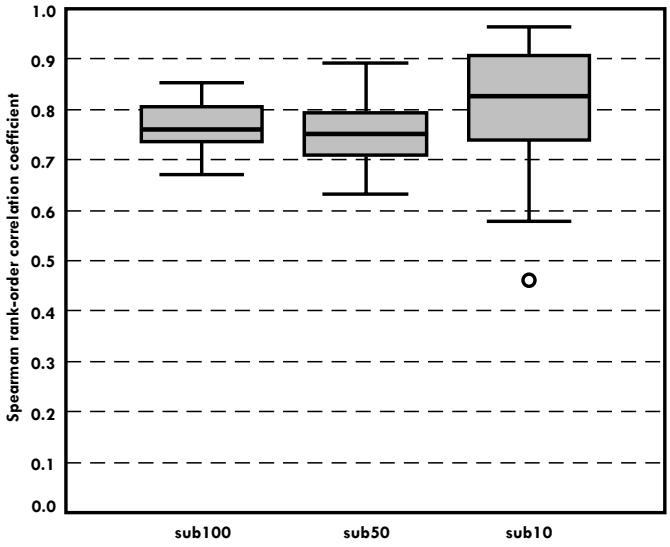


Figure 4.5 Box-and-whiskerplots illustrating the cross-correlations between G and AVQI for thirty subgroups of one-hundred, fifty and ten randomly chosen voice samples.

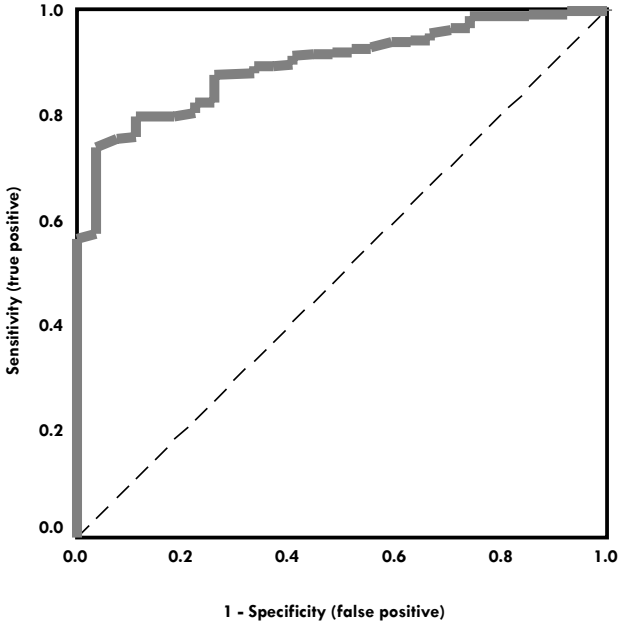


Diagnostic accuracy of AVQI

To evaluate the diagnostic accuracy of AVQI and its ability to distinguish vocally normal from voice disordered participants, a ROC-curve was constructed (Figure 4.6). The A_{ROC}, with the AVQI-scores as the test variable and the G-scores as the state variable, was 0.895, revealing relatively high discriminatory power to distinguish normal and pathological voices (with statistical significance at p=0.000, under the assumption of an asymptotic distribution). The ROC-curve was also used to identify which cut-off point achieved the best balance between sensitivity and specificity, would provide optimal discrimination between the normal and pathological groups. In this regard, an AVQI cutoff score of 2.36, produced sensitivity and specificity estimates of 91% and 59% respectively. Therefore, using this threshold, 91% of patients with dysphonia were correctly classified as being dysphonic (i.e., pathological). However, only 59 % of the normal subjects were correctly categorized as non-voice disordered (i.e., as having normal voice quality). This AVQI-score is accompanied with intermediate-range likelihood ratios: LR^{+}=2.23 and LR^{-}=0.15. In contrast, using an AVQI cutoff criterion of 2.95, produced estimates of sensitivity of 74%, and specificity of 96%. Only 74 % of the dysphonic patients were classified correctly, but almost all subjects with normal voice quality were correctly classified as such. Likelihood-analysis for this AVQI cutoff score resulted in much improved discriminatory power: LR^{+}=19.98 and LR^{-} =0.27. To assist in the interpretation of likelihood ratios in this specific study, the higher the LR^{+}, the more confident the clinician can be that a person with a higher

AVQI-score is voice disordered/dysphonic. A $LR^{+} \geq 10$ indicates that a positive AVQI-score (i.e., >2.95) is very likely to have come from a dysphonic person. The lower LR^{-}, the more confident the clinician can be that a person with a low AVQI-score (i.e., <2.95) is normophonic. A $LR^{-} \leq 0.10$ indicates that a low AVQI-score is very likely to have come from a person without dysphonia (Dollaghan, 2007).

Figure 4.6 ROC-curve to illustrate the diagnostic validity of AVQI (solid line).



DISCUSSION

We investigated the utility of combining sustained vowels and continuous speech in dysphonia severity measurement. Although sustained vowels have long been preferred, both stimulus types are important in perceptual and acoustic measurement, and both contexts would seem necessary to improve ecologically validity in voice assessment (Hammarberg et al., 1980; Murry & Doherty, 1980; Askenfelt & Hammarberg, 1986; Yiu et al., 2000; Parsa & Jamieson, 2001; Zraick et al., 2005; Eadie & Baylor, 2006). For this reason, the first two sentences of a commonly used Dutch text were concatenated with three seconds of a sustained /a/ vowel. The 251 concatenated samples were perceptually rated on a 4-point equal-appearing interval scale of overall voice quality (i.e. "G" from GRBAS). An acoustic analysis protocol which contained a diverse set of acoustic variables, consisting of thirteen frequency perturbation, amplitude perturbation as well as

spectral and cepstral measures, was applied. Absolute correlation coefficients between these acoustic variables and G-scores were the highest for the cepstral, HNR and amplitude perturbation measures.

The finding that cepstral measures were the most powerful predictor of G is compatible with existing reports in the literature. For instance, Heman-Ackah et al. (2002), completed a similar study, but on separated samples of sustained vowels and continuous speech, and they reported that for sustained vowels, Cepstral Peak Prominence (CPP) measures were the best predictor of G (r_s=-0.80), as compared to perturbation and glottal noise measures. The outcome was even more impressive for CPPs applied to continuous speech (r_s=-0.86). However, these study results were based on a rather small group of 18 subjects. Similarly, Eadie & Baylor (2006) also reported a strong association (r_s=0.806) between CPPs and overall severity of dysphonia on sustained vowels, but again this finding was derived from a small number of subjects. In a study by Awan & Roy (2006), based on 134 subjects (see Table 4.1), a cepstral measure called CPP/EXP (similar to the cepstral measures used in this study) was the most powerful contributor to their acoustic model for voice quality prediction. Finally, the cepstral peak magnitude, as reported in the study of Dejonckere & Wieneke (1996), was also the best predictor of hoarseness, compared to spectral and perturbation measures. In conclusion, the results from this investigation also confirm the strength of cepstral-based measures in predicting dysphonia severity, and demonstrate the inferiority of specific perturbation measures such as jitter (Parsa & Jamieson, 2000; Parsa & Jamieson, 2001; Kreiman & Gerratt, 2005; Awan & Roy, 2006).

In addition to bivariate analyses, a multiparametric analysis approach was employed to construct a weighted algorithm that would identify the most robust acoustic predictors of judgements of "G" or overall dysphonia severity (Eskenazi et al., 1990; Wolfe et al., 1995; Giovanni et al., 1996; Wolfe et al., 1997; Piccirillo et al., 1998; Wuyts et al., 2000; Yu et al., 2001; Bhuta et al., 2004; Awan & Roy, 2006; Ma & Yiu, 2006). Stepwise multiple regression analysis resulted in a model (i.e., AVQI) consisting of six acoustic measures. With an initial r_s-value of 0.78 between G and AVQI, i.e. a high degree of concurrent validity (Frey et al., 1991) (Table 4.2), this model can be considered a strong predictor of overall voice quality. The AVQI model appears to perform better than other acoustic models reported by Eskenazi et al. (1990), Wolfe et al. (1995), Wolfe et al. (1997), Piccirillo et al. (1998) and Bhuta et al. (2004), but not as favorably as the results reported by Awan & Roy (2006). In contrast to the other models which analyzed sustained vowels only, the performance of the AVQI is particularly compelling, given that AVQI incorporates continuous speech as well as sustained vowels.

Although the performance of the AVQI is very respectable (r_s^2=0.61), there still remains 39% of variance in G not accounted for by AVQI. This finding is confirmed by the rather wide 95% confidence interval illustrated in Figure 4.4. A narrower confidence interval would mean that there is less overlap in AVQI scores between adjacent perceptual levels of dysphonia severity, and the AVQI would better discriminate among or between these levels of severity. In this regard, one

factor which likely attenuates the ability of *any* acoustic model to account for true variance in listener ratings of dysphonia severity, is the "unreliability" of those perceptual judgements. Inter- and intra-judge unreliability ultimately contributes to increased error variance in the regression analysis, leaving less true variance to be explained or accounted for by the acoustic model. In this study, our inter-rater reliability was moderately low, perhaps reflecting differences in training and experience of the listeners. In light of the increased error variance related to only moderate levels of listener reliabilty, the amount of variance accounted for by the acoustic model is actually quite respectable. Perhaps more intensive training with external perceptual standards, as promoted by Chan & Yiu (2002) and Kreiman et al. (2007), could potentially improve the reliability of the ratings, and thus the predictive power of the multivariate acoustic model.

This investigation represents the first attempt to investigate concurrent validity and diagnostic precision in the same study. Parsa & Jamieson (2001), for example, only used ROC analysis to investigate the diagnostic accuracy of several acoustic measures, and in contrast, Awan & Roy (2006) focused only on the correlation between a multivariate acoustic model and the severity of dysphonia. In this investigation, diagnostic precision was studied using conventional estimates of diagnostic accuracy. In other words we determined how accurate the AVQI was in determining whether someone does or does not have dysphonia? A ROC-curve was constructed (Figure 4.6), with an impressive A_{ROC}=0.895. Since this statistic equals the probability of correctly discriminating between normal (normophonic) state and abnormal (dysphonic) states (Portney & Watkins, 2000)[24], the result of this study indicates that, based on the AVQI, a clinician could correctly identify almost 90% of the cases. Unfortunately, Giovanni et al. (1996), Wuyts et al. (2000), Yu et al. (2001) and Ma & Yiu (2006) did not use A_{ROC} to investigate the classification accuracy of their multivariate constructs. Thus, it is difficult to compare our results with the results from these previous studies. In addition, whereas the A_{ROC} in this study describes the discriminatory performance between two conditions (normophonic versus dysphonic), the values provided in Table 4.1 are based on the classification accuracy between more than two states (e.g. normophonic versus slightly dysphonic, moderately dysphonic, severely dysphonic, etc.) which also complicates comparisons. In general however, it seems that based upon the wide prediction interval around the regression line (Figure 4.4), the classification ability of AVQI between intermediate levels of dysphonia may be no better than reported by other acoustic models.

The ROC-curve can also be used to decide which AVQI cutoff points would determine optimum diagnostic performance. For example, AVQI=2.36 yields a sensitivity of 91% and a specificity of 59%. The high sensitivity indicates that the AVQI correctly identifies the majority of dysphonic subjects whereas, the lower specificity means that AVQI is less able to correctly identify normophonic subjects (controls). In a diagnostic setting, where one is especially interested in correctly labelling subjects as being dysphonic, this AVQI cutoff point could be proposed as a diagnostic threshold. However, once the results are adjusted for base-

rate differences as in the likelihood ratio analysis, the likelihood ratios associated with AVQI cutoff criterion suggest only intermediate-range LR^+ and LR^-, indicating weaker evidence of diagnostic accuracy. In contrast, if a clinician is primarily interested in correctly identifying normals or non-dysphonics, such as in a screening test, a higher AVQI cutoff score of 2.95 may be more appropriate, since this score yields a sensitivity of 0.740 and a specificity of 0.963. The improved specificity is reflected in the likelihood ratio analysis for this particular cutoff score which revealed excellent discriminatory accuracy for subjects who test positive (i.e., AVQI>2.95). However, even at this threshold level, the LR- results suggest that a clinician still cannot be sufficiently certain that subjects who test negative (i.e., AVQI<2.95) are indeed normophonic. In the final analysis, the results of the various indices of diagnostic precision are respectable and encouraging, but it is clear that the AVQI requires further refinement as a diagnostic index to distinguish vocally normal individuals from those with dysphonia.

Limitations and future directions

There are a number of limitations related to the analysis method employed and the results reported. First, the cross-validation was internally investigated on numerous subgroups of the same sample on which AVQI was originally modeled. Future investigations should externally confirm the validity of AVQI with new clinical voice samples and ratings. Second, although the perceptual ratings were made by experienced voice clinicians and external standards regarding normophonia were equalized across raters, there was only moderate inter-rater reliability, which likely attenuated the predictive power of the acoustic model. In order to increase the reliability of perceptual ratings, future methods should include multiple anchor stimuli representing different levels in the dysphonia continuum (Chan & Yiu, 2002; Kreiman et al., 2007). Instead of working with equal-appearing interval scales, future studies could probably benefit from the use of visual analog scales or a hybrid scale such as CAPE-V (Kempster et al., 2008), incorporating both equal-appearing and visual analog scales. Third, this study was the first to combine sustained vowels and continuous speech in the perceptual as well as the acoustic methodology. However, information regarding what precisely influences the final rating when combined stimuli are presented in this manner is unknown, and deserves further attention. It is possible that the perceptual rating of a concatenated voice sample is primarily determined by one of the speaking tasks, for instance by the most dysphonic speaking task, or by an average of the two speaking tasks; or alternatively by a recency or primacy effect, to mention a few possibilities only. Future research should explore the influence of such variables when employing such concatenated samples. Fourth, like other studies (Qi et al., 1999; Eadie & Doyle, 2002; Heman-Ackah et al., 2002), this study used only the first two sentences of a reading passage. However, it is possible that longer

samples of continuous speech will provide improved validity of acoustic and perceptual analysis results.

CONCLUSION

Voice quality assessment traditionally relies on measurement of sustained vowels. To improve ecological validity, acoustic and perceptual assessment of continuous speech should also be considered. The aim of this study was to investigate the feasibility and diagnostic precision of combining both voice contexts into one concatenated sample upon which auditory-perceptual ratings and acoustic measures could be completed. The results supported the viability of such an approach, with respectable bivariate associations between listener ratings of dysphonia severity and specfic acoustic variables. The diagnostic accuracy of a multivariate acoustic model (AVQI) was assessed, revealing respectable estimates of diagnostic precision. Further refinement of the acoustic algorithm is necessary.

REFERENCES

AKG Acoustics (2000). *C420: user instruction. MicroMic series II.* München: AKG Acoustics Harman Pro.

Askenfelt, A.G., & Hammarberg, B. (1986). Speech waveform perturbation analysis: a perceptual-acoustical comparison of seven measures. *Journal of Speech and Hearing Research, 29,* 50-64.

Awan, S.N., & Roy, N. (2006). Toward the development of an objective index of dysphonia severity: a four-factor acoustic model. *Clinical Linguistics & Phonetics, 20,* 35-49.

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

Bhuta, T., Patrick, L., & Garnett, J.D. (2004). Perceptual evaluation of voice quality and its correlation with acoustic measurement. *Journal of Voice, 18,* 299-304.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glottal International, 5,* 341-345.

Boersma, P., & Weenink, D. (2006). *Praat: doing phonetics by computer (Version 4.6.15) [Computer program].* Amsterdam: Institute of Phonetic Sciences, available from: http://www.praat.org.

Buder, E.H. (2000). Acoustic analysis of voice quality: a tabulation of algorithms 1902-1990. In R.D. Kent & M.J. Ball (Eds.), *Voice quality measurement (pp. 119-244).* San Diego: Singular Publishing Group.

Chan, K.M., & Yiu, E.M. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language and Hearing Research, 45,* 111-126.

Cohen, J.A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

De Bodt, M., Van de Heyning, P.H., Wuyts, F.L., & Lambrechts, L. (1996). The perceptual evaluation of voice disorders. *Acta Otorhinolaryngologica Belgica, 50,* 283-291.

De Bodt, M. (1997). *A framework for voice assessment, the relation between subjective and objective parameters in the judgment of normal and pathological voice.* Unpublished doctoral dissertation. Antwerp: University of Antwerp.

De Bodt, M.S., Wuyts, F.L., Van de Heyning, P.H., & Croux, C. (1997). Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice, 11,* 74-80.

Dejonckere, P.H., Obbens, C., de Moor, G.M., & Wieneke, G.H. (1993). Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatrica, 45,* 76-83.

Dejonckere, P.H., & Wieneke, G.H. (1996). Cepstra of normal and pathological voices: correlation with acoustic, aerodynamic and perceptual data. In M.J. Ball, & M. Duckworth (Eds.), *Advances in Clinical Phonetics* (pp. 217-226). Amsterdam: John Benjamins Publishing Co.

de Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech and Hearing Research, 37,* 985-1000.

Dollaghan CA. *The handbook for evidence-based practice in communication disorders.* Baltimore: MD Brookes; 2007.

Eadie, T.L., & Doyle, P.C. (2002). Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *Journal of the Acoustical Society of America, 112,* 3014-3021.

Eadie, T.L., & Baylor, C.R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice, 20,* 527-544.

Eskenazi, L., Childers, D.G., & Hicks, D.M. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research, 33,* 298-306.

Frey, L.R., Botan, C.H., Friedman, P.G., & Kreps, G.L. (1991). *Investigating communication, an introduction to research methods.* Englewood Cliffs: Prentice-Hall.

Giovanni, A., Robert, D., Estublier, N., Teston, B., Zanaret, M., & Cannoni, M. (1996). Objective evaluation of dysphonia: preliminary results of a device allowing simultaneous acoustic and aerodynamic measurements. *Folia Phoniatrica et Logopaedica, 48,* 175-185.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngologica, 90,* 441-451.

Heman-Ackah, Y.D., Michael, D.D., & Goding, G.S. (2002). The relationship between cepstral peak prominence and selected parameters of dysphonia. *Journal of Voice, 16,* 20-27.

Heman-Ackah, Y.D., Heuer, R.J., Michael, D.D., Ostrowski, R., Horman, M., Baroody, M.M., Hillenbrand, J., & Sataloff, R.T. (2003). Cepstral peak prominence: a more reliable measure of dysphonia. *Annals of Otology, Rhinology and Laryngology, 112,* 324-333.

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

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

Hillenbrand, J. (2006). *SpeechTool, version 1.56 [Computer program].* Retrieved April 10, 2006, from http://homepages.wmich.edu/~hillenbr/.

Hillman, R. (2003). Overview of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), instrument developed by ASHA Special Interest Division 3. *Paper presented at 32nd Symposium: Care of the Professional Voice*, Philadelphia, USA.

Hirano, M. (1981). Psycho-acoustic evaluation of voice. In G.E. Arnold, F. Winckel & B.D. Wyke (Eds.), *Disorders of Human Communication 5, Clinical examination of voice (pp. 81-84).* Wien: Springer-Verlag.

Jacobson, B.H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., and Benninger, M.S. (1997). The Voice Handicap Index (VHI): development and validation. *American Journal of Speech-Language Pathology, 6,* 66-70.

Karnell, M.P., Melton, S.D., Childes, J.M., Coleman, T.C., Dailey, S.A., & Hoffman, H.T. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice, 21,* 576-590.

KayPentax (2004). *Multi-Speech and CSL software: software instruction manual.* Lincoln Park, NJ: KayPentax.

Kempster, G.B., Gerratt, B.R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R.E. (2008). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology,* epub ahead of print.

Kreiman, J., Gerratt, B.R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research, 33,* 103-115.

Kreiman, J., Gerratt, B.R., Precoda, K., & Berke, G. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research, 35,* 512-520.

Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A., & Berke, G.S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for research. *Journal of Speech and Hearing Research, 36,* 21-40.

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

Kreiman, J., & Gerratt, B. (2005). Perception of aperiodicity in pathological voice. *Journal of the Acoustical Society of America, 117,* 2201-2211.

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

Ma, E., & Yiu, E. (2006). Multiparametric evaluation of dysphonic severity. *Journal of Voice, 20,* 380-390.

Maryn, Y., De Bodt, M., Van Cauwenberge, P., Roy, N., & Corthals, P. (2008). Acoustic measurement of overall voice quality: a meta-analysis. *Journal of the Acoustical Society of America,* submitted for publication.

Murry, T., & Doherty, E.T. (1980). Selected acoustic characteristics of pathological and normal speakers. *Journal of Speech and Hearing Research, 23,* 361-369.

Orlikoff, R.F., Dejonckere, P.H., Dembowski, J., Fitch, J., Gelfer, M.P., Gerratt, B.R., Haskell, J.A., Kreiman, J., Metz, D.E., Schiavetti, N., Watson, B.C., & Wolfe, V. (1999). The perceived role of voice perception in clinical practice. *Phonoscope, 2,* 89-104.

Parsa, V., & Jamieson, D.G. (2000). Identification of pathological voices using glottal noise measures. *Journal of Speech, Language, and Hearing Research, 43,* 469-485.

Parsa, V., & Jamieson, D.G. (2001). Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. *Journal of Speech, Language, and Hearing Research, 44,* 327-339.

Piccirillo, J.F., Painter, C., Fuller, D., Haiduk, A., & Fredrickson, J.M. (1998). Assessment of two objective voice function indices. *Annals of Otology, Rhinology and Laryngology, 107,*396-400.

Portney, L.G., & Watkins, M.P. (2000). *Foundations of clinical research: applications to practice (2nd Ed.).* Upper Saddle River: Prentice-Hall.

Prosek, R.A., Montgomery, A.A., Walden, E., & Hawkins, D.B. (1987). An evaluation of residue features as correlates of voice disorders. *Journal of Communication Disorders, 20,* 105-117.

Qi, Y., Hillman, R.E., & Milstein, C. (1999). The estimation of signal-to-noise ratio in continuous speech for disordered voices. *Journal of the Acoustical Society of America, 105,* 2532-2535.

Rabinov, C.R., Kreiman, J., Gerratt, B.R., & Bielamowocz, S. (1995). Comparings reliability of perceptual ratings of roughness and acoustic measures of jitter. *Journal of Speech and Hearing Research, 38,* 26-32.

Revis, J., Giovanni, A., Wuyts, F., & Triglia, J.M. (1999). Comparison of different voice samples for perceptual analysis. *Folia Phoniatrica et Logopaedica, 51,* 108-116.

Roark, R.M. (2006). Frequency and voice: perspectives in the time domain. *Journal of Voice, 20,* 325-354.

Roy, N., Gouse, M., Mauszycki, S.C., Merrill, R.M., & Smith, M.E. (2005). Task specificity in adductor spasmodic dysphonia versus muscle tension dysphonia. *Laryngoscope, 115,* 311-316.

Sheskin, D.J. (1997). *Handbook of parametric and nonparametric statistical procedures.* Boca Rotan: CRC Press LLC.

Umapathy, K., Krishnan, S., Parsa, V., & Jamieson, D.G. (2005). Discrimination of pathological voices using a time-frequency approach. *IEEE Transactions on Biomedical Engineering, 52,* 421-430.

Van de Weijer, J.C., & Slis, I.H. (1991). Nasaliteitsmeting met de nasometer. *Tijdschrift voor Logopedie en Foniatrie, 63,* 97-101.

Van Lierde, K. (2001). *Nasalance and nasality in clinical practice.* Unpublished doctoral dissertation. Ghent: University of Ghent.

Wolfe, V., Cornell, R., & Fitch, J. (1995). Sentence/vowel correlation in the evaluation of dysphonia. *Journal of Voice, 9,* 297-303.

Wolfe, V., Fitch, J., & Cornell, R. (1995). Acoustic prediction of severity in commonly occurring voice problems. *Journal of Speech and Hearing Research, 38,* 273-279.

Wolfe, V., Fitch, J., & Martin, D. (1997). Acoustic measures of dysphonic severity across and within voice types. *Folia Phoniatrica et Logopaedica, 49,* 292-299.

Wolfe, V.I., Martin, D.P., & Palmer, C.I. (2000). Perception of dysphonic voice quality by naïve listeners. *Journal of Speech, Language, and Hearing Research, 43,* 697-705.

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,* 508-517.

Wuyts, F.L., De Bodt, M.S., Molenberghs, G., Remacle, M., Heylen, L., Millet, B., Van Lierde, K., Raes, J., and Van de Heyning, P.H. (2000). The Dysphonia Severity Index: an objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language and Hearing Research, 43,* 796-809.

Yiu, E., Worrall, L., Longland, J., & Mitchell, C. (2000). Analysing vocal quality of connected speech using Kay's computerized speech lab: a preliminary finding. *Clinical Linguistics & Phonetics, 14,* 295-305.

Yu, P., Ouaknine, M., Revis, J., & Giovanni, A. (2001). Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements. *Journal of Voice, 15,* 529-542.

Yu, P., Revis, J., Wuyts, F.L., Zanaret, M., & Giovanni, A. (2002). Correlation of instrumental voice evaluation with perceptual voice analysis using a modified visual analog scale. *Folia Phoniatrica et Logopaedica, 54,* 271-281.

Zraick, R.I., Wendel, K., & Smith-Olinde, L. (2005). The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *Journal of Voice, 19,* 574-581.

Appendix 4.1 Script for the detection and extraction of voiced segments in continuous speech, as scripted by the second author (P.C.) and to be used in the program Praat (version 4.6.15).

```
Resample... 22050 50
Rename... original
samplingRate = Get sampling frequency
intermediateSample = Get sampling period
Create Sound... onlyVoice 0 0.001 'samplingRate' 0
select Sound original
To TextGrid (silences)... 50 0.003 -25 0.1 0.1 silence sounding
select Sound original
plus TextGrid original
Extract intervals where... 1 no "does not contain" silence
Concatenate
select Sound chain
Rename... onlyLoud
globalPower = Get power in air
select TextGrid original
Remove
select Sound onlyLoud
signalEnd = Get end time
windowBorderLeft = Get start time
windowWidth = 0.03
windowBorderRight = windowBorderLeft + windowWidth
globalPower = Get power in air
voicelessThreshold = globalPower*(30/100)
select Sound onlyLoud
extremeRight = signalEnd - windowWidth
while windowBorderRight < extremeRight
        Extract part... 'windowBorderLeft' 'windowBorderRight' Rectangular 1.0 no
        select Sound onlyLoud_part
        partialPower = Get power in air
        if partialPower > voicelessThreshold
                call checkZeros 0
                if (zeroCrossingRate <> undefined) and (zeroCrossingRate < 3000)
                        select Sound onlyVoice
                        plus Sound onlyLoud_part
                        Concatenate
                        Rename... onlyVoiceNew
                        select Sound onlyVoice
                        Remove
                        select Sound onlyVoiceNew
                        Rename... onlyVoice
                endif
        endif
        select Sound onlyLoud_part
        Remove
        windowBorderLeft = windowBorderLeft + 0.03
        windowBorderRight = windowBorderLeft + 0.03
        select Sound onlyLoud
endwhile
select Sound onlyVoice
procedure checkZeros zeroCrossingRate
```

```
        start = 0.0025
        startZero = Get nearest zero crossing... 'start'
        findStart = startZero
        findStartZeroPlusOne = startZero + intermediateSample
        startZeroPlusOne = Get nearest zero crossing... 'findStartZeroPlusOne'
        zeroCrossings = 0
        strips = 0
        while (findStart < 0.0275) and (findStart <> undefined)
                while startZeroPlusOne = findStart
                        findStartZeroPlusOne = findStartZeroPlusOne + intermediateSample
                        startZeroPlusOne = Get nearest zero crossing... 'findStartZeroPlusOne'
                endwhile
                distance = startZeroPlusOne - startZero
                strips = strips +1
                zeroCrossings = zeroCrossings +1
                findStart = startZeroPlusOne
        endwhile
        zeroCrossingRate = zeroCrossings/distance
endproc
```

Appendix 4.2 Scripts for the calculation of eleven acoustic measures in the program Praat (version 4.6.15) used in this study (as scripted by Y.M.). The original sounds object received the name 'Analysis'.

Slope of LTAS

```
select Sound Analysis
To Ltas... 1
ltasSlope = Get slope... 0 1000 1000 10000 energy
```

Tilt of trendline through LTAS

```
select Sound Analysis
To Ltas... 1
Compute trend line... 1 10000
ltasTrendlineTilt = Get slope... 0 1000 1000 10000 energy
```

Frequency perturbation measures

```
select Sound Analysis
To Pitch (cc)... 0 75 15 no 0.03 0.45 0.01 0.35 0.14 600
select Sound Analysis
plus Pitch Analysis
To PointProcess (cc)
percentJitter = Get jitter (local)... 0 0 0.0001 0.02 1.3
percentJitter = percentJitter*100
relativeAveragePerturbation = Get jitter (rap)... 0 0 0.0001 0.02 1.3
relativeAveragePerturbation = relativeAveragePerturbation*100
pitchPerturbationQuotient = Get jitter (ppq5)... 0 0 0.0001 0.02 1.3
pitchPerturbationQuotient = pitchPerturbationQuotient*100
```

Amplitude perturbation measures

```
select Sound Analysis
To PointProcess (periodic, cc)... 50 400
select Sound Analysis
```

```
plus PointProcess Analysis
percentShimmer = Get shimmer (local)... 0 0 0.0001 0.02 1.3 1.6
percentShimmer = percentShimmer*100
absoluteShimmer = Get shimmer (local_dB)... 0 0 0.0001 0.02 1.3 1.6
amplitudePerturbationQuotient = Get shimmer (apq11)... 0 0 0.0001 0.02 1.3 1.6
amplitudePerturbationQuotient = amplitudePerturbationQuotient*100
```

Glottal noise measures

```
select Sound Analysis
To Pitch (cc)... 0 75 15 no 0.03 0.45 0.01 0.35 0.14 600
select Sound Analysis
plus Pitch Analysis
To PointProcess (cc)
select Sound Analysis
plus Pitch Analysis
plus PointProcess Analysis
voiceReport$ = Voice report... 0 0 75 600 1.3 1.6 0.03 0.45
meanAutocorr = extractNumber (voiceReport$, "Mean autocorrelation: ")
nhr = extractNumber (voiceReport$, "Mean noise-to-harmonics ratio: ")
hnr = extractNumber (voiceReport$, "Mean harmonics-to-noise ratio: ")
```

THE ACOUSTIC VOICE QUALITY INDEX: TOWARD IMPROVED TREATMENT OUTCOMES ASSESSMENT IN VOICE DISORDERS

Youri Maryn
Marc De Bodt
Nelson Roy

This chapter has been accepted for publication in:
Journal of Communication Disorders.

ABSTRACT

Voice practitioners require an objective index of dysphonia severity as a means to reliably track treatment outcomes. To ensure ecological validity however, such a measure should survey both sustained vowels and continuous speech. In an earlier study, a multivariate acoustic model referred to as the Acoustic Voice Quality Index (AVQI), consisting of a weighted combination of 6 time-, frequency- and quefrency-domain metrics, was developed to measure dysphonia severity in both speaking tasks. In the current investigation, the generalizability and clinical utility of the AVQI is evaluated by first assessing its external cross-validity and then determining its sensitivity to change in dysphonia severity following surgical and/or behavioral voice treatment. The results, based upon a new set of normal and disordered voices compared favorably with outcomes reported earlier, indicating acceptable external validity. Furthermore, the AVQI was sensitive to treatment-related changes, validating its role as a potentially robust and objective voice treatment outcomes measure.

INTRODUCTION

Measuring treatment outcomes is a fundamental component of evidence-based practice in speech-language pathology. An outcome is simply the result of an intervention. In the area of voice disorder management, many methods have been proposed as potential voice treatment outcomes measures (i.e., laryngoscopic, vibratory, aerodynamic, auditory-perceptual, etc.). However, acoustic measurement of voice has received substantial attention as a potential objective treatment outcomes measure due to its relatively low cost, non-invasiveness, ease of application, and numerical output (Awan & Roy, 2009; Maryn, Corthals, Van Cauwenberge, Roy & De Bodt, in press; Parsa & Jamieson, 2001). The general need for objective outcomes measures of voice treatment, has motivated clinicians and researchers to develop and investigate the clinical value of specific acoustic measures of dysphonia severity. As a consequence, there has been a proliferation of acoustic analysis algorithms sensitive to measures of F_0 perturbation, amplitude perturbation, waveform perturbation, spectral configuration and cepstral characteristics in radiated and inverse filtered soundwaves (Buder, 2000). However, the validity and clinical utility of many of these acoustic measures, especially the more popular time-based, perturbation measures, has been strongly debated over the past two decades (e.g., De Bodt, 1997; Kreiman & Gerratt, 2005; Parsa & Jamieson, 2001; Titze, 1995). Moreover, Carding et al. (2004) confirmed inadequate sensitivity of any one of these time-domain perturbation measures, used in isolation, to treatment-related changes in voice and voice quality. To overcome the limited validity of single acoustic parameters, and also recognizing the multidimensionality of voice, several researchers have explored a multiparameter approach to measuring voice quality, and/or to distinguish among different voice types and levels of dysphonia severity (e.g., Awan & Roy, 2006; Ma & Yiu, 2006;

Wuyts et al., 2000; Yu, Ouaknine, Revis & Giovanni, et al., 2001). One of these multivariable models, a 6-factor model referred to as the 'Acoustic Voice Quality Index' (i.e., 'AVQI'), was developed by Maryn et al. (in press). Although the AVQI was constructed in a similar manner to other models, it is unique in that it permits objective assessment of dysphonia severity on sustained vowels as well as continuous speech.

In practice, acoustic measures are traditionally derived from sustained mid-vowel samples and not from continuous speech samples for several reasons. First, a sustained vowel represents relatively stable phonation whereas continuous speech involves fast and frequent glottal and supraglottal changes. Second, in contrast to continuous speech, sustained mid-vowel segments do not contain non-voiced phonemes, rapid voice on- and offsets and prosodic variations in F_0 and amplitude. Third, sustained vowels are not infuenced by speech rate and stress, vocal pauses, and phonetic context. Fourth, classic F_0 or T_0 perturbation and amplitude perturbation measures strongly rely on pitch detection and extraction algorithms; and consequently they become imprecise in continuous speech analyses, in which perturbation is significantly increased by intonational patterns, voice onsets and offsets, and unvoiced segments. Fifth, sustained vowels can be elicited and produced with less effort and in a more standardized manner than continuous speech. Sixth, there is no linguistic loading in a sustained vowel, resulting in relative immunity from influences related to language, dialect and region, etc. (Askenfelt & Hammarberg, 1986; Maryn et al., in press; Parsa & Jamieson, 2001; Zraick, et al., 2005).

The inclusion of both speaking tasks or stimulus types (i.e., continuous speech and sustained vowel) in voice analysis is important for several reasons however. First, vocal inconstancies typically observed in continuous speech rather than in sustained vowels (e.g., voice onset/offset, prosodic modulations, voice breaks, etc.) can be decisive in auditory-perceptual voice quality evaluation (Hammarberg, Fritzell, Gauffin, Sundberg, & Wedin, 1980). Second, the two stimulus types can express different types/degrees of vocal dysfunction and, consequently, result in different perceptual ratings (Wolfe, Cornell & Fitch, 1995a; Zraick et al., 2005). Adductor spasmodic dysphonia, for example, can often be characterized by relatively normal voice during sustained vowels, whereas voice in continuous speech is often more severely disrupted (Roy, Gouse, Mauszycki, Merrill & Smith, 2005). Third, dysphonia symptoms commonly emerge in conversational voice production instead of sustained vowels (except for singing voice) and they are usually revealed to patients in connected speech (Yiu, Worrall, Longland & Mitchell, 2000). Therefore, if a voice treatment outcomes measure is to be considered "ecologically valid" (i.e., one that is truly representative of daily speech and voice use patterns), then the acoustic measure ideally should be calculated on recordings of both speaking tasks.

To our knowledge, the AVQI is the first measure to incorporate samples of continuous speech, in addition to the sustained vowel samples used in other measurement protocols. To calculate the AVQI, a weighted combination of the

output of 6 acoustic time- (i.e., shimmer local, shimmer local dB and harmonics-to-noise ratio), frequency- (i.e., general slope of the spectrum and tilt of the regression line through the spectrum) and quefrency-domain (i.e., smoothed cepstral peak prominence) measures is modeled in a linear regression formula. The concurrent validity of the AVQI-model, i.e. its ability to measure overall severity of dysphonia (also known as Grade or G), is promising; and, a study of the diagnostic precision, validated its ability to generally discriminate between normophonia and dysphonia (Maryn et al., in press). The discriminatory power of the AVQI was considered to be at least equivalent to other multivariate models (Eskenazi, Childers & Hicks, 1990; Wolfe, Fitch & Cornell, 1995b; Wolfe, Fitch & Martin, 1997; Wuyts et al., 2000), but additional evaluation was considered essential to confirm and extend these early reports.

The present investigation examines further the validity and clinical utility of the AVQI as a potential treatment outcomes measure. In this regard two experiments were conducted. The first experiment examined the external cross-validity of the AVQI. When evaluating a predictive instrument like the AVQI, data are initially collected on an experimental sample (i.e., the sample originally used in Maryn et al. [in press]) and the mutiparameter algorithm associated with these data (e.g., predictive equation, cutoff score, etc.) is then applied later to a larger population. Data gathered on the initial experimental sample can sometimes differ substantially from data obtained from a subsequent sample containing different subjects and voices (Portney & Watkins, 2000). Consequently, it is unknown whether the AVQI's performance (i.e., its validity and accuracy) will necessarily be the same ason the original sample. Therefore, cross-validation is necessary to establish generalizability, and valid treatment outcomes measures demand such validation (Frey, Botan, Friedman & Kreps, 1991; Portney & Watkins, 2000). This generalizability or "external validity" of a measure can be investigated via two methods: internal and external cross-validity. Internal cross-validity works with subsets of subjects that are randomly drawn from the original sample. Maryn et al. (in press) have already explored the internal cross-validity of the AVQI via random selection from the original sample of 251 subjects, the authors formed 30 groups of 10, 50 and 100 subjects. For every group, a correlation coefficient between the AVQI and the auditory-perceptual ratings of dysphonia severity was then calculated. With mean correlations of 0.77 for the groups with 100 subjects, 0.75 for the groups with 50 subjects, and 0.80 for the groups with 10 subjects, it was concluded that the AVQI demonstrated stable validity across the different samples. External cross-validation on the other hand, is typically accomplished by assessing a measure's performance on a totally new group of subjects, under the assumption that the original and new groups are both good representatives of the population under consideration (i.e., vocally normal and voice-disordered subjects with various degrees of dysphonia). To generalize the AVQI's accuracy across samples, it is thus important to cross-validate it on a new sample (i.e., external cross-validity). The first experiment of this study represents the external cross-validity of the AVQI using a new sample of 39 subjects.

The second experiment focused on the AVQI's responsiveness to change. The ability of a measure to detect change over time is a crucial issue if this measure is to be used to assess the potential effects (i.e., outcome) of an intervention, and to serve as a viable treatment outcomes measure (Portney & Watkins, 2000). An acoustic measure is considered to be "change-responsive" when a change in its score is proportional to a change in the score of its clinical criterion (traditionally the auditory-perceptual rating of dysphonia severity). As such, there are only a few reports exploring whether or not a particular acoustic parameter is a valid outcomes measure of voice therapy. However, the design of these studies does not always permit conclusions regarding the responsiveness to change. Plant, Hillel and Waugh (1997), for example, examined the change in perceptual ratings and an acoustic measure called 'pitch amplitude' before and after medialization thyroplasty in 16 subjects with unilateral vocal fold immobility. 'Pitch amplitude' is the amplitude of the maximum correlation in the autocorrelation function of the inverse filtered voice signal (Davis, 1979). Although thyroplasty resulted in a statistically significant difference in both the perceptual rating and the pitch amplitude, no statistics regarding the proportionality between the differences/changes of the 2 measures were provided. Whether or not a change in pitch amplitude truly represented a proportionally equivalent change in perceived degree of dysphonia, i.e., the degree to which pitch amplitude is a valid outcomes measure of medialization thyroplasty, remains unclear. A similar study was recently reported by Jin et al. (2008). They monitored 40 patients with laryngopharyngeal reflux and investigated the changes in the reflux symptom index, the reflux finding score, and 'jitter' following medical therapy. 'Jitter' measures the mean difference in fundamental frequency or duration of adjacent periods, relative to the mean fundamental frequency or duration of all periods in the voice recording (Buder, 2000). Jin et al. (2008) found statistically significant differences for all measures before and after 1 month of treatment and concluded that "acoustic parameters can be used as indicators of treatment efficacy in the patients with laryngopharyngeal reflux" (p. 940). However, the proportional relationship between the magnitude in change of jitter and the magnitude of change in the reflux symptom index and the reflux finding score was not investigated, thus the use of jitter as a viable indicator of treatment (i.e., anti-reflux medication) efficacy also remains unknown. Awan & Roy (2009), on the other hand, investigated the relationship between the change scores associated with their 4-factor acoustic model and change scores observed on auditory-perceptual ratings of dysphonia severity before and after manual circumlaryngeal therapy in 88 patients with muscle tension dysphonia. They found a correlation (i.e., proportional relationship) of 0.75 between acoustically predicted change scores and perceived dysphonia severity change scores, and concluded that the acoustic model was a sensitive outcomes measure. Interestingly, although the acoustic model of Awan & Roy (2009) and the AVQI of Maryn et al. (in press) were constructed independently, they share some similarities. For instance, except for the noise-to-harmonics ratio in the AVQI and the pitch sigma in the model of Awan & Roy

(2009), both models contained measures of the cepstral peak, the ratio of low to high spectral energy, and the period-to-period amplitude perturbation. While the Awan & Roy (2009) acoustic model has shown respectable responsivity to change, the AVQI's sensitivity to change and thus its potential as a treatment outcomes measure awaits experimental verification. Therefore, the second experiment of the present study investigated the AVQI's responsiveness to change using a new set of 33 patients with various voice pathologies who underwent various kinds of intervention.

METHODS

Subjects

Voice samples from 72 patients recruited from the ENT department of the Sint-Jan General Hospital in Bruges, Belgium were employed in the 2 experiments. For the first experiment (i.e. the external cross-validation of AVQI), we used the recordings of 6 vocally normal and 33 voice-disordered subjects. Their voices were recorded at the beginning of a standard voice assessment. The group of subjects with a voice disorder consisted of 19 females and 14 males, ranging in age from 16 to 86 years (mean=49.2 years, SD=20.1 years). There also were 6 females without any voice disorder, ranging in age from 27 to 63 years (mean=36.1 years, SD=8.3 years).

For the second experiment (i.e. the AVQI's responsiveness to change), the pre- and post-therapy recordings of another 33 subjects with organic and/or non-organic voice disorders were used (Table 5.1). This group consisted of 22 females and 11 males, with a mean age of 40.9 years (SD=18.9 years, range 7–68 years). To examine the AVQI's ability to track different degrees of change in dysphonia severity following treatment, recordings of subjects experiencing varying degrees of therapeutic progress and change in overall voice quality were included in this experiment. The type of voice treatment and the combination of behavioral techniques used to improve voice is described in Table 5.1.

Voice treatment

Table 5.1 provides an overview of the interventions employed for all the individual subjects included in the second experiment. All 33 subjects received an eclectic treatment program, with an individualized combination of behavioral voice therapy techniques and/or surgery. Behavioral voice therapy included indirect strategies (i.e., counseling and advice concerning vocal hygiene, and healthy voice use) as well as direct strategies (i.e., combined exercises on speech-breathing, resonance, pitch, loudness, phonatory facilitation and voice onset) to improve the voice and/or to decrease the number and severity of voice-related complaints. There were 6 patients (i.e., 18.2%) who were treated primarily with surgery,

Table 5.1 Overview of relevant information on the subjects (gender, age and voice disorders) and their voice treatment (number of sessions, interval between the 2 assessments and the type and subtype of therapy).

Subject	Gender	Age at pre-therapy recording (years)	Number of therapy sessions	Interval between pre- and post-therapy (days)	Voice disorder[a,b]	Surgery[c]		Behavioral voice therapy						
						PS	LFS	Indirect	Breathing	Resonance	Pitch	Loudness	Facilitation	Onset
1	M	7	9	104	Bilateral VF nodules			▪	▪	▪	▪	▪		▪
2	F	65	21	221	Left VF paralysis				▪	▪	▪	▪	▪	
3	F	33	3	39	Polipoid mucosa				▪	▪	▪		▪	
4	F	24	1	3	Type 4 MTD					▪	▪		▪	
5	M	46	15	151	Type 2b MTD				▪	▪	▪		▪	▪
6	F	49	5	154	Polipoid mucosa			▪	▪	▪			▪	
7	M	24	16	112	Type 6 MTD						▪		▪	▪
8	F	40	1	47	Left VF cyst	▪		▪						
9	M	16	8	107	Type 6 MTD						▪	▪	▪	
10	F	26	5	35	Polipoid mucosa			▪	▪	▪	▪	▪		
11	M	11	15	252	Bilateral VF nodules			▪	▪	▪	▪		▪	▪
12	M	51	7	133	Left VF paralysis, type 2b MTD		▪			▪	▪	▪	▪	
13	F	63	1	119	Polipoid mucosa	▪		▪						
14	M	45	1	42	Right VF paralysis		▪	▪						
15	F	68	1	1	Type 2b MTD						▪		▪	
16	F	51	2	483	Right VF cyst, type 1 MTD	▪			▪	▪			▪	
17	F	56	5	97	Type 3 MTD					▪	▪		▪	
18	F	58	1	9	Type 3 MTD			▪	▪				▪	
19	M	62	1	10	Type 2a MTD			▪		▪				
20	F	65	49	286	Left VF paralysis				▪	▪	▪	▪	▪	

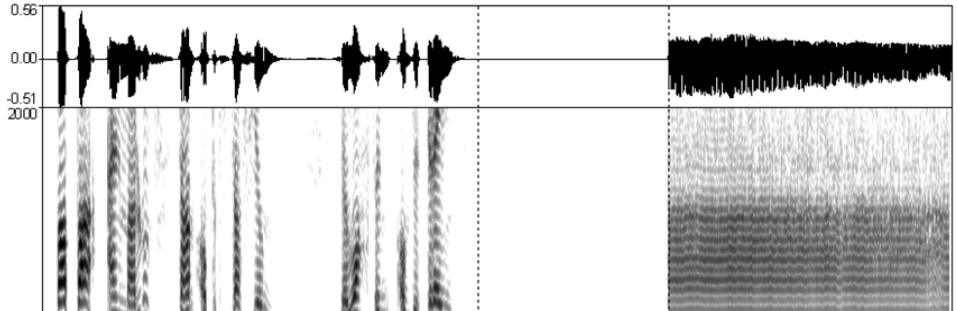
25	F	49	2	9	Type 4 MTD			▪	▪	▪	▪	
26	M	10	12	186	Bilateral VF nodules	▪	▪	▪	▪		▪	▪
27	F	26	8	113	Bilateral VF nodules	▪	▪	▪	▪	▪	▪	
28	F	35	1	126	Bilateral VF nodules	▪						
29	F	49	7	66	Left VF paresis			▪	▪	▪	▪	
30	F	43	1	1	Type 2b MTD			▪			▪	
31	F	23	1	21	Type 4 MTD						▪	
32	M	65	14	38	Left VF paralysis		▪	▪	▪		▪	
33	F	64	14	108	Left VF paralysis		▪	▪	▪	▪	▪	

[a] *VF = vocal fold*

[b] *MTD = muscle tension dysphonia (types of MTD according to Rammage et al., 2001), Type 1 MTD = laryngeal isometry with posterior incomplete adduction, Type 2a MTD = glottal compression or hyperadduction, Type 2b MTD = supraglottal hyperadduction, Type 3 MTD = anteroposterior compression, Type 4 MTD = longitudinal incomplete adduction, Type 6 MTD = transitional falsetto register*

[c] *PS = phonosurgery, LFS = laryngeal framework surgery*

Figure 5.1 Oscillographic display (y-axis: sound pressure in Pascal) and narrowband-spectrogram (y-axis: frequency in Hz; window length = 0.03 s) of a concatenated voice sample, as used for the perceptual evaluations of this study.



including phonosurgery (i.e., PS) in 3 subjects with a vocal fold cyst and 1 subject with severe polypoid degeneration, and laryngeal framework surgery (i.e., LFS) in 2 subjects with unilateral vocal fold paralysis who underwent type I thyroplasty.

The number of sessions of behavioral voice therapy ranged from 1 to 49, with an average of 7.4 sessions. Recordings of continuous speech and sustained vowels were made before and after the voice treatment, with a mean interval of 97.9 days (range 1 to 483 days).

Voice recordings

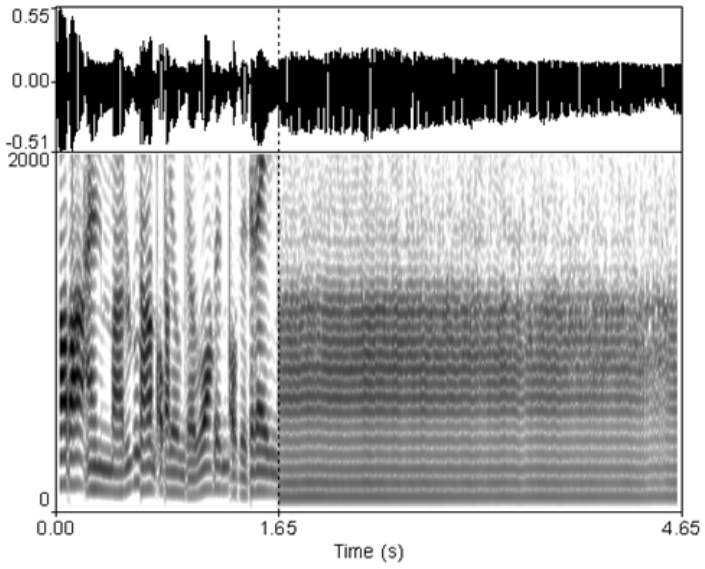
For all 105 recordings (i.e., 39 recordings in experiment 1, 33 pre-treatment recordings in experiment 2 and 33 post-treatment recordings in experiment 2), subjects were asked to sustain the vowel /a/ for at least 5 seconds and to read aloud a Dutch phonetically balanced text (Van de Weijer & Slis, 1991) at comfortable pitch and loudness. Both voice samples were recorded using an AKG C420 head-mounted condenser microphone (AKG Acoustics, München, Germany), digitized at a sampling rate of 44.1 kHz and a resolution of 16 bits using the "Computerized Speech Lab model 4500" (Kay Elemetrics Corp., Lincoln Park, NJ, USA), and saved in WAV-format. For this study, a copy of every vowel sample was edited to include only the central 3 seconds and a copy of the read text samples was trimmed to contain only the first two sentences. In the program 'Praat' (Paul Boersma, Institute of Phonetic Sciences, University of Amsterdam, The Netherlands), the text segment, a pause of two seconds and the vowel segment were concatenated for the perceptual evaluation of both speech types into a single rating. The resulting concatenated waveform is represented in Figure 5.1. Since certain acoustic measures employed in this study are only valid for voiced segments of the continuous speech samples, an algorithm for detection, segmentation, and concatenation of these voiced segments was used. This algorithm was originally based on Parsa & Jamieson (2001, pp. 332) and customized in Praat by Maryn et al. (in press). The resulting waveform is shown in Figure 5.2.

Voice quality ratings

Five experienced speech-language pathologists were asked to rate each of the 105 concatenated voice samples. These raters were the same as in Maryn et al. (in press) for the 2 experiments. With exception of the first author (who collected all recordings), all raters were blinded regarding the identity, diagnosis and disposition of the 72 subjects (i.e. normal, pre-treatment, post-treatment, etc.). There were 2 rating sessions (i.e. 1 per experiment). The listening environment and procedures were comparable to those described in the previous study. All concatenated voice samples were presented in random order and judged on overall severity of dysphonia (Grade, G) with a 4-point equal-appearing interval scale, as suggested by the Japan Society of Logopedics and Phoniatrics (Hirano, 1981).

Before judging the samples, G was described using the definition provided by Kreiman & Gerratt (2000). Furthermore, as recommended by Chan & Yiu (2002), an attempt was made to establish an external standard, ostensibly to increase the reliability of listener ratings. For the purpose of establishing an external standard, twelve samples were selected from the database from the previous study, i.e. three samples per level of G (0 = normal, 1 = slight dysphonia, 2 = moderate dysphonia, 3 = severe dysphonia). These samples were selected based upon prior unanimous agreement across raters regarding the degree of dysphonia, thus these samples were considered to be highly representative of a specific level of G.

Figure 5.2 Oscillographic display (y-axis: sound pressure in Pascal) and narrowband-spectrogram (y-axis: frequency in Hz; window length = 0.03 s) of a concatenated voice sample, as used for the acoustic measures in this study.



Acoustic measures

Objective measurement of overall voice quality consisted of determining the 6 acoustic parameters for calculating AVQI: smoothed cepstral peak prominence (CPPs) with the computer program 'SpeechTool' (James Hillenbrand, Western Michigan University, Kalamazoo, MI, USA) and harmonics-to-noise ratio (HNR), shimmer local, shimmer local dB, general slope of the spectrum (slope) and tilt of the regression line through the spectrum (tilt) with Praat. The method for the determination of the six acoustic measures was identical to the method of Maryn et al. (in press). Consequently, the Acoustic Voice Quality Index (AVQI)

was calculated according to the regression formula: 2.571 [3.295 – 0.111 (CPPs) – 0.073 (HNR) – 0.213 (shimmer local) + 2.789 (shimmer local dB) – 0.032 (slope) + 0.077 (tilt)]. From an initial set of thirteen acoustic measures, this combination of six weighted measures best predicted dysphonia severity; and this formula was constructed on the unstandardized coefficients of the statistical model after stepwise multiple linear regression analysis.

To assess the test-retest reliability of the acoustic analysis protocol, the first author later reanalyzed 20 samples (i.e., >25%) selected randomly. The samples were coded and deidentified prior to randomization and re-analysis. AVQI measures were recalculated and compared to the original analyses. Pearson's correlation (i.e., r_p) was used to estimate acoustic remeasurement reliability and revealed an r_p=0.991 (significant at the p=.000 level), indicating excellent test-retest reliability.

Statistics

All statistical analyses were completed using SPSS for Windows version 12.0 (SPSS Inc., Chicago, Illinois, USA). First, the *inter-rater reliability* of the five raters was investigated using the intraclass correlation coefficient (i.e., ICC). The ICC is a typical reliability index that measures the degree of consistency among ratings. Like other reliability coefficients, the ICC ranges from 0.00 (i.e., total absence of reliability) to 1.00 (i.e., perfect reliability). Although there are no standard values for the interpretation of the ICC, a general guideline suggests that values above 0.75 indicate good reliability, and values below 0.75 poor to moderate reliability. For many clinical measurements, however, reliability should exceed 0.90 to ensure reasonable reliability (Portney & Watkins, 2000). Both the single-measures ICC (i.e., reliability based on comparison of ratings from individual listeners) and the average-measures ICC (i.e., reliability based on averaged ratings) were calculated to estimate inter-rater reliability. Using averaged ratings has the effect of increasing reliability estimates, as means are considered better estimates of true scores, theoretically reducing error variance (Portney & Watkins, 2000, p. 562).

To assess the *external cross-validity* of the AVQI (i.e., how well can the AVQI measure the severity of dysphonia in a new set of clinical voice samples?), the Spearman rank-order correlation coefficient (i.e., r_s) and the coefficient of determination (i.e., r_s^2) between AVQI and mean G (as averaged over the five raters) served to estimate the *criterion-related concurrent validity* of the AVQI. Furthermore, several estimates were calculated to appraise the *diagnostic precision* of the AVQI: how well can AVQI discriminate between normal and dysphonic voices? The diagnostic accuracy of a measure is commonly evaluated by its sensitivity and specificity. Sensitivity is a test's ability to detect a condition or disease when present. In this case, sensitivity would be the percent of correctly identified dsyphonics who test positive on the AVQI. Specificity is an estimate of a test's ability to correctly identify non-cases (i.e, normophonics) when they test

negative on the AVQI. However, depending on the cutoff point chosen to define a positive result, different sensitivity and specificity values can be found. This trade-off between sensitivity and specificity can be examined using a graphic representation called the receiver operating characteristic (ROC) curve. This ROC-curve is created by plotting a point for each cutoff score that represents the true positive score (i.e., sensitivity) on the ordinate and the false positive score (i.e., 1–specificity) on the abscissa. In the present study, a voice was considered normophonic only when all five judges rated it as normal (i.e., mean G=0.0). On the other hand, a voice was considered dysphonic at the moment one judge evaluated it as slightly dysphonic or G1 (i.e., 0.2≤mean G≤3). The ability of the AVQI to discriminate between normal and dysphonic voices was represented by the "area under ROC-curve" (i.e., A_{ROC}). An A_{ROC}=1.0 is found for measures that perfectly distinguish between normal and dysphonic voices. An A_{ROC}=0.5 corresponds with chance-level diagnostic accuracy (Portney & Watkins, 2000). To provide additional evidence regarding the value of a diagnostic measure and to help reduce problems with sensitivity/specificity related to the base-rate differences in the samples (i.e., the uneven number of 6 normophonic and 33 dysphonic subjects in the sample), likelihood ratios should also be calculated (Dollaghan, 2007). The "likelihood ratio for a positive result" (i.e., LR^+) yields information regarding how the odds of the disease increase when the test is positive. It is calculated by LR^+=[(sensitivity)/1-specificity] and gives information regarding the likelihood that an individual is dysphonic when testing positive. The "likelihood ratio for a negative result" (i.e., LR^-) is an estimate that helps to determine if an individual does not have a particular disorder when they test negative on the diagnostic test. It is calculated by LR^-=[(1-sensitivity)/specificity] and gives information regarding the likelihood that an individual has a normal vocal quality when testing negative. As a general guideline, the diagnostic value of a measure is considered to be high when $LR^+\geq 10$ and $LR^-\leq 0.1$. Because the LR statistics consider sensitivity and specificity simultaneously, they are less vulnerable to sample size characteristics and base-rate differences in the sample between vocally normal and voice disordered participants (Dollaghan, 2007). Based on the results of our previous study (Maryn et al., in press), the diagnostic statistics A_{ROC}, LR^+ and LR^- were calculated using an AVQI cutoff point of 2.95.

Third, AVQI's *responsiveness to change* was investigated by means of the "standardized change score". A change score is obtained after subtracting the pre-therapy score from the post-therapy score. However, the auditory-perceptual rating of dysphonia severity (i.e., G) and the acoustic measure (i.e., the AVQI) are based on distinct scales, and thus their change scores cannot be directly compared. A unit free standardized score is necessary to make such a comparison. In each case, for the auditory-perceptual rating and the acoustic measure alike, the standardized change score (i.e., SCS) was obtained by dividing the change score by the standard deviation for the pre-therapy scores (Portney & Watkins, 2000). Subsequently, the change responsiveness of AVQI was analyzed by correlating the standardized change score in mean G (i.e., $SCS_{mean\ G}$) with the standardized change score in

AVQI (i.e., SCS_{AVQI}). The higher r_s observed between $SCS_{mean\ G}$ and SCS_{AVQI}, the more the AVQI is considered to be a responsive treatment outcomes measure which is sensitive to changes in perceived dysphonia severity. Furthermore, the means of $SCS_{mean\text{-}G}$ and SCS_{AVQI} were compared and the difference between the values of these two paired variables was computed with the Student t test (α-level=0.05).

RESULTS

Consistency of the auditory-perceptual ratings

The single-measures ICC of 0.698 (95% confidence interval: 0.609 – 0.779) indicated a moderate *inter-rater reliability* between individual raters. As expected, however, this reliability estimate increased to a very acceptable level with the average-measures ICC of 0.920 (95% confidence interval 0.886 – 0.946), because ratings of multiple listeners have been averaged and error variance decreased before calculating the ICC. These ICC's indicate that the 5 listeners consistently rated the overall severity of dysphonia, and thus the reliability of the G-ratings can be considered acceptable for the purposes of the two experiments in this study.

Table 5.2 Descriptive data of the AVQI values of the 2 experiments in this study.

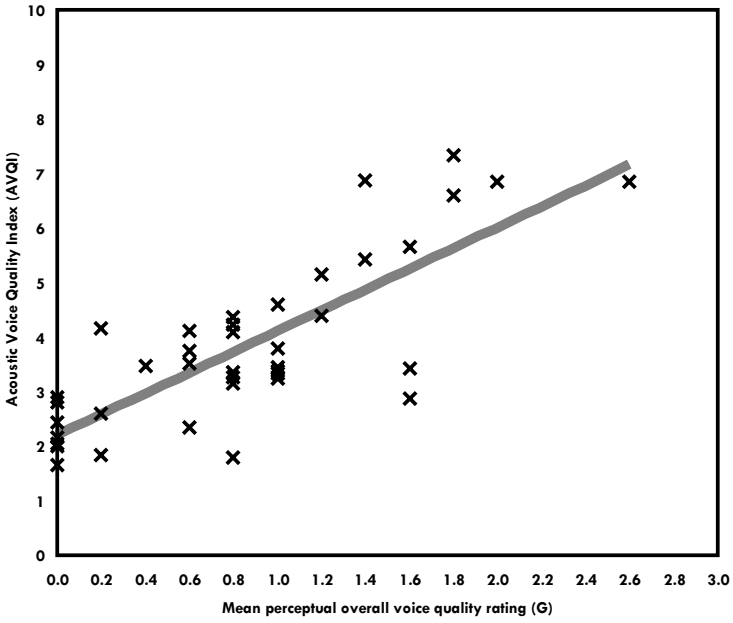
	Experiment 1		Experiment 2	
	Normophonic (N=6)	Dysphonic (N=33)	Pre-therapy (N=33)	Post-therapy (N=33)
Min	1.66	1.78	2.75	0.70
Max	2.89	7.33	10.16	4.07
Range	1.23	5.55	7.41	3.37
Mean	2.29	4.16	5.29	2.79
SD	0.45	1.48	2.09	0.65

Experiment 1: external cross-validity of AVQI

The first 2 columns of Table 5.2 list the descriptive data for AVQI in the group of 6 vocally normal cases and the 33 dysphonic subjects. The first item in the external cross-validation of AVQI was the *criterion-related concurrent validity*. This is expressed as the bivariate correlation of r_s=0.796 between mean G and the AVQI across the 39 subjects with or without various voice disorders. This correlation corresponds with a r_s^2=0.634, designating that more than 60% of the variance of mean G was accounted for by the AVQI, and confirms strong concurrent validity (Portney & Watkins, 2000). The proportional relationship

between mean G and the AVQI is illustrated by the scatterplot and regression line in Figure 5.3.

Figure 5.3 Scatterplot and linear regression line [y=1.9062(x)+2.2049] illustrating the proportional relationship between AVQI and mean G.



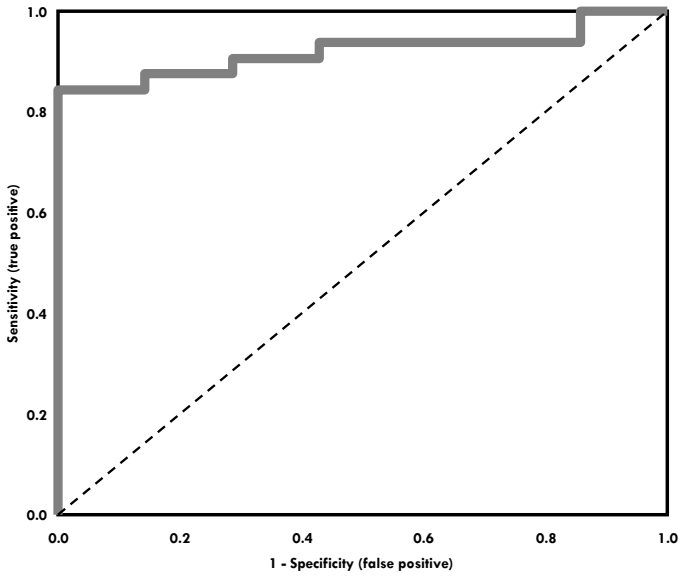
The second issue surrounds the *diagnostic accuracy* of the AVQI. To evaluate the AVQI's ability to distinguish vocally normal from voice disordered participants, a ROC-curve was constructed (see Figure 5.4). The A_{ROC}, with the G-scores as the state variable (i.e., mean G=0 indicating normal voice quality, mean G>0 indicating disrupted voice quality) and the AVQI-scores as the test variable, was 0.920. This result confirms excellent discriminatory power of the AVQI in distinguishing normal from pathological voices (with statistical significance at p=0.000, under the assumption of a nonparametric distribution). Furthermore, an AVQI cutoff score of 2.95 was chosen based on our previous results as a demarcation point between normal and dysphonic voices. In the present study, this cutoff score was associated with a sensitivity=0.85 and specificity=1.00, and an extremely impressive LR^{+}>2000 and a respectable LR^{-}=0.16.

Experiment 2: AVQI's responsiveness to change

The descriptive data of AVQI in the group of 33 dysphonic subjects who received treatment are also listed in the third and fourth column of Table 5.2. To

compare the outcomes before and after voice therapy, a standardized change score [i.e., SCS=(post-therapy score–pre-therapy score)/standard deviation for pre-therapy score] was first calculated from both the mean G data and the AVQI data. The size of the SCS indicated the degree of change in dysphonia severity: the higher the SCS, the greater the change in the voice quality following treatment. The sign of the SCS indicated the direction of the change in overall voice quality: a minus-sign corresponded with improvement or decrease in dysphonia severity, a plus-sign was associated with worsening or increase in dysphonia severity. For example, before voice therapy, subject nr° 2 had a moderate dysphonia, as evidenced by a mean G=1.6 and an AVQI=6.31. However, following voice therapy, a normal voice quality was achieved which resulted in a mean G=0.0 and an AVQI=1.83. The $SCS_{mean\ G}$=–2.48 was reflected in a comparable SCS_{AVQI}=–2.14. Subject nr° 23 on the other hand, showed a slightly increased dysphonia severity after voice therapy, which was reflected in a positively valanced $SCS_{mean\ G}$=0.31 (from mean G of 0.8 to 1.0), and again a comparable SCS_{AVQI}=0.20.

Figure 5.4 ROC-curve illustrating the diagnostic accuracy of AVQI (the dashed line resembles a virtual chance-level discrimination between normophonia and dysphonia).

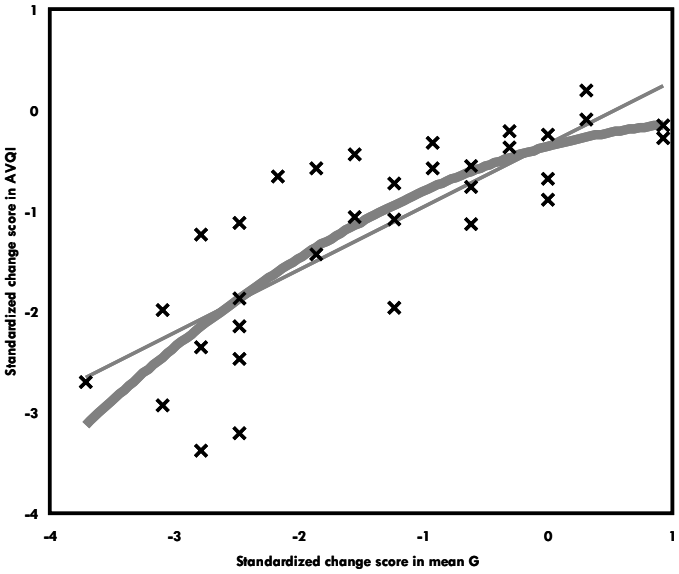


The correlation observed between $SCS_{mean\ G}$ and SCS_{AVQI} showed a strong proportional relationship (r_s=0.80). Figure 5.5 shows a scatterplot of the $SCS_{mean\ G}$ values and the SCS_{AVQI} values for the 33 pre- and post-therapy voice samples. Higher $SCS_{mean\ G}$ values are mostly associated with proportionally higher SCS_{AVQI} values, and vice versa. The linear regression line reflects a r_s^2=0.64. However,

inspection of the scatterplot in Figure 5.5 raised suspicion that a nonlinear relationship might exist. Because the lowest and the highest SCS_{AVQI} scores digressed somewhat to the lower side from what was expected based purely on the linear regression line, we also investigated a second-order polynomial relationship (with r_s^2=0.670). However, there was no statistically significant difference (2-tailed p=0.842) between the outcome of the linear model and the outcome of the curvilinear model.

Additionally, Figure 5.5 demonstrates that the data range for the $SCS_{mean\ G}$ (from –3.71 to 0.93) was very similar to the data range for SCS_{AVQI} (from –3.38 to 0.20). Statistical analysis of these data showed no significant difference (t=1.322; 2-tailed p=0.195), confirming the similarity in the data distributions of $SCS_{mean\ G}$ and SCS_{AVQI}.

Figure 5.5 Scatterplot, linear regression line [y=0.623(x)–0.3425] and curvilinear regression line [y=–0.1106(x²)+0.3332(x)–0.3612] illustrating the degree in which standardized changes in perceived overall voice quality are predicted by standardized changed in AVQI scores.



DISCUSSION

To improve ecological validity, assessment of dysphonia severity and tracking of treatment outcomes should employ procedures which survey both sustained vowels and continuous speech (Hammarberg et al., 1980; Parsa & Jamieson, 2001; Yiu et al., 2000; Zraick et al., 2005). To our knowledge, Maryn et al. (in press) were the first to construct a multivariable model based on acoustic

measures derived from both speaking tasks. This 'Acoustic Voice Quality Index' (a.k.a. 'AVQI') model incorporated the middle 3 seconds of a sustained /a/ and the first 2 sentences of a phonetically balanced Dutch text and concatenated them into 1 single sound file, upon which the following 6 acoustic measures were determined. Measures of harmonic versus noise energy (as 'harmonics-to-noise ratio' in the present study) and spectral tilt (as 'tilt of the trend line through the long-term average spectrum' and 'slope of the long-term average spectrum' in the present study) have traditionally been associated with insufficient glottal closure and breathiness (Sodersten & Lindestad, 1990; Awan & Roy, 2006). Measures of amplitude perturbation (as 'shimmer local' and 'shimmer local dB' in the present study) have classically been related to irregular vocal fold vibration and roughness (Awan & Roy, 2006). Although cepstral measures (as 'smoothed cepstral peak prominence' in the present study) have mainly been associated with breathiness and less with roughness (Hillenbrand et al., 1994; Hillenbrand & Houde, 1996; Heman-Ackah et al., 2002), it is assumed that all factors causing deviations in the voice signal decrease the prominence of the first rahmonic (Ferrer, de Bodt, Maryn, Van de Heyning and Hernández-Diaz, 2007). Consequently, the combination of these 6 parameters subsumes several measures sensitive to potential vibratory distortions at the level of the glottis, and would serve as a metric of overall dysphonia severity.

Although Maryn et al. (in press) concluded that their results supported the feasibility, predictive validity and diagnostic accuracy of this multiparameter model, confirmation of these results through replication was needed to support the generalizability of AVQI. We therefore investigated the *external cross-validity* of AVQI using a new set of 39 normophonic and/or dysphonic subjects. While the initial study (Maryn et al., in press) reported respectable concurrent validity (i.e., r_s=0.78 between AVQI and mean G) and diagnostic precision (i.e., A_{ROC}=0.90, sensitivity=0.74, specificity=0.96, LR^+=19.98, LR^-=0.27), the present study provided the requisite external cross-validation of the AVQI by essentially replicating the results of the original study, but this time using a different set of voices. The r_s=0.796 confirmed a strong relationship between the AVQI and mean G (see Figure 5.3), and substantiated the feasibility of AVQI as a measure sensitive to the continuum of dysphonia severity. Furthermore, as evidenced by impressive estimates of diagnostic precision (i.e., A_{ROC}=0.920, sensitivity=0.85, specificity=1.00, LR^+>2000, LR^-=0.16) the AVQI offers excellent discriminatory power, sufficient to distinguish normal from pathological voices. These statistics related to concurrent validity and diagnostic accuracy are comparable and in some cases superior to outcomes reported in the original study. This highlights the robustness of the AVQI, as a clinical measure of dysphonia severity and supports its external validity.

In order to serve as a valid treatment outcomes measure however, the AVQI needs to be sensitive to different degrees of change in dysphonia severity. Changes in perceived dysphonia severity (i.e., in 'G') should be reflected in proportionally equivalent changes in predicted voice quality (i.e., in AVQI scores).

In a second experiment, we therefore evaluated the *responsiveness to change* of the AVQI using pre and post-treatment voice samples from 33 voice-disordered patients who underwent surgical and/or individually customized behavioral voice treatment (Table 5.1). To compare change scores of 2 variables with different units/scales, such as mean G and AVQI, change scores were normalized over the standard deviation for the pre-therapy scores, resulting in unit free standardized change scores (i.e., SCS). Results indicated a strong correlation of r_s=0.80, wherein 64% of the variance in $SCS_{mean\ G}$ was accounted for by SCS_{AVQI}. These results support the sensitivity of the AVQI to treatment-related changes in dysphonia severity, and thus it should be considered as a valid treatment outcomes measure in voice.

In a very recent study, Awan & Roy (2009) also investigated the external cross-validity and the responsiveness to change of a 4-factor acoustic model for the prediction of dysphonia severity in sustained vowels, developed previously by the same authors in 2006. Interestingly, this model was constructed independently from our model, yet consisted of a similar set of acoustic parameters as AVQI. Both models contain measures of shimmer: "shimmer local" and "shimmer local dB" in AVQI and "shimmer (dB)" in Awan & Roy (2006, 2009). Correlations between shimmer measures and ratings of overall severity of dysphonia or hoarseness have indicated poor to moderately strong associations: r=0.54 (Wolfe et al., 1995a), r_s=0.41 (De Bodt, 1997), and r=0.70 Halberstam (2004). Both models also contain measures of spectral energy distribution: "slope of long-term average spectrum" and "tilt of trend line through long-term average spectrum" in AVQI and "discrete Fourier transform ratio" in the model of Awan & Roy (2006, 2009). Poor to moderate correlations have been found between similar spectral measures and scores of hoarseness or overall dysphonia severity: r=0.52 (Dejonckere & Wieneke, 1996), and r=–0.47 (Eadie & Doyle, 2005). Most importantly however, both models also contained a measure of the cepstral peak as the most dominant feature: "smoothed cepstral peak prominence" in AVQI (Maryn et al., in press) and "ratio of the actual amplitude of the cepstral peak prominence to the expected amplitude" in the model of Awan & Roy (2006, 2009). Several reports exist in the literature emphasizing that measures of the relative magnitude of the first rhamonic (i.e., the first cepstral peak) are promising correlates of dysphonia severity. Cepstral measures, in particular, have received much attention since Hillenbrand and his colleagues reported that the "cepstral peak prominence" (i.e., CPP) and the "smoothed cepstral peak prominence" (i.e., CPPs) were the strongest acoustic correlates of breathiness in sustained vowels (r=–0.92; Hillenbrand, Cleveland and Erickson, 1994) as well as in continuous speech (r=–0.88; Hillenbrand & Houde, 1996). Their results have been confirmed by many other reports. Dejonckere & Wieneke (1996), for example, reported that the "cepstral magnitude" was the best predictor of hoarseness in sustained vowels (r=–0.80), compared to spectral (such as the "spectral noise above and under 6000 Hz") and perturbation measures (such as the "jitter ratio"). Heman-Ackah, Michael and Goding (2002) also found that the CPPs was strongly related to overall voice quality in sustained vowels (r_p=–0.80)

and especially in continuous speech (r_p=–0.86), compared to the more traditional perturbation measures (for example "relative average perturbation" and "amplitude perturbation quotient"). Halberstam (2004) also reported very strong correlations of r_p=–0.90 and r_p=–0.88 between the degree of hoarseness and the respective CPP and CPPs measures on continuous speech. As in Heman-Ackah et al. (2002), these cepstral measures demonstrated superior validity in continuous speech, and consequently proved to be the most ecologically valid measures. Eadie & Baylor (2006) reported r_s=0.79 between CPP and overall dysphonia severity in continuous speech and r_s=0.81 between CPPs and overall dysphonia severity on sustained vowels. On the other hand, only Wolfe & Martin (1997) and Halberstam (2004) reported poor correlations between hoarseness and CPP (r_p=–0.30) and between hoarseness and CPPs (r_p=–0.55) on sustained vowel samples, respectively. Collectively, these findings together with the results from Maryn et al. (in press) and Awan & Roy (2006, 2009) confirm the power of the cepstrum-based parameters to predict the degree of overall dysphonia, and emphasize the superiority of these measures as compared to more traditional time-domain perturbation measures.

In some ways, it comes as no surprise that we found highly comparable outcomes for the 2 models – in terms of initial validity, internal cross-validity, external cross-validity and sensitivity to change. The initial validity of the model of Awan & Roy (2006) indicated a strong r=0.88, whereas Maryn et al. (in press) found with r=0.78 a slightly lower initial validity. The internal cross-validation based on repeated correlations in different randomized subgroups revealed a strong mean r=0.88 in Awan & Roy (2006) and a less strong mean r=0.77 in Maryn et al. (in press). The external cross-validition of the models in a new set of voice samples showed a very strong r=0.91 in Awan & Roy (2009) and a strong r_s=0.80 for the AVQI. These internal and external cross-validations, and the fact that acceptable results were found for the both models, confirm the robustness and stability of this multivariable approach to the measurement of overall voice quality. Finally, the responsiveness to change was investigated for both models. Awan & Roy (2009) found a strong r=0.75 between the unstandardized change scores (i.e., post-therapy score minus pre-therapy score) of the perceived and the acoustically predicted dysphonia severity. The present study examined the standardized change scores of the perceptual ratings and the AVQI-data and indicated a strong r_s=0.80 (Figure 5.5). Since both studies incorporated a continuum in voice quality change (i.e., from absence of change to dramatic change), it can be concluded that both models are acceptably responsive/sensitive to voice changes after voice treatment.

Based on these results, we conclude that the AVQI possesses respectable concurrent validity, concurrent cross-validity, diagnostic precision, and responsiveness to change. All of these are desirable attributes for an objective treatment outcomes measure. When one considers that the AVQI incorporates not only sustained vowels, but also continuous speech, it also possesses another desirable attribute "ecological validity". Furthermore, because the AVQI can be computed using computer programs freely available as in Praat and SpeechTool, it

has the additional appeal of easy access and affordability. Collectively, as clinicians, third party payers, and administrators demand objective evidence of positive or negative treatment outcomes, the results of this investigation confirm that the AVQI deserves further attention as a promising objective treatment outcomes measure that could be included as an important part of a multidimensional assessment of treatment effects.

ACKNOWLEDGMENTS

The authors sincerely appreciate Dr. James Hillenbrand (from the Western Michigan University, Kalamazoo, MI, U.S.A.) for providing the SpeechTool software for the CPP and CPPs measures. The authors also thank Dr. Gwen Van Nuffelen, Dr. Bernadette Timmermans and Fons Mertens for their contributions in the perceptual rating of the many concatenated voice samples.

REFERENCES

Awan, S.N., & Roy, N. (2006). Toward the development of an objective index of dysphonia severity: a four-factor model. *Clinical Linguistics & Phonetics, 20,* 35-49.

Awan, S.N., & Roy, N. (2009). Outcomes measurement in voice disorders: application of an acoustic index of dysphonia severity. *Journal of Speech, Language, and Hearing Research, 52,* 482-499.

Buder, E.H. (2000). *Acoustic analysis of voice quality: a tabulation of algorithms 1902-1990.* In R.D. Kent and M.J. Ball (Eds.), Voice quality measurement (pp. 119-244). San Diego, CA: Singular Publishing Group.

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

Chan, K.M., & Yiu, E.M. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language, and Hearing Research, 45,* 111-126.

Davis, S.B. (1979). *Acoustic characteristics of normal and pathological voices.* In N.J. Lass (Ed.), Speech and language: advances in basic research and practice, volume 1 (pp. 271-335). New York: Academic Press.

De Bodt, M. (1997). *A framework of voice assessment: the relation between subjective and objective parameters in the judgement of normal and pathological voice.* University of Antwerp, Antwerp: unpublished doctoral dissertation.

Dejonckere, P.H., & Wieneke, G.H. (1996). *Cepstra of normal and pathological voices: correlation with acoustic, aerodynamic and perceptual data.* In M.J. Ball and M. Duckworth (Eds.). Advances in Clinical Phonetics (pp. 217-226). Amsterdam: John Benjamins Publishing Co.

Dollaghan, C.A. (2007). *The handbook for evidence-based practice in communication disorders*. Baltimore: MD Brookes.

Eadie, T.L., & Doyle, P.C. (2005). Classification of dysphonic voice: acoustic and auditory-perceptual measures. *Journal of Voice, 19,* 1-14.

Eadie, T.L., & Baylor, C.R. (2006). The effect of perceptual training on inexperienced listeners' judgments of dysphonic voice. *Journal of Voice, 20,* 527-544.

Eskenazi, L., Childers, D.G., & Hicks, D.M. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research, 33,* 298-306.

Ferrer, C.A., de Bodt, M.S., Maryn, Y., Van de Heyning, P., Hernández-Díaz, M.E. (2007). Properties of the cepstral peak prominence and its usefulness in vocal quality measurements. Proceedings of MAVEBA'07, Florence, 93-96.

Frey, L.R., Botan, C.H., Friedman, P.G., & Kreps, G.L. (1991). *Investigating communication, an introduction to research methods.* Englewood Cliffs: Prentice-Hall.

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,* 70-73.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngologica, 90,* 441-451.

Heman-Ackah, Y.D., Michael, D.D., & Goding, G.S. (2002). The relationship between cepstral peak prominence and selected parameters of dysphonia. *Journal of Voice, 16,* 20-27.

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

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

Jin, B.J., Lee, Y.S., Jeong, S.W., Jeong, J.H., Lee, S.H., & Tae, K. (2008). Change of acoustic parameters before and after treatment in laryngopharyngeal reflux patients. *Laryngoscope, 118,* 938-941.

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

Kreiman, J., & Gerratt, B. (2005). Perception of aperiodicity in pathological voice. *Journal of the Acoustical Society of America, 117,* 2201-2211.

Ma, E., & Yiu, E. (2006). Multiparametric evaluation of dysphonic severity. *Journal of Voice, 20,* 380-390.

Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N., & De Bodt, M. (in press). Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *Journal of Voice*.

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

Parsa, V., & Jamieson, D.G. (2001). Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. *Journal of Speech, Language, and Hearing Research, 44*, 327-339.

Plant, R.L., Hillel, A.D., & Waugh, P.F. (1997). Analysis of voice changes after thyroplasty using linear predictive coding. *Laryngoscope, 107*, 703-709.

Portney, L.G., & Watkins, M.P. (2000). *Foundations of clinical research, Applications to practice (2nd Ed.)*. Upper Saddle River: Prentice-Hall.

Rammage, L., Morrison, M., Nichol, H., Pullan, B., Salkeld, L., & May, P. (2001). *Management of the voice and its disorders (2nd Ed.)*. San Diego: Singular Thomson Learning.

Robey, R.R., & Dalebout, S.D. (1998). A tutorial on conducting meta-analyses of clinical outcome research. *Journal of Speech, Language, and Hearing Research, 41*, 1227-1241.

Roy, N., Gouse, M., Mauszycki, S.C., Merrill, R.M., & Smith, M.E. (2005). Task specificity in adductor spasmodic dysphonia versus muscle tension dysphonia. *Laryngoscope, 115*, 311-316.

Sodersten, M., & Lindestad, P.Ä. (1990). Glottal closure and perceived breathiness during phonation in normally speaking subjects. *Journal of Speech and Hearing Research, 33*, 601-611.

Titze, I.R. (1995). *Workshop on acoustic voice analysis: summary statement*. Iowa City: National Center for Voice and Speech.

Van de Weijer, J.C., & Slis, I.H. (1991). Nasaliteitsmeting met de nasometer. *Tijdschrift voor Logopedie en Foniatrie, 63*, 97-101.

Wolfe, V., Cornell, R., & Fitch, J. (1995a). Sentence/vowel correlation in the evaluation of dysphonia. *Journal of Voice, 9*, 297-303.

Wolfe, V., Fitch, J., & Cornell, R. (1995b). Acoustic prediction of severity in commonly occurring voice problems. *Journal of Speech and Hearing Research, 38*, 273-279.

Wolfe, V., Fitch, J., & Martin, D. (1997). Acoustic measures of dysphonic severity across and within voice types. *Folia Phoniatrica et Logopaedica, 49*, 292-299.

Wolfe, V., & Martin, D. (1997). Acoustic correlates of dysphonia: type en severity. *Journal of Communication Disorders, 30*, 403-416.

Wuyts, F.L., De Bodt, M.S., Molenberghs, G., Remacle, M., Heylen, L., Millet, B., Van Lierde, K., Raes, J., and Van de Heyning, P.H. (2000). The Dysphonia Severity Index: an objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language, and Hearing Research, 43*, 796-809.

Yiu, E., Worrall, L., Longland, J., & Mitchell, C. (2000). Analysing vocal quality of connected speech using Kay's computerized speech lab: a preliminary finding. *Clinical Linguistics & Phonetics, 14*, 295-305.

Yu, P., Ouaknine, M., Revis, J., & Giovanni, A. (2001). Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements. *Journal of Voice, 15,* 529-542.

Zraick, R.I., Wendel, K., & Smith-Olinde, L. (2005). The effect of speaking task on perceptual judgment of the severity of dysphonic voice. *Journal of Voice, 19,* 574-581.

6

SPECTRAL, CEPSTRAL AND MULTIVARIATE EXPLORATION OF TRACHEOESOPHAGEAL VOICE QUALITY IN CONTINUOUS SPEECH AND SUSTAINED VOWELS

Youri Maryn
Catherine Dick
Caroline Vandenbruaene
Tom Vauterin
Tinne Jacobs

This chapter has been published in:
Laryngoscope, 2009, Epub ahead of print.

ABSTRACT

Objectives: the quality of tracheoesophageal voice can vary substantially. Although previous research has identified acoustic differences between various types of voicing (i.e., laryngeal, tracheoesophageal, esophageal, etc.), acoustic analysis has failed to quantify the degree of alaryngeal voice quality. This study assessed the value of several cepstral, spectral, and perturbation measures in quantifying the overall quality of tracheoesophageal voice production.

Study Design: cross-sectional, correlational.

Methods: continuous speech and sustained vowel samples from 16 tracheoeso-phageal speakers were concatenated and perceptually rated in a paired comparison paradigm on overall voice quality by 4 experienced clinicians. After removing the non-voiced fragments within the continuous speech samples, the concatenated samples were analyzed with 47 perturbation, spectral and cepstral measures. Correlation between perceptual ratings and acoustic measures was assessed. Multiple regression analysis resulted in a 2-factor acoustic model for the measurement of overall voice quality of the concatenated samples.

Results: the reliability of the perceptual judgements was moderate to high. The prominence of the cepstral peak (CPP) and of the first 2 spectral harmonics appeared to be the strongest correlates of tracheoesophageal voice quality. A linear regression-based combination of CPP and the height of the second harmonic produced a correlation of 0.87 with listener judgments.

Conclusion: it is clinically feasible to investigate both continuous speech and sustained vowel samples of tracheoesophageal speakers with acoustic methods described and assessed in this report. Results are discussed in the context of existing literature.

INTRODUCTION

Traditionally, there are 3 options for voice rehabilitation following total laryngectomy: esophageal voice, voice with an artificial larynx and tracheoesophageal voice. For esophageal voice production it is important that air from the mouth and the pharynx is first insufflated (inhaled or injected) into the esophagus. When the air then is forced upwards to the pharynx, it causes audible vibration in the pharyngeal-esophageal segment (a.k.a. neoglottis) upon which speech sounds can be produced (Ward & van As, 2007). Voice with an artificial larynx (e.g. an electrolarynx) requires that a pneumatically or electronically generated tone is transmitted through tissue surrounding the vocal tract to set the air within the vocal tract into vibration, again upon which speech sounds can be produced (Ward & van As, 2007). Tracheoesophageal voice requires a fistula between the trachea and esophagus, enabling expiratory air to be diverted into the esophagus when the tracheostoma is occluded, and to produce audible vibration in the pharyngeal-esophageal segment (Ward & van As, 2007). Differences in voice quality have been found between and within these 3 methods (Merwin et al., 1985;

Crevier-Buchmann et al., 1991; Ward & van As, 2007), mostly depending on the efficiency of the expiratory airflow and the vibratory characteristics of the pharyngeal-esophageal segment (Debruyne et al., 1994; Bertino et al., 1996). Currently, tracheoesophageal voice is the most frequently used method for vocal rehabilitation after a total laryngectomy (Ward & van As, 1997; Lundström et al., 2008).

There has been considerable research interest in describing the perceptual and acoustic characteristics of alaryngeal voice production. Similar to the study of normal and disordered voice in laryngeal speakers, acoustic parameters have been used in research and in the clinic to: (1) discriminate between different types of voice production (laryngeal-vocal, esophageal, tracheoesophageal and electrolaryngeal), (2) quantify the degree and desirability/acceptability of the alaryngeal speaker's voice, and (3) objectively monitor progress during therapy (Robbins, 1984; Debruyne et al., 1994; Bertino et al., 1996; van As et al., 1998; Arias et al., 2000; Moerman et al., 2004; Kazi et al., 2006; Štajner-Katušić et al., 2006; MacCallum et al., 2008). In this regard, acoustic measurement of post-laryngectomy voice quality typically employed measures of fundamental frequency, intensity, fundamental frequency perturbation (e.g. percent jitter, jitter ratio, etc.), amplitude perturbation (e.g. absolute shimmer, directional shimmer, etc.), spectral noise (e.g. harmonic-to-noise ratio), vibratory irregularities (e.g. unvoiced segments, subharmonic components, etc.) and nonlinear dynamic properties (e.g. second-order entropy). Although acoustic measures often correctly distinguished between different types of post-laryngectomy voice production (MacCallum et al., 2008), studies investigating the association between acoustic measures and auditory-perceptual ratings of laryngectomees' overall voice quality have yielded rather unsatisfactory correlations. For example, Bertino et al. (1996) examined the correlations between 2 acoustic measures (fundamental frequency and harmonics-to-noise ratio) and the acceptability of 2 seconds of a sustained vowel obtained from 8 tracheoesophageal and 10 esophageal. They found r=0.59 for fundamental frequency and r=0.63 for harmonics-to-noise ratio. van As et al. (1998) investigated sustained vowels of 21 tracheoesophageal speakers with 29 different acoustic voice parameters from the Multi-Dimensional Voice Program and the correlations of these parameters with several auditory-perceptual dimensions (for example steady-unsteady, clear-dull, deep-shrill, normal-abnormal, etc.). The perceptual dimension "normal vs. abnormal" was most significantly associated with the following acoustic parameters noise-to-harmonics ratio (r=0.63), coefficient of variation of fundamental frequency (r=0.64) and degree of unvoiced segments (r=0.67). For the other acoustic measures however, no correlation exceeding the 0.60 level was found. In a more recent study, Moerman et al. (2004) studied the association between 7 acoustic parameters (proportion of voiced frames, proportion of voiced speech frames, average voicing evidence of the voiced frames, 90th percentile of the voicing length distribution, jitter, corrected jitter and percentage of frames with unreliable fundamental frequency) and the overall impression of 53 tracheoesophageal, 14 esophageal, 5 hemilaryngectomized

and 6 normal speakers. They used recordings of sustained vowels as well as different syllables and continuous speech samples. None of the correlations exceeded the r=0.50 level, and the lowest absolute correlations were found for the jitter and corrected jitter measures. Based on these results, it appears that traditional, time-based acoustic measures that rely upon accurate identification of cycle boundaries in their calculation (e.g. jitter), are rather insensitive to different levels of overall voice quality disruption in the speech of laryngectomized individuals.

The poor performance of time-based, perturbation measures has been substantiated by a vast body of research aimed at establishing acoustic correlates of dysphonia severity in laryngeal speakers as well (Maryn et al., in press). In contrast, spectral or cepstrum-based measures which do not rely on cycle boundary identification have demonstrated more promising results with superior correlations between measures like cepstral peak prominence and auditory-perceptual ratings of overall voice quality or dysphonia severity (Maryn et al., in press). In this regard, it can be assumed that the more the Fourier spectrum of the laryngectomized speaker resembles the spectrum of a vocally normal speaker (in terms of prominence of the fundamental harmonic and the higher harmonics, relative to the interharmonic or noise level), the better the laryngectomee's voice quality will be. For example, Dejonckere & Lebacq (1987) investigated the harmonic emergence in vocally normal and dysphonic subjects and found it suitable for the measurement of dysphonia. Debruyne et al. (1994) adopted this measure of harmonic emergence and applied it to recordings of sustained vowels produced by 12 esophageal and 12 tracheoesophageal speakers. However, they did not report inferential statistics with regard to voice quality.

Another consideration in the acoustic analysis of laryngeal or alaryngeal voice production is to determine the type of voice context/sample to be analyzed. Acoustic measures are most frequently derived from sustained vowel samples, in contrast to continuous speech samples. Sustained vowels are indeed attractive for a variety of reasons (i.e. they are time-invariant, unaffected by intonation, easy to elicit, etc.), but relying solely on a sustained vowel does not provide the most representative voice assessment. Continuous speech, for example reading a standardized text, is less artificial and provides voice samples that are more representative of daily speech, and is ostensibly more ecologically valid. So it is essential for perceptual and acoustic analyses to be based upon both sustained vowels and continuous speech samples (Maryn et al., in press).

To our knowledge, the voice and the degree of voice quality in laryngectomees has never been analysed with measures of cepstral peak prominence and spectral peak prominence. Furthermore, with the exception for Moerman et al. (2004), alaryngeal voice quality has never been investigated using continuous speech, a putatively more representative context. The present exploratory investigation therefore examined the clinical utility and the validity of 2 measures of the cepstral peak and 36 measures of geometrical spectral properties for assessing overall voice quality in both sustained vowels and continuous speech

of 16 tracheoesophageal speakers. This study secondarily also explored the validity of the more traditional perturbation and noise measures, such as jitter, shimmer and noise-to-harmonics ratio. Therefore, the following research questions were addressed: (1) Is it feasible to employ continuous speech as well as sustained vowels in the acoustic measurement of tracheoesophageal voice quality? (2) Are measures of cepstral peak prominence, spectral peak geometry and perturbation of the neoglottal signal valid measures of overall voice quality of tracheoesophageal speakers?

METHODS

Subjects

Voice samples were provided by 2 female and 14 male laryngectomees (with or without concomitant radiation therapy) currently using tracheoesophageal voice. Fourteen patients were treated with total laryngectomy and postoperative radiation. Two patients underwent total laryngectomy as a secondary treatment after radiotherapy, and in the 2 remaining patients only a total laryngectomy was performed. No cricopharyngeal myotomy was performed in these patients. The ages at the time of laryngectomy, ranged from 45 to 76 years, with a mean of 60.2 years. At the time of the voice recordings mean age was 62.4 years. There was a mean period of 26 months (ranging between 1 and 86 months) between surgery and the recordings. All patients underwent a primary tracheoesophageal puncture with implantation of a Provox® voice prosthesis. In 14 patients, the stoma was digitally occluded by pressing on a Provox® normal or high-flow HME stomafilter. One patient used a Provox® Free-Hands stomafilter, and 1 patient directly digitally closed the stoma.

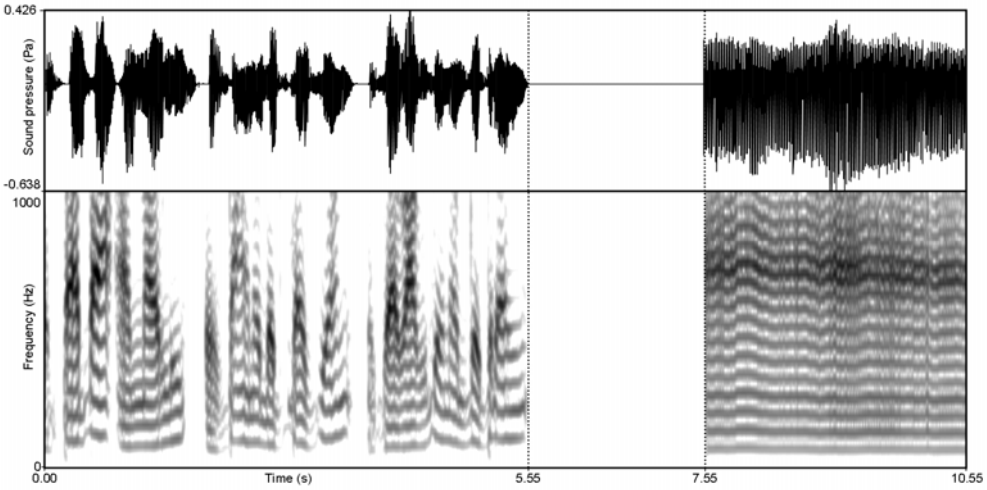
Voice samples

All participants were asked to sustain the vowel /a/ for at least 5 seconds and to read aloud a phonetically balanced text using a comfortable pitch and loudness. Both voice samples were recorded using an AKG C420 head-mounted condenser microphone (AKG Acoustics Harman Pro, München, Germany) and digitized at 44,100 samples per second and 16 bits of resolution using the Computerized Speech Lab model 4500 (Kay Elemetrics Corporation, currently known as KayPentax, Lincoln Park, USA). The samples were immediately saved in .wav format. The sustained vowel samples were trimmed to include only the middle 3 seconds. The continuous speech samples (read text) were edited to include only the first two sentences.

Auditory-perceptual evaluation

For the perceptual ratings, the voice samples were concatenated in the following order using "Praat" (computer program developed by Paul Boersma at the Institute of Phonetic Sciences, Amsterdam, The Netherlands): the text segment followed by a pause of 2 seconds and then the 3 second sustained vowel segment. An example of the resulting concatenated waveform is given in Figure 6.1.

Figure 6.1 Oscillogram and narrowband-spectrogram (window length = 0.07 s) of a concatenated voice sample (derived from subject 8), as used for the perceptual evaluations of this study. The left area corresponds with the first two sentences of the 'Papa en Marloes' text. The right portion reflects the middle three seconds of a sustained /a/. The area in the middle reflects two seconds of silence.



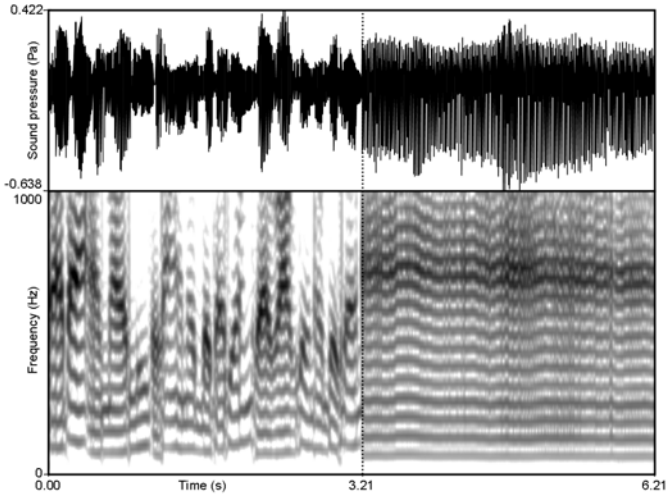
Four listeners (2 laryngologists and 2 speech-language pathologists) with at least 5 years experience in working with laryngectomees participated in a paired comparison task. The paired comparison paradigm has proven to be a reliable method for perceptual voice quality evaluation purposes (Kacha et al., 2005; Yiu et al., 2007; Kreiman et al., 2008). In this task, all listeners had to compare perceptually the overall voice quality (similar to 'Grade' or 'G') of every tracheoesophageal voice sample with the other 15 samples. In order to avoid comparison of 2 identical samples and double comparison of 2 identical (however reversed) sample pairs, the number of pairs to be compared (N) was multiplied with N-1 and divided by 2, respectively. This resulted in a total of 120 pairs that were randomly presented for the auditory-perceptual comparisons. Every time a sample was judged to have the best overall voice quality, it acquired 1 point. The sample with the worst overall voice quality received 0 points. When 2 samples were judged to have equal overall voice quality, they both acquired 0.5 point. To orient the listeners to the rating task (Chan & Yiu, 2002), listerners rated three pairs

of voice samples similar to the experimental samples, however these samples were not included in the actual listening experiment. Once all comparisons were made, the points per voice sample were tallied. The higher the total number of points (i.e., the more a voice sample was judged to have better overall voice quality as compared to another voice sample), the better the overall voice quality of a tracheoesophageal voice sample was relative to the other voice samples in this study. The result of this paired comparison task was a ranking of the 16 voice samples, from worst to best overall voice quality. At the end of the perceptual experiment, 24 randomized pairs of tracheoesophageal voice samples (i.e. 20 % of all pairs) were repeated a second time in order to determine intra-rater reliability.

Acoustic measures

Before computing the acoustic analysis, a custom-made voicing detection algorithm was used to extract and concatenate the voiced segments from the continuous speech files. The specifications and programming script of this algorithm, as implemented in the program "Praat", are described in Maryn et al. (in press). After this, the voiced continuous speech samples were concatenated with the sustained vowel samples. The resulting waveform of all these actions is represented in Figure 6.2. In total, 47 acoustic measures were computed.

Figure 6.2 Oscillogram and narrowband-spectrogram (window length = 0.07 s) of a concatenated voice sample (derived from subject 8), as used for the acoustic measures in this study. The left area corresponds with the first two sentences of the 'Papa en Marloes' text. The right portion reflects the middle three seconds of a sustained /a/.



First, the concatenated tracheoesophageal voice recordings were analyzed on the basis of the cepstrum. To create a cepstrum, a Fourier transformation of a complex acoustic waveform (i.e. the oscillogram composed of simple sine waves) is first executed to create a spectrum. This means transitioning from Figure 6.3.A (time-domain) to Figure 6.3.B (frequency-domain). Executing a new Fourier transformation of the spectrum, as if this spectrum itself were a complex waveform, then creates the unsmoothed cepstrum. This is essentially a Fourier transform of a Fourier transform. The result is transitioning from Figure 6.3.B (i.e., frequency-domain) to Figure 6.3.C (i.e., quefrency- or 1/frequency-domain). The cepstrum is often described in relation to the dominant peak displayed on the x-axis (quefrency axis). The amplitude of this peak, also known as the dominant rahmonic or cepstral peak, reflects the strength of the fundamental frequency of the voice as it emerges out of the background competing frequencies. Hillenbrand & Houde (1996) developed an automated computer program to derive the unsmoothed cepstrum from the oscillogram, called "SpeechTool" (James Hillenbrand, Western Michigan University, Kalamazoo, MI, USA). However, a simple modification in the computer algorithm produced an appreciable methodological improvement. This modification averaged across time and then across quefrency, resulted in a smoothed cepstrum, as illustrated in Figure 6.3.D. The rationale behind the cepstrum, as a domain for voice quality assessment, is that a highly periodic signal shows a well-defined harmonic structure in the spectrum and, subsequently, a more prominent cepstral peak as compared to a less periodic signal. To define the cepstral peak prominence (CPP), a linear regression line is fitted through the unsmoothed cepstrum and the difference in amplitude between the cepstral peak and the corresponding point on the regression line is calculated. The smoothed CPP (a.k.a. CPPs) implies the same action in the smoothed cepstrum. Both measures are illustrated in Figure 6.4. A detailed outline of the calculation of CPP and CPPs can be found in Hillenbrand & Houde (1996).

Second, 36 specific geometric properties of the presumed harmonic peaks in the long-term average spectrum (LTAS, i.e., an averaging of the spectral energy over a window with a specified duration) were measured. The following steps were undertaken in "Praat" via a custom-made programming script to straightforwardly mark the first 3 harmonics in the LTAS of tracheoesophageal voice samples:

a. Derive a LTAS with frequency steps of 1 Hz from the voiced concatenated waveform.
b. Indicate the maximum sound pressure level in the frequency range between 0 and 300 Hz. This frequency interval was chosen because it was expected to be the zone in which the first (a.k.a. fundamental) harmonic (H_1) can definitely be found (Debruyne et al., 1994; van As et al., 1998; Arias et al., 2000; Kazi et al., 2006; Lundström et al., 2008) and the maximum sound pressure level in this zone was considered to be the peak of the first harmonic ($H_{1\text{-}MAX}$).
c. Fix the ultimate frequency range in which the minima of the first harmonic can be found. The lower end of this range is typically defined by: $H_{1\text{-}MAX}/2$. The upper end of this range is typically defined by: $H_{1\text{-}MAX}+(H_{1\text{-}MAX}/2)$.

Figure 6.3 From oscillogram (or waveform) to smoothed cepstrum, for illustrative purposes derived from a sustained /a/ of 0.2 seconds produced by a normophonic male.

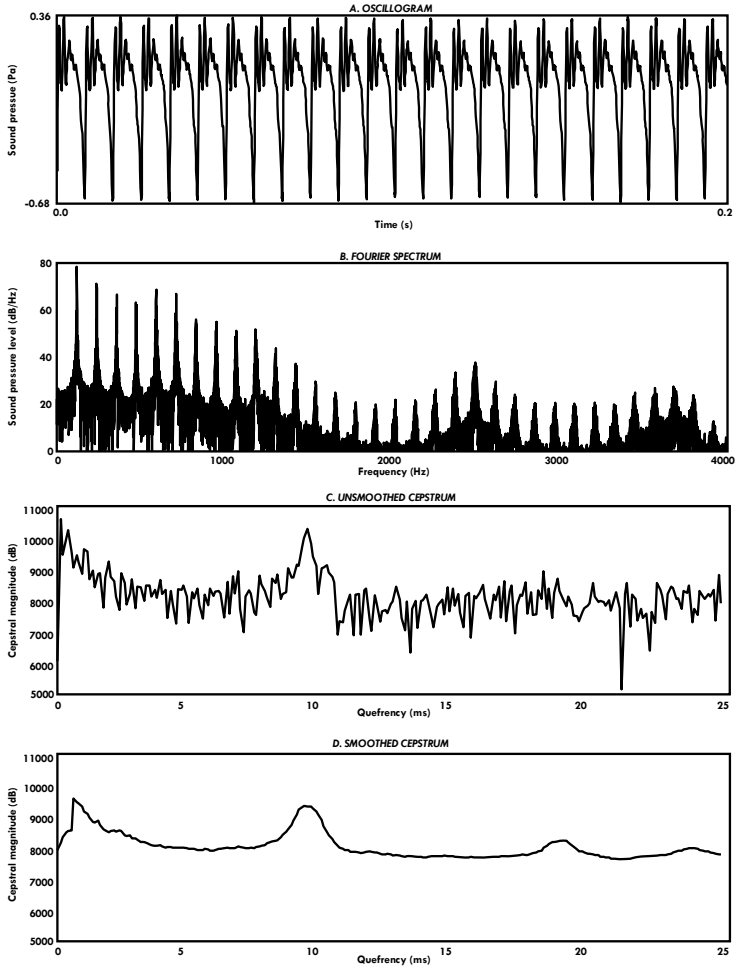
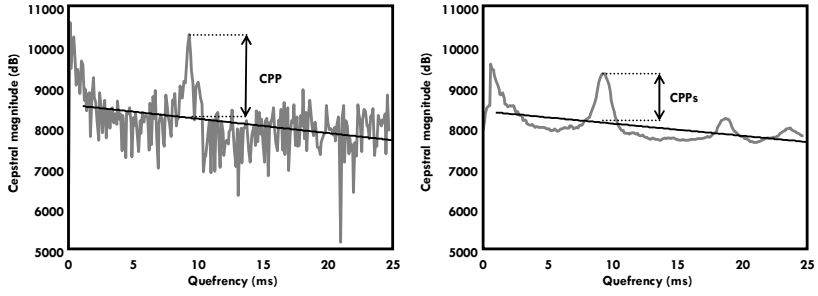


Figure 6.4 LEFT: Illustration of the unsmoothed cepstrum with indication of how CPP is measured. RIGHT: Illustration of the smoothed cepstrum with indication of how CPPs is measured.



d. Indicate the left (L) minimum sound pressure level of the first harmonic ($H_{1\text{-}MIN\text{-}L}$) in this frequency range where it is expected, namely between $H_{1\text{-}MAX}/2$ and $H_{1\text{-}MAX}$. Indicate the right (R) minimum sound pressure level of the first harmonic ($H_{1\text{-}MIN\text{-}R}$) in the frequency range where it is expected, namely between $H_{1\text{-}MAX}$ and $H_{1\text{-}MAX}+(H_{1\text{-}MAX}/2)$.
e. Draw the LTAS and connect $H_{1\text{-}MIN\text{-}L}$ with $H_{1\text{-}MAX}$, and $H_{1\text{-}MAX}$ with $H_{1\text{-}MIN\text{-}R}$.
f. Indicate the maximum and the minima of the second harmonic ($H_{2\text{-}MAX}$, $H_{2\text{-}MIN\text{-}L}$, $H_{2\text{-}MIN\text{-}R}$) in the same ultimate frequency range as for H_1 around the first integer multiple of the first harmonic's peak ($H_{1\text{-}MAX}\times 2$).
g. Connect $H_{2\text{-}MIN\text{-}L}$ with $H_{2\text{-}MAX}$, and $H_{2\text{-}MAX}$ with $H_{2\text{-}MIN\text{-}R}$ on the LTAS.
h. Indicate the maximum and the minima of the third harmonic ($H_{3\text{-}MAX}$, $H_{3\text{-}MIN\text{-}L}$, $H_{3\text{-}MIN\text{-}R}$) in the same ultimate frequency range as for H_1 around the second integer multiple of the first harmonic ($H_{1\text{-}MAX}\times 3$).
i. Connect $H_{3\text{-}MIN\text{-}L}$ with $H_{3\text{-}MAX}$, and $H_{3\text{-}MAX}$ with $H_{3\text{-}MIN\text{-}R}$ on the LTAS.

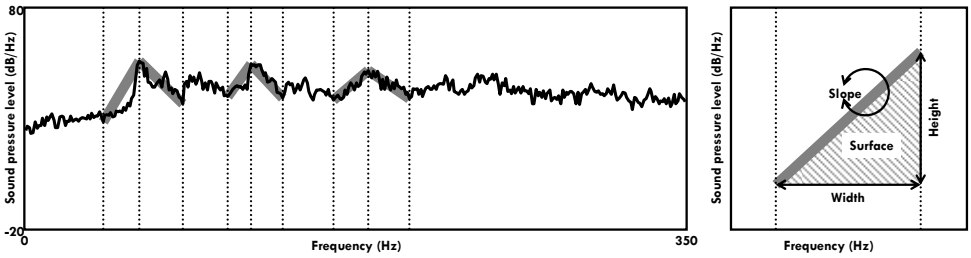
These actions resulted in a LTAS with 6 lines that connect the minima with the maxima of the first 3 spectral harmonics. The left graph in Figure 6.5 represents an example of such a LTAS. From all 6 lines, the following geometrical properties were measured:

a. The height (He) of the lines: the sound pressure level of H_{MAX} minus the sound pressure level of $H_{MIN\text{-}L}$ or $H_{MIN\text{-}R}$ – 6 measures: $H_{1\text{-}L\text{-}He}$, $H_{1\text{-}R\text{-}He}$, $H_{2\text{-}L\text{-}He}$, $H_{2\text{-}R\text{-}He}$, $H_{3\text{-}L\text{-}He}$, $H_{3\text{-}R\text{-}He}$.
b. The width (Wi) of the lines: the frequency of H_{MAX} minus the frequency of $H_{MIN\text{-}L}$, or the frequency of $H_{MIN\text{-}R}$ minus the frequency of H_{MAX} – 6 measures: $H_{1\text{-}L\text{-}Wi}$, $H_{1\text{-}R\text{-}Wi}$, $H_{2\text{-}L\text{-}Wi}$, $H_{2\text{-}R\text{-}Wi}$, $H_{3\text{-}L\text{-}Wi}$, $H_{3\text{-}R\text{-}Wi}$.
c. The surface (Su) of the area under the lines: height×width/2 – 6 measures: $H_{1\text{-}L\text{-}Su}$, $H_{1\text{-}R\text{-}Su}$, $H_{2\text{-}L\text{-}Su}$, $H_{2\text{-}R\text{-}Su}$, $H_{3\text{-}L\text{-}Su}$, $H_{3\text{-}R\text{-}Su}$.
d. The slope (Sl) of the lines: according to the formula slope=(y2-y1)/(x2-x1) – 6 measures: $H_{1\text{-}L\text{-}Sl}$, $H_{1\text{-}R\text{-}Sl}$, $H_{2\text{-}L\text{-}Sl}$, $H_{2\text{-}R\text{-}Sl}$, $H_{3\text{-}L\text{-}Sl}$, $H_{3\text{-}R\text{-}Sl}$.

Thus, this resulted in a set of 24 measures. These 4 properties are illustrated in the right graph of Figure 6.5. Furthermore, mean (M) height and slope and total (T) width and surface were calculated for all 3 harmonics. This resulted in an additional set of 12 measures: $H_{1\text{-}M\text{-}He}$, $H_{1\text{-}T\text{-}Wi}$, $H_{1\text{-}T\text{-}Su}$, $H_{1\text{-}M\text{-}Sl}$, $H_{2\text{-}M\text{-}He}$, $H_{2\text{-}T\text{-}Wi}$, $H_{2\text{-}T\text{-}Su}$, $H_{2\text{-}M\text{-}Sl}$, $H_{3\text{-}M\text{-}He}$, $H_{3\text{-}T\text{-}Wi}$, $H_{3\text{-}T\text{-}Su}$, $H_{3\text{-}M\text{-}Sl}$.

Third, also with the program "Praat", 3 fundamental frequency perturbation measures (jitter local or J_{local}, jitter rap or J_{rap}, jitter ppq5 or J_{ppq5}), 3 amplitude perturbation measures (shimmer local or S_{local}, shimmer local dB or S_{dB}, shimmer apq11 or S_{apq11}) and 3 noise measures (mean autocorrelation or mACF, noise-to-harmonics ratio or NHR, harmonics-to-noise ratio or HNR) of the neoglottal vibrations were derived from the 16 concatenated voice samples. The programming scripts that were used to obtain these 9 measures in "Praat" are provided in Maryn et al. (in press).

Figure 6.5 LEFT: Long-term average spectrum derived from the voice sample presented in Fig. 3. The spectral measures in this study are based on the geometrical properties of the thicker black lines that connect the 2 minima with the maximum of the spectral contour where the first 3 harmonics are expected. RIGHT: 4 geometrical properties of a line through the rising portion of a spectral harmonic.



Statistics

All statistical analyses were completed using SPSS for Windows version 12.0 (SPSS Inc., Chicago, Illinois, USA). The intra-listener and inter-listener agreement of the auditory-perceptual paired comparison task was investigated using 2 non-parametric coefficients. The Cohen kappa coefficient (κ) assesses the amount of agreement between evaluations by multiple pairs of raters when they are rating the same object. Guidelines for the interpretation of the κ statistic are provided by De Bodt et al. (1997). The Spearman rank-order correlation coefficient (r_s) reflects the degree to which a monotonic relationship exists between variables. Interpretation guidelines for r_s are provided by Frey et al. (1991).

The concurrent validity (i.e. the ability of one metric to measure the outcome of another) of the acoustic measurement of overall tracheoesophageal voice quality was investigated with the following statistics. First, r_s and the coefficient of determination (r_s^2) between the perceptual score and the 48 acoustic markers were calculated as typical measures of concurrent validity. Second, stepwise multiple linear regression analysis was performed to create a statistical model representing the best combination of acoustic parameters for the measurement of overall tracheoesophageal voice quality. A multiple regression equation was created based on the unstandardized coefficients of the statistical model. Third, r_s and r_s^2 were calculated between the outcomes of the statistical model and the perceptual scores, again as measures of concurrent validity.

RESULTS

Reliability of the perceptual evaluation

The final outcome of the paired comparison task was based on the sum of the points accumulated by each of the concatenated tracheoesophageal voice

samples and resulted in a ranking of the 16 samples from relatively best (ranked number 1) to relatively worst (ranked number 16) overall voice quality (Table 6.1).

The intra-listener reliability statistics, based on repeating 24 of the 120 voice samples, were mean κ=0.65 and a mean r_s=0.62. These outcomes reflect a moderate to good level of agreement within the listeners. Estimates of inter-listener reliability were mean κ=0.58 and a mean r_s=0.87, which indicate a moderate to high degree of concordance between listeners.

Table 6.1 The 16 concatenated tracheoesophageal voice samples ranked following the outcome of the paired comparison task.

Sample number	Total number of points	Rank
3	66.0	1
12	64.5	2
8	63.5	3
13	56.0	4
5	54.0	5
4	49.5	6
11	40.5	7
2	40.0	8
6	33.5	9
7	27.5	10
14	18.5	11
15	18.5	12
10	16.5	13
1	14.5	14
9	13.0	15
16	0.0	16

Acoustic measures

In Table 6.2, the descriptive data (i.e., minimum, maximum, mean and standard deviation) for the 47 acoustic variables are listed. It was possible to obtain all acoustic measures (excepted for S_{apq11}) from all 16 voice samples. For the 5 samples with rank number 7, 9, 10, 13 and 14, S_{apq11} could not be determined.

Relation between perceptual evaluation and acoustic measures

The correlation coefficients (r_s) and the coefficients of determination (r_s^2) between the overall voice quality ranking and the 47 acoustic measures are shown in Table 6.3. Statistically significant correlations were found for 14 measures derived from all groups of acoustic parameters. These measures included the cepstral rahmonic (CPP and CPPs), the height of the first harmonic ($H_{1\text{-}L\text{-}He}$, $H_{1\text{-}R\text{-}He}$ and $H_{1\text{-}M\text{-}He}$), the height of the second harmonic ($H_{2\text{-}L\text{-}He}$, $H_{2\text{-}R\text{-}He}$ and $H_{2\text{-}M\text{-}He}$), the fundamental frequency perturbation quotient (J_{ppq5}), the amplitude perturbation (S_{local} and S_{dB}) and the noise measures (mACF, NHR and HNR). With a r_s=-0.78, CPP yielded the highest absolute r_s, explaining approximately 60% of the variation

(r_s^2=0.61) of the overall voice quality ranking. With the exception of $H_{2\text{-}R\text{-}He}$, J_{Jppq5}, S_{dB} and NHR, these correlations were significant at the α=0.01 level. No other statistically significant associations were found.

Stepwise multiple regression analysis identified only 2 acoustic variables (CPP and $H_{2\text{-}M\text{-}He}$), when combined were the best predictors of overall tracheoesophageal voice quality. The equation, based on the unstandardized coefficients of the regression, is:

$$-48.413 + (2.352 \times H_{2\text{-}M\text{-}He}) + (4.304 \times CPP) \quad \text{(Eq. 1)}$$

The proportional relationship between the outcome of this combination of acoustic variables and the overall voice quality ranking was r_s=0.87 (statistically significant at α=0.01), revealing high concurrent validity. This association between the acoustic model and the perceptual score is illustrated in Figure 6.6. The coefficient of determination was 0.76 (explaining 76% of the variation of the perceptual score).

Table 6.2 Mean (M), standard deviation (SD), minimum (Min) and maximum (Max) of the outcomes of the 47 acoustic measures.

Measures°	M	SD	Min	Max	Measures°	M	SD	Min	Max
CPP	11.45	2.08	8.66	15.54	**$H_{2\text{-}R\text{-}Sl}$**	-0.93	0.92	0.20	-3.57
CPPs	3.94	2.16	1.21	8.20	**$H_{2\text{-}M\text{-}Sl}$**	1.05	1.02	0.24	3.47
$H_{1\text{-}L\text{-}He}$	25.82	7.96	8.70	39.80	**$H_{3\text{-}L\text{-}He}$**	11.75	3.89	6.78	20.43
$H_{1\text{-}R\text{-}He}$	18.52	5.83	11.35	31.89	**$H_{3\text{-}R\text{-}He}$**	11.68	5.19	0.00	19.82
$H_{1\text{-}M\text{-}He}$	22.17	6.42	10.02	33.39	**$H_{3\text{-}M\text{-}He}$**	11.72	3.65	4.95	19.16
$H_{1\text{-}L\text{-}Wi}$	34.56	18.83	4.00	78.00	**$H_{3\text{-}L\text{-}Wi}$**	19.81	15.21	2.00	54.0
$H_{1\text{-}R\text{-}Wi}$	38.13	20.91	3.00	87.00	**$H_{3\text{-}R\text{-}Wi}$**	42.19	37.51	0.00	108.0
$H_{1\text{-}T\text{-}Wi}$	72.69	36.22	7.00	165.00	**$H_{3\text{-}T\text{-}Wi}$**	62.00	32.80	7.00	113.0
$H_{1\text{-}L\text{-}Su}$	469.0	347.3	17.4	1552.1	**$H_{3\text{-}L\text{-}Su}$**	135.9	127.6	6.8	399.8
$H_{1\text{-}R\text{-}Su}$	348.0	208.6	17.0	967.5	**$H_{3\text{-}R\text{-}Su}$**	285.0	251.6	0.0	715.0
$H_{1\text{-}T\text{-}Su}$	817.1	526.1	34.4	2519.5	**$H_{3\text{-}T\text{-}Su}$**	420.9	235.1	25.0	749.9
$H_{1\text{-}L\text{-}Sl}$	0.99	0.60	0.24	2.34	**$H_{3\text{-}L\text{-}Sl}$**	1.05	0.85	0.27	3.39
$H_{1\text{-}R\text{-}Sl}$	-0.79	0.87	-0.20	-3.78	**$H_{3\text{-}R\text{-}Sl}$**	-0.75	1.00	-0.12	-3.99
$H_{1\text{-}M\text{-}Sl}$	0.89	0.67	0.23	2.98	**$H_{3\text{-}M\text{-}Sl}$**	0.92	0.72	0.41	3.05
$H_{2\text{-}L\text{-}He}$	13.70	8.27	0.00	28.55	**J_{local}**	5.05	3.00	1.43	11.23
$H_{2\text{-}R\text{-}He}$	14.85	5.09	5.81	23.42	**J_{rap}**	2.57	1.54	0.68	4.81
$H_{2\text{-}M\text{-}He}$	14.27	5.48	5.41	24.55	**J_{ppq5}**	2.87	1.76	0.81	5.90
$H_{2\text{-}L\text{-}Wi}$	31.19	28.95	0.00	92.00	**S_{local}**	16.93	6.51	6.03	31.73
$H_{2\text{-}R\text{-}Wi}$	28.75	19.57	3.00	74.00	**S_{dB}**	1.50	0.58	0.57	2.82
$H_{2\text{-}T\text{-}Wi}$	59.94	34.19	4.00	128.00	**S_{apq11}**	11.14	5.47	3.39	23.60
$H_{2\text{-}L\text{-}Su}$	268.6	309.0	0.0	1155.3	**mACF**	0.73	0.11	0.56	0.89
$H_{2\text{-}R\text{-}Su}$	232.1	192.5	8.7	722.6	**NHR**	0.48	0.24	0.14	0.86
$H_{2\text{-}T\text{-}Su}$	500.7	352.9	11.2	1362.0	**HNR**	5.80	3.38	1.17	11.30
$H_{2\text{-}L\text{-}Sl}$	1.15	1.36	0.16	5.01					

° *L: left line; R: right line; M: mean of left and right lines; T: total of left and right lines; He: height (in dB/Hz); Wi: width (in Hz); Su: surface (in dB/Hz*Hz); Sl: slope (no unit).*

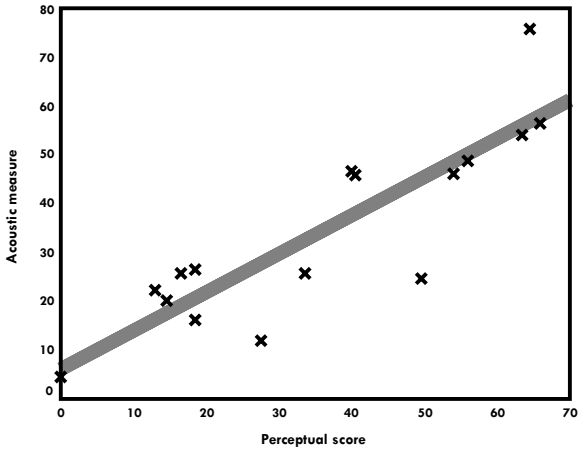
Table 6.3 Correlation coefficients (r_s) and coefficients of determination (r_s^2) between the auditory-perceptual overall voice quality ranking and the 47 acoustic measures.

Measures°	r_s	r_s^2	Measures°	r_s	r_s^2
CPP	-0.78**	0.61	$H_{2\text{-}R\text{-}Sl}$	-0.02	0.00
CPPs	-0.77**	0.59	$H_{2\text{-}M\text{-}Sl}$	0.15	0.02
$H_{1\text{-}L\text{-}He}$	-0.63**	0.40	$H_{3\text{-}L\text{-}He}$	-0.40	0.16
$H_{1\text{-}R\text{-}He}$	-0.66**	0.43	$H_{3\text{-}R\text{-}He}$	-0.29	0.08
$H_{1\text{-}M\text{-}He}$	-0.76**	0.58	$H_{3\text{-}M\text{-}He}$	-0.35	0.12
$H_{1\text{-}L\text{-}Wi}$	-0.04	0.00	$H_{3\text{-}L\text{-}Wi}$	-0.13	0.02
$H_{1\text{-}R\text{-}Wi}$	-0.21	0.05	$H_{3\text{-}R\text{-}Wi}$	-0.16	0.02
$H_{1\text{-}T\text{-}Wi}$	-0.01	0.00	$H_{3\text{-}T\text{-}Wi}$	-0.14	0.02
$H_{1\text{-}L\text{-}Su}$	-0.11	0.01	$H_{3\text{-}L\text{-}Su}$	-0.23	0.05
$H_{1\text{-}R\text{-}Su}$	-0.46	0.21	$H_{3\text{-}R\text{-}Su}$	-0.16	0.03
$H_{1\text{-}T\text{-}Su}$	-0.27	0.07	$H_{3\text{-}T\text{-}Su}$	-0.30	0.09
$H_{1\text{-}L\text{-}Sl}$	-0.29	0.09	$H_{3\text{-}L\text{-}Sl}$	0.02	0.00
$H_{1\text{-}R\text{-}Sl}$	0.04	0.00	$H_{3\text{-}R\text{-}Sl}$	-0.21	0.04
$H_{1\text{-}M\text{-}Sl}$	-0.19	0.04	$H_{3\text{-}M\text{-}Sl}$	0.00	0.00
$H_{2\text{-}L\text{-}He}$	-0.66**	0.43	J_{local}	0.45	0.20
$H_{2\text{-}R\text{-}He}$	-0.57*	0.32	J_{rap}	0.48	0.23
$H_{2\text{-}M\text{-}He}$	-0.73**	0.53	J_{ppq5}	0.50*	0.25
$H_{2\text{-}L\text{-}Wi}$	-0.14	0.02	S_{local}	0.67**	0.44
$H_{2\text{-}R\text{-}Wi}$	-0.15	0.02	S_{dB}	0.62*	0.38
$H_{2\text{-}T\text{-}Wi}$	0.05	0.00	S_{apq11}	0.36	0.13
$H_{2\text{-}L\text{-}Su}$	-0.34	0.12	**mACF**	-0.63**	0.39
$H_{2\text{-}R\text{-}Su}$	-0.34	0.11	**NHR**	0.54*	0.29
$H_{2\text{-}T\text{-}Su}$	-0.21	0.04	**HNR**	-0.63**	0.40
$H_{2\text{-}L\text{-}Sl}$	-0.02	0.00			

° *L: left line; R: right line; M: mean of left and right lines; T: total of left and right lines; He: height (in dB/Hz); Wi: width (in Hz); Su: surface (in dB/Hz×Hz); Sl: slope (no unit).*

**: significant at the 0.05 level (2-tailed); **: significant at the 0.01 level (2-tailed).*

Figure 6.6 Scatterplot to illustrate the concurrent validity of the statistical acoustic model, based on the formula [-48.413+(2.352×$H_{2\text{-}M\text{-}He}$)+(4.304×CPP)], with the perceptual score of the paired comparison task.



DISCUSSION

For the most ecologically valid assessment of voice quality, it has been argued analysis should include both sustained vowels and continuous speech (Maryn et al., in press). However such an analysis approach has rarely been applied in laryngeal or alaryngeal speakers. Therefore, as one aim, the present study examined the feasibility of combining sustained vowels and continuous speech in the acoustic measurement of overall voice quality in 16 tracheoesophageal speakers. Given that almost all acoustic measures could be determined for all samples (with the exception of S_{apq11}), such analysis is clinically feasible by dividing tracheoesophageal continuous speech into voiced and unvoiced segments, to extract only the voiced segments and to concatenate these voiced segments with sustained vowel samples (as demonstrated in Figure 6.2). This method for the editing and concatenation was adopted from Maryn et al. (in press) and made it possible to incorporate both sample types in the auditory-perceptual as well the acoustic measurement of overall voice quality (laryngeal and tracheoesophageal). Next, the level of overall voice quality was determined via an auditory-perceptual paired comparison paradigm in which all samples were compared with all other samples, resulting in a ranking of all 16 samples were ranked (from relatively worst to relatively best overall voice quality). In total, 47 acoustic measures were obtained from the concatenated samples and the concurrent validity between the perceptual rankings and the acoustic measures was examined. Finally, a multivariate model was constructed to a combination of acoustic variables that best predicted the auditory-perceptual rankings of overall the tracheoesophageal voice quality.

Interestingly, the 2 cepstral parameters in this study, CPP and CPPs (Hillenbrand & Houde, 1996), were the most robust correlates of tracheoesophageal voice quality, and most sensitive to differences among the various rankings. This finding is congruous with many reports in the literature. Maryn et al. (in press), for example, found CPP and CPPs to be the most powerful measures of "G" (a measure of overall dysphonia severity) in a study similar to the one reported but with a much larger number of laryngeal-vocal voice samples and with another panel of experienced listeners. As far as we know, this investigation was the first to apply cepstral analysis to alaryngeal speakers, and the results extend the validity of the cepstral analysis to tracheoesophageal voice assessment.

In addition, it was hypothesized that the more the Fourier spectrum of a tracheoesophageal voice sample resembled the Fourier spectrum of a normal voice sample, the better its overall voice quality would be, especially in terms of the presence of emergent harmonics and/or the relative absence of interharmonic noise levels. We therefore measured 4 geometric properties (width, height, surface and declivity of the slopes) in frequency bands where the first 3 harmonics were expected in long-term average spectra (LTAS) of the concatenated voice samples. The peak (i.e. the center) of the first harmonic was expected to correspond with the maximum sound pressure level between 0 and 300 Hz. The center of the following

2 harmonics were identified as the maximum sound pressure level in the vicinity of integer multiples of the peak of the first harmonic. This is illustrated in Figure 6.5. From the 36 geometric attributes, only the height of the first and the second harmonic were significantly correlated with the rankings of overall voice quality. We found r_s=0.76 for the prominence of the first harmonic ($H_{1\text{-}M\text{-}He}$) and r_s=0.73 for the amplitude of the second harmonic ($H_{2\text{-}M\text{-}He}$). The prominence of harmonic peaks is negatively affected by aperiodicities in the voice signal and/or by increased noise levels (Dejonckere & Lebacq, 1987; Hillenbrand & Houde, 1996). Because the pharyngeal-esophageal segment has a larger mass and is a less vibratory source as compared to natural vocal folds, laryngectomees are especially at risk for irregularities/aperiodicities (MacCallum et al., 2008). Additionally, they are especially susceptible to increased noise because of air turbulence generated at the pharyngeal-esophageal segment or emitted via the tracheostoma. Indeed, the LTAS in Figure 6.5 barely resembles the LTAS of vocally-normal laryngeal speakers. However, the more regular the tracheoesophageal voice signal and the less it is contaminated by additive noise, the more the harmonics will emerge in the spectrum or LTAS of laryngectomees (van As-Brooks et al., 2006), as confirmed by our results regarding the height of the first 2 harmonics.

In contrast to cepstral and spectral geometry measures, many reports have discussed the putative value of frequency and amplitude perturbation measures (i.e. jitter and shimmer, respectively). First, the ability of these measures to discriminate between different types of voicing (i.e. laryngeal-vocal, esophageal, tracheoesopharyngeal, electrolaryngeal, hemilaryngeal) has frequently been investigated (Robbins, 1984; Debruyne et al., 1994; Bertino et al., 1996; Arias et al., 2000; Kazi et al., 2006; Štajner-Katušić et al., 2006; MacCallum et al., 2008). For example Robbins (1984) reported that jitter ratio and mean shimmer significantly separated esophageal voices from laryngeal and tracheoesophageal voices. However, these measures could not significantly discriminate between tracheoesophageal voices and laryngeal voices. However, directional shimmer did differ significantly for these 2 voicing types. Bertino et al. (1996) and Arias et al. (2000) independently found that percent jitter and absolute shimmer discriminated between laryngeal voices and alaryngeal voices. But, only the harmonics-to-noise-ratio in Bertino et al. (1996) could significantly separate tracheoesophageal voices from esophageal voices. MacCallum et al. (2008) also found that percent jitter, percent shimmer and signal-to-noise ratio significantly differentiated normal voices from esophageal voices. Collectively, perturbation measures have proven to be useful in separating speakers according to their type or source of voice production. However, perturbation measures of jitter and shimmer vitally rely on accurate detection of the fundamental period. This is problematic in the analysis of disordered laryngeal voice signals (Roark, 2006) and is particularly problematic in highly aperiodic voices (Titze, 1995) like tracheoesophageal and esophageal speakers. Indeed, MacCallum et al. (2008) demonstrated that perturbation measures are insufficiently reliable in quantifying the aperiodic esophageal voice signal. It is likely that this lack of reliability explains the modest correlations between

perturbation measures and perceptual ratings of acceptibility, normality/abnormality, overall impression and overall quality of the voice of laryngectomees, reported by Bertino et al. (1996), van As et al. (1998), Moerman et al. (2004) and in the present study. Interestingly, the correlations in the present study were actually higher for shimmer than for jitter. Indeed, frequency perturbation is far more dependent on the exact placement of period/cycle boundaries than amplitude perturbation (Bielamowicz et al., 1993) and thus is much easier influenced by aperiodicities in the voice signal. Even minor errors in identifying these boundaries complicate jitter calculation and accuracy.

To our knowledge, this investigation represents the first attempt to apply multivariable statistical methods to the acoustic measurement of voice quality following total laryngectomy. Investigators have traditionally reported bivariate correlations between specific acoustic parameters and auditory-perceptual ratings of overall voice quality. However, like laryngeal voice quality, tracheoesophageal voice quality is a multidimensional phenomenon depending on volume and pressure quantities related to the driving force, intensity of the sound, resistance to airflow caused by the voice prosthesis (Grolman et al., 2008), the physical and physiological properties of the pharyngeal-esophageal segment (Lundström et al., 2008), etc. Given its multidimensionality, several investigators of laryngeal voice production have therefore studied the voice signal multiparametrically. A summary of the outcomes of such studies can be found in Maryn et al. (in press). In the present study, the stepwise multiple linear regression indicated the combination of $H_{2\text{-}M\text{-}He}$ and CPP to be the best measure of the overall tracheoesophageal voice quality. The outcome of the equation based on the unstandardized coefficients of the 2-factor regression model boosted their solitary correlations of respectively 0.73 and 0.78 to a highly respectable r_s=0.87. Consequently, the two factor multivariate model appears to be extremely sensitive to differences in overall quality of tracheoesophageal voice, and thus holds promise as a potentially important clinical tool to objectively track change in response to various interventions. For instance, one could employ this 2-factor acoustic model to assess the relative superiority of different types of voice prostheses in an individual tracheoesophageal speaker.

Limitations and suggestions for future research

One limitation of this study was the relatively low number of participants (n=16) and subsequently the restricted generalizability of the results. Although the results are promising, caution is warranted until they are confirmed by research employing a larger population of laryngectomees. Future investigations should externally cross-validate the utility of CPP, the prominence of the second harmonic, the other acoustic measures, as well as the 2-factor model with a larger number of new tracheoesophageal voice samples and ratings. A second limitation is the moderate reliability of some of the auditory-perceptual ratings, despite the experience of the listeners and the paired comparison paradigm. This plausibly

attenuated the concurrent validity of some of the acoustic measures. Future methods should therefore implement more training samples of alaryngeal voices with various degrees of overall voice quality disruption. A third limitation is related to the fact that we only investigated concurrent validity via correlation coefficients. An important and related question regarding whether or not a tracheoesophageal speaker has an "acceptable" voice quality, was not under study. Future research should therefore combine statistics for concurrent validity with statistics of diagnostic precision to separate acceptable from unacceptable speakers (e.g. sensitivity, specificity, area under receiver operating characteristic curve and likelihood ratio), for example as in Maryn et al. (in press). A final limitation is related to the impact of tracheoesophageal voice quality on the quality of life of laryngectomees. Quality of speech and voice, among other factors (e.g. appearance and nutrition), can be important determinants in the degree the quality of life after total laryngectomy (Eadie & Doyle, 2004). It would therefore be interesting to investigate how much the outcome of objective voice quality measures contributes to and explains the variance in self-rated measure of quality of life.

CONCLUSION

This exploratory study addressed 2 primary research questions. First, is it feasible to combine continuous speech as well as sustained vowels in the acoustic measurement of tracheoesophageal voice quality? Second, are measures of cepstral peak prominence, spectral peak geometry and pharyngeal-esophageal perturbation valid measures of overall voice quality of tracheoesophageal speakers? Based upon the results, our method of concatenating continuous speech with sustained vowel samples via a Praat programming script is feasible in a clinical setting. Furthermore, cepstral measures, as well as measures of harmonic emergence were the most sensitive to different levels of alaryngeal voice quality, especially when they were combined in a 2-factor multivariate statistical model. Both measures can easily be applied in the clinic. Finally, perturbation measures and the other geometrical properties of the spectral harmonics were least sensitive to degrees of voice quality calling into question their clinical utility and applicability.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Nelson Roy (from the University of Utah) for his editing support in the present manuscript. The authors also want to express their gratitude to Dr. James Hillenbrand (from the Western Michigan University) for providing the software for the CPP and CPPs measures.

REFERENCES

Arias, M.R., Ramón, J.L., Campos, M., & Cervantes, J.J. (2000). Acoustic analysis of the voice in phonatory fistuloplasty after total laryngectomy. *Otolaryngology & Head and Neck Surgery, 122,* 743-747.

Bertino, G., Bellomo, A., Miani, C., Ferrero, F., & Staffieri, A. (1996). Spectrographic differences between tracheal-esophageal and esophageal voice. *Folia Phoniatrica et Logopaedica, 48,* 255-261.

Bielamowicz, S., Kreiman, J., Gerratt, B.R., Dauer, M.S., & Berke, G.S. (1993) Comparison of voice analysis systems for perturbation measurement. *Journal of Speech and Hearing Research, 39,* 126-134.

Chan, K.M., & Yiu, E.M. (2002). The effect of anchors and training on the reliability of perceptual voice evaluation. *Journal of Speech, Language and Hearing Research, 45,* 111-126.

Crevier-Buchmann, L., Pfauwadel, M.C., Chabardes, E., Laccourreye, O., Brasnu, D., & Laccourreye, H. (1991). Comparative study of temporal parameters of alaryngeal voices. Esophageal and tracheo-esophageal voices. *Annales d'Oto-Laryngologie et de Chirurgie Cervico-Faciale, 108,* 261-265.

De Bodt, M.S., Wuyts, F.L., Van de Heyning, P.H., & Croux, C. (1997). Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *Journal of Voice, 11,* 74-80.

Debruyne, F., Delaere, P., Wouters, J., & Uwents, P. (1994). Acoustic analysis of tracheo-oesophageal versus oesophageal speech. *Journal of Laryngology and Otology, 108,* 325-328.

Dejonckere, P., & Lebacq, J. (1987). Harmonic emergence in formant zone of a sustained [a] as a parameter for evaluating hoarseness. *Acta Otorhinolaryngologica Belgica, 41,* 988-996.

Eadie, T.L., & Doyle, P.C. (2004). Auditory-perceptual scaling and quality of life in tracheoesophageal speakers. *Laryngoscope, 114,* 753-759.

Frey, L.R., Botan, C.H., Friedman, P.G., & Kreps, G.L. (1991). *Investigating communication, an introduction to research methods.* Englewood Cliffs: Prentice-Hall.

Grolman, W., Eerenstein, S.E.J., Tange, R.A., Canu, G., Bogaardt, H., Dijkhuis, J.P., Dreschler, W.A., & Schouwenburg, P.F. (2008). Vocal efficiency in tracheoesophageal phonation. *Auris Nasus Larynx, 35,* 83-88.

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

Kacha, A., Grenez, F., & Schoentgen, J. (2005). *Voice quality assessment by means of comparative judgments of speech tokens.* Proceedings of Interspeech. Lisbon, Portugal.

Kazi, R., Kiverniti, E., Prasad, V., Venkitaraman, R., Nutting, C.M., Clarke, P., Rhys-Evans, P., & Harrington, K.J. (2006). Multidimensional assessment of

female tracheoesophageal prosthetic speech. *Clinical Otolaryngology, 31,* 511-517.

Kreiman, J., Gerratt, B.R., & Ito, M. (2008). When and why listeners disagree in voice quality assessment tasks. *Journal of the Acoustical Society of America, 122,* 2354-2364.

Lundström, E., Hammarberg, B., Munck-Wikland, E., & Edsborg, N. (2008). The pharyngoesophageal segment in laryngectomees – videoradiographic, acoustic, and voice quality perceptual data. *Logopedics Phoniatrics Vocology, 33,* 115-125.

MacCallum, J.K., Cai, L., Zhou, L., Zhang, Y., Jiang, J.J. (2008). Acoustic analysis of aperiodic voice: perturbation and nonlinear dynamic properties in esophageal phonation. *Journal of Voice,* Epub ahead of print.

Maryn, Y., Corthals, P., Van Cauwenberge P, Roy, N., & De Bodt, M. (in press). Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *Journal of Voice*.

Merwin, G.E., Goldstein, L.P., & Rothman, H.B. (1985). A comparison of speech using artificial larynx and tracheoesophageal puncture with valve in the same speakers. *Laryngoscope, 95,* 730-734.

Moerman, M., Pieters, G., Martens, J.P., Van der Borgt, M.J., & Dejonckere, P. (2004). Objective evaluation of the quality of substitution voices. *European Archives of Otorhinolaryngology, 261,* 541-547.

Roark, R.M. (2006). Frequency and voice: perspectives in the time domain. *Journal of Voice, 20,* 325-354.

Robbins, J. (1984). Acoustic differentiation of laryngeal, esophageal, and tracheoesophageal speech. *Journal of Speech and Hearing Research, 27,* 577-585.

Štajner-Katušić, S., Horga, D., Mušura, M., & Globlek, D. (2006). Voice and speech after laryngectomy. *Clinical Linguistics & Phonetics, 20,* 195-203.

Titze, I.R. (1995). *Workshop on acoustic voice analysis: summary statement.* Iowa City, IA: National Center for Voice and Speech.

van As, C.J., Hilgers, F.J.M., Verdonck-de Leeuw, I.M., & Koopmans-van Beinum, F.J. (1998). Acoustical analysis and perceptual evaluation of tracheoesophageal prosthetic voice. *Journal of Voice, 12,* 239-248.

van As-Brooks, C.J., Koopmans-van Beinum, F.J., Pols, L.C., & Hilgers, F.J. (2006). Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. *Journal of Voice, 20,* 355-368.

Ward, E.C., & van As, C.J. (2007). *Head and neck cancer: treatment, rehabilitation and outcomes*. San Diego, CA: Plural Publishing.

Yiu, E.M., Chan, K.M., & Mok, R.S. (2007). Reliability and confidence in using a paired comparison paradigm in perceptual voice quality paradigm. *Clinical Linguistics & Phonetics, 21,* 129-145.

7

PROPERTIES OF THE CEPSTRAL PEAK PROMINENCE AND ITS USEFULNESS IN VOCAL QUALITY MEASUREMENTS

Carlos Ferrer
Marc De Bodt
Youri Maryn
Paul Van de Heyning
Maria Hernández-Diaz

This chapter is a slightly adapted version of an oral presentation from which the submission text has been published in the *Proceedings of the 5th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA) (pp. 93-96), 2007, December 13-15, Florence, Italy*.

ABSTRACT

Unlike many acoustic measures, cepstral peak prominence (CPP) has shown consistently high correlations with subjective voice quality ratings. However, this superiority of the CPP index is reported based on empirical results, with its theoretical advantages not always clearly stated. In this chapter, the properties of the CPP which makes it a good predictor for vocal quality are addressed, as well as how it differs from other measures. The reported experimental setups of the previous studies are analyzed, and reasons for the observed variability in the results are given. After this discussion, the clinical usefulness of CPP is addressed.

INTRODUCTION

Many acoustic measures have been proposed to correlate with overall voice quality or one of its dimensions (e.g., breathiness, roughness, etc.). An extensive tabulation of acoustic methods can be found in Buder (2000). In spite of the large number of measures available, there is a lack of consistent results across different studies for most of the measures (e.g., jitter, shimmer, harmonics-to-noise ratio, etc.) (Kreiman & Gerratt, 2000). Recent work (Heman-Ackah et al., 2002; Awan & Roy, 2005; Maryn et al., 2007) has shown the cepstral peak prominence (CPP) or its smoothed version (CPPs) to correlate highly with vocal quality dimensions and overall grade. The high correlation in these latter studies has been consistent and notably superior to the other acoustic measures considered. However, the theoretical advantages of CPP over the rest of the acoustic measures are not always clearly stated. Some of the experimental methods also favour abnormally high values for the amount of variance explained.

In this chapter, we address the properties of the CPP that makes it a good correlate for vocal quality, and how it differs from other measures. Besides, the reported experimental methods of the previous studies are analyzed, and reasons for the observed variability in the results are given. After this discussion, the clinical usefulness of the CPP is addressed.

CPP PROPERTIES

The CPP measure is originally a basic pitch detector, as its companion measure Pearson r at autocorrelation peak (RPK) (Hillenbrand et al., 2004). Both measures were devised to appraise the prominence of the peak that should occur at the pitch value in the cepstrum and autocorrelation functions, respectively. As such, CPP is sometimes erroneously believed to be a measure of signal periodicity, when in fact, it only measures the periodicity of the signal spectrum. It is precisely this subtle difference (i.e., measuring spectral harmonic periodicity instead of strict

periodicity) what makes it particularly suited for vocal quality measures, superior to many other measures.

Following is a categorization of measures in five groups, according to the signal characteristics, which have been most correlated with different vocal quality dimensions. The first two are the more common amplitude and fundamental frequency perturbations (i.e., shimmer and jitter, respectively), absent in the original studies on CPP (Hillenbrand et al., 1994; Hillenbrand & Houde, 1996), while the other three groups are the ones actually included in those studies. The sensibility of CPP to the signal characteristics is commented, as well as its possible advantages and drawbacks compared to other measures.

Amplitude perturbation

Signals with amplitude perturbations have frequently been related to roughness (Heman-Ackah et al., 2002), and sometimes with breathiness. The traditional measures of shimmer are obtained in the time-domain, relying on a pitch detection algorithm (i.e., PDA). The CPP is sensitive to shimmer, since shimmer affects the spectral harmonic structure (Schoentgen, 2003). Although CPP diminishes as shimmer increases, it is more robust than time-domain techniques that rely on a PDA. It has been shown that shimmer, jitter and time-domain harmonics-to-noise ratios (i.e., HNR) are quite sensitive to even small errors in the pulse boundaries (Hillenbrand, 1987).

Frequency perturbation

Jitter and shimmer share a similar condition, in that they have been mostly related to roughness, and less frequently to breathiness. Jitter affects spectral structure to a greater extent than shimmer (Schoentgen, 2003) and CPP can therefore also be regarded as a good measure of this perturbation. The same advantage regarding the sensibility of time-domain measures of jitter to errors in the PDA holds in this case in favor of CPP.

Additive noise

The presence of additive noise has been related mainly with breathiness. The prominence of the first rahmonic is also affected by increasing levels of noise, since it reduces the dip between harmonics. In fact, several studies have focused on this property to develop HNR measures (de Krom, 1993; Murphy, 2006). The CPP holds the advantage with respect to time-domain HNR measures, because it does not require an accurate PDA. Furthermore, CPP also holds the advantage with respect to many frequency-domain HNR measures, because it does require the determination of the harmonic frequencies. Existing HNR measures have been regarded as overal dysperiodicity measures, since they have been shown sensitive also to jitter and shimmer (de Krom, 1993; Schoentgen, 2003; Murphy, 2006).

However, the CPP also shares this feature, namely being sensitive to these three groups.

First harmonic amplitude

A high amplitude of the first harmonic, relative to the amplitude of the second harmonic (Hillenbrand & Houde, 1996) or the first formant (Shrivastav, 2003), has been related to breathiness. The underlying assumption is that breathy voices do not produce abrupt glottal closures, resulting in a more rounded and almost sinusoidal excitation. First harmonic amplitude prominences are closely related to glottal flow measures like the amplitude quotient or the speed quotient (Airas & Alku, 2007). Here the CPP is superior to its companion RPK (Hillenbrand et al., 1994) and to other HNR measures. The CPP will produce no prominent peak for a perfect sinusoid, since a sinusoid consists of only one harmonic (no spectral periodic structure). That is the main difference with other periodicity measures: a perfectly periodic signal not necessarily produces a high CPP. This lack of higher harmonics is also typical of nasal phonemes (Buder, 2000), maybe extending the validity of CPP to the nasality dimension.

Spectral tilt

An increment in the energy content in the higher portion of the spectrum has been related to breathiness (Fukazawa et al., 1988). The CPP is not able to measure spectral tilt changes, which would be reflected in the lower part of the cepstra, irrespective of its calculation. On the other hand, spectral tilt measures have been reported to be the worst correlates of auditory-perceptual breathiness ratings (Hillenbrand et al., 1994; Hillenbrand & Houde, 1996), and therefore can CPP's inability to follow spectral tilt changes be considered to be negligible in its prediction of breathiness.

As seen, CPP can produce an adequate response to most of the signal characteristics which have been related to many vocal quality dimensions (breathiness, roughness, hoarseness and nasality). If an orthogonal representation of the GRBAS scale is accepted (Bonastre et al., 2007), the CPP can be expected to be a better predictor of overall dysphonia, than of any individual dimension, because the selective response of CPP to one particular dimension is affected by its response to the others. The next section explains the results of the CPP index in several reported studies in terms of the previous discussion.

REPORTED STUDIES

This section covers five studies that assessed the correlation between the CPP and CPPs measures and dysphonia severity or specific voice qualities.

Hillenbrand et al. (1994)

In this study, fifteen normophonic volunteers were asked to produce four vowels with three different breathiness levels. Twenty listeners rated the breathinees in these recordings on a unrestriced visual-analog scale. The different acoustic indices were calculated over three types of signals: the original signal, a band-pass filtered signal and a high-pass filtered signal. With a correlation of 0.90 and 80% of the variance in breathiness explained, the CPP emerged as the best predictor of the breathiness ratings. The RPK measure showed similar results on the band-pass filtered signals.

This study intentionally used recordings with breathiness as the only dimension of dysphonia. This has, according to the previous discussion on the CPP properties, two consequences. First, the breathiness ratings coincide with overall dysphonia since it is the only deviant dimension. Second, the obtained correlations can be high because CPP was not affected by interference with other distortions. The possible influence of using non-pathological speakers is addressed in the analysis of the next study.

Hillenbrand & Houde (1996)

In this study, a broad database of pathological voices was screened to select twenty recordings presenting mainly breathiness and five recordings derived from normophonic subjects. The recordings included a sustained vowel as well as continuous speech. The degree of breathines was rated on a unrestricted visual-analog scale by twenty judges. Again, the cepstral measures CPP and CPPs were the best predictors with similar results (i.e., up to 85% and 92% of the variance of breathiness explained in the continuous speech and sustained vowel samples, respectively).

The RPK (with 72% of the variance in the breathiness ratings accounted for) could not match the results of its equivalent cepstral measures. A possible cause is that the dysphonic voices showed a stronger influence of first harmonics amplitude and spectral measures than the normophonic voices from the previous study, and CPP is better suited than RPK to reflect at least the former factor. Again, the restriction to only include breathy voices can explain the extremely high correlations that were obtained.

Heman-Ackah et al. (2002)

Pre- and post-surgery voice recordings of both sustained vowels and continuous speech were derived from nineteen voice-disordered patients. These recordings were rated on grade, breathiness and roughness on a visual-analog scale by two raters.

The CPP and CPPs mearsures showed lower correlations with the breathiness ratings than in Hillenbrand et al. (1994) and Hillenbrand & Houde

(1996). This was caused by the absence of a selective screening of the deviant dimensions, which is more likely to be the case in clinical practice. Here, the results for grade (65-75% of the variance explained) were better than the results for breathiness (50% of the variance accounted for) and roughness (20-25% of the variance explained). These results agree with the discussion on the CPP properties.

Awan & Roy (2005)

Voice recordings from eighty-three dysphonic and fifty-one normal female subjects were allocated into four types of voice/dysphonia (i.e., normal, breathy, rough, and hoarse) by twelve judges. The degree of the dysphonia dimension was not the goal of the study, only the type.

The study found a CPP-like measure to be good at discriminating normal from dysphonic voices, but it was not relevant for the separation among the different dysphonia types. A logarithmic shimmer measure was found best suited for the latter purpose. This also agrees with the discussion on the CPP properties. The CPP is similarly sensitive to breathy and rough signal characteristics, and can not be a reliable separator among them.

Maryn et al. (2007)

Both a sustained vowel and continuous speech were recorded from 229 voice-disordered and twenty-two normal subjects. Five listeners rated these voice samples on grade, roughness and breathiness.

The CPP ranked again the best among all acoustic measures considered, and again the correlation was strong with the overall dysphonia and breathiness rating. The results are the lowest reported (50% of the variance in grade explained), but the size of the database is also the largest, thus including more variability than previous studies.

DISCUSSION AND CONCLUSION

According to the previous sections, CPP can be expected to appraise overall dysphonia better than any other acoustic measure of vocal quality previously reported. If proper screening of samples in performed, i.e., with signal deviation limited to a single dimension, the CPP can produce extremely high correlations with this dimension.

A significant reduction in the percent of explained variance occurs when considering signals with a wide range of variability. But even in that case, CPP can still perform as the best single predictor of overall dysphonia severity. Another point in favor of CPP is its similar performance on sustained vowels and running speech. The desirability of using continuous speech (i.e., ecological validity) for acoustic measures has been pointed out in several studies (Hillenbrand & Houde,

1996; Maryn et al., 2007), and only a small fraction of the existing measures can work on continuous speech.

However, for the purpose of separating different dimensions of voice quality/dysphonia, the usefulness of CPP is limited. Its sensitivity to most of the relevant distortions found in pathological voices (i.e., breathiness and roughness) makes CPP better suited to predict overall dysphonia than any solitary dimension. Since the use of solitary dimensions is usually the case in clinical practice, complementary acoustic measures are needed to perform an accurate and exhaustive description of voice quality in terms of objective measures.

REFERENCES

Airas, M., & Alku, P. (2007). Comparison of multiple voice source parameters in different phonation types. *Paper presented at 8th Annual Conference of the International Speech Communication Association (Interspeech 2007) (pp. 1410-1413)*, Antwerp, Belgium.

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

Bonastre, J.F., Fredouille, C., Ghio, A., Giovanni, A., Pouchoulin, G., Revis, J., Teston, B., & Yu, P. (2007). *Paper presented at 8th Annual Conference of the International Speech Communication Association (Interspeech 2007) (pp. 1194-1197)*, Antwerp, Belgium.

Buder, E.H. (2000). Acoustic analysis of voice quality: a tabulation of algorithms 1902-1990. In R.D. Kent & M.J. Ball (Eds.), *Voice quality measurement (pp. 119-244).* San Diego: Singular Publishing Group.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research, 36,* 254-266.

Fukazawa, T., El-Assuooty, A., & Honjo, I. (1988). A new index for evaluation of the turbulent noise in pathological voice. *Journal of the Acoustical Society of America, 83,* 1189-1193.

Heman-Ackah, Y.D., Michael, D.D., & Goding, G.S. (2002). The relationship between cepstral peak prominence and selected parameters of dysphonia. *Journal of Voice, 16,* 20-27.

Hillenbrand, J. (1987). A methodological study to perturbation and additive noise in synthetically generated voice signals. *Journal of Speech and Hearing Research, 30,* 448-461

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

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

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

Maryn, Y., Corthals, P., De Bodt, M., & Van Cauwenberge, P. (2007). Cepstral peak prominence as a measure of overall voice quality in sustained vowel as well as continuous speech segments. *Paper presented at 7th Pan-European Voice Conference (PEVOC7)*, Groningen, The Netherlands.

Murphy, P.J. (2006). Periodicity estimation in synthesized phonation signals using cepstral rahmonic peakls. *Speech Communication, 48,* 1704-1713.

Schoentgen, J. (2003). Spectral models of additive and modulation noise in speech and phonatory excitation signals. *Journal of the Acoustical Society of America, 113,* 553-562.

Shrivastav, R. (2003). The use of an auditory model in predicting perceptual ratings of breathy voice quality. *Journal of Voice, 17,* 502-512.

EFFECTS OF ACOUSTIC BIOFEEDBACK IN PHONATORY DISORDERS AND PHONATORY PERFORMANCE: A SYSTEMATIC LITERATURE REVIEW

Youri Maryn
Marc De Bodt
Paul Van Cauwenberge

This chapter is an adapted version of an article that has been published in:
Applied Psychophysiology and Biofeedback, 2006;31:65-83.

ABSTRACT

The purpose of this article was to systematically review the literature on the effects of biofeedback therapy in the domain of phonatory disorders and phonatory performance, using studies in peer-review journals. An extensive definition of biofeedback is given and its place in voice treatment is defined. Eighteen group or case studies or reports considering the effects of electromyographic, laryngoscopic and acoustic biofeedback in dysphonic patients (hyperfunctional voice disorders, hypofunctional voice disorders, psychogenic voice disorder, laryngeal trauma, total laryngectomy, vocal cord dysfunction) and participants with normal voices are included and an analysis of procedure as well as research design and results is presented. The usefulness of biofeedback in phonatory disorders and performance was to be interpreted based on tendencies, since there is a lack of randomized controlled efficacy studies. In only 3 of 18 studies (16.7 %) did biofeedback therapy fail to improve voice quality or not result in better results than other forms of therapy. Recommendations for improved methodologies are made. In addition, recommendations for future study of biofeedback include the use of acoustic voice quality parameters.

INTRODUCTION

With hardware and software becoming commonly available, the use of biofeedback in the treatment of phonatory disorders is becoming increasingly important. Real-time spectra and spectrograms, for example, are now freely available through the internet (Huckvale, 2003; Huckvale, 2004). With this increase of biofeedback applications, questions arise about factors underlying clinical effectiveness.

According to Mathieson (2001), behavioral treatment approaches to voice disorders can be divided in seven categories: education and explanation, vocal tract care, vocal hygiene, improved vocal techniques, facilitating techniques, indirect treatments and finally pedagogic strategies. Biofeedback therapy belongs to the category of pedagogic strategies. Biofeedback and its clinical application has a long history. Beginning in the late 1950's, applied biofeedback arose from several clinical and scientific fields and has continued to evolve (Schwartz & Olson, 2003). Learning theorists provided theoretic models and scientific evidence that, for example, autonomic nervous system responses can be instrumentally or operantly conditioned. Behavioral therapists provided the principles of applying operant and classical conditioning models as well as observational learning models and cognitive (information-processing) models. Biofeedback became a major specialty in the field of behavioral medicine, and in this context, has proven to have various applications such as stress management, relaxation therapy and pain management. In addition to these fields, and as stated by Schwartz & Olson (2003, pp. 8), there would be no biofeedback without high-quality instrumentation that accurately and reliably measures physiological events. Biomedical engineers have developed

instruments for real-time monitoring of physiological activity. Through the use of the modern computer, there can be very rapid signal processing and analyses, multiple channel recordings and user-friendly displays. It seems like technological progress, in speech and language therapy as well as in other applied sciences, provides many options, with biofeedback as one of them.

Many definitions and descriptions of biofeedback have been given. But a thorough definition of biofeedback, considering its process and its objectives, can be found in Schwartz & Schwartz (2003, pp. 34-35). As a process, biofeedback is a group of therapeutic procedures that uses electronic and electromechanical instruments to accurately measure, process and feed back, to persons and their therapists, information with educational and reinforcing properties, about their neuromuscular and autonomic activity, both normal and abnormal, in the form of analog or binary, auditory, and/or visual feedback signals. Best achieved with a competent biofeedback professional, the objectives are to help persons develop greater awareness of, confidence in, and an increase in voluntary control over physiological processes that are otherwise outside awareness and/or under less voluntary control, by first controlling the external signal, and then by using cognitions, sensations, or other cues to prevent, stop, or reduce symptoms (Schwartz & Schwartz, 2003). Summarized and simplified, when humans are given real-time (instant and continuous) electronic displays of their internal physiologic events (using meters, lightbars, etc.), they can be taught to manipulate otherwise unsensed events voluntarily (Basmajian, 1981). Synonyms for the term biofeedback are artificial proprioception, electromyographic feedback, audiovisual neuromuscular reeducation, neuromyometry and sensory integration (Fernando & Basmajian, 1978; Basmajian, 1981).

Biofeedback has been applied to various conditions. Therapy with electromyographical biofeedback has been used in upper motor neuron lesions (for example hemiplegia due to stroke or spastic muscle function in cerebral palsy), lower motor neuron lesions (for example Bell's facial palsy), hysterical paralysis, and dyskinesias (for example spasmodic torticollis and Parkinson's disease) (Basmajian, 1978). Electroencephalography has been used for conscious manipulation of brain waves in patients with epileptic seizures. Skin temperature, through its relation with bloodflow, was applied in vascular headaches. Direct plethysmography is related with blood pressure and was applied in essential hypertension (Basmajian, 1981). It is obvious that biofeedback has different faces covering the same idea: by accurately detecting a physiological event and converting the resulting electronic signal into auditory, visual, tactile, or kinesthetic feedback, the individual can be made immediately and continuously aware of the level of the physiological event.

In contrast to research in biofeedback around medical specialties [urinary incontinence (Tries & Eisman, 2003), essential hypertension (McGrady & Linden, 2003), diabetes mellitus (McGrady & Bailey, 2003), etc.] and the fact that the concept of biofeedback is mentioned in many handbooks about voice disorders (examples are Aronson, 1990; Boone & McFarlane, 1988, 1994; Mathieson, 2001;

Rammage, Morrison & Nichol, 2001), there seem to be relatively few data-based biofeedback studies in speech and language pathology and therapy, and in the specific field of voice disorders. As for speech and language therapy, biofeedback has been used in several disorders: hearing disorders and deafness (Stark, 1971), resonance disorders (Fletcher, 1972; Moller et al., 1973; Shelton et al., 1975; Künzel, 1982; Brunner et al., 1994; Goldstein et al., 1994; Whitehill, Stokes & Man, 1996), neurogenic speech disorders including facial paralysis (Netsell & Cleeland, 1973; Finley et al., 1977; Daniel & Guitar, 1978; Jankel, 1978; Hand et al., 1979; Netsell & Daniel, 1979; Nemec & Cohen, 1984; Rubow et al., 1984; Schram & Burres, 1984; Rubow & Swift, 1985; Hammerschlag et al., 1987; Gentil et al., 1994; Goldstein et al., 1994), fluency disorders (Guitar, 1975; Hanna, Wilfling & McNeill, 1975; Craig & Cleary, 1982), articulation disorders (Brooks et al., 1981; Michi, 1993), swallowing disorders (Denk & Kaider, 1997; Huckabee & Pelletier, 1999) and respiratory disorders (Murdoch et al., 1999). Also in the field of voice disorders, there have been studies regarding the use and effects of biofeedback, as will be discussed in the next section.

In this article, the authors sought for an answer to the following two questions. First, is biofeedback an effective tool in the treatment of voice disorders and voice performance? Second, what are the challenges for future biofeedback therapy in patients with voice disorders? Because statistical pooling concealed the most important clinical relevance issues, analysis was made from a descriptive literature review rather than a meta-analysis.

LITERATURE REVIEW ON ACOUSTIC BIOFEEDBACK

In order to answer the above mentioned questions, a Medline/Pubmed search for articles was conducted, using a combination of key words such as 'biofeedback', 'feedback', 'voice' and 'therapy'. Further information and publications were also sought via references in texts. All articles were thoroughly analysed and data were summarized in a spreadsheet. Based on information in the title and abstract, the usefulness of the articles potentially to be included in the literature review was evaluated. Inclusion criteria were: (1) peer-reviewed articles considering (2) the use of real-time biofeedback as an additional sensory source to monitor the vocal or vocally related (for example general laryngeal tension) function (excluding feedback applications as auditory masking, delayed auditory feedback, etc.) and (3) in patients with an organic or nonorganic voice disorder (excluding biofeedback studies in the fields of respiration, deglutition, articulation, resonance and buccofacial expression) or (4) in normal subjects who were vocally trained (including speech and language therapy students, singers, etc.). There were no restrictions regarding type of study design (group and single-subject designs as well as case studies were included) or language. Only microphone-based (i.e., acoustic measurement-based) biofeedback studies were included. Because systematic review of randomized trials has become the gold standard for judging whether or not a certain therapy works (Sackett et al., 1996), the authors were

hoping to find such study designs. Based upon Frey et al. (1991), Portney & Watkins (2000) and Van Borsel (2004), degree of evidence (true, pre-, quasi- or non-experimental) and kind of design were identified in the various studies. In true or full experiments, the independent variable is manipulated, participants are randomly assigned to equivalent experimental groups and there is a high control of extra-experimental variables. Quasi-experiments manipulate or observe the independent variable and when there is a comparison group, the experimental groups are not equivalent because the participants are systematically but not randomly allocated (quasi-experimental groups). There is only moderate control of extra-experimental variables. Pre-experiments manipulate or observe the independent variable and when there is a comparison group, the design lacks random or systematic allocation of subjects, eliminating any control of extra-experimental variables (Frey et al., 1991). Finally, non-experiments do not manipulate an independent variable. There is no experiment (Van Borsel, 2004).

Some articles identified in this search describe a biofeedback application or device rather than consider treatment efficacy (Davis & Drichta, 1980; Horii, 1983; Volin, 1991); more relevant to the purposes of this paper are 18 studies evaluating the outcome of biofeedback therapy in phonatory problems and phonatory performance, which comprise the literature review. Table 7.1 summarizes the procedure aspects of all the mentioned effect studies. Table 7.2 condensates the design aspects and results of all those studies. In the following text, these studies will be grouped and discussed according to the input modality of the applied biofeedback device.

The microphone as input-device

Seven studies worked with acoustic energy and used a microphone in the biofeedback therapy. Holbrook, Rolnick & Bailey (1974) used the Vocal Intensity Controller in 32 patients with hyperfunctional voice disorders, presenting vocal cord lesions and/or dysphonia related to vocal hyperfunction. In this non-experimental one-group pretest-posttest design, 27 patients completed the therapy with pre- and post-treatment laryngoscopy. In 88.9 % a positive result was accomplished with complete resolution of vocal fold mucosal lesions (such as vocal nodules, polyps or contact ulcers) (40.8 %), important reduction of lesion-size (29.6 %) or, in cases without initial lesion but with dysphonia, resolution of dysphonia and improved voice quality (18.5 %). Only three patients (11.1 %) were not helped by the treatment and underwent microlaryngoscopy and vocal cord stripping.

Brody, Nelson & Brody (1975) described 2 case studies with a multiple measurement design (without baseline). Both participants were adults with mild mental retardation, diagnosed with hypofunctional dysphonia. In the first participant, difficulty in professional integration was mainly due to his insufficient vocal intensity. After traditional voice therapy of one and a half years, he had inconsistent control over increased intensity but there was no transfer to

nontherapy situations. Biofeedback therapy was then initiated for two months, resulting in 23 half-hour sessions. As long as the participant produced a minimal intensity level of 60 dB_{SPL}, a white light was lit on a control unit, endorsing the participants' effort. This therapy had 3 phases: (1) in all sessions, the control unit was placed in front of him during 50 conversational utterances (biofeedback phase installing the louder voice); (2) starting in session 6 and always after phase 1, 50 conversational utterances were produced with the control unit out of his view (knowledge-of-results phase in which the percentage of louder utterances was told after their production); (3) starting in session 13, use of louder voice outside therapy with staff members, friends, etc. was monitored (transfer phase). Results show that at least 74 % of utterances produced with biofeedback reached 60 dB_{SPL}. Outside therapy, the presence of three or more instances of louder voice was confirmed by staff reports and clinician information. There was incomplete consolidation of increased intensity after five months. The second participant had a rapid speech rate and habitual soft voice, resulting in decreased intelligibility. Braking speech rate failed to increase intelligibility and biofeedback therapy was implemented. Differences with the biofeedback therapy from participant one included 1) the use of a less distracting blue night-light, 2) 25 half-hour sessions, 3) a voice intensity threshold of 65 dB_{SPL}, 4) utterances which were elicited by pictorial stimuli, and 5) extra external reinforcement in the form of a cup of coffee. During the initial five sessions, an average of 58 % of the utterances was produced at at least 65 dB_{SPL}. Gradual progress resulted in an average of 85 % during the final five sessions. There was no generalization into spontaneous conversion. Clearly, biofeedback was successful in increasing vocal intensity but no transfer of this ability occurred.

A pre-experimental one-group pretest-posttest study with patients who had undergone a total excision of the larynx (laryngectomy) was presented by Till, England & Law-Till (1987). Typical in these patients is the presence of acoustic noise through the tracheostoma in the neck. This stomal noise depreciates the acceptibility of the speech of laryngectomized patients (Shipp, 1967). In 7 laryngectomees with immoderate stomal noise, they used a design with (1) one initial baseline session, (2) one auditory biofeedback session in which the patients who had undergone a laryngectomy received monaurally their own, amplified stomal noise and (3) one final baseline session without biofeedback. Every session contained the same set of stimuli (isolated vowels, consonant-vowel-consonant words, two syllable words, etc.) Where traditional procedures failed to decrease stomal noise, the averaged group results of this single-session biofeedback therapy showed significant lower levels (reduction of 5-10 dB) of stomal noise. Individual data demonstrated an abrupt onset and offset of the biofeedback-effect. This strengthens the idea that the reduction of stomal noise was due to the auditory biofeedback. It also shows the absence of transfer of the experimental effect, although the poor transfer may have been the result of a low number of treatment sessions.

Table 7.1 Procedure aspects in effect studies on phonatory biofeedback.

Source	Biofeedback Goal	Biofeedback Device		Baseline	Instruction	Adaptable Threshold	Extrinsic Reinforcem. / Punishm.
		Input (Sensor)	*Output*				
Holbrook, Rolnick & Bailey (1974)	Decreasing abusive intensity	Throat microphone (VIC)	A – tone in earphone (when excessively loud speech)	Y	Y	Y	N
Brody, Nelson & Brody (1975)	Increasing hypofunctional intensity	Microphone	V – light bulb going on when voice exceeded 60 dB_{SPL} (as reinforcement)	N	Y	N	N
Till, England & Law-Till (1987)	Reducing stomal noise	Microphone (custom-built stomal noise transducer)	A – stomal noise in monoaural headphone	Y	U	N	N
Howard & Welch (1989)	Increasing pitch matching ability	Microphone SINGAD	V – dotted pitch line	Y	N	Y	N
McGillivray, Proctor-Williams & McLister (1994)	Decreasing abusive intensity	Microphone	A – loud tone going off when voice exceeded 65 dB_A (as punishment)	N	N	N	Y*
Rossiter, Howard & DeCosta (1996)	Increasing voice performance	Microphone	V – ALBERT displayed phonatory parameters (CQ and Ratio %)	Y	N	N	N
Laukkanen et al. 2004)	Increasing overtones at 3-5 kHz (for ringing voice quality)	Microphone	V – FFT and LPC spectrum	Y	Y	N	N

*VIC: Voice Intensity Controller – sEMG: surface electromyography – A: auditory – V: visual – Y: yes / present – N: no / absent – SINGAD: Singing Assessment and Development – CAFET: Computer-Aided Fluency Establishment Training – *: punishment through verbal prompting (for example: 'You were talking too loudly.') – ALBERT: Acoustic and Laryngeal Biofeedback Enhancement Real-Time – CQ: closed quotient (electroglottography) – FFT: Fast Fourier Transform – LPC: Linear Predictive Coding – U: undefined*

Table 7.2 Design aspects in and results of effect studies on phonatory biofeedback.

Source	Biofeedback Goal	Research Design						Results	
		Type	T	Q	P	N	Exp.	Contr.	+/–	LT
Holbrook, Rolnick & Bailey (1974)	Decreasing abusive intensity	One-group pretest-posttest design			*		27	0	+ In 89 % total or partial resolution of VF pathology and/or dysphonia	U
Brody, Nelson & Brody (1975)	Increasing vocal intensity	Two case studies (with multiple measurement)				*	2	0	+ Consistent increase of vocal intensity in both subjects	–
Till, England & Law-Till (1987)	Reducing stomal noise	One-group pretest-posttest design			*		7	0	+ Statistically significant reduction of stomal noise (without transfer)	–
Howard & Welch (1989)	Increasing pitch matching ability	Pretest-posttest matched control group design	*				U	U	+ Statistically significant improvement for exp. group, not for contr. group	U
McGillivray, Proctor-Williams & McLister (1994)	Decreasing abusive intensity	One case study (with multiple measurement)				*	1	0	+ Important decrease of instances of abusive vocal intensity	U
Rossiter, Howard & DeCosta (1996)	Increasing voice performance	One-way repeated measures design		*			1	1	+ Consistent increase CQ and Ratio % with small reduction in final values	U
Laukkanen et al. (2004)	Increasing overtones at 3-5 kHz (for ringing voice quality)	Pretest-posttest randomized control group design	*				6	6	– Statistically non-significant differences in voice quality and spectral slope	U

T: true experimental design – Q: quasi-experimental design – P: pre-experimental design – N: non-experimental design – VF: vocal fold(s) – Y: yes / present – N: no / absent – +: succesful – –: not succesful or not better than other forms of therapy – LT: long-term effects – U: undefined

Singers can also benefit from biofeedback applications. Howard & Welch (1989) employed the SINGAD (Singing Assessment and Development; Howard et al., 1987) system in a pretest-posttest matched control group design in order to train singing ability in primary school children. SINGAD estimates fundamental frequency and gives a visual pitch display. With the SINGAD assessment module, a person's pitch matching ability is measured with an absolute mean semitone difference as result. In the SINGAD development module, a dotted pitch line is visually displayed in real-time, enabling non-singers to develop and dilate their vocal pitch skills. Frequency threshold can be adjusted by setting the fundamental frequency range on the monitor (with complexity increasing from wide frequency range to narrow frequency range). After a baseline SINGAD assessment session, Howard & Welch (1989) divided 32 children in three groups: (a) traditional singing program (singing with guitar accompaniment) for the control group, (b) SINGAD with adult interaction (discussing the feedback and switching to further development screens), and (c) SINGAD without adult interaction. Every child received 7 ten-minute sessions across the school term, followed by a second SINGAD assessment. Whereas the control group showed no significant improvement, both SINGAD groups had significant advancement and those with adult interaction had slightly but not significantly better results than those without adult interaction.

Another multi-measurement case study was presented by McGillivray, Proctor-Williams & McLister (1994). A four-year-old child with vocal nodules spoke with a consistently loud voice and could not decrease her loudness for more than one utterance. An application was conceived which interrupted the conversation with a loud tone (with frequencies between 500 Hz and 10000 Hz) whenever speech outdid a previously determined intensity level (65 dB_A). Further, whenever this tone was produced, all therapy participants stopped talking, with the exception of verbal prompting telling the child about her excessive loudness. In order to determine whether the loudness decrease could be maintained during spontaneous speech, the device was turned on and off for five-minute intervals. In six weekly sessions of 20 to 30 minutes, her speech was followed during playing. Results show a remarkable decrease in instances with vocal intensity above 65 dB_A, from 109 instances in session 1 to 5 instances in session 6.

Rossiter, Howard & DeCosta (1996) used the ALBERT (Acoustic and Laryngeal Biofeedback Enhancement Real Time; Rossiter & Howard, 1995) system in a one-way repeated measures design. ALBERT (Rossiter & Howard, 1995) was designed to develop professional voices and gives visual biofeedback based on the following parameters: fundamental frequency, closed quotient, spectral distribution (Ratio %), sound pressure level, amplitude perturbation (shimmer) and frequency perturbation (jitter). It is able to provide one-, two- and three dimensional graphics in a user-friendly and easily configured interface with the possibility to combine parameters and to work with different visual displays (bar, graph, colour scheme). The study of Rossiter, Howard & DeCosta (1996) consisted of 2 participants without previous experience in voice therapy or vocal

tuition. One participant was conventionally trained, while the other was trained with ALBERT providing EGG biofeedback (CQ or closed quotient), acoustic amplitude singer's formant biofeedback [Ratio % or (summed amplitude levels between 2-4 kHz * 100) / summed amplitude levels between 0-5 kHz], and a combination of CQ and Ratio %. Both participants were assessed on speaking voice and singing voice and sound pressure level measures were taken. Results indicate generally increased sound pressure levels during speaking and singing for both participants. Consistently increased CQ levels with decrease in speaking and singing for one participant in the final values was reported. Consistently increased CQ levels with little decrease in speaking and singing was demonstrated by the second participant (trained with biofeedback). Both participants showed consistently increased Ratio % values during speaking and singing. Rossiter, Howard & DeCosta (1996) did not find clear differences between the participants (thus between conventional techniques and biofeedback). An important observation made from this article is that biofeedback apparently seems to be effective only on the parameter displayed (for example, CQ biofeedback gave much more increase of CQ levels than of Ratio % levels and the reverse).

Finally, Laukkanen et al. (2004) studied the effect of real-time spectrum display on the presence of overtones in the range of 3–5 kHz and the ringing quality of voice in student actors. Both the experimental group (6 participants) and control group (6 participants) were classically trained with the Niilo Kuuka voice exercises. The only difference was that the experimental group could watch the spectrum of their voices while performing the exercises. An important result of this pretest-posttest randomized control group design was the absence of a statistically significant difference in spectral slope and voice quality between the two groups. Another finding is based on the comments of the subjects, who stated that exercising with visual feedback was motivating and seemed to add efficacy to voice treatment.

Obviously, the microphone was therapeutically used for changing acoustical features (related to vocal intensity and frequency). This means that there is no discrepancy with the goal for which it was used, since microphones are developed for acoustic waveform measurements (intensity and pitch). Microphones (and the real-time feedback on acoustical properties) seem to be effective in decreasing abusive vocal intensity (Holbrook, Rolnick & Bailey, 1974; McGillivray, Proctor-Williams & McLister 1994), increasing hypofunctional intensity (Brody, Nelson & Brody, 1975), increasing vocal performance (Howard & Welch, 1989; Rossiter, Howard & DeCosta 1996) and reducing stomal noise (Till, England & Law-Till, 1987), but seems unable to increase overtones (Laukkanen et al., 2004).

DISCUSSION AND CONCLUSION

Some limitations on the scope of this literature review require explicit mention. Information was mainly retrieved through a Medline/Pubmed search;

further information and publications were obtained from references in those articles. Other databases exist in the present health and communication-related outlets (e.g., internet) relevant to biofeedback treatment that were not consulted. Furthermore, this literature review is limited to papers in peer-reviewed journals and therefore, although the review is rather extensive, might be viewed as lacking completeness. A final limitation is, in the attempt to present a realistic review of the current status of the literature, older studies are included that used methodologies no longer seen as optimal.

The purpose of this manuscript was to evaluate acoustic biofeedback as an effective tool for treatment of phonatory disorders and optimalization of phonatory performance. One method potentially available to this review was the use of a meta-analysis. However, in order to use this technique, a statistical pooling of the results of individual studies is necessary (Portney & Watkins, 2000). Due to heterogeneity in measuring unit, methodology and clinical aspects (differences in etiology, age, number of treatment sessions, duration of follow-up, etc.), statistical pooling and thus meta-analysis was determined to be more harmful than useful, so the determination was made to perform a systematic and descriptive literature review. Analysis of studies was hampered by the absence of true experimental research designs. In addition, previous meta-analysis or systematic literature reviews of the effects of an approach in voice treatment were hard to find, and thus, a model from which to begin was not readily available. The extensive literature reviews of Ramig & Verdolini (1998) about the efficacy of voice treatment and of Pannbacker (1999) on treatment for vocal nodules are exceptions to this general statement. The review by Ramig and Verdolini (1998) also identified the need for large group and continued single-subject experimental designs. It is important to reiterate that the real importance of true experimental research (with randomized subject groups and controlled variables) lies in its capability to confirm a causal connection between two phenomena, for example, between the use of biofeedback and the improvement of voice quality. These concerns are evident in the present context. Only 2 of the 8 studies (28.6 %) had a true experimental design. Other problems are also evident; the number of participants in the control and experimental groups are not always given (c.f., Howard & Welch, 1989), and statistical evaluation of the significance of a finding was not always made (Laukkanen et al., 2004). Instead, clinical description sometimes replaced scientific evaluation (for example, in the latter study, the authors pointed out that the biofeedback group tried harder to achieve their goal, suggesting that biofeedback may add some efficacy in voice training).

One study (14.3 %) was quasi-experimental, 2 studies (28.6 %) were pre-experimental and 2 studies (28.6 %) were non-experimental. Although 6 studies (85.7 %) report a positive result (in terms of decrease of laryngeal tension, improvement of voice quality and/or resolution of dysphonia), difficulty exists in confidently attributing the treatment effect to the use of biofeedback. Instead of having clear evidence, there is only a tendency in the advantage of biofeedback. Thus, the question whether or not acoustic biofeedback leads to better voices

should be cautiously answered, and the results of these studies need to be critically interpreted. Biofeedback was almost always useful, and analysis sometimes resulted in a statistically significant finding of reduced vocal intensity and/or better pitch matching. But, as more importance is placed on evidence based practice, there is the need to establish clinical relevance – for example as expressed in terms of number needed to treat (the number of patients who need to be treated to prevent one adverse outcome; Cook & Sackett, 1995) – which is ideally based on randomized controlled trials. Since the latter studies are not available, and because results of individual studies could not be plotted in 2x2 tables, the magnitude of effect and consequently the clinical relevance of the described studies cannot be determined. Still, based on the number of studies in which a positive impact on a voice-related aspect was demonstrated, the authors conclude that acoustic biofeedback appears to be a valuable adjunctive treatment of phonatory disorders.

There have been numerous applications of biofeedback in phonatory disorders (Davis & Drichta, 1980; Volin, 1991). Surface EMG has often been used for reducing laryngeal tension (Stemple et al., 1980; Watanabe et al., 1982; Andrews, Warner & Stewart, 1986; Sime & Healey, 1993; Pettersen & Westgaard, 2002). Regarding the effect of sEMG on voice quality, conflicting results are reported. Whereas sEMG biofeedback appeared to result in an improvement of voice quality in the above mentioned studies, it failed to demonstrate such an effect in other studies. Perhaps noteworthy in this regard was the finding reported by Schliesser (1987), who found no correspondence between voice quality and EMG values. Biofeedback of acoustical features (vocal intensity, pitch, stomal noise, etc., as assessed by the CAFET, SINGAD and ALBERT), on the other hand, resulted in better vocal performance and/or voice quality. Although there are many quantitative biofeedback tools (for training intensity, pitch, pitch matching, etc.), applications of biofeedback therapy using an acoustic voice quality parameter are seldom reported (voice quality is a term that includes all the leftover perceptions after pitch, loudness and phonetic category have been identified; Titze, 1994). Consequently, instead of quantifying pitch, amplitude or phonetic category, some acoustic parameters [as periodicity measures for breathiness (Hillenbrand, Cleveland & Erickson, 1994), harmonics-to-noise ratio for voice quality (Eskenazi, Childers & Hicks, 1990), vocal frequency perturbation for roughness and vocal amplitude perturbation for breathiness (Dejonckere et al., 1996), frequency perturbation for degree of hoarseness (Wuyts et al., 1996), etc.] quantitatively describe aspects of hoarseness (Schoentgen, 2004), and can therefore be related to voice quality. In this context, the therapeutic use of real-time presentation of such an acoustic parameter (such as HNR, jitter, shimmer, etc.) could be situated in the direct treatment of voice quality. Rossiter & Howard (1995), for example, reported on the use of real-time visual jitter display, using the ALBERT-system. Since improvement of voice quality is a very important voice therapy goal (Mathieson, 2001), one would think that direct voice quality biofeedback (using a voice quality parameter, rather than indirect through the use of quantitative pitch training or physiological sEMG training) and its effect on voice quality and vocal pathology

would already been thoroughly investigated. To our knowledge, no such efficacy study has been published through peer-review outlets.

Some clinically-directed or practically-directed disadvantages and advantages regarding biofeedback therapy can be identified. A first disadvantage is that the use of biofeedback in children can be compromised by the fact that they are not always intrinsically motivated in trying to alter physiology. Having produced the correct electrophysiological response (even when realised) does not always work as a reinforcement (Finley et al., 1977). Volin (1991) also mentions the fact that children require stimuli with motivational power in addition to informational content (for example, by working with animated cartoons). A second disadvantage is the high cost of biomedical tools and applications. Biofeedback devices require hardware and software, thus financial investment. A third disadvantage is difficulty with generalization. As shown in Table 7.2, only 2 studies (28.6 %) mentioned long-term results: there was no long-term effect of biofeedback after 5 months (Brody, Nelson & Brody, 1975), or, in one case, even in the second baseline period directly after biofeedback therapy (Till, England & Law-Till, 1987). Several advantages can be enumerated. Today, user-friendly and simple displays, in which the patient is provided with precise information concerning the specifically observed activity involved in speech and voice production, can be offered (Gentil et al., 1994). According to Basmajian (1981), it is not necessary for participants to have any knowledge about the biofeedback modality or instrument. Biofeedback tasks can be easily treated as games, in children as well as in adults (Booker, Rubow & Coleman, 1969). Moreover, dynamic presentations have a motivating effect and modern technology enables digital storage and thus easy therapy follow-up (Sime & Healey, 1993; Gentil et al., 1994) and the voice therapist can add objectivity to therapy techniques (Stemple et al., 1980).

Finally, what are the challenges for future biofeedback therapy in patients with voice disorders and performance concerns? The most important challenge, in phonatory biofeedback therapy as well as in other fields of speech and language pathology, is to expand the body of high-level research studies and to find evidence meeting current standards that justifies a treatment approach. The protocol of the European Laryngological Society for standardization of voice assessment (Dejonckere et al., 2001), for example, could serve as an ideal research tool to evaluate the effect of biofeedback therapy on phonatory disorders and/or vocal performance. The long-term and precise effects of biofeedback therapy on voice disorders should also be explored. A third challenge lies in direct rather than analogue evidence voice quality. Instead of working with quantitative measurements – as for example Holbrook, Rolnick & Bailey (1974), Brody, Nelson & Brody (1975), Till, England & Law-Till (1987), Howard & Welch (1989), and others – this might be best achieved through the use of real-time continuous display of a voice quality related parameter.

REFERENCES

Andrews, S., Warner, J., & Stewart, R. (1986). EMG biofeedback and relaxation in the treatment of hyperfunctional dysphonia. *British Journal of Disorders of Communication, 21,* 353-369.

Aronson, A.E. (1990). *Clinical voice disorders: an interdisciplinary approach.* New York: Thieme.

Basmajian, J.V. (1977). Motor learning and control: a working hypothesis. *Archives of Physical and Medical Rehabilitation, 58,* 38-41.

Basmajian, J.V. (1981). Biofeedback in rehabilitation: a review of principles and practices. *Archives of Physical and Medical Rehabilitation, 62,* 469-475.

Bastian, R.W., & Nagorsky, M.J. (1987). Laryngeal image biofeedback. *Laryngoscope, 97,* 1346-1349.

Booker, H.E., Rubow, R.T., & Coleman, P.J. (1969) Simplified feedback in neuromuscular retraining: an automated approach using electromyographic signals. *Archives of Physical and Medical Rehabilitation, 50,* 621-625.

Boone, D.R., & Mc Farlane, S.C. (1988). *The voice and voice therapy (4th Ed.).* Englewood Cliffs: Prentice-Hall.

Boone, D.R., & Mc Farlane, S.C. (1994). *The voice and voice therapy (5th Ed.).* Englewood Cliffs: Prentice-Hall.

Brody, D.M., Nelson, B.A., & Brody, J.F. (1975). The use of visual feedback in establishing normal vocal intensity in two mildly retarded adult. *Journal of Speech and Hearing Disorders, 40,* 502-507.

Brooks, S., Fallside, F., Gulian, E., & Hinds, P. (1981). Teaching vowel articulation with the Computer Vowel Trainer: methodology and results. *British Journal of Audiology, 15,* 151-163.

Brunner, M., Stellzig, A., Decker, W., Strate, B., Komposch, G., Wirth, G., & Verres, R. (1994). Video-Feedback-Therapie mit dem flexiblen Nasopharyngoskop. *Fortschritte in Kieferorthopädie, 55,* 197-201.

Carding, P.N., Horsley, I.A., & Docherty, G.J. (1999). A study of the effectiveness of voice therapy in the treatment of 45 patients with nonorganic dysphonia. *Journal of Voice, 13,* 72-104.

Cook, R.J., & Sackett, D.L. (1995). The number needed to treat: a clinically useful measure of treatment effect. *British Medical Journal, 310,* 452-454.

Craig, A.R., & Cleary, P.J. (1982). Reduction of stuttering by young male stutterers using EMG feedback. *Biofeedback and Self-Regulation, 7,* 241-255.

Daniel, B., & Guitar, B. (1978). EMG feedback and recovery of facial and speech gestures following neural anastomosis. *Journal of Speech and Hearing Disorders, 43,* 9-20.

D'Antonio, L., Lotz, W., Chait, D., & Netsell, R. (1987). Perceptual-physiologic approach to evaluation and treatment of dysphonia. *Annals of Otology, Rhinology and Laryngology, 96,* 187-190.

Davis, S.M., & Drichta, C.E. (1980). Biofeedback theory and application in allied health: speech pathology. *Biofeedback and Self-Regulation, 5,* 159-174.

Dejonckere, P.H., Remacle, M., Fresnel-Elbaz, E., Woisard, V., Crevier-Buchmann, L., & Millet, B. (1996). Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Revue de Laryngologie-Otologie-Rhinologie, 117,* 219-224.

Dejonckere, P.H., Bradley, P., Clemente, P., Cornut, G., Crevier-Buchman, L., Friedrich, G., Van de Heyning, P., Remacle, M., & Woisard, V. (2001). A basic protocol for functional assessment of voice pathology, especially for the investigating efficacy of (phonosurgical) treatments and evaluating new assessment techniques: guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). *European Archives of Otorhinolaryngology, 258,* 77-82.

Denk, D.M., & Kaider, A. (1997). Videoendoscopic biofeedback: a simple method to improve the efficacy of swallowing rehabilitation of patients after head and neck surgery. *ORL, 59,* 100-105.

Earles, J., Kerr, B., & Kellar, M. (2003). Psychophysiologic treatment of vocal cord dysfunction. *Annals of Allergy, Asthma and Immunology, 90,* 669-671.

Eskanazi, L., Childers, D.G., & Hicks, D.M. (1990). Acoustic correlates of voice quality. *Journal of Speech and Hearing Research, 33,* 298-306.

Fernando, C.K., & Basmajian, J.V. (1978). Biofeedback in physical medicine and rehabilitation. *Biofeedback and Self-Regulation, 3,* 435-455.

Finley, W.W., Niman, C.A., Standley, J., & Wansley, R.A. (1977). Electrophysiologic behavior modification of frontal EMG in cerebral-palsied children. *Biofeedback and Self-Regulation,2,* 59-79.

Fletcher, S.G. (1972). Contingencies for bioelectronic modification of nasality. *Journal of Speech and Hearing Disorders, 37,* 329-346.

Frey, L.R., Botan, C.H., Friedman, P.G., & Kreps, G.L. (1991). *Investigating communication: an introduction to research methods.* Englewood Cliffs: Prentice-Hall.

Garber, S.R., Burzynski, C.M., Vale, C., & Nelson, R. (1979). The use of visual feedback to control vocal intensity and nasalization. *Journal of Communication Disorders, 12,* 399-410.

Gentil, M., Aucouturier, J.L., Delong, V., & Sambuis, E. (1994). EMG biofeedback in the treatment of dysarthria. *Folia Phoniatrica et Logopaedica, 46,* 188-192.

Goebel, M. (1986). *A computer-aided fluency establishment trainer (CAFET).* Falls Church: Annadale Fluency Clinic.

Goldstein, P., Ziegler, W., Vogel, M., & Hoole, P. (1994). Combined palatal-lift and EPG-feedback therapy in dysarthria: a case study. *Clinical Linguistics & Phonetics, 8,* 201-218.

Guitar, B. (1975). Reduction of stuttering frequency using analog electromyographic feedback. *Journal of Speech and Hearing Research, 18,* 672-685.

Hammerschlag, P.E., Brudny, J., Cusumano, R., & Cohen, N.L. (1987). Hypoglossal-facial nerve anastomosis and electromyographic feedback rehabilitation. *Laryngoscope, 97,* 705-709.

Hand, C.R., Burns, M.O., & Ireland, E. (1979). Treatment of hypertonicity in muscles of lip retraction. *Biofeedback and Self-Regulation, 4,* 171-181.
Hanna, R., Wilfling, F., & McNeill, B. (1975). A biofeedback treatment for stuttering. *Journal of Speech and Hearing Disorders, 40,* 270-273.
Henschen, T.L., & Burton, N.G. (1978). Treatment of spastic dysphonia by EMG biofeedback. *Biofeedback and Self-Regulation, 3,* 91-96.
Hillenbrand, J., Cleveland, R.A., & Erickson, R.L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research, 37,* 769-778.
Hirano, M. (1981). *Clinical examination of voice.* New York: Springer-Verlag.
Holbrook, A., Rolnick, M.I., & Bailey, C.W. (1974). Treatment of vocal abuse disorders using a vocal intensity controller. *Journal of Speech and Hearing Disorders, 39,* 298-303.
Horii, Y. (1983). Automatic analysis of voice fundamental frequency and intensity using a visi-pitch. *Journal of Speech and Hearing Research, 26,* 467-471.
Howard, D.M., Welch, G.F., Gibbon, R.R., & Bootle, C. (1987). The assessment and development of singing ability: initial results with a new system. *Proceedings of the Institute of Acoustics, 9,* 159-166.
Howard, D.M., & Welch, G.F. (1989). Microcomputer-based singing ability assessment and development. *Applied Acoustics,* 27, 89-102.
Huckabee, M.L., & Pelletier, C.A. (1999). *Management of adult neurogenic dysphagia.* San Diego: Singular.
Huckvale, M. (2005). SFS/RTSPECT version 2.1: Windows tool for real-time waveforms and spectra. University College London. Retrieved in June 2005, http://www.phon.ucl.ac.uk/resource/sfs/rtspect/.
Huckvale, M. (2005). SFS/RTGRAM version 1.1: Windows tool for real-time speech spectrogram display. University College London. Retrieved in June 2005, http://www.phon.ucl.ac.uk/resource/sfs/rtgram/.
Jankel, W.R. (1978). Bell palsy: muscle reeducation by electromyograph feedback. *Archives of Physical and Medical Rehabilitation, 59,* 240-242.
Künzel, H.J. (1982). First applications of a biofeedback device for the therapy of velopharyngeal incompetence. *Folia Phoniatrica, 34,* 92-100.
Laukkanen, A.M., Syrjä, T., Laitala, M., & Leino, T. (2004). Effects of two-month vocal exercising with and without spectral biofeedback on student actors' speaking voice. *Logopedics Phoniatrics Vocology, 29,* 66-76.
Mathieson, L. (2001). *Greene and Mathieson's The voice & its disorders (6th ed.).* London: Whurr.
McGillivray, R., Proctor-Williams, K., & McLister, B. (1994). Simple biofeedback device to reduce excessive vocal intensity. *Medical & Biological Engineering & Computing, 32,* 348-350.
McGrady, A., & Bailey, B. (2003). *Diabetes mellitus.* In Schwartz M.S. and Andrasik F. (Eds.), Biofeedback, a practitioner's guide. New York: Guilford Press.

McGrady, A., & Linden, W. (2003). *Biobehavioral treatment of essential hypertension*. In Schwartz M.S. and Andrasik F. (Eds.), Biofeedback, a practitioner's guide. New York: Guilford Press.

Michi, K., Yamashita, Y., Imai, S., Suzuki, N., & Yoshida, H. (1993). Role of visual feedback treatment for defective /s/ sounds in patients with cleft palate. *Journal of Speech and Hearing Research, 36,* 277-285.

Moller, K.T., Path, M.P., Werth, L.J., & Christiansen, R.L. (1973). The modification of velar movement. *Journal of Speech and Hearing Disorders, 38,* 323-334.

Murdoch, B.E., Pitt, G., Theodoros, D.G., & Ward, E.C. (1999). Real-time continuous visual biofeedback in the treatment of speech breathing disorders following childhood traumatic brain injury: report of one case. *Pediatric Rehabilitation, 3,* 5-20.

Nemec, R.E., & Cohen, K. (1984). EMG biofeedback in the modification of hypertonia in spastic dysarthria: case report. *Archives of Physical and Medical Rehabilitation, 65,* 103-104.

Netsell, R., & Cleeland, C.S. (1973). Modification of lip hypertonia in dysarthria using EMG feedback. *Journal of Speech and Hearing Disorders, 38,* 131-140.

Netsell, R., & Daniel, B. (1979). Dysarthria in adults: Physiologic approach to rehabilitation. *Archives of Physical and Medical Rehabilitation, 60,* 502-508.

Newman, K.B., & Dubester, S.N. (1994). Vocal cord dysfunction: masquerader of asthma. *Seminars in Respiratory Critical Care Medicine, 15,* 161-167.

Pannbacker, M. (1999). Treatment of vocal nodules: options and outcomes. *American Journal of Speech-Language Pathology, 8,* 209-217.

Pettersen, V., & Westgaard, R.H. (2002). Muscle activity in the classical singer's shoulder and neck region. *Logopedics Phoniatrics Vocology, 27,* 169-178.

Portney, L.G., & Watkins, M.P. (2000). *Foundations of clinical research: applications to practice (2nd Ed.).* Upper Saddle River: Prentice-Hall.

Prosek, R.A., Montgomery, A.A., Walden, B.E., & Schwartz, D.M. (1978). EMG biofeedback in the treatment of hyperfunctional voice disorders. *Journal of Speech and Hearing Disorders, 43,* 282-294.

Ramig, L.O., & Verdolini, K. (1998). Treatment efficacy: voice disorders. *Journal of Speech, Language and Hearing Research, 41,* S101-S116.

Rammage, L., Morrison, M., & Nichol, H. (2001). *Management of the voice and its disorders (2nd ed.)*. San Diego: Singular.

Rossiter, D., & Howard, D.M. (1995). ALBERT: a real-time visual feedback computer tool for professional vocal development. *Journal of Voice, 10,* 321-336.

Rossiter, D., Howard, D.M., & DeCosta, M. (1996). Voice development under training with and without the influence of real-time visually presented biofeedback. *Journal of the Acoustical Society of America, 99,* 3253-3256.

Rubow, R.T., Rosenbek, J.C., Collins, M.J., & Celesia, G.G. (1984). Reduction of hemifacial spasm and dysarthria following EMG biofeedback. *Journal of Speech and Hearing Disorders, 49,* 26-33.

Rubow, R., & Swift, E. (1985). A microcomputer-based wearable biofeedback device to improve transfer of treatment in Parkinsonian dysarthria. *Journal of Speech and Hearing Disorders, 50,* 178-185.

Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B., & Richardson, W.S. (1996). Evidence-based medicine: what it is and what it isn't. *British Medical Journal, 312,* 71-72.

Schliesser, H.F. (1987). EMG biofeedback as a function of voice-quality change. *Perceptual & Motor Skills, 64,* 719-724.

Schoentgen, J. (2004). Perspectives in acoustic analysis of voice disorders. *Proceedings of VOX 2004.* Nijmegen: University Medical Center St. Radboud.

Schram, G., & Burres, S. (1984). *Non surgical rehabilitation after facial paralysis.* In Portmann M. (Ed.), Proceedings of the Vth International Symposium on the Facial Nerve. New York: Masson.

Schwartz, M.S., & Olson, R.P. (2003). *A historical perspective on the field of biofeedback and applied psychophysiology.* In Schwartz M.S. and Andrasik F. (Eds.), Biofeedback, a practitioner's guide. New York: Guilford Press.

Schwartz, N.M., & Schwartz, M.S. (2003). *Definitions of biofeedback and applied psychophysiology.* In Schwartz M.S. and Andrasik F. (Eds.), Biofeedback, a practitioner's guide. New York: Guilford Press.

Shelton, R.L., Paesani, A., McClelland, K.D., & Bradfield, S.S. (1975). Panendoscopic feedback in the study of voluntary velopharyngeal movements. *Journal of Speech and Hearing Disorders, 40,* 232-244.

Shipp, T. (1967). Frequency, duration, and perceptual measures in relation to judgments of alaryngeal speech and acceptability. *Journal of Speech and Hearing Research, 10,* 412-427.

Sime, W.E., & Healey, E.C. (1993). An interdisciplinary approach to the treatment of a hyperfunctional voice disorder. *Biofeedback & Self-Regulation, 18,* 281-287.

Stark, R.E. (1971). The use of real-time visual displays of speech in the training of a profoundly deaf nonspeaking child: a case report. *Journal of Speech and Hearing Disorders, 36,* 397-409.

Stemple, J.C., Weiler, E., Whitehead, W., & Komray, R. (1980). Electromyographic biofeedback training with patients exhibiting a hyperfunctional voice disorder. *Laryngoscope, 90,* 471-476.

Stoyva, J. (1976). Self-regulation: A context for biofeedback. *Biofeedback & Self-Regulation, 1,* 1-6.

Till, J.A., England, K.E., & Law-Till, C.B. (1987). Effects of auditory feedback and phonetic context on stomal noise in laryngectomized speakers. *Journal of Speech and Hearing Disorders, 52,* 243-250.

Titze, I.R. (1994). *Principles of voice production.* Englewood Cliffs: Prentice-Hall.

Tries, J., & Eisman, E. (2003). Urinary incontinence: Evaluation and biofeedback treatment. In Schwartz M.S. and Andrasik F. (Eds.), Biofeedback, a practitioner's guide. New York: Guilford Press.

Van Lierde, K.M., Claeys, S., De Bodt, M., & Van Cauwenberge, P. (2004). Outcome of laryngeal and velopharyngeal biofeedback treatment in children and young adults: & pilot study. *Journal of Voice, 18,* 97-106.

Van Borsel, J. (2004). *Wetenschappelijk onderzoek in de logopedie. [Scientific research in speech-language therapy.]* Leuven: Acco.

Volin, R.A. (1991). Microcomputer-based systems providing biofeedback of voice and speech production. *Topics in Language Disorders, 11,* 65-79.

Watanabe, H., Komiyama, S., Ryu, S., Kannae, S., & Matsubara, H. (1982). Biofeedback therapy for spastic dysphonia. *Auris Nasus Larynx (Tokyo), 9,* 183-190.

Whitehill, T.L., Stokes, S.F., & Man, Y. (1996). Electropalatography treatment in an adult with late repair of cleft palate. *Cleft Palate-Craniofacial Journal, 33,* 160-168.

Wuyts, F.L., De Bodt, M., Bruckers, L., & Molenberghs, G. (1996). Research work of the Belgian Study Group on Voice Disorders: results. *Acta Otorhinolaryngologica Belgica, 50,* 331-341.

Wuyts, F.L., De Bodt, M.S., Molenberghs, G., Remacle, M., Heylen, L., Millet, B., Van Lierde, K., Raes, J., & Van de Heyning, P.H. (2000). The Dysphonia Severity Index: an objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language and Hearing Research, 43,* 796-809.

9

GENERAL CONCLUSIONS

ACOUSTIC MEASUREMENT OF VOICE QUALITY IN SUSTAINED VOWELS AND CONTINUOUS SPEECH

Voice quality is that attribute of auditory sensation that includes all the leftover perceptions after pitch, loudness and phonetic category have been identified. In cases with an organic or a non-organic voice disorder, there can be a deterioration of the vibratory patterns of the vocal folds, causing disruption of the voice quality (i.e., dysphonia). In these cases, the voice is typically perceived as rough and/or breathy. Information on voice quality can be traced in the acoustic waveform. This waveform can be captured and processed by the auditory system of the listener. It can also be recorded by a computerized data acquisition system. Both options are commonly used in the clinical voice assessment, with the subjective auditory-perceptual evaluation and the objective acoustic analysis, respectively. This study addressed several methodological issues related to both perceptual rating and acoustic parameterization of overall voice quality. We concentrated on the implementation of sustained vowels as well as continuous speech in clinical assessment of dysphonia severity. This research leads to the following conclusions.

1 Study of the agreement and differences between two computer programs commonly used for voice perturbation measures

In Chapter 2, we examined correlations and differences between two common computer programs on seven perturbation measures (absolute jitter, percent jitter, relative average perturbation, pitch perturbation quotient, shimmer in dB, percent shimmer, and amplitude perturbation quotient) in 50 subjects with various voice disorders. Sustained vowels were recorded with the Computerized speech lab (i.e., CSL) and subsequently analyzed with two computer programs: Multi-dimensional voice program (i.e., MDVP) and Praat. This method allowed keeping data acquisition-related items (e.g., microphone type and placement, recording environment, external hardware, internal sound card, etc.) invariant, which provided an opportunity to investigate the influence of software-related items (mainly the fundamental period detection algorithm) on perturbation outcomes.

We found weak to moderate correlations for the four fundamental frequency perturbation measures and moderate to strong correlations for the amplitude perturbation measures. This implies that F_0 perturbation measures are far more susceptible for small errors in the determination of cycle boundaries than amplitude perturbation measures. We also found statistically significant differences for all the perturbation measures, with Praat data being consistently lower than MDVP data. Again, this was explained by the difference in the T_0 detection algorithm of both programs. However, the relatively low correlations and the absence of a prominent linear relationship precluded a mathematical transformation of one measure to its equivalent in the other program.

2 Study of the agreement and differences between two computer systems commonly used for voice perturbation measures

We also studied correlations and differences between data acquisition with two common hardware systems for the seven perturbation measures in Chapter 2. Computer system 1 consisted of CSL equipped with MDVP. Computer system 2 consisted of a common desktop personal computer (i.e., PC) installed with Praat.

The results of this study showed statistically significant differences for all the perturbation measures, with CSL-system measures being consequently higher than the measures of the PC-system. Furthermore, we only found weak to moderate correlations for all perturbation measures of the two computer systems. Stated otherwise, there is no proportional relationship between the outcomes of both systems. Based on these results, we concluded that data acquisition environment, microphone placement and software for acoustic analysis interfere with the outcome of the perturbation measures.

The clinical consequence of these two studies, is that for the time being one should not directly compare perturbation measures across computer systems and programs (e.g., between different voice clinics). The only feasible comparison is the comparison of perturbation measures within computer systems and programs (e.g., from pre- to post-treatment).

3 Meta-analytic study of the concurrent validity of acoustic algorithms to measure overall voice quality in sustained vowels

In the literature, there is a plethora of acoustic algorithms claiming the sensitivity to adequately measure overall voice quality in sustained vowels. Through meta-analysis (i.e., the "analysis of analyses" for amalgamating, summarizing, and reviewing previous quantitative research) we were able to reconsider in a comprehensive way the correlational findings of twenty-one reports. Individual correlation coefficients between specific acoustic metrics and perceptual ratings of overall voice quality in a sustained vowel were weighted on the basis of sample size and then averaged. Furthermore, homogeneity of these correlation coefficients was investigated.

Results of this meta-analysis revealed four promising measures out of a set of sixty-nine, yielding homogeneous weighted average correlation coefficients of at least 0.60 in sustained vowels: smoothed cepstral peak prominence, spectral flatness of the residue signal, Pearson r at autocorrelation peak and pitch amplitude. All other acoustic measures, including popular and frequently investigated perturbation measures, did not reach the required level of weighted averaged correlation.

4 Meta-analytic study of the concurrent validity of acoustic algorithms to measure overall voice quality in continuous speech

Acoustic measures have also been investigated as metrics of overall voice quality in continuous speech. Following the same meta-analytic approach as for sustained vowels, we integrated the correlational results of seven reports.

This meta-analysis included a set of twenty-six acoustic measures, only three of which yielded a homogeneous weighted average correlation of at least 0.60 in continuous speech: signal-to-noise ratio from Qi, cepstral peak prominence and smoothed cepstral peak prominence. The required level of weighted averaged correlation was not reached by the other acoustic measures, including the popular and most commonly investigated perturbation measures.

Collectively, the two cepstral measures, and smoothed cepstral peak prominence in particular, seem to be good correlates of overall voice quality, giving rise to the potentially most accurate predictive acoustic algorithms.

5 Study of the feasibility of concatenating samples of sustained vowels and continuous speech for perceptual and acoustic measurements of overall voice quality

One of the main statements in this thesis is that perceptual and acoustic assessment of disordered voice quality should ideally include both sustained vowels and continuous speech, in order to be representative of daily speech and voice use patterns (i.e., in order to be ecologically valid). We therefore concatenated the first two sentences of a frequently used Dutch text ('Papa en Marloes staan op het station. Ze wachten op de trein') with a three second midvowel extract of a sustained /a./. This sound chain was used for perceptual ratings. To obtain a sample to derive objective measures from, we used a customized programming script to automatically extract all voiced segments from these two sentences. These segments were than concatenated with the three second midvowel extract. As such, both the perceptual ratings and the acoustic measures pertained to the same voice samples.

Based on our experiences during three studies in which this script was applied to 323 laryngeal-vocal and 16 tracheoesophageal voice samples (see Chapters 4, 5 and 6), we can conclude that is it clinically feasible to work with concatenated voice samples. An important caveat, however, is related to the reliability of the perceptual ratings of these samples. Inspite of the experience of the five listeners in rating dysphonia severity in a clinical setting notwithstanding, the novelty of rating these 'new' concatenated samples may have decreased the reliability of their ratings, which may have interfered with the experimental outcome of the acoustic measure. Nevertheless, the perceptual ratings of the three experiments in Chapters 4, 5 and 6 yielded acceptable reliability levels.

6 Study of the criterion-related concurrent validity of thirteen acoustic metrics to measure overall voice quality in sustained vowels and continuous speech

The ability to predict the perceived degree of dysphonia severity in concatenated voice samples was investigated for thirteen acoustic measures. Eleven traditional acoustic metrics were obtained via the Praat computer program: slope of the long-term average spectrum (Slope), tilt of the trend line through the long-term average spectrum (Tilt), jitter local (a.k.a. percent jitter), jitter rap (a.k.a. relative average perturbation), jitter ppq5 (a.k.a. pitch perturbation quotient), shimmer local (a.k.a. percent shimmer), shimmer local dB (a.k.a. shimmer in dB), shimmer apq11 (a.k.a. amplitude perturbation quotient), mean autocorrelation (mACF), noise-to-harmonics ratio (NHR) and harmonics-to-noise-ratio (HNR). Two less common metrics were obtained in the SpeechTool computer program: cepstral peak prominence (CPP) and smoothed cepstral peak prominence (CPPs).

The correlation coefficients between perceptual ratings and the thirteen acoustic measures revealed that smoothed cepstral peak prominence was the best predictor of overall voice quality. This finding confirmed the literature on cepstral measures and thus corroborated the main conclusion of the meta-analysis in Chapter 3. Furthermore, cepstral peak prominence, harmonics-to-noise ratio and the amplitude perturbation measures yielded a $r>0.60$. The inferiority of the fundamental frequency perturbation measures also confirmed literature and the meta-analysis.

7 Construction of a multivariate model of acoustic markers to measure overall voice quality in sustained vowels and continuous speech

Prompted by the multidimensional nature of voice and by the limited predictive power of single acoustic markers, we explored multivariate statistics for the prediction of the level of voice quality disruption and for the discrimination among different perceptual categories/levels of dysphonia severity. A stepwise multiple linear regression procedure resulted in a six-factor model, called 'Acoustic Voice Quality Index' (i.e., 'AVQI'). The equation for AVQI, after linear rescaling to obtain values between 0 and 10, is:

$$\begin{aligned}\text{AVQI} = (&3.295 - 0.111 \times \text{CPPs} - 0.073 \times \text{HNR} - 0.213 \times \\ &\text{shimmer local} + 2.789 \times \text{shimmer local dB} - 0.032 \times \text{slope} + \\ &0.077 \times \text{tilt}) \times 2.571\end{aligned} \qquad \text{(Eq. 1)}$$

Based on our experience and as concluded in Chapter 4, the clinical feasibility of AVQI can be supported, especially because its calculation can easily be implemented in a customized AVQI Praat programming script.

8 Study of the concurrent validity and the diagnostic precision of the Acoustic Voice Quality Index

An important issue in the validation of a measurement tool is the assessment of its concurrent validity. We therefore calculated a correlation coefficient between the AVQI and the perceptual ratings of the initial 251 subjects. A second issue is the evaluation of its diagnostic accuracy, and we therefore determined the area under the ROC curve and the positive and negative likelihood ratios. Both diagnostic accuracy and concurrent validity yielded acceptable statistics. Results (e.g., predictive equation, cutoff score, etc.) based on an experimental sample (i.e., the sample described in Chapter 4) are designed to be used in different and larger populations.

9 Study of the internal and external concurrent and diagnostic cross-validity of this statistical model for acoustic measurement

However, data gathered on the initial experimental sample can differ from data that would be obtained from a sample with different subjects. Consequently, the "error" variance in AVQI is expected to be greater in another set of subjects. This means that AVQI's validity and accuracy will not necessarily be as prominent as it is for the original experimental sample. We therefore investigated the internal cross-validity (see Chapter 4) and the external cross-validity (see Chapter 5) of AVQI. Results from these additional studies confirmed the strong relationship between perceptual ratings and the AVQI-scores and the strong power to distinguish normal (AVQI<2.95) from pathological voices (AVQI≥2.95). We therefore can conclude that the AVQI is a valid measure of dysphonia severity.

10 Study of the responsiveness to change of the Acoustic Voice Quality Index

Another important issue of the AVQI is its sensitivity to detect change and its suitability as a voice treatment outcomes measure. We therefore compared changes in perceived and acoustically measured overall voice quality from pre-treatment to post-treatment voice recordings. Recordings from 33 subjects, who underwent surgical and/or individually customized behavioral voice treatment, were carefully selected to represent various degrees of change in dysphonia severity (i.e., from clear improvement to absence of change). In order to obtain unit free outcome scores for the comparison, we calculated standardized change scores for both the perceptual ratings and the AVQI-scores. Subsequently, the change responsiveness of AVQI was analyzed by correlating the standardized change score in perceived dysphonia severity with the standardized change score in AVQI. The higher the correlation between these standardized change scores, the more AVQI can be considered to be a responsive treatment outcome measure, sensitive to changes in perceived dysphonia severity.

A strong proportional relationship was found between the standardized change scores in perceived dysphonia severity and the standardized change scores in AVQI. This result supports the susceptibility of the AVQI to quantify treatment-related changes in dysphonia severity. It should therefore be regarded as a valid measure of voice treatment outcomes.

11 Study of the concurrent validity of forty-seven acoustic measures of overall tracheoesophageal voice quality in sustained vowels and continuous speech

Tracheoesophageal voice production is the preferred method of speech rehabilitation after total laryngectomy. This tracheoesophageal voice, however, can vary substantially in terms of voice quality. For the clinical management of tracheoesophageal voice (e.g., decision-making regarding prolongation of voice and speech therapy, monitoring of therapy effectiveness, comparison between laryngectomees, etc.) it is important to quantify the degree of alaryngeal voice quality. We therefore explored the validity of forty-seven acoustic time-, frequency- and quefrency-domain markers as measures of auditorily perceived overall tracheoesophageal voice quality in sixteen laryngectomees (see Chapter 6). To our knowledge, tracheoesophageal voice quality has never been analysed with measures of cepstral peak prominence and spectral peak prominence. We also investigated a multivariate approach to boost the correlation between perceived and acoustically measured voice quality.

First, the results of this study demonstrate that the cepstral measures (cepstral peak prominence and smoothed cepstral peak prominence) correlate best with perceived tracheoesophageal voice quality. This finding agrees with the results of many other studies in the literature and in this thesis that indicate the cepstral measures as promising measures of laryngeal voice quality. Second, the height of the first two harmonics, as well as the amplitude perturbation measures and the harmonics-to-noise ratio, can also be associated with the quality of tracheoesophageal voice. Third, other geometrical properties than the height of the harmonics and the fundamental frequency perturbation measures are not related to perceived tracheoesophageal voice quality. This finding also corroborates with other studies in which jitter proved to be insufficiently associated with laryngeal voice quality. Fourth, stepwise linear regression analysis revealed that a combination of the cepstral peak prominence and the height of the second harmonic best predicted of overall tracheoesophageal voice quality. Based on this report, we can conclude that it is viable to implement specific time-, frequency- and quefrency-domain properties in the assessment of sustained vowels and continuous speech from laryngectomees.

12 Systematic literature review of the effects of acoustic biofeedback in the management of voice disorders and vocal performance

Acoustic biofeedback is sometimes used in voice clinics to provide immediate information on the performance during vocal tasks/exercises. Does this biofeedback cause the voice to perform better? Answers to this question were sought in the literature. Seven reports on acoustic biofeedback were considered to be eligible for this literature review. They were scrutinized on the type of biofeedback device (i.e., input and output modalities), the type of biofeedback protocol (e.g., instruction, threshold, etc.), on their research design and on reported therapy outcomes.

The results of this review underscore the usefulness of acoustic biofeedback in decreasing/increasing vocal intensity, reducing stomal noise and improving pitch matching. However, the most important limitation in this review was the absence of randomized controlled studies, hampering conclusions regarding the causative relation between acoustic biofeedback and therapy outcome. Nevertheless, given the number of reports in which a positive impact on a voice-related aspect was demonstrated, we cautiously concluded that acoustic biofeedback is a valuable adjunctive tool in the treatment of phonatory disorders. Furthermore, the behavioral voice therapy of almost all patients in our study regarding AVQI's responsiveness to change (see Chapter 5) also consisted of real-time acoustic (mainly narrowband-spectrographic) biofeedback as one of the treatment tools/techniques. This indirectly emphasizes the feasibility of acoustic biofeedback in the treatment of dysphonia. However, given the nature of the study of AVQI's sensitivity to change (i.e., evaluating of an acoustic model instead of evaluating treatment outcome), the question whether or not this biofeedback lead to the measured improvement in overall voice quality remains unanswered.

SUMMARY

Clinical measurement of dysphonia severity typically involves auditory-perceptual evaluations and acoustic analyses of the sound wave. Meta-analysis of the proportional association between these two methods (Chapter 3) showed that many of the popular perturbation metrics and noise-to-harmonics and others ratios do not yield sufficiently strong correlations with perceptual overall dysphonia ratings. However, this meta-analysis also demonstrated that the validity of specific autocorrelation- and cepstrum-based measures of 'periodicity prominence' (i.e., that do not rely on pitch detection) was much more convincing, and appointed 'smoothed cepstral peak prominence' as the most promising and valuable metric of overall voice quality. Original research of the correlation between auditory-perceptual ratings and many pitch detection-based as well as cepstrum-based measures confirmed the inferiority of the perturbation measures. Interestingly, the smoothed cepstral peak prominence yielded the highest correlation, boosting the superiority of the cepstral indices in dysphonia measurement of laryngeal-vocal

voice samples (Chapter 4). The importance of the relative dominance of the first rahmonic in voice quality measurement was additionally bolstered in original research with tracheoesophageal voice samples (Chapter 6), in which the 'cepstral peak prominence' was reported to be the best correlate of overall dysphonia.

To be truly representative for daily voice use patterns (i.e., to be considered ecologically valid), clinical measurement of overall voice quality is ideally founded on the analysis of both sustained vowels and continuous speech. A customized method for the extraction of voiced segments in continuous speech, the concatenation with the mid-portion of a sustained vowel, and the calculation of the multivariate Acoustic Voice Quality Index (i.e., AVQI) was constructed for this purpose. The main contributor to this 6-factor model was the smoothed cepstral peak prominence. Original methodological study of the AVQI revealed acceptable results in terms of initial concurrent validity, diagnostic precision, internal and external cross-validity and responsiveness to change (Chapters 4 and 5). It thus was concluded that the AVQI is a clinically feasible method to track changes in dysphonia severity across the voice therapy process.

There are many freely and commercially available computer programs and systems that provide acoustic metrics of dysphonia severity. However, the data across these programs and systems cannot always be compared. We therefore investigated the agreements and differences between two commonly available programs (i.e., Praat and Multi-Dimensional Voice Program) and systems (Chapter 2). The results indicated that clinicians better not compare frequency perturbation data across systems and programs and amplitude perturbation data across systems.

Finally, acoustic information (i.e., regarding fundamental frequency, intensity and vocal quality) can also be therapeutically utilized as a biofeedback modality during voice exercises. Based on a systematic literature review (Chapter 8), it was cautiously concluded that acoustic biofeedback appears to be a valuable adjunctive tool in the treatment of phonatory disorders and performance.

It can generally be concluded that, when applied with caution, acoustic algorithms (particularly the cepstrum-based measures and the AVQI) have merited a special role in the assessment and/or in the treatment of dysphonia severity.

WEAKNESSES

Considering the limitations and caveats as expressed in the different chapters, we iterate the shortcomings in this thesis.

- The meta-analysis, as well as the studies with the AVQI and with the tracheoesophageal voice recordings, only concentrated on overall voice quality and dysphonia severity. It is important to recognize, however, that clinical assessment of voice also focuses on particular dysphonia types, namely breathiness and roughness. This thesis draws no conclusion nor gives any recommendation on the acoustic measurement of these specific vocal qualities.

▶ The meta-analysis only focused on the correlation coefficient as effect size of the validity of the acoustic measures. It should however be emphasized that there are other valuable statistics to investigate the validity, such as the statistics of diagnostic precision (e.g., sensitivity, specificity, area under the ROC curve, etc.).

▶ This thesis identified measures such as cepstral peak prominence and pitch amplitude as most promising acoustic markers of dysphonia severity. Furthermore, it indicated the most commonly used perturbation measures (and specifically the fundamental frequency perturbation measures) as measures with insufficient validity. It has been evidenced that data acquisition and analysis algorithms have a statistically significant influence on the outcome of these perturbation measures. This thesis however never examined the influence of the data acquisition environment on the accuracy of the more promising cepstrum- or autocorrelation-based measures.

▶ An essential issue in many of the studies covered by the meta-analyses or raised by the original studies in this thesis, is the reliability/unreliability of auditory-perceptual ratings. Although it was concluded that the raters mostly reached levels of reasonable inter- and intra-rater reliability (e.g., in the AVQI studies), we believe that moderate raters' reliability might have limited the correlation coefficient between the acoustic measures and the perceptual evaluations of dysphonia severity, and consequently lessened the predictive and diagnostic potential of the acoustic measures and models.

▶ In some of the original studies in this thesis we concatenated sustained vowel with continuous speech samples. These samples were not only used to obtain the acoustic measures, but also to be auditory-perceptually rated by a panel of experienced raters. Since it was this panel's judgment that was to be predicted by the acoustic measures/models, it is crucial to find out what determines the final perceptual evaluation. Is it the dysphonia severity in the sample type with the heaviest dysphonia? Or is it, on the contrary, the dysphonia severity of the speaking tasks with the slightest dysphonia? Is there a primacy or recency effect involved or is the final judgment determined by the average dysphonia severity in both sample types? An answer to these questions has not been stated in this thesis.

▶ In addition to the sustained vowels, we recorded oral readings of a Dutch text from all participants. For practical reasons (e.g., computer analysis time) and in accordance with the protocol of other studies on continuous speech, we only analyzed the first two sentences. This implies that the results of our studies only apply to these 'partial' samples of continuous speech and that longer samples of continuous speech (e.g., readings of the complete text) might have resulted in different and maybe even more valid findings.

▶ The number of subjects was sufficient in all our original studies, except for the study on tracheoesophageal voice quality. In this specific study, we only examined sixteen subjects, which limits generalizability of these specific results.

▶ Quality of life has become the key feature in healthcare and improved qualify of life is the primary goal in the treatment. It is therefore important to understand the impact of dysphonia on the quality of life and to investigate the relations between outcomes on quality of life-related questionnaires (e.g., Voice Handicap Index and Voice-Related Qualife of Life) and voice quality-related measures (e.g., AVQI). However, these items have not been addressed in this thesis.

STRENGTHS

▶ The research in this thesis and the conclusions resulting from it are based on a strong foundation of literature review and meta-analysis. Because literature repetitively indicated cepstral metrics as superior measures of dysphonia severity, we continued to investigate their validity in the concatenated voice samples. Our research confirms the superiority of cepstral peak prominence and smoothed cepstral peak prominence when compared to time-domain measures (e.g., jitter local, shimmer local dB and harmonics-to-noise ratio) and frequency-domain measures (e.g., slope of the long-term average spectrum and tilt of the trend line through the long-term average spectrum). It is important to state that this superiority not only prevailed in our research on laryngeal voice samples, but that it was also confirmed in our exploratory investigation of tracheoesophageal voice samples. As such, our research is in agreement with many reports on cepstral and other measures and it amplifies the conclusions from the meta-analysis.

▶ It came as no surprise that the cepstral metrics forms the most important item in our multivariate models of dysphonia severity (i.e., the AVQI and the two-factor model on tracheoesophageal voice samples). We have examined different aspects of AVQI's validity: initial concurrent validity, initial diagnostic accuracy, internal and external concurrent cross-validity, external diagnostic cross-validity and responsiveness to change. All these investigations repetitively confirmed the AVQI as a highly valuable voice treatment outcomes measuring tool. Furthermore, the prominence of the first rahmonic was also the best correlate of tracheoesophageal voice quality. The latter finding is especially interesting, because it proves the robust feasibility of the computer algorithm for the cepstral analysis and for the determination of the cepstral peak prominence and smoothed cepstral peak prominence and the other AVQI-measures.

▶ The ultimate goal of this thesis was to create a protocol with which continuous speech could be implemented in the objective assessment of voice quality and dysphonia severity. With the presented protocol (i.e., decomposition of the continuous speech in voiced and unvoiced segments, extraction and concatenation

of the voiced segments, final concatenation of this voiced speech with the central three seconds of a sustained vowel, and acoustic analysis of the concatenated samples) we succeeded in remodeling the continuous speech samples and implementing them in a clinical voice assessment. This protocol, as we used it to obtain the AVQI and to analyze the tracheoesophageal voice samples, showed to be very feasible in almost all voice samples across the different original studies in this thesis.

▶ The AVQI, with a scale from 0 to 10 and with a threshold value of 2.95, is a readily interpretable measure of dysphonia severity. Furthermore, because the AVQI consists of six acoustic measures that can be obtained using freely available software (the programs Praat and SpeechTool) and customized programming scripts, it can rather easily and directly be inserted in a voice assessment protocol across voice clinics.

FUTURE PERSPECTIVES

The following areas of future research have been identified based on the strengths and weaknesses of the presented research.

▶ To use the meta-analytic approach for the appraisal of the acoustic-phonetic predictors of breathiness and the establishment of population relationship estimates for several acoustic measures.

▶ To use the meta-analytic approach for the appraisal of the acoustic-phonetic predictors of roughness and the establishment of population relationship estimates for several acoustic measures.

▶ To study the influence of different aspects of the data acquisition system (e.g., microphone placement, external hardware, etc.) and environment (e.g., signal-to-noise ratio in the recording room) on the outcome of the cepstral measures and the AVQI.

▶ To study the influence of the reliability of the auditory-perceptual ratings of voice quality on their correlation with the acoustic measures.

▶ To study what determines the final rating of the concatenated voice samples: the sample type with the best perceived voice quality, the sample type with the worst perceived voice quality, the sample type that was presented first (i.e., primacy-effect), the sample type that was presented last (i.e., recency-effect), or the mean perceived voice quality in both sample types.

▶ To expand the presented research methods by experimenting with longer continuous speech fragments (e.g., oral readings of the whole text) and studying their validity and feasibility in clinical voice quality assessment.

▶ To proceed with the research on tracheoesophageal voice quality and to apply the presented methods on a larger group of subjects (i.e., to externally cross-validate the present results).

▶ To study the relationship between voice quality and quality of life and to investigate the impact of perceived and acoustically measured dysphonia severity on the outcome of general and voice-related quality of life questionnaires.

Additional ideas for future research can be proposed as following.

▶ To physiologically validate the clinical use of the cepstral and other acoustic measures, and to study the influence of glottal phenomena (i.e., irregularities in the vocal fold vibrations, additive noise due to inadequate vocal fold closure), as parameterized in synthetic voice samples or as kymographically measured in laryngeal high-speed videorecordings, on these cepstral and other acoustic measures.

▶ To study the effectiveness of acoustic biofeedback and other biofeedback modalities in behavioral voice treatment.

▶ To create Dutch training material (based on a database currently consisting of recordings of both speaking tasks of about 750 normophonic and dysphonic subjects and laryngectomees) for standardization in the auditory-perceptual evaluation of overall voice quality and specific voice quality dimensions, for implementation of rating ‘anchors’ in auditory-perceptual evaluation sessions and, finally, for increased reliability of the auditory-perceptual evaluations of voice quality in students as well as professionals.

SCIENTIFIC CURRICULUM

Publications[1]

Maryn, Y., De Bodt, M., Willockx, V., & Van Lierde, K. (1999). Velofaryngale stoornissen: terminologie en logopedische protocollering. *Logopedie, 12*, 21-36.

Maryn, Y., & De Bodt, M. (2000). Adenotomie, tonsillectomie en de gevolgen met betrekking tot nasale resonantie. *Stem-, Spraak- en Taalpathologie, 9,* 260-269.

Maryn, Y., De Bodt, M.S., & Van Cauwenberge, P. (2003). Ventricular dysphonia: clinical aspects and therapeutic options. *Laryngoscope, 113*, 859-866.

Maryn, Y., Van Lierde, K., De Bodt, M., & Van Cauwenberge, P. (2004). The effects of adenoidectomy and tonsillectomy on speech and nasal resonance. *Folia Phoniatrica et Logopaedica, 56*, 182-191.

Dejaeger, E., Delesie, C., Maryn, Y., & Simpelaere, I. (2006). Logopedische therapie van verbale apraxie en dysartrie: enquête, interactieve behandeling en casus. *Logopedie, 19,* 70-80.

De Graeve, C., Deketelaere, I., Maryn, Y., Dick, C., Beernaert, A., Caenen, M., Verhoye, C., De Moor, S., Verstraete, J., Michaux, I., & Deklerck, J. (2006). Fonochirurgie met koude instrumenten of $C0_2$ laser: vergelijkende stemkwaliteitsbeoordeling bij goedaardige larynxaandoeningen. *Stem-Spraak-Taalpathologie, 14,* 213-223.

Geldof, R., Lefevere, S., & Maryn, Y. (2006). Onderzoek naar de invloed van het afnametijdstip (ochtend – avond) op het resultaat van een automatisch fonetogram bij 18- tot 22-jarige studenten handelswetenschappen en bedrijfskunde. *Logopedie, 19,* 43-51.

Maryn, Y., De Bodt, M., & Van Cauwenberge, P. (2006). Effects of biofeedback on phonatory disorders and phonatory performance: A systematic literature review. *Applied Psychophysiology and Biofeedback, 31*, 65-83.

Maryn, Y., De Bodt, M., & Van Cauwenberge, P. (2006). Stemstoornissen en vocale performantie na logopedische behandeling met biofeedback. *Stem-Spraak-Taalpathologie, 14,* 224-246.

Maryn, Y., De Bodt, M., & Van Cauwenberge, P. (2006). Als valse stemplooien waarheid spreken … Ventriculaire dysfonie ontleed. *Logopedie, 19,* 17-27.

Mazure, A., & Maryn, Y. (2006). De invloed van het soort stimulus op het resultaat van nasometrisch onderzoek. *Logopedie, 19,* 65-69.

Vandenbruaene, C., Dick, C., & Maryn, Y. (2007). Logopedische slikevaluatie en –behandeling bij patiënten met een tracheacanule. *Logopedie, 20,* 35-41.

Ferrer, C.A., De Bodt, M.S., Maryn, Y., Van de Heyning, P., & Hernández-Diaz, M.E. (2007). Properties of the cepstral peak prominence and its usefulness in vocal quality measurements. *Proceedings of 5th international workshop of*

[1] References of A1-publications are printed in underlined style.

Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), 13-15 December, Firenze, Italy.

Maryn, Y. (2008). Training van toonhoogte en luidheid. In De Bodt M., Mertens F. & Heylen L. (Red.): *Werken aan stem (pp. 125-156)*. Antwerpen: Garant.

Alpan, A., Maryn, Y., Grenez, F., & Schoentgen, J. (2008). Vocal dysperiodicities in connected disordered speech. *Proceedings of 3rd international workshop on Advanced Voice Function Assessment (AFVA)*, 12-14 May, Aachen, Germany.

Alpan, A., Maryn, Y., Grenez, F., Kacha, A., & Schoentgen, J. (2008). Multi-band and multi-cue analyses of disordered connected speech. *Proceedings of INTERSPEECH*, 22-26 September, Brisbane, Australia.

Alpan, A., Schoentgen, J., Maryn, Y., Grenez, F., & Murphy, P. (2009). Cepstral analysis of vocal dysperiodicities in disordered connected speech. *Proceedings of INTERSPEECH*, 6-10 September, Brighton, United Kingdom.

Maryn, Y., Corthals, P., De Bodt, M., Van Cauwenberge, P., & Deliyski, D. (2009). Perturbation measures of voice: a comparative study between Multi-dimensional voice program and Praat. *Folia Phoniatrica et Logopaedica, 61*, 217-226.

Maryn, Y., Corthals, P., Roy, N., Van Cauwenberge, P., & De Bodt, M. (2009). Toward improved ecological validity in the acous-tic measurement of overall voice quality: combining continuous speech and sustained vowels. *Journal of Voice, Epub ahead of print.*

Maryn, Y., Dick, C., Vandenbruaene, C., Vauterin, T., & Jacobs, T. (2009). Spectral, cepstral and multivariate exploration of tracheoesophageal voice quality in continuous speech and sustained vowels. *Laryngoscope, Epub ahead of print.*

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

Pollet, E., Vanderhaeghe, L., Verstraete, J., Deklerck, J., & Maryn, Y. (2009). Stemanalyse bij 58 vrouwelijke tweedejaarsstudenten kleuteronderwijs: evolutie na één jaar opleiding. *Logopedie, 22*, 13-26.

Maryn, Y., De Bodt, M., & Roy, N. (in press). The Acoustic Voice Quality Index: toward improved treatment outcomes assessment in voice disorders. *Journal of Communication Disorders*.

Oral presentations[2,3]

Maryn, Y., & De Bodt, M. (10/01/2003). Ventriculaire dysfonie: Klinische aspecten en therapeutische opties. *Postacademic course 'Stemstoornissen', University of Antwerp, Antwerp, Belgium.*

[2] The presented list consists of a selection of oral presentations concerning voice-related topics.

[3] Oral presentations on an international forum are printed in underlined style.

Maryn, Y., & De Bodt, M. (07/09/2003). Working with ventricular dysphonia: literature versus practice. *5th CPLOL Congress 'Evidence-based practice in speech-language therapy', Edinburgh, Schotland.*

Maryn, Y. (28/05/2004). Real-time spectrography as a biofeedback tool in voice therapy. *Postacademic course 'Stemstoornissen', University of Antwerp, Antwerp, Belgium.*

Maryn, Y. (08/10/2004). Microperturbaties in de glottale vibratie. *Intensive course 'Meten moet … gratis!', University College Gent Vesalius, Ghent, Belgium.*

Maryn, Y., De Bodt, M., & Van Cauwenberge, P. (17/12/2004). Biofeedback in stemstoornissen, lang gekend … maar ongewild. *Symposium 'Practical tools in speech and language therapy', University of Ghent, Ghent, Belgium.*

Maryn, Y. (13/05/2005). Real-time analyse als therapeutisch instrument. *Postacademic course 'Stemstoornissen', University of Antwerp, Antwerp, Belgium.*

Maryn, Y. (22/07/2006). Variability in acoustic voice quality measurements: MDVP versus Praat. *3rd World Voice Conference, Istanbul, Turkye.*

Maryn, Y. (05/08/2006). Forty years of acoustic voice quality measurement: prediction and discrimination. *5th International Voice Symposium, Austrian Voice Institute, Salzburg, Austria.*

Maryn, Y. (16/03/2007). Spraak in grafieken: akoestische (bio)feedback in spraaktherapie. *Postacademische course 'Spraak en spraakverstaanbaarheid', University of Antwerp, Antwerp, Belgium.*

Maryn, Y., Corthals, P., De Bodt, M., & Van Cauwenberge, P. (30/08/2007). Cepstral peak prominence (CPP) for measuring overall voice quality in continuous speech and sustained vowels. *7th Pan-European Voice Conference (PEVOC7), Groningen, The Netherlands.*

Maryn, Y. (07/12/2007). Duur en frequentie van stemtherapie: kort en krachtig of langzaam en zeker. *Symposium 'Dilemma's in stemtherapie', University of Ghent, Ghent, Belgium.*

Maryn, Y. (24/04/2008). Meten aan spraak: grafieken voor (bio)feedback. *Postacademic course 'Neurologische taal- en spraakstoornissen', Arteveldehogeschool, Ghent, Belgium.*

Maryn, Y. (16/05/2008). Praat'en is meten van spraak. *Intensive course, CIOOS, Antwerpen, Belgium.*

Maryn, Y. (30/05/2008). Feedback in stemtherapie: anders of beter. *Postacademic course 'Best practice in stemtherapie', University Hospital Antwerp, Antwerp, Belgium.*

Maryn, Y. (12/12/2008). De waarde van akoestische metingen in stemtherapie: biofeedback en effectmeting. *30ste VVL congress 'Therapie', Vlaamse Vereniging voor Logopedisten, Antwerp, Belgium.*

Maryn, Y. (16/05/2009). Perception of overall voice quality: combining continuous speech and sustained vowels. *7th CPLOL Congress 'Speech-language therapy in Europe: sharing good clinical practice', Ljubljana, Slovenia.*

Maryn, Y. (16/05/2009). Influence of encouragement on automatic voice range profiles. *7th CPLOL Congress 'Speech-language therapy in Europe: sharing good clinical practice', Ljubljana, Slovenia.*

Maryn, Y. (16/05/2009). Acoustic voice quality index: external cross-validation. *7th CPLOL Congress 'Speech-language therapy in Europe: sharing good clinical practice', Ljubljana, Slovenia.*

Maryn, Y. (11/12/2009). Acoustic Voice Quality Index: een maat voor de outcome van stemtherapie. *31ste VVL congress 'Upgrading logopedie', Vlaamse Vereniging voor Logopedisten, Antwerp, Belgium.*

Poster presentations[4]

Maryn, Y., De Bodt, M., & Van Cauwenberge, P. (2005). Biofeedback in voice disorders, results of a literature review. *6th Pan-European Voice Conference (PEVOC6), London, United Kingdom.*

Geldof, R., Lefevere, S., & Maryn, Y. (2006). Onderzoek naar de invloed van het afnametijdstip op het resultaat van een automatisch fonetogram. *28ste VVL congress 'Leerstoornissen', Vlaamse Vereniging voor Logopedisten, Antwerp, Belgium.*

Maryn, Y., Corthals, P., Vanwynsberge, E., Vanderbeke, J., Deklerck, J., De Bodt, M., & Van Cauwenberge, P. (2007). Voice perturbation measures: a comparative study between Praat and MDVP. *7th Pan-European Voice Conference (PEVOC7), Groningen, The Netherlands.*

Maryn, Y., Corthals, P., De Bodt, M., & Van Cauwenberge, P. (2007). Overall voice quality in sustained vowels and continuous speech: a preliminary study with perceptual and acoustic markers. *7th Pan-European Voice Conference (PEVOC7), Groningen, The Netherlands.*

Caelenberghe, E., Trauwaen, I., Maryn, Y., Verstraete, J., & Deklerck, J. (2009). Normatieve studie inzake akoestische stemkwaliteit. *31ste VVL congress 'Upgrading logopedie', Vlaamse Vereniging voor Logopedisten, Antwerp, Belgium.*

Reviews for journals

- American Journal of Medical Genetics
- Head & Face Medicine
- Laryngoscope
- Logopedie

[4] Poster presentations on an international forum are printed in underlined style.

UNIVERSITEIT
GENT

ISBN 978-90-9024980-3