INTRA- AND INTER-RATER RELIABILITY OF VOICE SAMPLES BY SPEECH-LANGUAGE PATHOLOGISTS USING GRBAS SCALE

Renjini Annie Mathew

Register Number: 12SLP020

A Dissertation Submitted in Part Fulfilment of Final Year

Master of Science (Speech Language Pathology)

University of Mysore, Mysore

ALL INDIA INSTITUTE OF SPEECH AND HEARING

MANASAGANGOTHRI, MYSORE – 570006

MAY, 2014

CERTIFICATE

This is to certify that this dissertation entitled "**Intra- and Inter-rater reliability of voice samples by Speech-Language Pathologists using GRBAS scale**" is a bonafide work submitted in part fulfilment for the Degree of Master of Science (Speech Language Pathology) of the student (Registration No.: 12SLP020). This has been carried out under the guidance of a faculty of this institute and has not been submitted earlier to any of the University for the award of any other Diploma or Degree.

Mysore

May, 2014

Dr. S. R. Savithri
Director
All India Institute of Speech and Hearing
Manasagangothri, Mysore -570 006.

CERTIFICATE

This is to certify that this dissertation entitled "**Intra- and Inter-rater reliability of voice samples by Speech-Language Pathologists using GRBAS scale**" has been prepared under my supervision and guidance. It is also certified that this has not been submitted earlier in other University for the award of any Diploma or Degree.

Mysore

May, 2014

Dr.K.Yeshoda
Guide
Reader & H.O.D
Department of Clinical Services
All India Institute of Speech and Hearing
Manasagangothri, Mysore - 570 006.

DECLARATION

This is to certify that this dissertation entitled "**Intra- and Inter-rater reliability of voice samples by Speech-Language Pathologists using GRBAS scale**" is the result of my own study under the guidance of Dr.K.Yeshoda, Reader & H.O.D, Department of Clinical Services, All India Institute of Speech and Hearing, Mysore, and has not been submitted earlier in other University for the award of any Diploma or Degree.

Mysore

May, 2014.

Register No.: 12SLP020

Table of contents

List of tables

ACKNOWLEDGEMENTS

- **'Almighty God'**....u hav alwys been with me throughout as my silent comforter..... thanku so much for helping me come out from all the emotional outbursts I went through during this period....when I thought I'l break down and should just quit, u gave me the strength, motivation to finish off what I had started through my beloved parents and sisters

- **'Ammachi'**.....you are the most important person in my lyf....only becoz of ur constant prayers and blessings I have come this far.....Lov u loads ammachi

- **'Pappa & Mamma'**.....u ppl have been my pillars of support...which does not even bulge with the strongest wind coming against it....shielding me and helping me face all my troubles

- **'Roshini ch, Bittucha and ethan & Rashmi ch, Anish acha'** (luv you loads for giving me the mental support and encouragement. Blessed to hav u guys in my lyf. One smile from my ethu and my smile comes back even though I had a hell of a day)

- **'Savithri Ma'am'**...thanku so much for giving us an opportunity....you have been a wonderful teacher...and we had the best classes ever

- **'Yeshoda ma'am'**....thanku for ur guidance and support throughout this period

- Thanks a lot to all my '**judges** and **participants**'....without your time and support.. guys I would never have completed my study

- A very big thanks to '**Jayashree ma'am**'....u have alwys been concerned about us and tried to show us the right way....thanku so much for being such a wonderful teacher....

- Special thnks to '**Jayakumar sir**' and '**Rajasudhakar sir**'....thanku for helping me out

- '**Irfana chechi**'...U have been a constant support and a good friend.....thank a lot....blessed to have seniors lyk u...

- Thanks to all my teachers who directly and indirectly helped us a lot.... '**Pushpavathi ma'am, Sreedevi ma'am, Manjula ma'am, Swapna ma'am & Gopi kishore Sir**'.....

- '**Classmates**' ....had a very good tym with you guys....a new experience in the past two years...

- A big hug and thanks to '**Raju appacha, Banu ammachi & Asha chechi**'....

- Last but not the least.....'**Offbeats**'...cannot forget u guys....!! (loads of luv to.....**Anusha, Sindhu, Preethi, Suze, Varsha, Gagana, Amrutha, Anjali, Steby**.....miss u guys..!!)

CHAPTER I

INTRODUCTION

In human speech production the major elements considered are voice, articulation and language. In ancient times the voice production were considered as magical whereas presently it is a powerful communication tool and as an artistic medium. Voice provides expression, feeling, intent, and mood to our daily articulated thoughts and it gives melody to our speech (Stemple, 2010).

A voice to be considered as good should be clear to the listener, have a resonant quality, stable and well supported by adequate breath control. The voice should have appropriate pitch and rate of speech which will in turn help in delivering the message to the listener adequately and also the messages are clearly understood.

There are many characteristics that an effective speaking voice should have such as; adequate loudness, clearness and purity of tone, a pleasing and effective pitch level, ease and flexibility, a vibrant sympathetic quality and an ease of diction (Anderson, 1961).

"A voice disorder exists when a person's quality, pitch and loudness differ when compared with those of similar age, gender, cultural background and geographic location" (Aronson, 1980). The overall quality of a sound is formally defined as that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly present and having the same loudness and pitch are dissimilar (American National Standards Institute, 1960).

What is voice quality?

"Voice quality" refers to the perceived signal which is comparable to the difference between the vibratory function and perception of the listener, i.e. the "frequency" and the "pitch" of the signal respectively.

"The term *voice quality* refers to how the voice sounds to a listener". The terms "voice" and "voice quality" can be discussed using either a narrow or a broad manner, and each one is compared to a two sided coin, in which one side of the coin represents the characteristics of perception and the other represents the characteristics of production.

Voice quality can be explained in a narrow manner and here it represents a specific domain of the process of phonation such as the perceived quantity of airflow present in the voice signal which is not modulated in any manner; in a less narrow manner, the term can mean the perceived end result of the process of phonation; or in a broad manner, it can refer to the overall speech perception of the listener. These terms appear in different contexts, and therefore proper use of the terms is based on the need and the view point and therefore a proper definition of both the terms are difficult (Kreiman & Sidtis, 2011).

ANSI Standard definition, defines the quality (or timbre) of a sound as "that component 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" (ANSI Standard, 1960; Helmholtz, 1885). The strength of ANSI standard definition is that it considers quality of voice as the result of a perceptual process and also includes the different nature of voice quality rather than understanding it as a fixed quantity thereby highlighting the importance of both listeners and the signal for determining the quality of voice (Kreiman & Sidtis, 2011).

Consistent with narrow definitions of voice, vocal quality may be defined as the perceptual impression created by the vibration of the vocal folds. More broadly, and parallel to broad definitions of voice, voice quality may be defined as the perceived result of coordinated action of the respiratory system, resonatory system and articulatory system. Abercrombie (1967), viewed voice quality as "those characteristics which are present more or less all the time that a person is talking: It is a quasi permanent quality running through all the sound that issues from his mouth". Similarly, Laver (1980) referred to voice quality as "a cumulative abstraction over a period of time of a speaker-characterizing quality, which is gathered from the momentary and spasmodic fluctuations of short-term articulations used by the speaker for linguistic and paralinguistic communication".

Fairbanks (1960) tried to categorize voice quality defects into three categories-hoarseness, harshness, and breathiness. In clinical practice it is rarely seen that single dimensions of voice such as quality, loudness, pitch or flexibility contributing on abnormal voice separately. It is often seen that among the dimensions, any one may predominate whereas the other dimensions usually present in different combination and proportion.

The type of voice quality perceived due to irregularity in the vibrations of vocal fold was considered perceptually as harshness and acoustically such type of voice quality was associated with variations in both amplitude and time period (Wendahl, 1966 & Moore, 1975). The perception of breathy voice quality was associated with the escape of air through partially closed glottis and which resulted in turbulence noise that reduced the harmonic to noise ratio. Noise of relatively high frequency produced by transient, highly unstable variations was characteristic of the hoarse voice quality (Moore, 1971).

Assessment of voice

In the area of measuring voice there are many methods or approaches in the literature. The main two methods are subjective and objective methods of evaluation. Objective method of assessment includes the usage of instruments in order to get the results whereas subjective assessment depends on the perceptual assessment of voice without any instruments.

The primary objectives of the diagnostic voice evaluation are to discover the etiologic factors associated with the development of the voice disorder, describe the deviant vocal symptoms and develop an understanding of how the disorder is affecting the subsystems of voice production, respiration, phonation and resonance (Stemple, 2010).

The commonly used method to find out the type of voice quality was by using perceptual which simply includes the method of listing different terms used to describe the different impressions of various listeners and this helps in breaking down the overall voice quality into its different components followed by which quality of each voice is rated based on the extent to which it has the particular feature. In another way, the listeners can simply note down the features present in each voice sample. It can be difficult to determine the basis on which terms in such lists have been selected, and it was seen that each voice quality had a varied set of description which overlaps with other voice qualities. The description of voice can be made based on visual (e.g., brilliant, dark), kinaesthetic (strained, tight), physical (heavy, thin), aesthetic (pleasing, faulty), anatomic features (pectoral, nasal) etc (Orlikoff, 1999).

Objective evaluation

Objective evaluation is considered as one of the best methods of assessment and it includes both invasive and non invasive methods for assessment. Non invasive methods of investigating voice quality are widely used to assist perceptual analysis and laryngoscopic findings in dysphonic patients. These methods serve as an expert system allowing standardized assessment of voice quality for comparison of pathological conditions, patients, and therapeutic techniques. A wide variety of measurement techniques have been presented in the literature to analyze stability of the signal (jitter for frequency instability and shimmer for intensity instability) and to analyze signal to noise ratio (SNR) (Baken & Orlikoff, 2000).

Perceptual evaluation (Subjective evaluation)

The perceptual importance of different aspects of voice depends on context, attention, a listener's background and the listening task (Kreiman, Geratt & Kempster, 1993). Perceptual evaluation was fundamental in assessing voice quality, the relevance of defects and their impact on the subject's ability to communicate and was found that perceptual evaluation was easier, less time consuming compared to objective evaluation and it helped in validation of the diagnosis along with objective evaluation.

Perceptual voice evaluation by clinically well trained listeners may be reliable if based on standardized rating procedure and that training for voice therapists could be more effective if perceptual acoustic relationships were identified (Hammarberg, Fritzell, Gauffin & Sundberg, 1986).

Types of rating scales

Different perceptual protocols were developed based on different rating scales in the literature and such perceptual protocols were used for the perceptual evaluation of different voices. The various rating scales are mentioned below:

- Categorical ratings: A particular voice sample may be assigned a discrete category like mild, moderate and severe
- Equal appearing intervals: Perceived voice character is assigned a numerical value to denote severity. Most commonly used scale is 1-7 with the higher numbers representing increased perception of quality disruption
- Visual analog scale: Provides options for the judge to be unbiased as an undifferentiated line is provided and a mark has to be placed to indicate the severity of vocal quality. The extremes of the line are labelled as minimal v/s extreme.
- Direct magnitude estimation: a numerical value is assigned in an unrestricted manner to indicate the degree of voice deviation
- Paired comparison: Two voice samples are judged for extent of difference/ similarity using single or multiple dimensions of vocal parameters.

Different perceptual protocols

There are various perceptual protocols based on the above mentioned rating scales with their own advantages and disadvantages. The following perceptual protocols have been developed by various authors:

- Voice Profile (Wilson, 1987)
- The Vocal Profile Analysis Protocol (Laver, 1980)

- The GRBAS scale (Committee of Phonatory Function tests of the Japan Society and Logopedics and Phoniatrics (Hirano, 1981)
- Buffalo III Voice Profile (Wilson, 1987)
- The Consensus Auditory-Perceptual Evaluation of Voice (ASHA, 2002)

Advantages

- Perceptual evaluation using standardized scale is an inexpensive, readily available and practical tool for evaluation purposes
- Found to be reliable in its findings in both intra- and inter judge reliability
- For reliable assessment, objective measures are always correlated with perceptual evaluation of voice by researchers.

Disadvantages

- Variability in the perception (internal references/ standards can vary)
- Proficiency of judge has an effect in the perception of voice

Need for the study

In general it was found that voice quality was a difficult parameter to measure and with many variations in the diagnosis due to varied perception of contributing factors. It was also noted that the Speech Language Pathologists (SLP) professionals found it difficult to diagnose based on the perceptual scales. Therefore a widely used, reliable and easy perceptual scale (GRBAS) was chosen to check the reliability in voice quality diagnosis. In this regard, the present study was planned to understand extent of intra-rater and inter-rater reliability/ variations across SLPs in different clinical set ups in India.

Objective of the study

To check the intra- and inter-rater reliability in assessing voice quality using GRBAS scale by Speech- Language Pathologists.

CHAPTER II

REVIEW OF LITERATURE

Perception of voice disorders, including pitch, loudness, and quality observations, was most commonly used across speech language and voice clinics with regard to diagnosis and evaluation of progress in voice therapy (Colton & Casper, 1996).

'Voice quality is fundamentally perceptual in nature. Patients seek treatment for voice disorders because they do not sound normal, and they often decide the success of the treatment based on whether they sound better or not. For this and other reasons, speech clinicians' use and value perceptual measures of voice and speech far more than instrumental measures (Gerratt, Till, Rosenbek, Wertz, & Boysen, 1991)'.

'Voice quality, primarily measured through the use of perception, can be defined as ''an interaction between an acoustic voice stimulus and a listener'' and has been referred to as the ''gold standard'' for relating the quality of voice to acoustical measurements (Kreiman et al, 1993)'.

The following table compares different labels of voice quality used in literature over the years.

Table 2.1.

Venerable and Modern Labels for Voice Quality (cited in Austin, 1806, Pannbacker, 1984 and adapted from Kreiman & Sidtis, 2011)

*After Julius Pollux, 2nd century AD*a	*Moore, 1964*b	*Gelfer, 1988*
High (altam)	-	High
Powerful (excelsam)	Ringing	Strong, intense, loud
Clear (claram)	Clear, light, white	Clear
Extensive (latam)	Rich	Full
Deep (gravam)	Deep	Resonant, low
Brilliant (splendidam)	Bright, brilliant	Bright, vibrant
Pure (mundatam)	-	-
Smooth (suavam)	Cool, smooth, velvety	Smooth
Sweet (dulcem)	-	-
Attractive (illecebrosam)	Pleasing	Pleasant
Melodious, cultivated (exquisitam)	Mellow	Mellow, musical
Persuasive (persuasibilem)	-	-
Engaging, tractable (pellacem, tractabilem)	Open, warm	Easy, relaxed
Flexible (flexilem)	-	Well-modulated
Executive (volubilem)	-	Efficient
Sonorous, harmonious (stridulam)	Chesty, golden, harmonious, orotund, round, pectoral	Balanced, open
Distinct (manifestam)	-	-
Perspicuous, articulate (perspicuam)	-	-
Obscure (nigram)	Dark, guttural, throaty	Husky, guttural, throaty
Dull (fuscam)	Dead, dull, heavy	Dull, heavy, thick

Unpleasing (injucundam)	-	Unpleasant
Small, feeble (exilem, pusillam)	Breathy	Breathy, soft, babyish
Thin (angustam)	Constricted, heady, pinched, reedy, shallow, thin	Thin
Faint (difficilem auditu, molestam)	Whispery	Weak
Hollow, indistinct (subsurdam, obscuram)	Covered, hollow	Muffled
Confused (confusam)	-	-
Discordant (absonam)	Blatany, whiney	Strident, whining
Unharmonious, uncultivated (inconcinnam, neglectam)	Coarse, crude	Coarse, gruff
Unattractive, unmanageable (intractabilem)	-	Shaky
Uninteresting (inpersuasibilem)	Blanched, flat	-
Rigid (rigidam)	Hard, tight	Monotonous, constricted, flat
Harsh (asperam)	Harsh, strident, twangy	Harsh, gravelly
Cracked (distractam)	Pingy, raspy	Strained, raspy, grating, Creaky
Doleful (tristem)	-	-
Unsound, hoarse (infirmam, raucam)	Faulty, hoarse, poor, raucous, Rough	Hoarse, rough, labored, Noisy
Brassy (aeneam)	Buzzy, clangy, metallic	Metallic

Shrill, sharp (acutam)	Cutting, hooty, piercing, pointed, sharp, shrill	Shrill, sharp
-	Nasal	Nasal
-	Denasal	Denasal
-	Toothy	-

It can be observed that from 2nd century AD to 1964 to 1988, labelling became more specific and less descriptive.

The following table lists different types of voice qualities along with their descriptions which helped in categorizing the voice qualities and also in comparing the physiological component with the perceptual attribute (Titze, 1994).

Table 2.2.

Summarizes Different Voice Qualities and Its Description

Voice Quality	Perception	Physiologic component
Aphonic	No sound or a whisper	Inability to set vocal folds into vibration, caused by lack of appropriate power (air pressure) or a muscular/tissue problem of the folds
Biphonic	Two independent pitches	Two sources of sound (e.g., true folds and false folds, or two folds and whistle due to vortex in air)
Bleat	Similar as flutter	-
Breathy	Sound of air is apparent	Noise is caused by turbulence in or near glottis, caused by loose valving of laryngeal muscles (lateral cricoarytenoid, interarytenoid and posterior cricoarytenoid).
Covered	Muffled or 'darkened' sound	Lips are rounded and protruded or larynx is lowered to lower all formants so a stronger fundamental is obtained
Creaky	Sounds like two hard surfaces rubbing against one another	A complex pattern of vibrations in the vocal folds creates a intricate formation of subharmonics and modulations
Diplophonic	Pitch supplemented with another pitch one octave lower, roughness usually apparent	A period doubling, or F0/2 subharmonic
Flutter	Often called bleat because it sounds like a lamb's cry	Amplitude changes or frequency modulations in the 8-12 Hz range
Glottalized	Clicking noise heard during voicing	Forceful adduction or abduction of the vocal folds during speech
Hoarse (raspy)	Harsh, grating sound	Combination of irregularity in vocal fold vibration and glottal noise generation
Honky	Excessive nasality	Excessive acoustic energy couples to the nasal tract
Jitter	Pitch sounds rough	Fundamental frequency varies from cycle to cycle

Nasal	Similar as honky	-
Pressed	Harsh, often loud (strident) quality	Vocal processes of the arytenoid cartilages are squeezed together, constricting the glottis, and causing low airflow and medial compression of the vocal folds
Pulsed (fry)	Sounds similar to food cooking in a hot frying pan	Sound gaps caused by intermittent energy packets below 70 Hz and formant energy dies out prior to re-excitation
Resonant (ringing)	Brightened or 'ringing' sound that carries well	Epilaryngeal resonance is enhanced, producing a strong spectral peak at 2500-3500 Hz; in effect, formants F3, F4 and F5 are clustered
Rough	Uneven, bumpy sound appearing to be unsteady short-term, but persisting over the long-term	Modes of vibration of the vocal folds are not synchronized
Shimmer	Crackly, buzzy	Short-term (cycle-to-cycle) variation in a signal's amplitude
Strained	Effortfulness apparent in voice, hyperfunction of neck muscles, entire larynx may compress	Excessive energy focused in laryngeal region
Strohbass	Popping sound; vocal fry during singing	Sound gaps caused by intermittent energy packets below 70 Hz and formant energy dies out prior to re-excitation
Tremerous	Affected by trembling or tremors	Modulation of 1-15 Hz in either amplitude or pitch due to a neurological or biomechanical cause
Twangy	Sharp, bright sound	Often attributed to excessive nasality, but probably also has an epilaryngeal basis
Ventricular	Very rough (Louis Armstrong-type voice)	Phonation using the false folds anterior rather than the vocal folds; unless intentional due to damage to the true folds, considered an abnormal muscle pattern dysphonia
Wobble	Wavering or irregular variation in sound	Amplitude and/or frequency modulations in the 1-3 Hz range

Yawny	Quality is akin to sounds made during a yawn	Larynx is lowered and pharynx is widened, as people do when yawning - hence the name

(adapted from https://www.ncvs.org/)

Comparing both the tables of different labelling of voice quality, it can be noted that this table gave clear idea about the type of voice qualities that were more simplified in nature and self explanatory. This table continues to help in comparing the perceptual and physiologic domains and also in assisting objective evaluation for assessment of voice disorders.

Need to care about voice quality

All voices are different in nature and when spoken it conveys information about people as different individuals. The voice of speakers may be perceived very differently i.e. whenever one speaks, voices may sound young, tired etc. Voice helps in understanding ideas, emotions, etc. Other than this factor, voice also helps in distinguishing gender, cultural background, etc. (Kreiman & Sidtis, 2011).

The impression of familiarity versus unfamiliarity could also be done with the help of an individual's voice and it also can help one to compare the voice characteristics between different people. The knowledge gained by the listeners by listening to different voices may not imply the correct situation; for example, the surprise of meeting a person after speaking on the phone may not match the mental picture framed about the individual just by listening to his/ her voice. Even though the mismatches occur, voice quality is considered as one of the main parameter by which

the individuals project their identity – their "physical, psychological, and social characteristics" (Laver, 1980).

Even from the time of birth, human infants are able to recognize their mother's voice (DeCasper & Fifer, 1980) and also the responses to the maternal voices can be measured in-utero suggesting that abilities to recognize voice is developed much earlier (Hepper, Scott, and Shahidullah, 1993). It is also noted that as voice can convey emotional attitude, the change in voice quality relative to the speaker's normal voice quality can deliver changed emotions such as, sarcasm, etc. (Breitenstein, Van Lancker, and Daum, 2001, Van Lancker, Canter, and Terbeek, 1981).

The other changes can be altered rate of speech and fundamental frequency which affect the credibility of voice (Geiselman and Bellezza, 1977) and in the pragmatic skills, voice cues the order of turn taking in a conversation (Wells and Macfarlane, 1998) and helps resolves any ambiguities present in a sentence (Schafer, Speer, Warren, and White, 2000). Judgement of the nativity of language among the speakers can also be done based on voice quality cues (Piske, MacKay, and Flege, 2001).

With all the above mentioned characteristics and uses of voice quality changes there comes a requirement of measuring and analyzing the different voice quality changes. The analysis of voice can be done using various methods and these are mentioned below.

Analysis of voice

There are various methods of analysis of voice, developed by different researchers (Baken, 1987; Hirano, 1981). It can be done either subjectively or objectively.

Objective measures includes instruments for its evaluation such as acoustic measures, aerodynamic measures etc., whereas subjective measures use the human ears ability to recognize speaker's voice and compare between voices. Trained voice clinicians are often able to determine the causative pathologies on the basis of psycho acoustic impression of voice (Hirano, 1975).

Relation between objective and subjective measures

"According to the European Laryngeal Society, an assessment of voice disorders should consist of (video) laryngostroboscopy, perceptual voice assessment, acoustic analysis, aerodynamic measurements, and subjective self evaluation of voice (Dejonckere, et.al, 2001)".

A study was conducted to determine the relation between a subjective measurement perceptual protocol, GRBAS scale and a scale for objective measurement of voice, the Multi-Dimensional Voice Program (MDVP) scale wherein the authors did a retrospective review of 37 voice patients (12 male and 25 females) and each voice sample was perceptually evaluated using the GRBAS scale and acoustically analyzed using the MDVP scale by an experienced Speech Language Pathologist. It was found that the Grade measure correlated with voice turbulence index (VTI), noise harmonic ratio (NHR), and soft phonation index (SPI). Roughness correlated with NHR only. Breathiness correlated with SPI only. Aesthenia also correlated with SPI only. Of the 19 acoustic variables measured by the MDVP

system, only three noise parameters significantly correlated with the GRBAS perceptual voice analysis and so the authors concluded that perhaps “noise” is the perceived acoustical quality of the dysphonic voice. Significant correlation was seen between the noise related parameters of Multi-Dimensional Voice Program (MDVP) and the components of GRBAS scale (Bhuta, Patrick & Garnett, 2004).

The score on Grade of the GRBAS scale was compared with Dysphonia Severity Index (DSI) in order to investigate the usefulness of DSI as an objective multi-parametric measurement in assessing dysphonia. The study included 294 patients with different voice pathologies and 118 in the control group and the comparison was done. With a DSI cut-off of 3.0, maximum sensitivity (0.72) and specificity (0.75) were found in this study (Hakkesteegt, Brocaar, Wieringa & Feenstra, 2006) and it was concluded that DSI is a useful instrument to measure the severity of dysphonia objectively. Therefore in this study the importance of DSI measure was found out by comparing it with the reliable GRBAS perceptual protocol.

A study was aimed by Reynolds, Buckland, Bailey, Lipscombe, Nathan, Vijayasekaran, Kelly, Maryn & French, (2012) to evaluate the Acoustic Voice Quality Index (AVQI), a multivariate acoustic measure of dysphonia in a pediatric population as there was lack of appropriate, validated acoustic measurements for use in the pediatric population. This was a prospective observational study of a sample of dysphonic and normophonic children and AVQI analysis was conducted on a prolonged vowel sample and a sample of continuous speech. Results showed that AVQI have diagnostic accuracy and specificity in this population of children with and without dysphonia. It was also moderately correlated with ratings of severity on the

GRBAS [overall grade of hoarseness (G), roughness (R), breathiness (B), aesthenicity (A), and strain (S)], a subjective rating scale.

To conclude the above mentioned studies helped in finding the relation between perceptual and objective measurements and how both these measurements play their roles in supplementing and validating the diagnosis of various voice disorders.

Different scales of perceptual evaluation

There a many scales for perceptual evaluation of voice mentioned in the literature. The perceptual protocols are usually presumed easy to administer and less time consuming when compared to objective evaluations. Perceptual protocols help in clinical diagnosis of various voice disorders and also in categorizing the deviated quality factors. GRBAS was one of the first and popularly used for clinical diagnostic assessment. A brief summary of the popular scales is attempted in the following paragraphs.

- Voice Profile (Wilson, 1987)

  It is an eight point rating scale (1: voice problem barely perceptible, 7: voice problem significantly affects communication) which helps in documenting abnormal voice in children and adults. It evaluates voice on various parameters like laryngeal qualities, resonation qualities, vocal range, loudness, rate etc.

- The Vocal Profile Analysis Protocol, (VPA) (Laver, 1980)

  In this scale both laryngeal and supra laryngeal aspects are included. The author charted the positions of labial, mandibular, lingual, velopharyngeal and laryngeal

structures to which he gave tension ratings. He provided phonetic description of voice quality. Phonation types are classified as harshness, whisper, breathiness, creaky, falsetto and modal.

Positive features:

- ✓ Detailed analysis of vocal tract configurations
- ✓ Suggests corresponding therapy interventions
- ✓ Profiles individual vocal characteristics
- ✓ Suitable for normal and abnormal voices
- ✓ Relates to physiological function
- ✓ Two day training programme needed

Limitations:

- ✓ Regular listening skills practice needed
- ✓ Time consuming compared with GRBAS and Buffalo III

- Buffalo III Voice Profile (Wilson, 1987)

  This scale rates the laryngeal tone, loudness, pitch, nasal resonance, oral resonance, breath supply, muscles, vocal abuse, rate, speech anxiety, speech intelligibility and overall voice proficiency on a five point rating scale, with appropriate descriptive terms listed for marking with each category. Speech samples should include connected speech, oral reading, individual phonemes and counting.

Positive features:

- ✓ Simple clinical measurement
- ✓ Broad range of categories

- ✓ Overall voice rating (1-5)
- ✓ Easy/ Quick to use, learn

Limitations:

- ✓ Includes non voice quality parameters
- ✓ No formalized training

Kreiman, Gerratt, Kempster, Erman & Berke (1993) reviewed 57 different papers selected from the literature that used various approaches to auditory perceptual analysis of voice in USA and suggested that the Buffalo Voice Profile is probably the most widely used rating scale in North America.

Munoz, Mendoza, Fresneda, Carballo & Ramirez (2002) conducted a study to estimate the agreement and reliability of voice evaluation by a group of expert listeners using the central portion of a sustained vowel and a fragment of connected speech as voice samples. Ratings were made using Wilson's Buffalo III Voice Screening Profile. Analysis showed that intra-individual listeners' agreement presented variability in the evaluation of both voice samples. In the evaluation of the central portion of the sustained vowel, inter-individual listener agreement was moderate for breathiness, hyponasal resonance, and overall voice rating; in connected speech, agreement was moderate for most voice qualities (breathy, rough, high/low pitch, and hyponasal resonance). Finally, Wilson's Buffalo III Voice Screening Profile presented good reliability values for both voice samples, with overall voice rating achieving higher values (.90) than any other voice-quality variable.

- The Consensus Auditory-Perceptual Evaluation of Voice, (CAPE-V) (ASHA, 2002)

The attributes in this scale are overall severity, roughness, breathiness, strain, pitch and loudness. This displays each attribute accompanied by 100 millimetres forming visual analog scale. The clinician indicates the degree of perceived deviance from normal for a parameter on the scale using a tick mark. Judgements may be assisted by referring to general region indicated below scale. “MI” refers to mildly deviant, “MO” refers to moderately deviant and “SE” refers to severely deviant; C indicates consistency and I indicate inconsistency.

Positive features:

- ✓ A sensitive and detailed tool
- ✓ Includes common features like pitch and loudness
- ✓ Scales are defined
- ✓ Includes consistency factor

Limitations:

- ✓ Time consuming when compared with GRBAS

Rater (intra- and inter-) reliability was studied using the perceptual rating scale CAPE-V in paediatric voice disordered cases post laryngotracheal reconstruction. Here the sentence portion of the Consensus Auditory Perceptual Evaluation- Voice (CAPE-V) rating scale was used and three experienced speech-language pathologists independently rated voice samples of 50 subjects in the age range 4-20 years on six salient perceptual vocal attributes. It was found that the estimates of inter-rater reliability were strongest for perceptual ratings of breathiness (intra-class correlation coefficient (ICC = 71%), roughness (ICC = 68%), pitch (ICC = 68%), and overall severity (ICC = 67%). Reliability was lower for ratings of loudness (ICC = 57%) and

strain (ICC = 35%). For each rater, the intra-rater reliability on all but one parameter (strain) was moderate to strong (ICC = 63–93%). There was a strong inter-rater reliability for four of six vocal parameters rated using the CAPE-V in a population of children and adolescents with marked dysphonia (Kelchner, Brehm, Weinrich, Middendorf, deAlarcon, Levin & Elluru, 2008).

- GRBAS scale - G: Overall Grade, R: Roughness, B: Breathiness, A: Aesthenic, S: Strain, developed by Committee of Phonatory Function tests of the Japan Society and Logopedics and Phoniatrics (Hirano, 1981).

GRBAS scale evaluates voice on five scales,

G- Overall Grade: degree of voice abnormality represents the degree of hoarseness or voice abnormality. It corresponds to the factors of evaluative nature obtained by the semantic differential technique.

R-Roughness: represents psycho acoustic impression of the irregularity of the vocal fold vibration. It corresponds to the irregular fluctuation in fundamental frequency or amplitude of the glottal source sound.

B-Breathiness: represents psychoacoustic impression of the extent of air leakage through glottis is related to the turbulence.

A-Aesthenic (weak): represents weakness or lack of power in the voice. This is related to the weak intensity of the glottal source sound/ lack of higher harmonics.

S-Strain: represents psychoacoustic impression of a hyper-functional state of phonation. It corresponds to abnormal high fundamental frequency, noise in high frequency range and richness in high frequency harmonics.

	0	1	2	3
G				
R				
B				
A				
S				

Each parameter is rated on four point rating scale ranging from 0 (non hoarse/ normal), 1 (slight), 2 (moderate) and 3 (extreme). The grading maybe as follows: G1R1B2A0S0, G3R3B3A0S3 etc.

Positive features:

- ✓ Simple clinical measurement
- ✓ Rates abnormality
- ✓ Overall severity rating (0-3)
- ✓ Rates pertinent laryngeal features
- ✓ Defined terminology
- ✓ Based on acoustic theory
- ✓ Easy/ quick to use/learn

Limitations:

- ✓ Rates laryngeal level only (no supra glottis parameters)
- ✓ No rating of commonly used parameters such as pitch and loudness
- ✓ No formalized training

Of the perceptual protocols, the GRBAS scale is reported to be reliable, easy and valid, and offers no discomfort or inconvenience to the judge. This scale is widely accepted in different parts of the world for judging disordered voice quality.

A study was aimed to assess the reliability of three common scales (The Buffalo Voice Profile, The Vocal Profile Analysis Scheme (VPA) and GRBAS. Sixty-five varyingly dysphonic and five normal voices were recorded onto CD in random order. Thirty voices were recorded twice. Seven experienced and trained speech and language therapists rated all voices on the three scales. Only the overall grade was found to be reliable for the Buffalo Voice Profile. The reliability of the VPA scheme was found to be poor to moderate. The VPA may have had use as a multi-dimensional and in-depth evaluation of voice types, but its greater scope was at the expense of reliability. The GRBAS was reliable across all parameters except Strain. The authors detailed reliability analysis comparing performance of three commonly used rating scales provided further evidence to support GRBAS as a simple reliable measure for clinical use (Webb, Carding, Deary, MacKenzie, Steen & Wilson, 2003).

Voice quality in European Portuguese was compared using GRBAS and CAPE-V scale and statistical significances were found between the perceptual subscale grade from GRBAS and subscales global and roughness from CAPE-V, roughness in GRBAS and global in CAPE-V, and breathiness in GRBAS and in CAPE-V. The correlation values were good, ranging from 0.60 to 0.87 (Jesus, Barney, Sa Couto, Vilarinho & Correia, 2009).

A study was conducted with an aim to provide mutual understanding between different evaluation scales for pathological voice quality by comparing analysis between the GRBASI (includes Grade, Roughness, Breathiness, Aesthenia, Strain and Instability) and RASATI (Pinho & Pontes, 2008) (includes Roughness, Harshness, Breathiness, Aesthenia, Strain, Instability) systems. Listeners rated 100 voice samples and these were analyzed to identify the significant interrelations between the scales,

with asthenia, roughness and instability as the common factors. It was found that Grade of hoarseness only included in GRBASI, corresponds to a combination of roughness, breathiness and instability. Harshness, included only in RASATI can be predicted by breathiness with strain in the GRBASI scale. Among all the three factors considered roughness was the most consistent and the easiest to identify by evaluators (Yamauchi, Imaizumi, Maruyama & Haji, 2010).

The above paragraphs were summary of various popular perceptual protocols and the studies related to them as described in the literature. The following table compares few perceptual protocols proposed by various authors, selection of one/ more protocol/s based on its features and the need for use.

Table 2.3.
A Guide to Select a Perceptual Voice Quality Evaluation Scheme

	GRBAS	VPA	Buffalo III
Terms based on theoretical framework	Yes (acoustic)	Yes (phonetic)	No
Training prerequisite	No	Yes	No
Applicable to normal voice	No	Yes	No
Abnormality rating	**Yes**	No	Yes
Audio tapes for listener training	Yes (Japanese)	Yes (English)	No
Laryngeal note rating	**Yes**	Yes	Yes
Vocal tract ratings	No	Yes	No
Prosodic features	No	Yes	Yes
Intra/inter-judge reliability evidence	**Yes**	Yes	Yes
Number of parameters	**5**	31	12
Rating range	**0-3**	Varies according to parameter	1-5
Protocol form	No	Yes	Yes
Time to administer (approximately)	**< 5min**	10 min	5-10 min
Applicable to voice/singing teacher	No	Yes	No

(*adapted from Carding, Carlson, Epstein, Mathieson & Shewell, 2000*)

Among these approaches, the GRBAS scale was reported to be widely used for judging disordered voice quality (Carding, Wilson, MacKenzie & Deary, 2009). It gives more objectivity regardless of the type of speech sample (Nemr, Zenari, Cordeiro, Tsuji, Ogawa, Ubrig & Menezes, 2012).

Reliability of perceptual evaluation

There are many studies in the literature which shows that the intra-rater and inter-rater reliability fluctuate among professionals but the contributing factors are not very clear. There are many scales available for the assessment of voice problem and among them the well accepted in the literature are the GRBAS, CAPE-V etc.

To select parameters on the base of reliability (low intra-judge and inter-judge variation) and clinical relevance (good discrimination between voices) a study was carried out in which 15 parameters were taken (comprising the GRBAS parameters) to assess 12 clearly dissimilar voices of different pathologies. These voice samples were assessed by 6 speech therapists as judges. It was found that on the basis of intra judge (low), inter judge (low) and inter voice (high) variances, the GRBAS scale parameters appear to be quite reliable and are of clinical relevance for evaluating the overall severity of hoarseness. The best correlation between judges (0.7) was found for the overall grade of severity and this seems to be mainly determined by the component breathiness. It was also found that the GRBAS profiles differ significantly between different pathological groups especially between primarily organic and primarily functional voice disorders (Dejonckere, Obbens, de Moor & Wieneke, 1993).

A strong correlation was found between the GRBAS scale and the acoustic measurements in which the perceptual scale for deviant voice quality (completed with a 'I' [GIRBAS]: I: Instability= Fluctuation of voice quality over time) was tested and GIRBAS scale seemed to be a valuable instrument for clinical practice (Dejonckere, Remacle, Elbaz, Woisard, Buchman & Millet, 1996).

GRBAS was also used as a scale to find out the influence of experience and professional background on the perceptual rating of voice samples. For this, nine voice samples were presented to a group of twenty three judges twice which included both otolaryngologists with and without experience along with speech pathologists. For the reliability check the time interval was taken as 14 days. Results indicated that the test re-test reliability was moderate using GRBAS scale and the best agreement was obtained for the G (grade) parameter and the worst agreement was for the S (Strained) parameter. This study was done to check the test-retest reliability of the GRBAS scale with the influence of experience and professional background on the perceptual rating of voice quality and it was found that professional background had a greater impact on perceptual rating than experience (De Bodt, Wuyts, Van De Heyning & Croux, 1997).

A study focused on the reliability in perceptual analysis in which a listening group of 10 listeners, 7 experienced speech therapists and 3 speech- language therapist students evaluated the voice samples (text reading- at two loudness levels and sustained vowel- at 3 levels) by 15 vocal characteristics using Visual Analog scales. The results indicated a high Inter-rater reliability for most perceptual characteristics. Connected speech was evaluated more reliably, especially at the normal level, but both types of voice signals were evaluated reliably, although the reliability for connected speech was higher than for sustained vowels. Experienced listeners tended to be more consistent in their ratings than did the student raters (Bele, 2004).

A prospective reliability study and retrospective chart review were carried out in a study to examine the reliability of two methods for documenting voice quality and

compared the methods for documenting patients' perception of voice quality. The two clinician based perceptual protocols taken were GRBAS and CAPE-V and they were compared after use in voice assessments of 42 males and 61 females which was performed by a speech pathologist specializing in assessment of voice disorders. Patient based scales such as Voice Related Quality of Life or V-RQOL, and Iowa Patient's Voice Index or IPVI were also obtained from the patients and were compared with each other and also with the clinician based scales. Reliability of clinicians' ratings of overall severity of dysphonia using GRBAS and CAPE-V scales was very good ($r > 0.80$). There was relatively weak agreement between patient based and clinician based scales and these differences supported the conclusion that clinicians and patients experience and consider dysphonia differently (Karnell, Melton, Childes, Coleman, Dailey & Hoffman, 2005).

Intra- and inter- rater reliability may vary for different voice samples and to see the effect of it on the reliability, a prospective study was carried out by Law, Kim, Lee, Tang, Lam, van Hasselt & Tong (2011). Their aim was whether different types of voice samples affected rater reliability and which type of sample could be rated more reliably among two types of connected speech- passage reading and conversational speech. In this study, 14 speech pathologists experienced in managing voice disorders were given one hundred and fifty samples from forty speakers for perceptual judgement. The voice samples included sustained vowels, passage reading and conversational speech and it was rated on four vocal parameters such as overall severity, roughness, breathiness and strain on a ten point equal appearing interval scale. Results revealed differences in intra-rater reliability across the three types of voice samples. Higher intra-rater reliability was seen in connected speech than with

sustained voice samples. Inter-rater reliability showed no statistically significant difference across the three types but increased with the severity of dysphonia.

GRBAS rating scale was also used as a base in order to validate a new tool used for rating voice parameters, namely, the Newcastle Audio Ranking (NeAR) test (Gould, Waugh, Carding & Drinnan, 2011). Effect of consensus training of listeners on intra- and inter-rater reliability and agreement of perceptual voice analysis was investigated using a four point equal appearing interval scales. The stimuli consisted of text reading by authentic dysphonic patients and thirteen students of audiologopedics served as judges. The consensus training for each perceptual voice parameter included definition, underlying physiology, presentation of carefully selected sound examples representing the parameter in three different grades followed by group discussions of perceived characteristics and practical exercises including imitation to make use of the listeners' proprioception. Results indicated that the intra-rater reliability and agreement showed a marked improvement for intermittent aphonia but not for vocal fry. Inter-rater reliability was high for most parameters before training with a slight increase after training. Inter-rater agreement showed marked increases for most voice quality parameters as a result of the training (Iwarsson & Peterson, 2011).

A study was conducted to evaluate the reliability and consensus of the GRBAS scale and the Consensus Auditory Perceptual Evaluation- Voice (CAPE-V) scale when applied to the same voice sample at different times. It was an observational cross-sectional study in which voice samples (i.e. phonation sample of 3-5 seconds, reproduction of six sentences and spontaneous speech sample) of sixty

subjects were recorded for the CAPE-V analysis whereas for GRBAS analysis the sustained vowel and reading tasks were used. Three expert speech therapists who were familiar with both the scales and who had more than 5 years experience carried out the auditory-perceptual voice analysis. A strong correlation was observed in the intra-judge consensus analysis, both for the GRBAS scale as well as for CAPE-V, with intra class coefficient values ranging from 0.923 to 0.985 and GRBAS was considered as the fastest and CAPE-V was considered as the most sensitive scale (Nemr, et. al, 2012).

To determine if clinical experience affects perceptual rating, a study was conducted in which five speech language clinicians and five naive listeners rated the similarity of pairs of normal and dysphonic voices and for this multidimensional scaling was used to determine the voice characteristics that were perceptually important for each voice set and listener group. Results indicated that naive and expert listeners attended to different aspects of voice quality when judging the similarity of voices, for both normal and pathological voices. All naive listeners used similar perceptual strategies; however, individual clinicians differed substantially in the parameters they considered important when judging similarity (Kreiman, Gerratt & Precoda, 1990).

To summarize voice quality is an important measure of the voice parameters and there are different types of voice qualities according to the literature. To analyze the voice quality many perceptual protocols were proposed by various authors each having its own advantages. According to the literature studies it was noted that

GRBAS was one of the most reliable and easy tool for the perceptual assessment of voice and its scores were also reliable with various objective measurements.

Therefore to conclude there is a need to understand the extent of intra-rater and inter-rater reliability/ variations across SLPs in different clinical set-ups in India. Hence, the present was planned with the following aims,

- To obtain the ratings for voice quality disorders using GRBAS from Speech-Language Pathologists (SLPs) working in different set-ups.
- To compare the GRBAS ratings across SLPs and obtain intra- and inter-rater reliability.

CHAPTER III

METHOD

The aims of the present study were to obtain the ratings for voice quality disorders using GRBAS from Speech-Language Pathologists (SLPs) working in different set-ups and to compare the GRBAS ratings across SLPs and obtain intra- and inter-rater reliability

Participants

A total of eight participants took part in the study. They were divided into two groups.

Group I: Six male participants were included in the study, who were diagnosed as having voice quality disorders during the regular diagnostic assessment at the institute. They the participants were native Kannada speakers in the age range of 21-40 years; mean age 29.33 years and SD: 6.8.

Group II: 2 male native speakers of Kannada, who possessed normal voice characteristics, were chosen for obtaining the control sample (normal sample). They were both aged 22 years.

Inclusion criteria:

- Diagnosed to have a voice quality disorder (hoarse/ harsh/ breathy) for participants in Group I.
- All participants in both groups were screened for normal hearing and oro-motor skills.

Ethical issues

- Consent forms were signed and taken from the participants before taking the voice samples

- The participants were informed about the aim of the study and its implications

Listeners/ Judges

Twenty nine Speech Language Pathologists in the age range of 24-55 years (mean age of 29.45 years and SD: 8.0) with ≥2 years of clinical experience in diagnosing voice quality disorders and working in different set-ups: hospitals (9), academic institutions (9), private clinical set-ups (9) and school set-up (2), formed the judges for the intra- and inter- judge reliability measures.

Procedure

Task: included a phonation sample of about 5 seconds, monologue of about 20 seconds and reading a standardized Kannada passage.

Recording of samples

The tasks were recorded individually from each participant. The recordings were carried out in a quiet environment. The participants were instructed to phonate after a deep inhalation followed by monologue and then reading the standardized Kannada passage.

Instrumentation

Tape recorder: A high quality portable digital audio recorder (Olympus L-100, Multi-track linear PCM) with in-built microphone was used to record the samples (phonation and speech) of each participants (both clinical and control participants).

Dr. Speech software: was used to confirm the diagnosis of the samples recorded. The quality of voice was confirmed using the Dr. Speech software (Tiger DRS Inc.) as hoarse, harsh and breathy respectively.

Laptop: The ASUS model (K55v series, intel CORE i3) with provision for external headphones (HP on ear headphones with microphone attached) was used to present the audio samples to the judges for perceptual listening task.

Perceptual Analysis

Tokens: All samples were checked for duration and clarity. The samples were edited to form tokens to include 3 seconds phonation and 20 seconds reading sample in sequence. A total of eight tokens were prepared. Four lists were made in four different orders of tokens for presentation to listeners. These lists of tokens were copied to an audio CD and played in random order during perceptual analysis. For the reliability check, the orders of the tokens were rearranged.

Perceptual tool: The GRBAS scale [Committee of Phonatory Function tests of the Japan Society and Logopedics and Phoniatrics, (Hirano, 1981)] was used in the study for rating the severity of the voice samples. GRBAS is a four point rating scale rating the voice based on five parameters namely: Overall grade of hoarseness, Roughness, Breathiness, Aesthenic quality and Strain. Each parameter is rated on four point rating scale ranging from 0 (non hoarse/ normal), 1 (slight), 2 (moderate) and 3 (extreme). A scoring sheet was prepared for the responses of the listeners in the perceptual experiment. Appendix shows the score sheet used for the analysis.

Perceptual analysis: It was carried out in a quiet environment on an individual basis. For the perceptual analysis, the listeners were instructed to listen to both the phonation and speech sample of each participant and rate the samples using the GRBAS scale. They were also instructed that they could listen to the same sample as many times required for them to rate it. The listeners were blind folded to the aims of

the study. They were also requested to describe the parameters that helped them to rate the sample in the space provided. The responses of the listeners were compiled and subjected to further statistical analysis.

Reliability Check: The tokens were given to the same participants (Speech Language Pathologists) in a random order after an interval of two weeks for checking the intra-rater reliability of the diagnosis of voice quality using GRBAS scale.

Statistical analysis: Statistical analysis was carried out by using the SPSS 16.0 version software. The mean, median, mode, minimum, maximum, standard deviations, frequency and percentage of each domain scores in various set-ups were found out using descriptive statistics in this software. For the reliability check between various parameters and measure of agreement in the SPSS software, both Cronbach's alpha coefficient and Kappa measure of agreement were used.

Cronbach's alpha coefficient was used to obtain the intra- and inter- rater reliability among the judges and across the settings in all the domains of the scale by considering the values as in a continuous scale. Kappa measurement of agreement was also used to obtain the trial agreement within set-ups across domains.

CHAPTER IV

RESULTS & DISCUSSION

The aim of the present study was to find out the intra- and inter-judge reliability of the voice samples by Speech Language Pathologists working in different set-ups using GRBAS scale.

The results are discussed as follows,

- Comparisons of rating by judges/ listeners working in different set-ups
- Reliability of ratings

Table 4.1.

Summary of the Mean, SD, Min, Max, Median and Mode of the Overall Grade

	First trial						Second trial					
	Mean	SD	Min	Max	Med	Mod	Mean	SD	Min	Max	Med	Mod
H	0.99	0.76	0	2	1.00	1	1.01	0.77	0	2	1.00	1
I	**1.28**	0.84	0	3	1.00	1	**1.21**	0.83	0	3	1.00	1
S	0.69	**1.13**	0	3	0.00	0	0.88	**1.08**	0	3	0.00	0
P	1.15	0.88	0	3	1.00	1	1.15	0.83	0	3	1.00	1

SD: standard deviation; Min: Minimum; Max: Maximum; Med: Median; Mod: Mode; H: hospital set-up; I: academic institute set-up; S: school set-up; P: private set-up

Table 4.2.

Summary of the Mean, SD, Min, Max, Median and Mode of Roughness

	First trial						Second trial					
	Mean	SD	Min	Max	Med	Mod	Mean	SD	Min	Max	Med	Mod
H	0.78	0.80	0	3	1.00	0	0.78	0.71	0	2	1.00	1
I	1.14	0.81	0	3	1.00	1	1.07	0.75	0	3	1.00	1
S	0.38	0.61	0	2	0.00	0	0.56	0.72	0	2	0.00	0
P	**1.15**	**1.03**	0	3	1.00	1	**1.12**	**0.88**	0	3	1.00	1

SD: standard deviation; Min: Minimum; Max: Maximum; Med: Median; Mod: Mode; H: hospital set-up; I: academic institute set-up; S: school set-up; P: private set-up

Table 4.3.

Summary of the Mean, SD, Min, Max, Median and Mode of Breathiness

	First trial						Second trial					
	Mean	SD	Min	Max	Med	Mod	Mean	SD	Min	Max	Med	Mod
H	0.65	0.82	0	3	0.00	0	0.89	0.76	0	2	1.00	0
I	0.93	0.82	0	3	1.00	1	1.07	0.77	0	3	1.00	1
S	0.69	**1.07**	0	3	0.00	0	0.75	0.77	0	2	1.00	0
P	**1.12**	1.00	0	3	1.00	0	**1.31**	**0.94**	0	3	1.00	2

SD: standard deviation; Min: Minimum; Max: Maximum; Med: Median; Mod: Mode; H: hospital set-up; I: academic institute set-up; S: school set-up; P: private set-up

Table 4.4.

Summary of the Mean, SD, Min, Max, Median and Mode of the Aesthenic Quality

	First trial						Second trial					
	Mean	SD	Min	Max	Med	Mod	Mean	SD	Min	Max	Med	Mod
H	0.58	0.68	0	2	0.00	0	0.60	0.64	0	2	1.00	0
I	0.85	**0.79**	0	3	1.00	0	0.78	**0.73**	0	2	1.00	1
S	**1.00**	0.73	0	2	1.00	1	**1.06**	0.68	0	2	1.00	1
P	0.55	0.64	0	2	0.00	0	0.69	0.62	0	2	1.00	1

SD: standard deviation; Min: Minimum; Max: Maximum; Med: Median; Mod: Mode; H: hospital set-up; I: academic institute set-up; S: school set-up; P: private set-up

Table 4.5.

Summary of the Mean, SD, Min, Max, Median and Mode of the Strain Quality

	First trial						Second trial					
	Mean	SD	Min	Max	Med	Mod	Mean	SD	Min	Max	Med	Mod
H	0.61	0.79	0	3	0.00	0	0.90	0.71	0	3	1.00	1
I	0.92	0.86	0	3	1.00	0	1.10	0.77	0	3	1.00	1
S	0.38	0.61	0	2	0.00	0	1.00	0.63	0	2	1.00	1
P	**1.03**	**0.91**	0	3	1.00	1	**1.17**	**0.78**	0	3	1.00	1

SD: standard deviation; Min: Minimum; Max: Maximum; Med: Median; Mod: Mode; H: hospital set-up; I: academic institute set-up; S: school set-up; P: private set-up

Tables 4.1.-4.5. summarizes the mean, SD, minimum, maximum, median and mode of the domains of the GRBAS scale in various set-ups both in both the trials.

In Table 4.1. it can be noted that both in the first and second trial, the institute and school set-up has the maximum and minimum mean respectively. The standard deviation is more in the school set-up and least in the hospital set-up whereas in the other set-ups (private and institute set-up) not much difference can be noted.

Table 4.2. represents the roughness, and here the highest mean and standard deviation was found in the private set-up in both the trials.

In table 4.3., breathiness measures were calculated and it was noted that in both the trials private set-up has the maximum mean but the high standard deviation measures varied between the trials, i.e., in the first trial, school set-up has the maximum deviation whereas in the second trial school set-up was found to have the maximum standard deviation.

In both the trials of table 4.4. representing aesthenic voice quality, it is observed that school set-up has the highest mean and institute set-up was found to have the highest standard deviation.

In the table 4.5., private set-up is seen to have the maximum mean and standard deviation when compared to the other set-ups for the strain quality domain of GRBAS scale. It can be also noted that in all the tables compared, the most frequently occurring score for all the domains was the '1' score which represents mild degree.

The above results can be attributed to the fact that in the school set-up, the exposure to varied voice disorder cases was limited and also the number of judges from this set-up was less (2) when compared to the other set-ups. Among all the other perceptual qualities, occurrence of roughness is common, easy to identify and strikingly different compared to the other quality disorders. After roughness,

breathiness quality was considered as the most common whereas harsh quality is considered as the most difficult to identify as it is the least common voice quality. Along with roughness in the literature it is mentioned that strain quality is also perceived better when compared to the other qualities (Yamauchi, Imaizumi, Maruyama & Haji, 2010)

Table 4.6.

Summary of Frequency and Percentage of the Responses for Overall Grade in the First and Second Trial

	0	**1**	**2**	**3**
G1hf	21	31	20	-
G1h%	29.20	43.11	27.82	-
G2hf	21	29	22	-
G2h%	29.20	40.3	30.6	-
G1if	14	28	26	4
G1i%	19.4	38.9	36.1	5.6
G2if	15	31	22	4
G2i%	20.8	43.1	30.6	5.6
G1sf	11	1	2	2
G1s%	15.3	1.4	2.8	2.8
G2sf	9	1	5	1
G2s%	12.5	1.4	6.9	1.4
G1pf	17	33	16	6
G1p%	23.6	45.8	22.2	8.3
G2pf	15	**36**	16	5
G2p%	20.8	**50**	22.2	6.9

Note: G1- overall grade in the first trial; G2: overall grade in the second trial; h: hospital set-up; i: academic institute set-up; s: school set-up; p: private set-up; f: frequency of occurrence; %: percentage of occurrence; responses of GRBAS scale- 0: normal; 1: mild; 2: moderate; 3: severe.

Table 4.7.

Summary of Frequency and Percentage of the Responses for Roughness in the First and Second Trial

	0	**1**	**2**	**3**
R1hf	31	28	11	2
R 1h%	43.1	38.9	15.3	2.8
R 2hf	28	32	12	-
R 2h%	38.9	44.4	16.7	-
R 1if	15	36	17	4
R 1i%	20.8	50	23.6	5.6
R 2if	16	**37**	17	2
R 2i%	22.2	**51.4**	23.6	2.8
R 1sf	11	4	1	-
R 1s%	15.3	5.6	1.4	-
R 2sf	9	5	2	-
R 2s%	12.5	6.9	2.8	-
R 1pf	23	25	14	10
R 1p%	31.9	34.7	19.4	13.9
R 2pf	19	30	18	5
R 2p%	26.4	41.7	25.0	6.9

Note: R1- roughness value in the first trial; R2: roughness value in the second trial; h: hospital set-up; i: academic institute set-up; s: school set-up; p: private set-up; f: frequency of occurrence; %: percentage of occurrence; responses of GRBAS scale- 0: normal; 1: mild; 2: moderate; 3: severe.

Table 4.8.

Summary of Frequency and Percentage of the Responses for Breathiness in the First and Second Trial

	0	**1**	**2**	**3**
B1hf	40	18	13	1
B 1h%	55.6	25.0	18.1	1.4
B 2hf	25	30	17	-
B 2h%	34.7	41.7	23.6	-
B 1if	23	35	10	4
B 1i%	31.9	48.6	13.9	5.6
B 2if	15	**41**	12	4
B 2i%	20.8	**56.9**	16.7	5.6
B 1sf	10	3	1	2
B 1s%	13.9	4.2	1.4	2.8
B 2sf	7	6	3	-
B 2s%	9.7	8.3	4.2	-
B 1pf	25	20	20	7
B 1p%	34.7	27.8	27.8	9.7
B 2pf	17	23	25	7
B 2p%	23.6	31.9	34.7	9.7

Note: B1- breathiness value in the first trial; B2: breathiness value in the second trial; h: hospital set-up; i: academic institute set-up; s: school set-up; p: private set-up; f: frequency of occurrence; %: percentage of occurrence; responses of GRBAS scale- 0: normal; 1: mild; 2: moderate; 3: severe.

Table 4.9.

Summary of Frequency and Percentage of the Responses for Aesthenic Quality in the First and Second Trial

	0	1	2	3
A1hf	38	26	8	-
A 1h%	52.8	36.1	11.1	-
A 2hf	35	31	6	-
A 2h%	48.6	43.1	8.3	-
A 1if	28	28	15	1
A 1i%	38.9	38.9	20.8	1.4
A 2if	29	30	13	-
A 2i%	40.3	41.7	18.1	-
A 1sf	4	8	4	-
A 1s%	5.6	11.1	5.6	-
A 2sf	3	9	4	-
A 2s%	4.2	12.5	5.6	-
A 1pf	**39**	27	6	-
A 1p%	**54.2**	37.5	8.3	-
A 2pf	28	38	6	-
A 2p%	38.9	52.8	8.3	-

Note: A1- aesthenic value in the first trial; A2: aesthenic value in the second trial; h: hospital set-up; i: academic institute set-up; s: school set-up; p: private set-up; f: frequency of occurrence; %: percentage of occurrence; responses of GRBAS scale- 0: normal; 1: mild; 2: moderate; 3: severe.

Table 4.10.

Summary of Frequency and Percentage of the Responses for Strain in the First and Second Trial

	0	**1**	**2**	**3**
S1hf	40	22	8	2
S 1h%	55.6	30.6	11.1	2.8
S 2hf	20	**41**	9	2
S 2h%	27.8	**56.9**	12.5	2.8
S 1if	28	24	18	2
S 1i%	38.9	33.3	25.0	2.8
S 2if	16	34	20	2
S 2i%	22.2	47.2	27.8	2.8
S 1sf	11	4	1	-
S 1s%	15.3	5.6	1.4	-
S 2sf	3	10	3	-
S 2s%	4.2	13.9	4.2	-
S 1pf	24	27	16	5
S1p%	33.3	37.5	22.2	6.9
S2pf	14	35	20	3
S2p%	19.4	48.6	27.8	4.2

Note: S1- strain value in the first trial; S2: strain value in the second trial; h: hospital set-up; i: academic institute set-up; s: school set-up; p: private set-up; f: frequency of occurrence; %: percentage of occurrence; responses of GRBAS scale- 0: normal; 1: mild; 2: moderate; 3: severe.

Table 4.6.- 4.10. summarizes frequency and percentage of the responses of the judges for all the scores obtained in various domains of the GRBAS scale in the first and second trials. It can be observed that apart from the aesthenic voice quality, all the other domains have the maximum frequency for the score '1' whereas, in the aesthenic voice quality measure, the maximum frequency was found to be for the score '0'.

Inter-rater reliability

Table 4.11.
Reliability Check Among All the Judges in The First and Second Trial for All The Domains

	First trial (α)	Second trial (α)
G (1-29)	0.93	0.96
R (1-29)	0.92	0.90
B (1-29)	0.95	0.96
A (1-29)	0.91	0.96
S (1-29)	0.89	0.94

Note: G: Overall grade; R: Roughness; B: Breathiness; A: Aesthenic quality; S: Strain (α): Cronbach's alpha coefficient measure

Cronbach's alpha coefficient measure (α) was used as the statistical measure to obtain the inter-rater reliability and the results were computed and it was seen that in each domain among all the judges the alpha ($\alpha > 0.70$) indicated good reliability both in the first and second trials. Here the reliability measure ranged between $0.891 \leq \alpha \leq 0.969$. This result can be supported by a study in which the inter-rater variances were found to be low and indicated that GRBAS can be considered as a reliable tool (Dejonckere, Obbens, de Moor & Wieneke, 1993).

Intra-rater reliability

Table 4.12.
Reliability Check Between the Same Judge in the First and Second Trial for All the Domains

	G (α)	R(α)	B(α)	A(α)	S(α)
J1	0.94	0.87	1.00	0.66	0.72
J2	1.00	0.87	0.85	0.87	0.95
J3	0.83	0.85	0.76	**- 0.09**	0.42
J4	0.88	0.85	0.88	0.87	0.92
J5	1.00	0.94	0.96	0.72	0.94
J6	1.00	0.94	1.00	1.00	1.00
J7	0.93	1.00	0.84	1.00	0.84
J8	1.00	0.90	0.76	**0.00**	**0.00**
J9	1.00	1.00	0.84	1.00	1.00
J10	1.00	0.93	0.97	0.95	1.00
J11	0.71	0.92	0.84	0.66	0.72
J12	0.85	0.97	0.97	0.87	0.71
J13	**0.41**	1.00	0.92	0.90	0.92
J14	1.00	0.93	0.93	0.85	0.94
J15	1.00	0.63	0.77	**- 0.72**	1.00
J16	**0.43**	0.72	1.00	0.96	0.00
J17	1.00	0.96	0.94	1.00	1.00
J18	1.00	0.92	1.00	0.90	1.00
J19	0.78	0.87	0.72	0.94	0.61
J20	0.89	0.78	0.91	1.00	0.75
J21	0.97	0.97	0.94	0.78	0.98
J22	1.00	0.67	0.97	0.72	**0.50**
J23	0.85	0.75	0.90	1.00	**0.56**
J24	1.00	0.90	0.81	1.00	1.00
J25	0.93	0.95	0.93	0.90	0.92
J26	1.00	1.00	0.97	0.85	0.96
J27	1.00	1.00	0.94	**0.51**	0.00
J28	1.00	1.00	1.00	0.85	1.00
J29	0.93	1.00	0.97	0.87	0.91

Note: J1-J29: Number of Judges; G: Overall grade; R: Roughness; B: Breathiness; A: Aesthenic quality; S: Strain; (α): Cronbach's alpha coefficient measure

For intra- rater reliability also the Cronbach's alpha coefficient measure (α) was used and the domain values for each judge were computed within the trials. It can be noted that among the domains of the GRBAS rating scale, the rough 'R' and breathy 'B' quality had the maximum reliability measure among all the judges across the trials. Overall grade 'G' and strained quality 'S' had better reliability scores when compared to the aesthenic 'A' quality. The least reliability scores among the judges for the various domains were the aesthenic 'A' quality reliability measures.

The above result was supported by a study which aimed at comparing voice quality by comparing 100 voice samples analysis between the GRBASI (includes Grade, Roughness, Breathiness, Aesthenia, Strain and Instability) and RASATI (includes Roughness, Harshness, Breathiness, Aesthenia, Strain, Instability) systems. Listeners rated 100 voice samples and these were analyzed to identify the significant interrelations between the scales, with asthenia, roughness and instability as the common factors. In this study also among all the three factors considered roughness was the most consistent and the easiest to identify by evaluators (Yamauchi, Imaizumi, Maruyama & Haji, 2010). This may be due to the fact that among all the voice quality measures, the 'R' rough quality can be easily discriminated from the other voice qualities.

The study by Dejonckere, Obbens, de Moor & Wieneke (1993) which supported good inter-judge reliability using GRBAS scale also gives good intra-judge reliability measures. This study involves 15 parameters which were taken (comprising the GRBAS parameters) to assess 12 clearly dissimilar voices of different pathologies. These voice samples were assessed by 6 speech therapists as judges. It was found that on the basis of intra judge (low), inter judge (low) and inter voice (high) variances, the GRBAS scale parameters appear to be quite reliable and are of clinical relevance

for evaluating the overall severity of hoarseness. The best correlation between judges (0.7) was found for the overall grade of severity and this seems to be mainly determined by the component breathiness.

Inter-rater reliability (within set-ups)

Table 4.13.

Reliability Check Among All the Judges Across Different Set-ups in the First and Second Trial for All the Domains

Set ups	First trial (α)					Second trial (α)				
	G	R	B	A	S	G	R	B	A	S
Hospital (9)	0.86	0.88	0.85	0.64	0.81	0.92	0.89	0.89	0.91	0.86
Institute (9)	0.73	0.81	0.83	0.62	0.72	0.88	0.83	0.84	0.93	0.82
School (2)	0.71	**-.09**	0.68	**0.06**	0.78	0.68	**0.22**	0.66	**0.00**	0.75
Private (9)	0.81	**0.47**	0.93	0.91	**0.43**	0.87	**0.49**	0.96	0.92	0.83

G: Overall grade; R: Roughness; B: Breathiness; A: Aesthenic quality; S: Strain
(α): Cronbach's alpha coefficient measure

To obtain the inter-rater reliability of the domains in the GRBAS scale across different set-ups Cronbach's alpha coefficient measure (α) was used. It was noted that in the school set-up rough and aesthenic quality of voice had the least reliability measure. This may be due to the less number of judges in the school set-up. Hospital and institute set-ups had the best reliability measure according to the Cronbach's alpha measure (α). Compared to the hospital and institute set-ups, the private setup had poor reliability measure ($\alpha < 0.5$) particularly for the rough quality in both the trials and strained quality in the first trial.

Trial agreement (Kappa measurement of agreement)

Table 4.14.
Summary of rater agreement within setups across domains

	Grade	Rough	Breathy	Aesthenic	Strained	**Total (K)**
Hospital	0.87	-	-	0.73	0.55	**0.70**
Institute	0.79	0.71	0.78	-	0.75	**0.75**
School	0.66	0.76	-	0.89	**0.20**	**0.61**
Private	0.91	0.82	0.73	0.59	0.75	**0.77**
Total (K)	**0.85**	**0.76**	**0.70**	**-**	**0.66**	**0.73**

Note: K: Kappa measurement of agreement

An agreement measure, Kappa measurement of agreement (K) was used to find the trial agreement of judges within set-ups across domains. The total reliability measures in all the domains of GRBAS scale within set-ups was good. In the aesthenic quality measure, the values obtained from the GRBAS scale were not symmetrical in both the trials and therefore a Kappa measure of agreement could not be obtained. The reliability measure of the strained quality in the school set-up was poor and the value was not significant ($p > 0.05$). The overall grade of severity had the best reliability measure and the strained quality had the most poor agreement score on Kappa measurement of agreement among all the domains in the GRBAS scale. The total trial agreement within set-ups and across domains had an agreement measure of 0.738 which was considered as a good measure.

Reliability within judges across settings

Table 4.15.
Summary of the Reliability Within Judges Between Trials in Various Settings

	Hospital (α)	Institute (α)	School (α)	Private (α)
Judge 1	0.95	-	-	-
Judge 2	0.96	-	-	-
Judge 3	0.72	-	-	-
Judge 4	0.96	-	-	-
Judge 5	0.98	-	-	-
Judge 6	0.99	-	-	-
Judge 7	0.98	-	-	-
Judge 8	0.97	-	-	-
Judge 9	0.99	-	-	-
Total	**0.96**			
Judge 10	-	0.99	-	-
Judge 11	-	0.84	-	-
Judge 12	-	0.90	-	-
Judge 13	-	0.98	-	-
Judge 14	-	0.98	-	-
Judge 15	-	0.75	-	-
Judge 16	-	0.86	-	-
Judge 17	-	1.00	-	-
Judge 18	-	0.98	-	-
Total		**0.94**		
Judge 19	-	-	0.93	-
Judge 20	-	-	0.88	-
Total			**0.89**	
Judge 21	-	-	-	0.97
Judge 22	-	-	-	0.73
Judge 23	-	-	-	0.90
Judge 24	-	-	-	0.94
Judge 25	-	-	-	0.91
Judge 26	-	-	-	0.98
Judge 27	-	-	-	0.94
Judge 28	-	-	-	0.99
Judge29	-	-	-	0.96
Total				**0.95**

(α): Cronbach's alpha coefficient measure

Cronbach's alpha coefficient measure (α) was used to obtain the reliability measure within judges between trials in various settings and it was noted that the judges in the hospital set-up had the best reliability score followed by the private, institute and school set-ups. The poor scores obtained for the judges in the school set-

up could be attributed to the less number of judges rating in the school set-up when compared to the other set-ups.

The description about what parameters helped the judges to rate the voice samples using GRBAS scale was not very informative as it was an open ended question. Different judges perceived the voice and the descriptions were varied. The description of the voice samples were majorly based on the pitch and loudness parameter.

CHAPTER V

SUMMARY & CONCLUSION

To summarize the intra- and inter-rater reliability is good when Speech Language Pathologists were taken as judges to rate voice samples using GRBAS scale. This study also suggests that GRBAS is a very reliable and easy to administer perceptual tool and can be used for everyday clinical diagnostic purposes. In the intra-rater reliability measure, the variations between the judges were less and the reliability was good in all the domains of the GRBAS except for the aesthenic 'A' voice quality measure which is discussed in the results above. As noted in this study, the various set-ups in which the Speech Language Pathologists in had an effect on the reliability measures. Even though the effect was different using the statistical measures (Kappa measurement of agreement & Cronbach's alpha reliability measure) it was found that the measures were having good reliability.

To conclude the study findings can be used to validate the results found in the existing literature as that intra- and inter-rater reliability scores were found to be good and therefore GRBAS perceptual tool can be considered as a reliable tool for everyday clinical investigations. The generalization of the study should be carried out with caution as the number of judges in each set-up was limited, therefore further studies can be carried out by considering large sample of judges in each set-up.

References

Abercrombie, D. (1967). *Elements of General Phonetics.* Chicago: Aldine.

ANSI (1960). Acoustical Terminology, ANSI S1.1.12.9, p. 45, American National Standard Institute, New York

Anderson, V. A. (1961). *The Effective Voice: Training the speaking voice*. 2nd edn, New York: Oxford University Press

Aronson, A. (1980). *Clinical Voice Disorders: An Interdisciplinary Approach*. 1st edn, New York, Brian . C. Decker

ASHA Special Interest Division 3, Voice and Voice Disorders 2002-2006. Consensus-Auditory Perceptual Evaluation of Voice (CAPE-V), The American Speech Language Hearing Association. From http//:www.asha.org. Retrieved on September 12, 2008

Austin, G. (**1806**). *Chironomia* (London: Cadell and Davies). Reprinted by Southern IllinoisUniversity Press, Carbondale, IL, 1996.

Baken, R. J., (1987). *Clinical Measurement of Speech and Voice*, Boston: College Hill

Baken, R., & Orlikoff, R. (2000). *Clinical Measurement of Speech and Voice*, San Diego, Singular publishing

Breitenstein, C., Van Lancker, D., & Daum, I. (2001). "The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample," *Cognition and Emotion ,15*, 57–79.

Coleman, R.F. (1971). Effect of waveform changes upon roughness perception. *Folia Phoniatrica*, *23*, 314-322.

Bele, I.V (2004). Reliability in Perceptual Analysis of Voice Quality. *Journal of Voice*, Vol 19, No 4, pp 555-573.

Bhuta, T., Patrick, L. & Garnett,J.D. (2004). Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice*, *18*, No 3, pp 299-304

Carding, P., Carlson, E., Epstein, R., Mathieson, L. & Shewell, C. (2000). Formal perceptual evaluation of voice quality in the United Kingdom: Voice Forum, *Logopedics, Phoniatrics Vocology*; *25*, 133-138

Carding, P.N., Wilson, J.A., MacKenzie, K., & Deary, I.J. (2009). Measuring voice outcomes: State of the science review. *Journal of Laryngology and Otology 123* (8): 823-829

Colton, R.H. & Casper, J.K. (1996*) Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment*. Baltimore, MD: Williams & Wilkins

De Bodt, M.S., Wuyts, F., 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*, No. I, pp. 74—80

DeCasper, A.J. & Fifer, W.P. (1980). "Of human bonding: Newborns prefer their mothers' voice," *Science, 208*, 1174-1176.

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., Remacle, M., Elbaz, E., Woisard, V., Buchman, L. & Millet, B. (1996) "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements," *Revue de Laryngolie Otologie Rhinologie*, *117*, pp. 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 investigating the 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

Fairbanks, G. (1960). *Voice and Articulation drill book*, New York: Harper and Bros

Geiselman, R.E. & Bellezza, F.S. (1977). "Incidental retention of speaker's voice," *Memory & Cognition ,5*, 658–665.

Gerratt, B. R., Till, J. A., Rosenbek, J.C., Wertz, R. T & Boysen, A. E. (1991). Use and perceived value of perceptual and instrumental measures in dysarthria management. In C. A. Moore, K. M. Yorkston, & D. R. Beukelman (Edrns), *Dysarthria and Apraxia of Speech* (pp 77-93), Baltimore

Gould, J., Waugh, J., Carding, P. & Drinnan, M. (2011). A new voice rating tool for clinical practice. *Journal of Voice*, *26*, No 4, 163-170

Hakkesteegt, M.M., Brocaar, M.P., Wieringa, M.H. & Feenstra, L. (2006). The relationship between perceptual evaluation and objective multiparametric evaluation of dysphonia severity. *Journal of Voice*, *22*, No 2, 138-145.

Hammarberg, B., Fritzell, B., Gauffin, T., Sundberg, J. & Wedin, L.(1980). Perceptual and Acoustic Correlates of Abnormal Voice Qualities. *Acta Orolaryngology*, *90*, 441-451

Hammarberg, B., Fritzell, B., Gauffin, T. & Sundberg, J.(1986). Acoustic and Perceptual analysis of vocal dysfunction. *Journal of Phonetics*. *14*, 533-547

Hepper, P.G., Scott, D., & Shahidullah, S. (1993). "Newborn and fetal response to maternal voice," *Journal of Reproductive and Infant Psychology, 11*, 147–153.

Hirano, M. (1975). Phonosurgery: Basic and clinical investigations. Otologia, 21, 239-240, Cited in Baken, R.J., (1987). *Clinical measurement of Speech and Voice*. Taylor & Francis Ltd

Hirano, M. (1981). *Clinical Examination of Voice: Disorders of Human Communication*, 5, NewYork: Springer

Iwarsson, J. & Peterson, N.R. (2011). Effects of consensus training on the reliability of auditory perceptual ratings of voice quality. *Journal of Voice*, *26*, No 3, pp 304-312

Jesus, L.M.T., Barney, A., Sa Couto, P., Vilarinho, H. & Correia, A.(2009). Publications In, *6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications,* Florence, Italy 61-64

Karnell, M.P., Melton, S.D., Childes, J.M., Coleman, T.C., Dailey, S.A. & Hoffman, H.T. (2006). Reliability of Clinician-Based (GRBAS and CAPE-V) and Patient-Based (V-RQOL and IPVI) Documentation of Voice Disorders. *Journal of Voice*, *21*, No. 5, pp. 576–590

Kelchner, L.N., Brehm, S.B., Weinrich, B., Middendorf, J., de Alarcon, A., Levin, L. & Elluru, R. (2008). Perceptual Evaluation of Severe Pediatric Voice Disorders: Rater Reliability using the Consensus Auditory Perceptual Evaluation of Voice . *Journal of Voice*, *24*, No. 4, pp. 441–449

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., Geratt, B.R. & Kempster, G.B. (1993). Perceptual evaluation of voice quality: Review, Tutorial and a framework for future research. *Journal of Speech & Hearing Research*, *36*, 21-40

Kreiman, J. & Sidtis, D. (2011). *Foundations of Voice Studies: An interdisciplinary approach to voice production and perception*. United Kingdom: Blackwell Publishing.

Laver (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Law, T., Kim, J.H., Lee, K.Y., Tang, E.C., Lam,J.H., van Hasselt, A.C. & Tong, M.C. (2011) Comparison of Rater's Reliability on Perceptual Evaluation of Different Types of Voice Sample. *Journal of Voice*, *26*, No. 5, pp. 666.e13-666.e21

Moore, G.P. (1971). Voice disorders organically based. In L. E. Travis (Ed), *Handbook of Speech Pathology and Audiology*, New Jersey: Prentice Hall, pp 539-569

Moore, G.P. (1975).Observation on the physiology of hoarseness- Proceedings of the fourth international congress of phonetic science. Helsinki: Finland pp 92-95. Cited in Childrens, D.G. and Lee, C.K. (1991). Voice quality factors:

Analysis, synthesis and perception. *Journal of Acoustical Society of America, 90* (5), 2394-2410.

Munoz, J., Mendoza, E., Fresneda, M.D., Carballo, G. & Lopez, P. (2003). Acoustic and Perceptual Indicators of Normal and Pathological Voice. *Folia Phoniatrica Logopedics; 55,* 102-114.

Munoz, J., Mendoza, E., Fresneda, M.D., Carballo, G. & Ramirez, I. (2002). Perceptual analysis of different voice samples: agreement and reliability. *Perceptual and Motor skills*, *94* (3 Pt 2), 1187-1195.

Nemr, K., Zenari, M.S., Cordeiro, G.F., Tsuji, D., Ogawa, A.I., Ubrig, M.T., & Menezes, M.H.M (2012) GRBAS and Cape-V Scales: High Reliability and Consensus When Applied at Different Times. *Journal of Voice*, *26*, No. 6, pp. 812.e17-812.e22

Pannbacker, M. (1984). "Classification systems of voice disorders: A review of the literature" *Language, Speech, and Hearing Services in Schools* 15, 169–174.

Piske, T., MacKay, I.R.A., & Flege, J.E. (2001). "Factors affecting degree of foreign accent in an L2: A review," *Journal of Phonetics ,29*, 191–215.

Reynolds, V., Buckland, A., Bailey, J., Lipscombe, J., Nathan, E., Vijayasekaran, S., Kelly, R., Maryn, Y. & French, N. (2012). Objective assessment of pediatric voice disorders with the acoustic voice quality index. *Journal of Voice*. *26* (5): pp 672.e1-672.e7

Schafer, A.J., Speer, S.R., Warren, P., & White, S.D. (2000). "Intonational disambiguation in sentence production and comprehension," *Journal of Psycholinguistic Research ,29*, 169–182.

Stemple, J.C., Glaze, L.E. & Klaben, B. (2010). *Clinical Voice Pathology: Theory and Management*. 4th edn, San Diego: Plural Publishing.

Titze, I.R. (1994). List of voice qualities. Proceedings of the 8th Vocal Fold Physiology Conference. Retrieved from http://www.ncvs.org/

Van Lancker, D., Canter, G.J., & Terbeek, D. (1981). "Disambiguation of ditropic sentences: Acoustic and phonetic cues," *Journal of Speech and Hearing Research,24*, 330–335.

Webb, A.L., Carding, P,N., Deary, I.J., MacKenzie, K., Steen, N. & Wilson, J.A. (2003). The reliability of three perceptual evaluation scales for dysphonia. *European archives of oto-rhino-laryngology,261* (8), 429-434

Wells, B. & Macfarlane, S. **(**1998**).** "Prosody as an interactional resource: Turn–projection and overlap," *Language and Speech ,41*, 265–294.

Wendahl, R.W. (1966). Laryngeal Analogy synthesis of jitter and shimmer auditory parameters of harshness. *Folia Phoniatrica*. 18, 98-108. Cited in Baken, R.J. (1987). *Clinical measurement of speech and voice*. Boston: College Hill

Wilson, F.B. (1987). *Voice Disorders*. Texas, Austin

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.

Yamauchi, E. J., Imaizumi, S., Maruyama, H. & Haji, T. (2010). Perceptual evaluation of pathological voice quality: A comparative analysis between the RASATI and GRBASI scales. *Logopedics, Phoniatrics Vocology*; *35* (3): 121-128.