RELIABILITY OF PERCEPTUAL EVALUATION OF VOICE USING CAPE-V RATING SCALE IN INDIAN CONTEXT

Register No. 07SLP008

A dissertation submitted in part fulfillment for the degree of

M.Sc., (Speech Language Pathology)

University of Mysore, Mysore.

ALL INDIA INSTITUTE OF SPEECH AND HEARING,

NAIMISHAM CAMPUS, MANASGANGOTHRI,

MYSORE – 570006.

MAY – 2009

CERTIFICATE

This to certify that this dissertation entitled **“RELIABILITY OF PERCEPTUAL EVALUATION OF VOICE USING CAPE-V RATING SCALE IN INDIAN CONTEXT”** is the bonafide work in part fulfillment for the degree of Masters in Science (Speech Language Pathology) of the student (Register No. 07SLP008). This has been carried out under the guidance of a faculty of this institute and has not been submitted earlier to any other university for award of any other degree or diploma.

Place – Mysore

May 2009

Dr. Vijayalakshmi Basavraj

Director

All India Institute of Speech & Hearing

Naimisham campus, Manasgangothri

Mysore – 570006

CERTIFICATE

This is to certify that this dissertation entitled **"RELIABILITY OF PERCEPTUAL EVALUATION OF VOICE USING CAPE-V RATING SCALE IN INDIAN CONTEXT"** has been prepared under my supervision and guidance.

Place – Mysore

May 2009

Dr M. Pushpavathi

Reader & Head of Department

Dept. of Clinical Services

All India Institute of Speech & Hearing

Naimisham Campus, Manasgangothri

Mysore – 570006

DECLARATION

This dissertation entitled **"RELIABILITY OF PERCEPTUAL EVALUATION OF VOICE USING CAPE – V RATING SCALE IN INDIAN CONTEXT"** is the result of my own study under the guidance of Dr M. Pushpavathi, Reader & HOD, Dept of Clinical Services, All India Institute of Speech and Hearing, Mysore, and has not been submitted earlier at any university for any other Diploma or Degree.

Place – Mysore **Reg. No. 07SLP008**

May 2009

DECLARATION

This dissertation entitled **"RELIABILITY OF PERCEPTUAL EVALUATION OF VOICE USING CAPE – V RATING SCALE IN INDIAN CONTEXT"** is the result of my own study and has not been submitted earlier at any university for any other Diploma or Degree.

Place – Mysore **Reg. No. 07SLP008**

May 2009


Dedicated to
My Lord,
My Maa
& Papaji…

AKNOWLEDGMENTS

It gives me immense pleasure in presenting this dissertation and I would like to express my sincere gratitude to all those who have directly or indirectly helped me in the accomplishment of the study.

The dictionary is full of words of every length and kind, but I am at a loss of words to convey my gratitude to my teacher and my guide , Dr. M Pushpavathi, Reader & HOD of Clinical Services, AIISH, Mysore. It was her able supervision and timed approach to work throughout the course of completion of the study and making this tortoise (me) move a bit fast in pace. Ma'am I am thankful to you not only for the invaluable guidance, constant support and encouragement you have given me but also for forgiving me for many of my mistakes & slip ups and always giving me a next chance to prove myself. Ma'am I am obliged for whatever you have done for me!

I extend my gratitude to Dr. Vijayalakshmi Basavraj, Director, AIISH, Mysore for giving me this opportunity to carry out the present study.

I am very grateful to Vasanthalakshmi Mam for her support and the time she gave me for the statistical work of the study. Thank you so much maam!

Maa and papa…..no words of mine will be sufficient to thank you. I thank you for believing in me and my capabilities. Mummy u are my closest friend… you understand me even when you don't hear me. I am indebted to you mummy for every thing. You both are God to me. Papaji you are the best father a girl can have. You have pampered your 'lado' like a small baby forever and you still make me feel like a small girl in papa's arms. Papa I wish I had patience, humbleness and soft heart like you have towards all. My world is full of smiles because of you both. Thanks for everything mummy – papa!!

Pintu Bhaiya …you are not only my elder brother but a friend, a father. The way you encourage me for my little successes, makes me feel that I have won the whole world. It brings smile on my face thinking of the days we both used to fight like hell for little hearts biscuits, cycle & what not! Pappu bhaiya I love you for being my closest friend who can understand me 'the best' and for always showering love & reverence upon your sis. Dare you both change ever! Thanks for being a part of my life. Dear grand parents (dadiji and naniji) …thank you for all your inspirations & prayers.

I am grateful to all my teachers from nursery till now for imparting knowledge, moral values and the right path in my life.

Friendship is a wine, tastes better when it's old. I run short of words to thank all my class mates there in Mumbai. Pooja, Bhavna, Sappu tai, Priya, Neeli, , Pragati, Shru, PP, Pallu, Taz, Sush, Meenu, Sami…not just for the good - bad times we have shared but because 'the family' we have made & will carry on this relationship forever. You all taught me so many things in your own strange ways and I'm so grateful to God that he blessed me with love & care from you all.

Priya I seriously missed your suggestions while doing the study. Hope you were here for me. Nili thanks for all the kick - fights and care! Bhavna and Pooja for your pampering love to me! Your friendship is a blessing to me!!

Sahil , Karry, Mohsin, Ani, Senu, Neelesh, Sanky you were the best friends any one would think off. Be it a small help to a big decision. You guys have always given me your shoulder to rest upon. I missed you all so much. It's all because of you all I feel the strength.

Prafful you are one of the best persons I have ever met. I am indebted for all the helps you have always given me. Your inspirations and support in everything is invaluable to me, forever in my life. I pray my God to fulfill all your "Desire" and to achieve and do best in life.

I also thank my school buddies Neha Srivastava, Neha Gupta, Srishti and Ritu for all the moral support they have forever showered on me.

Sasmi…if you were not here, with whom I would have shared my thoughts, my worries. Thanks for making me feel like a younger sis. I will miss our long chats in night. I know we will carry our friendship forever like this. Wish you a great future sweety…..

Pari… you know the best thing about is your super caring nature!! Thanks for all the care & love you have showered on me always. Meera ….you share a very special place in my heart. The time we have spent together is just unforgettable. Thanks for teaching 'this day dreamer' and being the best study buddies & jolly friends!! Devi …u rock!! Life at AIISH would not have been so if I didn't find you as my roomie. Thanks girlie!!

Riddhima…you taught me the importance of fun, dance and chill attitude in life. You are a perfect tension reliever with a bonus of beautiful heart and super duper brain! Pallu…you are

one of the most encouraging friend who always rendered me a listening ear and helping hand whenever I was depressed. Especial thanks to Pallu and Shuchi for taking care of my laptop in this crucial period.

Navitha and Aiishu…you both are just unparalleled to any one else. Navi..you are friend of all & Aiishu for me you are the best example of ''simple living & great thinking''. Thanks to u both for all the help you have given me. Prathi…thanks for keeping me awake in classes & m sorry for always irritating you!! Love you Mrs Nagesh!! Meenakshi...you are a great friend. Thanks dear for all the help and guidance you have given for my dissertation.

I thank all my friends Ramya, Pooja, Sinthia, Sweety, Samas and every one for making my AIISH stay a memorable one. I thank ALL MY CLASS MATES for giving me moral support & making my AIISH life the most memorable part of my life. Years might pass but thoughts will always remain in me. I will miss you all badly.

I extend my sincere gratitude towards all library staffs. Sir, without your cooperation I would have never completed my dissertation. Thanks a lot! I am very grateful to Dr Sundar Raju and Raghuna Sister. For if you were not I would have never completed my study on time. I also want to thank all the participants and judges of the study, for their patience and cooperation, in spite of the busy schedule.

Thanks to all hostel inmates, all juniors and seniors for their love and cooperation. And special thanks to akkas for the love and care they have given me!

Last but not the least to my God, who makes me believe in my self and to whom I owe everything I m today, everything I possess today in this world. Without you this girl is nothing. Please promise your presence with me….forever!!

~~~~
~~~~

TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE NO.
1	INTRODUCTION	1 - 9
2	REVIEW OF LITERATURE	10 - 38
3	METHOD	39 - 46
4	RESULTS AND DISCUSSION	47 - 73
5	SUMMARY AND CONCLUSION	74 - 81
	REFERENCES	
	APPENDIX I	
	APPENDIX II	
	APPENDIX III	
	APPENDIX IV	

LIST OF TABLES

Table No.	Title	Page No.
Table 1	Inter rater reliability of Phonation task	49
Table 2	Inter rater reliability of Sentences task	50
Table 3	Inter rater reliability of spontaneous speech task	51
Table 4	Overall inter rater reliability across tasks on both the scales of CAPE-V	52
Table 5	Correlation between ordinal and VA scale on phonation task.	55
Table 6	Correlation between ordinal and VA scale on Sentences Reading task.	56
Table 7	Correlation between ordinal and VA scale on spontaneous speech task.	57
Table 8	Scheme for transforming VA score to 10 point modified VA scale.	59
Table 9	Cross tabulation for across all samples.	60
Table10	Overall correlation between the two scales on all the three tasks.	64
Table 11	Overall correlation between the two scales on all the three tasks (using mVA and ordinal scale).	65

LIST OF FIGURES

Figure No.	Title	Page No.
Figure 1	Descriptive framework for perceptual evaluation of voice.	20
Figure 2	CAPE- V Form.	42
Figure 3	Visual representation of relationship between VA and ordinal scales.	63
Figure 4	Graph representing the rating patterns on the phonation task.	66
Figure 5	Graph representing the rating patterns on the sentences task.	67
Figure 6	Graph representing the rating patterns on the spontaneous speech task.	67
Figure 7	Visual representation of relationship between VA and ordinal scales.	78

CHAPTER – I

INTRODUCTION

"Every human Society, no matter how primitive has developed the ability to communicate through speech and our ability to communicate through spoken and written language has been cited as one single most important characteristic that sets the human apart from other animals"(Curtis, 1978).

The underlying basis of speech is voice. According to Green (1964) "voice plays the musical accompaniment to speech rendering it tuneful, pleasing, audible and coherent and is an essential feature of efficient communication by spoken word". The speaking voice conveys information about the speaking individual and the voice quality serves as a primary means by which speakers project their physical, psychological and social characteristics to the world.

A good voice is a clear, resonant, stable, well supported by adequate breath control. It is at a pitch level that is appropriate to the speaker and the message. Rate of speech is such that the messages are clearly understood. An effective speaking voice should have the following characteristics (Anderson, 1961),

- Adequate loudness
- Clearness and purity of tone
- A pleasing and effective pitch level
- Ease and flexibility
- A vibrant, sympathetic quality
- Ease of diction

Voice disorders arise when an individual’s quality, pitch or loudness differs from voice characteristics typical of speakers of similar age, gender, cultural background and geographical location. The range of etiologies of voice disorders is large and these differences may result from a variety of factors. Structural, medical or neurologic alterations of the respiratory, laryngeal, and vocal tract mechanisms may create a voice disorder.

Voice disorders can be classified based on pitch, loudness and quality. The classification based on quality includes hoarse voice, harsh voice, breathy voice and strained voice. Hoarse voice is a common term used to describe the voice disorder characterized by breathy and harsh voice. Voice quality is a term that subsumes a wide range of possible meanings covering both laryngeal and supra-laryngeal aspects. It is perception of the physical complexity of laryngeal tone modified by cavity resonation.

Fairbanks (1960) tried to distill voice quality defects into three categories – hoarseness, harshness and breathiness. Rarely in clinical practice, does abnormal voice vary along a single dimension of quality, loudness, pitch or flexibility. Most of the times, even though one may predominate, the others are usually present in different combination and proportion.

It is difficult to define hoarseness. But it is a psycho acoustic term used in broader sense to mean any abnormal voice quality due to laryngeal pathology. According to Bayens (1966) it is the quality of voice that is rough, grating, harsh and more or less dominant and a lower in pitch than the normal for the individual. Moore (1971) defines hoarse voice as a voice which is characterized by noise of relatively high frequency that is produced by transient, highly unstable variations. Sederholm, Mc Allister, Sundberg and Dalkwist (1992), reported with the help of factor analysis that the hyper-function, breathiness and roughness are good predictors of hoarseness.

Harshness and breathiness are two important components of hoarseness. Harshness is perceived due to irregularity in the vocal fold vibrations (Wendahl, 1966; Coleman, 1971 & Moore 1975) i.e., variations or perturbations in both amplitude and time period from cycle to cycle give the impression of harshness. Breathiness is perceived by escape of air through partially closed glottis and the resultant turbulence noise reduces the harmonic to noise ratio.

Assessment of Voice

Voice can be evaluated objectively as well as subjectively in many ways. Objective measurements include acoustic, aerodynamic and physiologic parameters using instruments. Similarly, subjective evaluation includes perceptual ratings of voice on various parameters like roughness, breathiness, resonance, loudness etc. But when objective acoustic measures alone are used to analyze vocal quality there appear to represent only a friction of the set of all the measures used by the human listener.

Acoustic analysis is the process of objective identification and description of the voice. Evaluation of the various parameters of the vocal signal can be carried out by individual instruments designed for the particular purpose or increasingly by software packages that can analyze each parameter and subsequently integrate the data acquired regarding these individual aspects. An acoustic analysis profile emerges from that, which indicates the extent to which each parameter deviates from normative values and which acts as a baseline for treatment progresses.

Perceptual Evaluation

Human ears have the ability to identify and recognize the speaker's voice. A trained voice clinician is often able to determine the causative pathologies on the basis of psychoacoustic impression of voice (Hirano, 1975). Perceptual voice evaluation is an integrated process of listening to and describing a particular voice. The clinician needs intensive training in voice dimensions that identify pathology most effectively. Rating voice quality perceptually is universally acknowledged as difficult task and one requires considerable experience in perceptual judgments. Voice quality may be considered as the perceived result of coordinated actions of the various systems. The perceptual importance of different aspects of voice depends on context, attention, a listener's background and the listening task (Kreiman, Garratt, Kempster, Erman & Berke, 1993).

In the literature, there are varieties of perceptual scales described and the reliability of the data varies from study to study. There are no reliable verbal terms defining vocal characteristics. Significant correlation between frequency perturbation and perceptual qualities such as instability, flutter, roughness, diplophonia and creakiness/ vocal fry were found. Hammerberg and Gauffin (1986) concluded that perceptual evaluation by well trained listeners is reliable and reproducible and can be used for systematic evaluation purposes, if handled with precaution. These authors further concluded that voice quality can be more precisely perceived, if professional terminologies were given to the listener.

The reliability of perceptual voice evaluation can be improved by (Sarita, 2000),

- ✓ Operationally defining the voice parameter to be evaluated.
- ✓ Illustrating the voice quality parameters by samples of audio recordings.
- ✓ Searching for acoustic and physiological correlates of perceptual parameters.

Hammerberg, Fitzell, Gauffin and Sundburg (1986) pointed out that perceptual voice evaluation by clinically well trained listeners can be reliable if based on standardized rating procedure and that training for voice therapists can be more effective if perceptual acoustic relationships are identified.

Advantages of Perceptual Evaluation

- The importance of perceptual measures is also demonstrated by their frequent use as a standard against which acoustic measures are validated or compared (Kreiman et al, 1993)
- Researchers proposing objective voice measures often demonstrate their measure's utility by reporting a correlation between the measures and ratings of perceived vocal quality (Kojima, Gould, Lambiase, Isshiki, 1980; Fukazawa &

El- Assuooty, 1988; Ladefoged, Maddieson & Jackson, 1988; Klattt & Klatt, 1990; Hillenbrand, 1994; Takahashi & Koike, 1975).

- Perceptual voice evaluation using any standardized scale is an inexpensive, readily available and practical tool for evaluation purposes.
- It has been found to be reliable in its findings in both inter-judge and intra-judge reliability.
- Perception of patient's voice is the heart of evaluating and treating patients with voice disorders. Thus, listener judgment is essential, both for clinical consideration and criterion validation of instrumental voice measures.
- For a reliable assessment, objective tests are always correlated with perceptual evaluation of voice.

Limitations of Perceptual Evaluation of Voice

The subjective evaluation of voice quality are not highly regarded as either clinical or research tools because of the following reasons, like

- They are considered to lack objectivity and do not require great technical sophistication (Weismer & Liss, 1981).
- There is no accepted set of perceptual scales used by the clinicians (Yumoto, Gould & Baer, 1982).

- Also, the factors like reliability and uncertainty regarding the use and meaning of various rating scales have led some to abandon perceptual measures in favor of instrumental approaches to voice assessment and because of inherent problems with inter-judge and intra-judge reliability (Cullinan, Prather & Williams, 1963; Ludlow, 1981). However, as pointed out by Moll (1964), if a measure of vocal quality has to be useful, it must be closely related to listener's judgment of that vocal quality dimension. Both clinical and research practices are built upon perceptual data, but these data have never been gathered in ways that foster the confidence of clinicians or researchers.

Because of these views the subjective assessment of voice has received back seat in evaluation of voice pathology with objective evaluation as a primary means of assessment.

Different Rating Scales

In literature, there are many types of perceptual scales available for the judgment of the voice disorders. They may be a Categorical rating scale, Equal Appearing Interval (EAI) scales, Visual Analog (VA), Direct Magnitude Estimation (DME) or Paired Comparison.

The following scales have been developed by several authors are -

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

The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) was developed as a tool for clinical auditory-perceptual assessment of voice from a consensus meeting sponsored by ASHA held in Pittsburg (2002). Its primary purpose is to describe the severity of auditory-perceptual attributes of a voice problem, in a way that can be communicated among clinicians. Its secondary purpose is to contribute to hypotheses regarding the anatomic and physiological bases of voice problems and to evaluate the need for additional testing. CAPE-V is not intended for use as the only means of determining the nature of the voice disorder. It is not to be used to the exclusion of other tests of vocal function. Finally, it is not expected to demonstrate a 1:1 relation to results from other tests of vocal function.

Need for the Study

Perception of a patient's voice is at the heart of evaluating and treating patients with voice disorders. Patients and their families decide whether treatment has been successful based largely on whether the patient sounds better. Similarly clinicians make many decisions about managing the speech and voice disorder based upon perceptual judgment. Thus, there is always a need of a reliable perceptual voice rating scale, which has good inter-judge and intra-judge reliability.

Since CAPE – V is relatively new in its coming, not many studies have been reported in literature on the reliability of the scale. Thus, a need was felt to assess the reliability of the perceptual evaluation of voice using the CAPE – V scale.

Aims of the Study

The present study aimed at investigating the reliability of perceptual evaluations of voice disorders using CAPE-V scale (Consensus Auditory Perceptual Evaluation of Voice, 2002) for different tasks like phonation, sentences and spontaneous speech, in Indian context. The aims of the study were, to evaluate the

- Reliability across judges on different speech tasks (phonation, sentences and spontaneous speech) on categorical ratings (mild, moderate, severe) and on 100 mm visual analog scale.
- Correlation between categorical naming (mild, moderate, and severe) and numerical value (on VAS) assigned across three tasks.
- Determination of the most suitable voice sample (phonation/ sentences/ spontaneous speech) for perceptual evaluation of voice using CAPE-V.

CHAPTER - II

REVIEW OF LITERATURE

Vocal quality is the perceptual correlate of harmonics, resonance and symmetry of vocal fold vibrations. Characterization of voice quality is one of the key facets of perceptual assessment of voice and an integral aspect of voice evaluation.

Quality is traditionally defined 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" (ANSI Standards S.1.12.9,1960; Helmholtz,1885). Normal voice quality encompasses many dimensions related to physical, physiological, acoustic, emotional and social factors. Normal voice is a voice with no apparent pathology and no unusual voice characteristics or habits.

A voice disorder exists when a person's voice quality differs from those of similar age, sex, background and geographical background (Moore, 1971; Aronson, 1980). In other words, when the acoustic and aerodynamic properties of voice are so deviant that they draw attention to the speaker's voice, then disorder of voice is considered to be present.

Common categorization of voice disorders includes hoarse, harsh and breathy (Fairbanks, 1960).

Breathy Quality: When pulmonary air stream is passed through the open glottis without the laryngeal modification, voice is perceived to be breathy. Hypofunctional conditions of vocal folds results in breathy voice quality.

Harsh Quality: Voice is perceived as harsh when vocal folds are hyper adducted and result in excessive low pitch productions.

Hoarse quality: Hoarseness is a combination of both harsh and breathy voice qualities (Van Riper and Irwin, 1978). Hoarseness is a psychoacoustic term used in broader sense to mean any abnormal voice quality due to laryngeal pathology. To a lay man it implies a sudden change in voice quality or an unpleasant voice. According to Casper et al (1981) hoarseness is a deviation in the tonal quality of the voice resulting when the vocal cords vibrate in an aperiodic or haphazard manner.

Terms such as creaky, tense, husky, guttural, strained etc. are also used to describe the vocal quality.

Clinical Evaluation of Voice

"The treatment of patient with voice disorder depends upon the ability to assess initially the type and degree of voice impairment and also to monitor the patient's subsequent progress through treatment" (Kelmen, 1981). "Diagnosis is intended to define the parameters of the problem, determine the etiology and outline a logical course of action" (Emerick and Hatten, 1974). The ultimate aim of the studies on normality or abnormality of voice assessment and diagnosis of the voice disorder is to enforce procedure which will eventually bring back the voice of an individual to normal or optimum level. There are various methods for analysis of voice, developed by different researchers (Baken, 1987; Hirano, 1981). It can be done either subjectively or objectively.

Objective Evaluation of Voice

With the advances in technology, the perspective of assessment and treatment of voice disorders have changed. These include methods like electromyography, aerodynamic measurement, acoustic analysis, electroglottography, photoglottography etc. Even though these techniques have been promising, there have been problems with instrumentation, methodology and analysis (Aparna, 2000). Acoustic analysis of voice has been considered as basic tool in the investigation of voice disorders. It has advantages over other objective methods as its non - invasive and provides quick, convenient, repeatable and objective data (Hirano, 1981).

Perceptual Evaluation of Voice

Human ears have the ability to identify and recognize speaker's voice. Well trained voice clinicians are often able to determine the causative pathologies on the basis of psycho acoustic impression of voice (Hirano, 1975). Hammarberg and Gauffin (1986) opined that the perceptual evaluation by well trained listeners is a reliable and reproducible and can be used for systematic evaluation purpose, if handled with precaution. They also reported that voice quality can be perceived more precisely, if professional terminology is given to the listener.

Clients seek treatment for voice disorders because as they do not sound normal and they often decide on whether the treatment has been successful based on whether they sound better or not. Hence speech clinicians' use and value perceptual measures of voice and speech far more than the instrumental measures (Gerratt, Till, Rosenbeck, Wertz & Boysen, 1991). Further the listeners' judgments are usually the standard against which other measures of voice i.e. acoustic, aerodynamic etc are evaluated (Wendler, Doherty & Hollien, 1980). Because voice is fundamentally perceptual in nature (Kreiman et al, 1993), perceptual evaluation of voice remains an important assessment tool for the assessment of voice in the voice clinic. Therefore it has maintained its place next to these more technical and objective evaluations. The power of the perceptual scale lies in their accessibility for any clinician and researcher involved in the study of voice (Wuyts et al, 1999).

Often the perceptual measures act as a standard against which acoustic measures are validated or compared. Researchers proposing objective voice measures often demonstrate their measure's utility by reporting a correlation between the measures and ratings of perceived vocal quality (Kojima, Gould, Lambiase, Isshiki, 1980; Fukazawa & El- Assuooty, 1988; Ladefoged, Maddieson, & Jackson, 1988; Klattt & Klatt, 1990). Thus listener's judgment is essential, both for clinical consideration and criterion validation of instrumental voice measures.

Traditionally, the clinicians use visual inspection of larynx and subjective perceptual evaluation of voice quality to diagnose the laryngeal pathology (Yanagihara, 1967). Subjective perceptual evaluations have had some degree of success in separating normal and pathological voice. However, it has its own limitation on test retest and inter – judge reliability (Yanagihara, 1967). Perceptual judgment of voice is an integrated process of listening to and describing a particular voice. The clinician needs intensive training in voice dimensions that identify pathology most effectively. Rating voice quality perceptually is universally acknowledged as difficult task and one requires considerable experience in perceptual judgments.

Eadie and Baylor (2006) carried out a study to determine inter judge and intra judge reliability changes in inexperienced listener's after two hours of listening training in rating normal and dysphonic voice samples. Thirty adults with dysphonia and six normal speakers' speech samples were audio recorded. Samples included 21 test stimuli and 15 training stimuli of both sustained vowel and connected speech. Sixteen inexperienced judges were rated all the samples for overall severity, roughness and breathiness on visual analog scale. This formed the pre training baseline ratings. Then these listeners were trained using 15 anchor voice samples and 15 training stimuli in which they were also provided with definitions of rating dimensions, accuracy feedback and anchor samples. Post training and pre training scores were compared and analyzed. Results indicated that intra judge reliability was least variable for judgment of overall severity but improved with training. Listeners' judgment of roughness and breathiness in vowel was least reliable at baseline but improved significantly after training. Thus, the study has implications for developing training programs in perceptual evaluation of voice in order to increase reliability of judgment.

Correlation across Perceptual and Acoustic Measures

Many studies have been done to find the correlation between the perceptual and acoustic measures. Most of these studies reveal that there is a good correlation between acoustic parameters studied and amount of hoarseness perceived.

i. "It is also known that perturbations with large magnitude give rise to perception of rough vocal quality" (Wendahl 1966; Coleman,1971).

ii. This connection between perceived roughness and waveform irregularities exists independent of whether the irregularities are caused by amplitude perturbation or frequency perturbations (Wendahl, 1966).

iii. Investigation of the acoustic waves of synthesized complex sounds and human phonations have revealed that rapid, random variations in the periods and the amplitudes of successive cycles are associated with perceived roughness of the signal. (Lively and Emanuel, 1970).

iv. Askenfelt and Hammerberg (1986) compared the perturbation measure with regard to acoustic perceptual correlation and their ability to discriminate between normal and pathological voice status and concluded that the standard deviation of the distribution of relative frequency differences was most useful acoustic measure for clinical application.

v. Huang (1995 b) based on an investigation between perceptual judgments and acoustic parameters made the following conclusions,

- Perception of hoarse voice quality should be considered as combination of breathiness and harshness.
- Vocal jitter appears to be related primarily to harsh vocal quality.
- Shimmer appears to be primary influence on hoarse voice quality.
- The spectral tilt of the glottal source is significantly related to perceived breathiness.

Reliability and Validity in Perceptual Evaluation of Voice

Reliability refers to the degree to which test scores are free from errors of measurement (American Psychological Association, 1985) and the construct of reliability then has to be defined as relatively free of random errors of measurement (Crocker and Algina, 1986). Random errors of measurement affect the score of a person because of purely chance happenings. These types of errors are not consistent and will smooth down over time if a test is repeated several times. Sources of such random errors may include "guessing, distraction in the testing situation, administration errors, content sampling, scoring errors, and fluctuations in the individual examinee's state" (Crocker and Algina, 1986).

By contrast, systematic measurement errors are those that consistently affect the score of a person because of particular characteristic of the person or the test that has nothing to do with the construct being measured (e.g., a rater who always uses the scale in the same manner). Such tendencies are supposed to persist across repeated ratings with the same instrument and affect the score of the rater in a consistent manner. Even if both error types are of concern in score interpretation, systematic measurement errors do not result in inconsistent measurement. Still, they may lead to low validity in the ratings and thereby reduce the utility. Random error, however, may reduce both the reliability and the validity and thereby the utility of the ratings (Crocker et al, 1986).

Theoretically, the observed score can be considered a function of three components (Kleven, 1995) - ‘The valid score + the systematic errors of measurement + the random errors of measurement’. Wherein, valid score signifies the score which was obtained out of valid/intended observation by the observer. Systematic errors of measurement are those which are caused by contamination of other constructs than the construct in relation to which we interpret the results. Random errors of measurement are errors which are due to chance happenings.

Although reliability is an important attribute, the most critical property of any test is its validity. Validity refers to what the measurement actually measures and how useful the measurement is to researchers. Reliability is a condition for validity and it places an upper limit on the validity of a test. Unreliable measures will allow tests to show little, if any, validity. Reliability is a necessary, but not a sufficient, prerequisite for the test to have validity (Crocker et al, 1986).

The standard rules do not exist for what constitutes a minimally accepted value for the reliability coefficient. According to Crocker et al (1986) many standardized achievement test manuals report coefficients ranging in the 0.80s and 0.90s (It is usual to refer to low, moderate, and high reliability). The reliability of the ratings of voice quality by listeners is a central issue in voice research. Studies show a large variation in both inter-judge and intra-judge reliability (Bassich & Ludlow, 1986; Hammerberg et al 1986; Kreiman et al 1993). A poorly operationalized construct of a voice characteristic, or whether a characteristic lacks perceptual reality will make the listeners unable to rate it consistently. According to Krieman et al (1993), a rated vocal characteristic may be systematically related to the listener. These factors include: listener's experience with voices, their perceptual habits and biases (Kreiman, Gerratt & Precoda, 1990; Kreiman, Gerratt, Precoda & Berke, 1992) and seemingly an overall sensitivity to the quality being judged.

These factors change slowly over time and thus hypothetically affect inter-judge reliability more than intra-judge reliability. Additional factors related to listeners include fatigue, attention lapses and mistakes. These error factors should affect both interrater and intra-rater reliability (Kreiman et al 1993). Systematic interactions among listener and task factors may also occur. Listener sensitivity may interact with scale resolution, which adds noise to the data or results in information loss.

Factors related to the task of rating that affect measurement of voice quality also exist. The two types of error variances as stated by Bele (2005) are as follows,

- First, the listeners may differ in their rating of the voices (which means that, they to a different degree agree on the relative rating of voices).

- Second, they use the scales differently (i.e., a difference exists between the listeners' in the rating of the same voice).

If all voices are rated by the same set of listeners, the differences in means between the listeners, as a source of error need not be taken into account in the study because, these differences will not influence the differences between the voices (i.e., they will influence all voices to an equal extent). This extra source of error, that some raters were stricter than others, will therefore not influence the data. The source of error concerned here is what influences the rating of the voices in relation to each other. The degree of rater agreement also includes the aspect of validity (Bele, 2005).

Krieman et al (1993) proposed a descriptive framework of specifying several sources of variability in voice ratings. They stated that when listeners rate the voice on one quality dimension (e.g. roughness) they compare and match the presented stimuli to an internal standard or scale. These internal standards are developed out of a listener's experiences with voices and are maintained in memory; accordingly they differ from listener to listener. Also, these internal standards are highly unstable and may be influenced by internal factors such as lapses in memory and attention, and external variables, such as acoustic contexts (Kreiman et al 1992) and listening task.

Further, the variability can be associated with the overall sensitivity of the listeners to the voice characteristic being rated and, and/or in response bias (Kreiman et al, 1990; Kreiman et al, 1992). Additionally, effects related to the specific rating task contribute to rating variability. These include the context effects and the number of points on the rating scale (Rossi, Pavlovic & Espesser, 1990; Krieman et al, 1993). Interactions between task and the listener factors may also occur.

Descriptive framework given by Krieman et al (1993) on factors involved in mapping an acoustic signal onto voice quality rating can be depicted as follows:

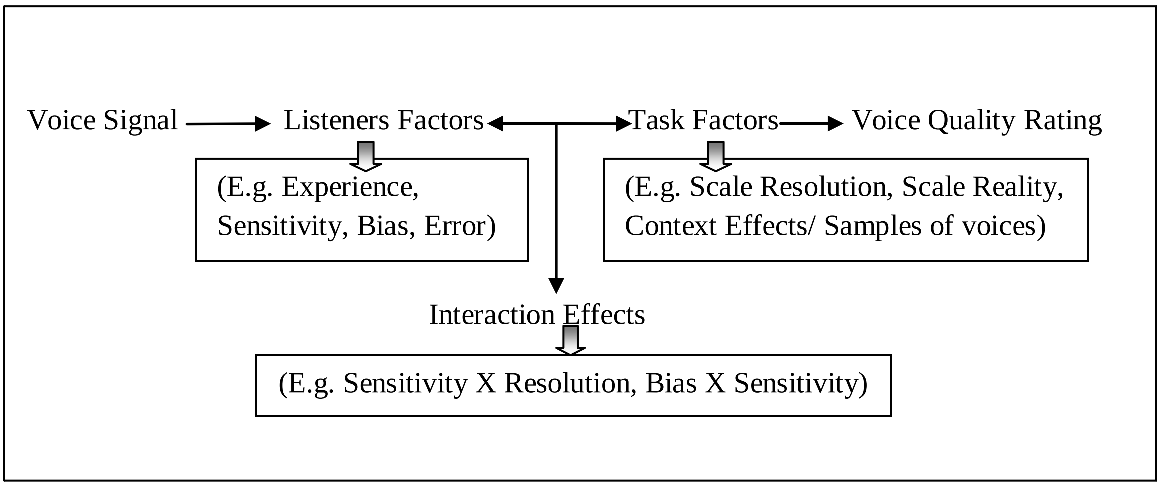

Figure 1: Descriptive framework for perceptual evaluation of voice (Krieman et al, 1993)

Listener's factors included factors like their experience, their individual perceptual habits and biases (Kreiman et al 1990, Kreiman et al, 1992) and overall sensitivity to the quality being judged. These factors change relatively slowly over time and so hypothetically affect inter- judge reliability more than intra- judge reliability. Other factors include listener's fatigue, attention lapses and mistakes. These 'error' terms will affect both inter and intra – judge reliability (Kreiman et al, 1993).

There are certain factors related to the task, like if the quality to be rated lacks perceptual reality, listeners will not be able to rate it consistently. Also, perceptual context can cause systematic drift in rating presumably because of its effect of altering a listener's internal standard. These factors can affect rating within a given session and thus affect both inter and intra – judge reliability. Several systematic interactions among listener and task may also occur. Listener's sensitivity may interact with scale resolution and mismatches may add noise to the data or result in information loss. Also, listener's bias may interact with 'scale specificity'. If the quality being rated is multidimensional in nature, listeners may selectively focus on one dimension or other, reducing apparent agreement levels (Kreiman et al, 1993). Kreiman et al (1992) demonstrated that listener's differential attention to various aspects of each quality is a significant source of inter – judge unreliability in voice quality ratings.

For perceptual ratings to be meaningful, listeners must use scales consistently. A given rater must rate a voice sample the same way every time he/she hears it. Additionally for ratings to be clinically useful, inter-rater agreement must be high. Each rater who hears a voice sample must rate it similarly. Thus, reliability of such judgment is a central issue in the study of voice disorders and voice quality (Kreiman et al, 1993).

Scales for Perceptual Evaluation of Voice

In literature, a variety of assessment tools have been given which have been widely used for perceptual evaluation of voice. Some of them are the formal and standardized tools for assessment of voice quality. Such tools enable the speech language pathologist to qualify the problem as well as describing the voice in an organized profile.

The formal assessment attempts to minimize the confusion that can arise from the plethora of synonyms and ambiguous descriptors. The perceptual evaluation does not produce hard data and no single protocol is universal mean that certain problems are intrinsic to the process.

Types of Rating Scales

Perceived voice quality can be measured using a variety of tasks. There are different types of scales and measures for rating voice. In general, for a scale to be a regarded as a valuable tool for clinical circumstances it should be robust, consistent and it has to have high inter-judge agreement. Hence these constraints imply certain points like, (Wuyts, De Bodt & Van de Heyning, 1999)

- ✓ Small changes between the voice samples should be reflected by small changes on the scale.
- ✓ A given judge must rate a voice sample the same way every time he or she evaluates it.
- ✓ Thirdly, different judges should assign similar ratings to the same voice sample.

Some of the most commonly used types are:

Categorical Rating: It involves assigning speech or voice samples to discreet, unordered categories (e.g. breathy, rough).

Equal Appearing Interval (EAI): Such scales require listeners to assign the numbers between 1 to n to a voice sample, where ‘n’ is the number of points in the scale. Points on

EAI are assumed to be equidistant, so measurements are generally treated as interval level and parametric statistics applied.

Direct Magnitude Estimation (DME): Listeners assign the number to voice sample to indicate the extent to which a voice possesses a given characteristic. The range of possible numbers is generally not restricted.

Paired Comparison (PC): The listeners compare the two stimuli. They may judge the extent of difference on some dimension, similarity/difference, relative roughness and so on.

Visual Analog Scale (VAS): These scales have undifferentiated lines, often 100mm long. Listeners rate voices on these scales by marking a mark on the line to indicate the extent to which a voice possesses a given characteristic.

The most common form of VAS consists of 100 mm continuous lines, with extremes corresponding to nonexistent and to extremely high occurrence of vocal characteristics. For some characteristics, these labels are not adequate, and a line of 200 mm has been proved to be more appropriate length. These scales are bipolar and go from one extreme via a neutral reference point to the opposite extreme (McAllister, 1997; Sederholm, 1996).

Some of the most popularly used scales of perceptual voice evaluations are:

- The Voice Profile, (Wilson,1987)
- The Voice Profile Analysis Protocol, (Laver,1980)
- The GRABS Scale, (Committee of Phonatory Function tests of Japan Society and Logopedics and Phoniatrics, Hirano, 1981)
- Consensus Auditory Perceptual Evaluation of Voice (CAPE- V), ASHA (2002)

The Voice Profile

It was described by Wilson (1970). It is a simple method of documenting the abnormal voice in adults and children. It is an eight point rating scale with ‘1’ indicating that the voice problem is barely perceptible while ‘7’ indicates that it significantly interferes with communication. It evaluates voice on various parameters like laryngeal qualities, resonation qualities, vocal range, loudness, rate etc.

The Voice Profile Analysis Protocol

It was developed by Laver (1980) in which he included laryngeal and supralaryngeal aspects. He charted the positions of different articulators 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. This presents a formidable list of items and large varieties of vocal features to be considered during assessment of dysphonic speakers.

The GRABS Scale

It was developed by committee of phonatory function tests of Japan Society and Logopedics and phoniatrics (Hirano, 1981). It evaluates voice on five parameters namely Grade- degree of voice abnormality, Roughness, Asthenia, Breathiness and Strain. Each parameter can be rated on 4 point rating scale with '0' representing normal voice and '3' representing extreme voice abnormality.

The Buffalo III Voice Profile

It was developed by Wilson (1987) and is one of the most commonly used scales. It rates the laryngeal tone, pitch, nasality, oral resonance, breath supply, muscles, voice abuse, rate speech anxiety, speech intelligibility and overall voice proficiency on a five point scale, with appropriate descriptive terms listed for marking with each category.

Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V)

It was developed by Consensus meeting sponsored by ASHA held at Pittsburgh (2002). It has additional feature of visual analog scale (100 mm scale) where the judge is required to mark the voice quality on the scale with left end of scale indicating no abnormality and the right end indicating severe voice problem. The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) was developed as a tool for clinical auditory-perceptual assessment of voice. The primary purpose of CAPE- V is to describe the severity of auditory-perceptual attributes of a voice problem, in a way that can be communicated among other clinicians. The secondary purpose is to contribute to hypotheses regarding the anatomic and physiological bases of voice problems and to evaluate the need for additional testing. It was recommended that CAPE-V is *not* intended for use as the only means of determining the nature of the voice disorder and is not to be used to the exclusion of other tests of vocal function. Finally, the authors report that it is not expected to demonstrate a 1:1 relation to results from other tests of vocal function.

Design Considerations

The consensus was that the clinical evaluation of auditory-perceptual characteristics of voice should be derived from a tool with the following attributes:

- Perceptual dimensions should reflect a minimal set of clinically meaningful, perceptual voice parameters, identified by a group of expert clinicians;
- Procedures and results should be obtainable expediently;
- Procedures and results should be applicable to a broad range of vocal pathologies and clinical settings;
- Ratings ultimately should be demonstrated to optimize reliability within and across clinicians, and
- Ultimately, exemplars should be available for training.

Description of the Tool

The CAPE-V indicates six salient perceptual vocal attributes. The preferred attributes were regarded as most identifiable and important attributes for the analysis of voice disorders. Also they reported them to be most commonly used and easily comprehendible ones. They are as follows

- Roughness
- Breathiness
- Strain
- Pitch
- Loudness
- Overall Severity

Each attribute is displayed and represented by a 100- millimeter line forming a visual analog scale (VAS). The clinician indicates the degree of perceived deviance from normal voice for each parameter on this scale, using a tic mark. Judgments may be assisted by referring to general regions indicated below each scale on the CAPE-V: “MI” refers to "mildly deviant," “MO” refers to “moderately deviant,” and “SE” refers to "severely deviant." A key issue is that the regions indicate gradations in severity, rather than discrete points. The clinician may place tick marks at any location along the line.” Ratings are based on the clinician’s direct observations of the patient’s performance during the evaluation, rather than patient report or other sources.

Rationale for the Quality Features

Despite much debate over the description, validity and independence of any list of voice quality features, these six have consistently appeared in both national and international voice literature for decades (Fairbanks, 1960; Hirano, 1981; Wilson, 1987; De Bodt & Wuyts, 1996). Thus, the rationale for including these six voice quality features is the belief that both clinicians and researchers find these attributes meaningful. Another common descriptor, “hoarse,” was excluded from the list of terms, because the authors agreed with Fairbanks (1960) that “hoarseness” is perceived by many as a combination of “roughness” and “breathiness.”

The CAPE-V form also includes two unlabelled scales. These allow the clinician to document other salient perceptual features of a patient’s voice, for example, degree of nasality, spasm, tremor, intermittent aphonia, falsetto, glottal fry, weakness, or other aspects that may best characterize individual features of a patient’s voice quality.

Scale

A 100 mm line scale with unlabelled anchors, commonly known as a visual analog scale, is used to assess each of the six quality features. The left most portion of the scale reflects normal voice (in the case of judging Severity, Pitch, or Loudness) or none of the quality being judged (in the case of Roughness, Breathiness, and Strain). The right end of the scale is to reflect the listener’s judgment of the most extreme example of deviance. Measurement from the left end of the scale to each tick mark, in millimeters, is denoted on the blank to the far right of the scale (___/100).

Rationale for the Scale

Marks recommends that auditory-perceptual judgments of voice quality be made on a visual analog scale (or set of scales), using open-ended anchor points at either end as a way to inhibit end effects of the scale. Visual analog scales are easy for raters to use and appear to have become more common place in voice research in the past two decades.

Verbal Descriptor Degree of Deviance

While the primary measurement index is an interval scale provided by the 100 millimeter visual analog line, the CAPE-V also includes the ordinal ratings of “mild,” “moderate,” and “severe,” printed below the measurement line, to serve as a supplemental severity indicator. These qualitative terms are positioned in a non-equidistant fashion and reflect the range of voice severity using terminology more familiar to clinicians than the discrete intervals measured on the 100 millimeter visual analog scale.

Additional CAPE-V Elements

A nominal rating judgment allows the clinician to classify the consistency or intermittent presence of the voice quality feature within and across evaluation tasks. Sections devoted to resonance or other features supplement the CAPE-V protocol by allowing other salient descriptors to document a patient’s voice quality. This flexibility is needed to capture the spectrum of voice disorders and associated conditions or features. The list of terms provided on the form is not inclusive, meant only as examples of specific features that may help describe auditory-perceptual attributes.

Rating Procedure

The CAPE-V judgments are intended to reflect the clinician's direct observations of the patient's performance during the evaluation and should not take into account patient report or other sources. Standard audio recording procedures should be used, such as recording in a quiet environment and using a standard mouth-to-microphone distance with the highest possible sampling rate for digital conversion. If a patient returns following an initial assessment, the clinician may compare the initial voice sample and CAPE-V ratings directly to any subsequent recordings, to optimize the internal consistency or reliability of repeated sequential ratings, particularly for assessing treatment outcomes. As always, clinicians are encouraged to minimize bias in all ratings.

Concurrent Validity and the CAPE-V

Berg and Eden (2003) compared aspects of the CAPE-V to the Stockholm Voice Evaluation Approach (SVEA) on patients with three different voice pathologies. This study involved a translation of the CAPE-V into Swedish. The authors determined that intra- and inter-rater reliability was acceptably high in both protocols, and no obvious differences were found between the two approaches in terms of listener variability. Both protocols were able to separate the three disorders from each other and showed significant pre-post treatment changes in voice quality.

Karnell, Melton, Childes, Coleman, Dailey, & Hoffman,(2006) published a preliminary report comparing the reliability of clinician based auditory-perceptual judgments using the CAPE-V to those made with the GRABS voice rating scheme

(Hirano, 1981) and two other quality of life scales (Voice Related quality of Life or V-RQOL and Iowa Patient's voice Index or IPVI). These protocols (CAPE-V and GRABS) were then compared after use in voice assessment of forty males and sixty one females by certified speech language pathologists. They found comparable estimates of inter-judge reliability for the two scales, both at high levels but suggested that the CAPE-V may offer "more sensitivity to small differences within and among patients than the GRABS scale".

However, three important factors discriminate the CAPE-V from the GRBAS scale. First, the GRABS has no published, standardized protocol to follow in English. Hirano's (1981) reference most often cited for the GRABS provides no guidelines for clinical administration, speech material, or rating calibration. In contrast, the CAPE-V includes a specific protocol that designates the tasks, procedures, and scaling routine, toward the larger goal of improving the consistency of clinical assessment from one clinician to another, without excessive demands on clinician time or learning. Second, the CAPE-V provides interval scale measures of voice quality by incorporating millimeter measures on visual analog scales. Such scales are shown to better accommodate the task of measurement of multidimensional features, such as vocal quality (Chan & Yiu, 2002; Gerratt et al., 1993). The GRABS scale, however, only allows ordinal judgments on a four-point scale of normal (1), mild (2), moderate (3), or severe (4) which severely limits its application to research design and statistical analysis. Finally, the CAPE-V attempts to document more voice quality features than the GRBAS, across more speech tasks, while allowing room for supplemental feature scales and comment areas.

Study done by Wuyts, De Bodt, Van de Heyning (1999) aimed at finding and comparing the reliability of visual analog scale and an ordinal scale by perceptual voice evaluation of fourteen pathological voices by twenty nine listeners using GRABS scale. Agreement was found to be higher with original 4 point scale than with the visual analog scale version for the scale items G, R, A, B and S. A tendency was noted to rate the voice on the middle of the visual analog scale and with increased freedom of judgment the inter judge agreement decreases considerably but it also seemed that finer judgment of voice quality is possible with VA scale. But, the authors reported that, it is logical to assume that the listener perceives the vocal characteristics along a continuum, rather than quantifying perception by intervals. The VAS appears to be advantageous for comparison with absolute acoustic measurements as it offers more detailed information.

Similar study done by Yu, Revis, Wuyts, Zanaret (2002) aimed at determining the most suitable scale for perceptual assessment of voice that is a visual analog scale or ordinal scale as a gold standard for validating the objective analysis protocols. The authors took seventy four female voices in which sixty eight females were diagnosed as having voice disorder and six served as controls. Panel of four experienced judges were asked to rate the voices according to Grade component of the GRABS scale. Two scales were used. One with conventional ordinal scale and second one with modified visual analog scale. Objective measurements included acoustic, aerodynamic and physiologic parameters. Instrumental measures were compared with the results obtained from perceptual analysis. Results demonstrated that correlation between perceptual and objective voice judgment is better using modified visual analog scale (r = 0.88) than a conventional ordinal scale (r = 0.64).

Karnell, Melton, Childes, Coleman, Dailey and Hoffman (2007) examined the reliability of documenting voice quality by clinicians and compared the method for documenting patient's perception of voice quality. Two clinicians based protocols i.e., GRABS and CAPE-V were evaluated. These protocols were then compared after use in voice assessment of forty males and sixty one females by certified speech language pathologists. In addition two patient based scales (Voice Related quality of Life or V-RQOL and Iowa Patient's Voice Index or IPVI) obtained from the same patients were compared with each other and with the clinician based scales. Reliability of clinicians rating of overall severity of dysphonia using GRABS and CAPE-V scales was very good (r > 0.80). Also the authors reported that CAPE-V system appeared to be more sensitive to small differences within and among the patients than the GRABS system. Overall there was a week agreement between the patient - based scales and the clinician - based scales.

A nonrandomized prospective study done by Kelchner, Brehm, Weinrich, Middendorf , Dealarcon , Levin and Elluru (2009) aimed to quantify the inter- and intra-rater reliability of experienced speech-language pathologist's perceptual ratings of voice in pediatric patient's post-laryngotracheal reconstruction (LTR). Using the sentence portion of the Consensus Auditory Perceptual Evaluation-Voice (CAPE-V) rating scale, three experienced speech-language pathologists independently rated randomized voice samples of fifty participants ages 4 – 20 years, who had acquired or congenital airway conditions requiring at least one LTR on the six salient perceptual vocal attributes. Estimates of interrater reliability were strongest for perceptual ratings of breathiness (intraclass 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. The parameter of strain, when rated by auditory sample alone and apart from the clinical context, was difficult to rate.

Effect of Speaking Task/Sample on Reliability of Perceptual Assessment

Less number of published information is available on the effect of the speech/voice sample on ratings of dysphonia. Most auditory-perceptual studies of voice have focused on either isolated vowels (usually /a/) or connected speech. The application of sustained vowels has been reported more frequently than the application of connected speech, for several reasons.

- First, vowels are easily elicited by the clinician.
- Second, vowels are easily produced by the client. As such, vowels may be more controlled and standardized than connected speech (Krom, 1994).
- Because sustained vowels are relatively stable and less affected by articulation and dialectal influences, the listener may focus more on the sound source.

However, others argue that connected speech is more representative of a person's daily voice (Hammerberg, 1980).

De Krom (1994) conducted a perceptual experiment in which six listeners used the GRABS scale to rate voice fragments from seventy eight dysphonic speakers. Four different types of stimuli were presented to each listener: one based on connected speech fragments and the other three on segments of a sustained vowel. Analyses focused on the consistency and reliability of ratings and results indicated that stimulus type had virtually no effect on either intra-rater or inter-rater reliability. When determined as a function of the overall degree of severity of a voice, the reliability of ratings for the breathiness and roughness parameters was slightly higher for vowel stimuli than for connected speech.

Revis, Giovanni, Wuyts and Triglia, (1999) examined agreement across seven experienced listeners, using the GRBAS scale. The listeners had to rate dysphonia in sixty adult speakers and twenty normals. Two fold purpose of this study was to validate the pertinence of sustained vowel for perceptual analysis in native speakers of French and second to test whether the use of only the sustained portion of any vowel (phonation) will cause underestimation of dysphonia. Three different sample materials were obtained from each participant i.e., connected speech, complete sustained vowel and only the stable portion of the sustained vowel (stabilized sustained vowel). Reliability across judges was measured as a percentage of agreeing judgments on the same subject. No difference in reliability (consistency) was observed using the three samples. Judgments on stabilized sustained vowel were confirmed as less severe than judgments on connected speech.

Judgments on complete sustained vowel were similar to that in connected speech. Hence, the results were similar to De Krom (1994) in that speaking sample did not affect reliability of ratings for most listeners.

Munoz, Mendoza, Fresneda, Carballo and Ramirez, (2002) examined the agreement and reliability of ratings made by thirty four expert listeners using the Buffalo Voice Profile System. A sustained vowel and a short utterance of connected speech were presented to each listener. Results revealed that for the evaluation of the sustained vowel, interrater agreement was moderate for judgments of breathiness, hyponasal resonance, and overall severity of dysphonia and for connected speech agreement was moderate for most voice qualities.

Zraick, Wendel, and Olinde (2005) investigated the effect of speaking task on auditory perceptual judgment of the severity of dysphonia. Three speech-language pathologists experienced in evaluating of disordered voices rated twenty nine recorded speakers, each of whom produced speech elicited via the same three tasks: sustained vowel /a/, oral reading of a standard passage, and connected speech describing a standard picture. Stimuli were presented in sound field and raters used direct magnitude estimation (DME) with a visual analog (VA) scale. A repeated measures analysis of variance (ANOVA) revealed a statistically significant ($P < 0.05$) effect of speaking task, with post hoc analyses that indicate a statistically significant difference between ratings for the sustained vowel versus connected speech elicited via picture description ($P < 0.05$). Between oral reading and picture description or between oral reading and the sustained vowel, no statistically significant difference in ratings was found. The ANOVA also revealed a statistically significant difference among raters ($P < 0.001$), but no statistically significant task by rater interaction. No statistical difference in rating of overall severity made from two types of connected speech samples was attributed to the fact that perceptual rating for the other vocal parameters, such as pitch, loudness, vary as function of the type of connected speech sample.

Methods to Improve the Perceptual Evaluation of Voice

Auditory perception is central to evaluating voice quality and requires strict methods to ensure consistency of judgments (Fex, 1992; Revis, Giovanni, Wuyts, Triligia, 1999). In review of literature on voice perception, Kreiman et al (1993) identified three areas of methodological concern:

i. Rating protocols
ii. Rater's performance
iii. The nature of speech/ voice sample

Listener's performance has been the focus of numerous studies and several conclusions can be drawn. First, it is generally agreed that inter-judge reliability is often more of a concern than intra-rater reliability. This concern is largely caused by two factors: (1) the variable internal standards developed by persons (Gerratt et al, 1993; Kreiman et al, 1993) that are easily affected by memory and acoustic context, (Gerratt et al, 1993; Kreiman et al, 1993; Kreiman et al, 2000) and (2) measurement error (Srivastava, 2004). In regard to the former, external standards or anchors have been proposed and studies with natural or synthesized voice samples as anchors have shown that they improve rater reliability (Gerratt et al, 1993; Kreiman et al, 1993). In regard to the latter, potential sources of measurement error must be controlled.

Second, experienced listeners are more reliable raters than inexperienced listener, or less-trained listeners (Hammerberg et al 1980; Gerratt et at 1993; Kreiman et al, 1993; De Bodt & Wuyts, 1996; Kreiman et al, 2000). Providing training to listeners is reported to improve the reliability of their ratings (Bassich et al, 1986; Shewell, 1998).

Finally, listeners' performance seems to be a function of the vocal parameter to be rated. It has been shown, for example, that when listeners use the GRBAS scale, they are most consistent in rating the Grade parameter compared with the other voice qualities (Revis et al, 1999).

The above review focuses on the different perceptual rating scales and the methodological concerns. The present study is focused on the reliability of CAPE-V. As this scale is introduced in 2002 the published studies using this scale are scanty. Hence the present study is aimed to study the reliability on using the CAPE – V across different speech tasks.

CHAPTER – III

METHOD

The study is aimed to evaluate the efficacy of CAPE - V rating scale for the reliability of perceptual evaluation of hoarseness of voice in an Indian context.

Subjects

Twenty one participants were included in the study who were diagnosed as having hoarse voice by a qualified speech language pathologist using acoustic analysis and perceptual assessment (without any use of standardized scale). Total fifteen native Kannada speakers and six English speakers were considered as subjects.

Inclusion Criteria

- Males with hoarse voice disorder were included in the study.
- All the participants were in the mean age range of 25 - 45 years.
- The participants were native speakers of Kannada or English.
- Subjects diagnosed as having hoarseness of voice were included.
- Based on informal evaluation, subjects with normal hearing, normal oro- motor structure and functions were included in the study

Exclusion Criteria

- Subjects, who had voice problems associated with neurological disorder were excluded from the study.
- The subjects with other types of voice problem like breathiness or harsh voice quality were excluded from the study.

Instrument and Environment

The audio recordings of the speech samples were done in a sound treated room, free from all distractions and minimum ambient noise. The room was well ventilated and well lighted. Recordings were done using Cool Edit (Version II) software in a Compaq Lap top with an 'hp Microphone'. The microphone was placed at a distance of 6 cm and slightly to the side of the subject's mouth to minimize breathing noise. Gain was adjusted to avoid saturation and ensure optimal use of recording dynamics. Subjects were instructed to read standardized sentences at a comfortable pitch and volumes as naturally as possible. The complete sample was audio recorded using the Cool Edit software.

Tool

Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) rating scale given by ASHA, 2002 was used in the study for rating the severity of voice samples. CAPE-V rates the voice based on six parameters (attributes) namely, Roughness, Breathiness, Strain, Pitch, Loudness and Overall Quality of voice.

CAPE-V displays each attribute accompanied by a 100- millimeter line forming a visual analog scale (VAS) where the rater (judge) can rate the degree of perceived deviance from normal for each parameter. For each dimension, scalar extremes are unlabeled. Judgments may be assisted by referring to general regions indicated below each scale on the CAPE-V:

“**MI**” refers to "Mildly deviant," “**MO**” refers to “Moderately deviant” and “**SE**” refers to "Severely deviant.” The regions indicate gradations in severity, rather than discrete points. Hence, it gives the rater freedom, to rate the voice in between mild, moderate or severe. “**C**” represents “Consistent” and “**I**” represent "Intermittent" presence of a particular voice attribute. A judgment of “consistent” indicates that the attribute was continuously present throughout the tasks. A judgment of “intermittent” indicated that the attribute occurred inconsistently within *or* across tasks. Ratings are based on the clinician’s direct observations of the patient’s audio recorded samples, rather than patient report or other sources.

Following Figure 2 is the complete CAPE – V form as given by ASHA, 2002.

Consensus Auditory – Perceptual Evaluation of Voice (CAPE-V)

Name:

Date:

The following parameters of voice quality will be rated on completion of the following tasks.

1. Sustained vowels /a/ and /i/ for 3-5 second duration each.
2. Sentence production
   a. The blue spot is on the key again.
   b. How hard did he hit him?
   c. We were away a year ago.
   d. We eat eggs every Easter.
   e. My mama makes lemon muffins.
   f. Peter will keep at the peck.
3. Spontaneous speech in response to "Tell me about your voice problem" or "tell me how your voice is functioning".

Legend

C = consistent, I = Inconsistent, MI = Mildly deviant, MO = Moderately deviant, SE = Severely deviant

Overall severity	MI MO SE	C		___ / 100
Roughness	MI MO SE	C	I	___ / 100
Breathiness	MI MO SE	C	I	___ / 100
Strain	MI MO SE	C	I	___ / 100
Pitch	Indicate nature of abnormality ________ MI MO SE	C	I	___ / 100
Loudness	Indicate nature of abnormality________ MI MO SE	C	I	___ / 100
	MI MO SE	C	I	___ / 100
	MI MO SE	C	I	___ / 100

Comments about resonance: Normal / other (provide description) ______________________

Additional features (for example, diplophonia, fry, falsetto, asthenia, aphonia, pitch instability, tremor, wet / gurgly, or other relevant terms

Figure 2: CAPE – V FORM

Procedure

The subjects were enrolled for the study after the completion of assessment. Informed consent was obtained from each subject.

Data Collection/ Sample Recording

The data (voice sample) was collected by asking the subjects to carry the following three tasks-

Task I - Maximum Phonation Duration: The subjects were asked to phonate vowels /a/, /i/ and /u/ after a deep inhalation, as long as they could in their comfortable pitch.

Task II - Reading Sentences (Repetition task if the subject was illiterate).

Three oral and three nasal, Kannada or English sentences were given to the participants who were native speakers of respective languages. Kannada sentences were taken from stimuli developed for nasometer assessment by Jaya Kumar, 2004 (Appendix II). English oral sentences were taken directly from the original CAPE-V rating scale given by ASHA, 2002. English nasal sentences were taken from the stimuli developed by Kummer (2008) (Appendix III).

Task III- Spontaneous Speech – The subjects were asked to describe about their voice problem at a comfortable loudness level. The participants were instructed to speak in response to questions like, “Describe your voice problem”? or “What problems are you facing because of the voice problem”?

The three tasks were recorded with an interval of twenty seconds between each task. The samples were recorded on Cool Edit (version II). After the speech recordings the samples from different subjects were randomized. As, there were twenty one participants, twenty one sets were made which had different speech tasks recording of different participants. This was done to avoid biasing by judges while rating the voice. These randomized sets were then copied to compact discs, which were given to the six judges of the study. These samples along with the rating sheets (Appendix I) were given to the judges for rating the voice quality based on CAPE- V (2002)

Judges

Six speech and language pathologists, who have an experience of 3-4 years in assessment and management of voice disorders, were selected as judges for the study. The randomized samples were played through the compact disc (CD) and they were asked to rate the samples based on CAPE-V rating scale. The analysis sheets (Appendix I) were given to the judges for perceptual judgment.

The complete perceptual analysis of all the twenty one samples by each judge was done in two sessions. Six samples were played in the first session, then followed by a week other fifteen samples were given to the judges for perceptual ratings on CAPE-V.

The rating sheets from judges (Appendix IV) were then rearranged based on randomization which was previously done. Each subject's speech samples were then grouped together. The obtained data was tabulated on to SPSS software (version -16). Appropriate statistical measures were applied to the get the reliability co-efficient between the parameters and judges.

Statistical Analysis

The following statistical analysis were done-

Cronbach's alpha coefficient was computed to obtain reliability between judges in the entire task and between all the voice parameters. Followed by *Spearman's coefficient* computation, which was carried out in order to obtain correlation between the markings given by judges on two scales of CAPE-V namely, ordinal scale (mild, mild- moderate, moderate, moderate- severe, severe) and numerical rating on 100 mm visual analog scale.

Ethical Concerns

- There was no specific ethical concern as none of the subjects had to undergo any invasive or experimental procedure.
- The participants selected for the study were clients with voice disorder voluntarily visiting the clinic for treatment.
- Participants selected did not undergo any additional test procedures apart from recording their voice samples.
- They were not required to visit repeatedly, the voice sample was taken in the assessment sessions or when they came for therapy sessions.

- Subjects were informed that they may not have any immediate benefit from the study; however the study will be helpful in predicting the accurate diagnostic procedure especially using the perceptual evaluation of voice using the CAPE-V rating scale for other clients with any type of voice disorder.

CHAPTER - IV

RESULTS AND DISCUSSION

The purpose of the study was to evaluate the reliability of perceptual evaluation of voice using Consensus Auditory Perceptual Evaluation of Voice Scale (CAPE- V) in Indian Context. The scale has been reported to be one of the reliable tools for perceptual evaluation of voice in voice clinics of Western Countries. However, in Indian Context not many published studies on this scale are available.

In the present study twenty one males with dysphonia were taken from AIISH clinic, who were diagnosed to have hoarseness of voice due to benign causes. The subjects considered were in the age range of 25 to 45 yrs of age. Three types of samples i.e. phonation of vowels /a/, /i/ and /u/, three oral and three nasal sentences and spontaneous speech from each individual was audio recorded using Cool Edit software. To avoid biasing the samples were randomized and twenty one sets were made which consisted of different voice samples of different individuals. These sets were given to six judges, who have an experience of 3 – 4 years in diagnosis and management of voice disorder. The judges were given CAPE-V rating scales (Appendix I) and were asked to rate the voice on six parameters i.e. roughness, breathiness, strain, pitch, loudness and overall quality on three different types of rating scales i.e. ordinal scale (mild, moderate, severe ratings), 100 mm visual analog scale and dichotomous scale for consistency or inconsistency of that voice parameter in that particular task. The data obtained after perceptual examination of voice by various judges was subjected to statistical analysis using SPSS 16 in order to determine if there is,

- Reliability across judges on different speech tasks (phonation, sentences and spontaneous speech) on categorical ratings (mild, moderate, severe).
- Correlation between categorical naming (mild, moderate, and severe) and numerical value (VA) assigned across three tasks.
- To find the most appropriate task for perceptual analysis of voice using CAPE-V.

I. Comparison of Reliability Across Judges on Different Speech Task

To evaluate reliability across judges the data was tabulated into SPSS software (Version 16) in which the ratings of all judges for each parameter were compiled. The twenty one participants were placed on the vertical column in the SPSS data sheet. Variables in the data sheet were based on (in order):

1. Task: Phonation/ sentences/ spontaneous speech
2. Voice parameter: Roughness, Breathiness, Strain, Pitch, Loudness, Overall
3. Judges: Judge1- 6
4. Type of scale in CAPE-V: Degree on ordinal scale (mild, mild-mod, mod, mod-severe, severe) and numerical rating on 100 mm VA scale.

The response obtained by each judge for each subject was compiled according to the various tasks of CAPE- V. Further these scores were compared across judges, tasks and parameters, for reliability.

To evaluate the reliability of judges in rating the voice samples Cronbach's alpha (α) coefficient was computed. Inter judge reliability was evaluated for each parameter of each task. Table 1 depicts degree of reliability on phonation task.

S.No	Parameters	Reliability coefficient (α) on Ordinal Scale	Reliability coefficient (α) on VA Scale
1.	Roughness	0.76	0.80
2.	Breathiness	0.84	0.84
3.	Strain	0.83	0.86
4.	Pitch	0.80	0.84
5.	Loudness	0.75	0.75
6.	Overall quality	0.83	0.81

Table 1: Inter rater reliability on phonation task

Table 1 illustrates inter judge reliability of six judges in ordinal scale (mild, mild- moderate, moderate, moderate- severe and severe) in assigning degree to each parameter of phonation task. Cronbach's alpha coefficient was computed for reliability between the judges in each parameter. All the six parameters of the scale received high α value with highest on breathiness (0.84), strain and overall quality (0.83), pitch (0.80), roughness (0.76) and relatively low on loudness parameter (0.75). Similarly, Cronbach's alpha coefficient was computed for reliability across the judges in each parameter on 100 mm VA scale. As seen in the table, strain received highest reliability (0.86), followed by breathiness and pitch (0.84), overall quality (0.81) and roughness (0.80). Lower reliability was seen on loudness parameter (0.75) which is the similar to that obtained on the ordinal scale. Also, in breathiness parameter same value of α coefficient (0.84) has been obtained in both the scales, with almost similar values of α in all the parameters of both ordinal and VA scale. This shows that there is a good correlation across the judges in rating phonation voice sample on ordinal scale and VA scale.

Reliability was estimated for sentences across judges. Table 2 depicts inter judge reliability of six judges in ordinal scale for assigning degree to each parameter of oral – nasal sentences voice sample and their equivalent rating on VA scale.

S.No	Parameters	Reliability coefficient (α) on Ordinal Scale	Reliability coefficient (α) on VA Scale
1.	Roughness	0.83	0.89
2.	Breathiness	0.78	0.73
3.	Strain	0.83	0.80
4.	Pitch	0.65	0.62
5.	Loudness	0.61	0.63
6.	Overall quality	0.81	0.85

Table 2: Inter rater reliability on sentences task

Cronbach’s alpha coefficient was computed for reliability across the judges in each parameter. On the ordinal scale, roughness and strain demonstrate maximum reliability between the judges (0.83), followed by overall quality (0.81), breathiness (0.78). Moderate reliability was observed for loudness (0.61) and pitch (0.65) parameters. Similarly on VA scale reliability was high for roughness (0.89), followed by overall quality (0.85), strain (0.80) and breathiness (0.73). Moderate reliability was observed for loudness and pitch parameter, which corresponds to the findings from ordinal scale.

Reliability was estimated for spontaneous speech across judges. Table 3 depicts inter judge reliability of six judges in ordinal scale for assigning degree to each parameter of oral – nasal sentences voice sample and their equivalent rating on VA scale.

S.No	Parameters	Reliability coefficient (α) on Ordinal Scale	Reliability coefficient (α) on VA Scale
1.	Roughness	0.80	0.83
2.	Breathiness	0.79	0.77
3.	Strain	0.82	0.78
4.	Pitch	0.62	0.65
5.	Loudness	0.73*	0.61
6.	Overall quality	0.78	0.79

Table 3: Inter rater reliability on spontaneous speech task

***** indicates reliability between 3rd & 4th judge

The above table depicts inter rater reliability of six judges in ordinal scale for assigning degree (mild, mild- moderate, moderate, moderate- severe and severe) to each parameter of spontaneous speech sample and also on VA scale. Cronbach's alpha coefficient was computed for reliability between the judges in each parameter. High reliability is observed for strain (0.82), roughness (0.80), breathiness (0.79) and for overall quality (0.78). Moderate reliability was observed for pitch parameter (0.62). For loudness parameter, relatively low reliability was obtained across the six judges. Hence, reliability was considered between 3rd and 4th judges. Similarly, on VA scale reliability for roughness was high (0.83) followed by overall quality (0.79), strain (0.78), and breathiness (0.77). Moderate reliability was observed for parameters like pitch (0.65) and loudness (0.61).

Table 4 depicts overall reliability across judges on different parameters across the three different tasks of the study.

		PHONATION		SENTENCES		SPEECH	
S.No.	*PARAMETERS*	*On ORD Scale*	*On VA Scale*	*On ORD Scale*	*On VA Scale*	*On ORD Scale*	*On VA Scale*
1.	Roughness	0.76	0.80	0.83	0.89	0.80	0.83
2.	Breathiness	0.84	0.84	0.78	0.73	0.79	0.77
3.	Strain	0.83	0.86	0.83	0.80	0.82	0.78
4.	Pitch	0.80	0.84	0.65	0.62	0.62	0.65
5.	Loudness	0.75	0.75	0.61	0.63	0.73*	0.61
6.	Overall quality	0.83	0.81	0.81	0.85	0.78	0.79

Table 4: Overall inter rater reliability across tasks on both the scales of CAPE-V

* indicates reliability between 3rd & 4th judge

In general high reliability was found across judges and across all the tasks (except for loudness parameter on ordinal scale in spontaneous speech). The reliability range was within acceptable limits of 0.05. This indicates that the judges were reliable in giving ratings on VA scale and on ordinal scale. The results of the present study indicate that high reliability for perceptual evaluation of voice across tasks and judges was obtained by using CAPE- V.

The present study revealed high reliability across judges and across the parameters on perceptual evaluation of voice using CAPE-V scale. The results obtained from the present study support the findings of Berg et al (2003), Karnell et al (2007), Zraick et al (2005) and Kelchner et al (2009) who have reported significant reliability for the perceptual evaluation using CAPE-V. High reliability of perceptual evaluation of voice was found using other scales like GRABS (Hirano, 1981) and Wilson's Voice Profile (Wilson, 1987) etc. Some of the studies like Wolfe, Ratusnik (1988) have reported reliability of $\alpha = 0.95$ using 7 point EAI scale, Moran and Gilbert (1984) had used Wilson's Voice Profile and reported high spearman's rank correlation for intra- judge reliability (r = 0.85 to 0.95) and ANOVA for inter judge reliability (0.77 - 0.95). Reliability for perceptual analysis of speech was studied by comparing the two different perceptual scales. Karnell et al (2006) who compared the CAPE-V scale to the GRABS scale obtained high reliability of the two perceptual scales and reported CAPE-V to be more sensitive than GRABS.

The difference among all the above studies are mainly with respect to the methodology, pertaining to the type of scale used (EAI, VA, Ordinal etc) and the sample selected (phonation, sentences and spontaneous speech) and judges' experience (trained Vs untrained). In the present study, the methodology used for assessing the reliability is more relevant as the samples considered were from simple phonation to spontaneous speech.

Zraick et al (2005) discussed high inter and intra judge reliability as a result of expertise of the listeners in identifying and describing dysphonia. Also, that experience and professional background may partially account for slightly higher intra judge reliability, which is consistent with the findings of De Bodt et al, (1996); Bassich et al (1986); Askenfelt et al (1986), who opined that more is the experience of the listener more is the reliability.

Kreiman et al (1992) have suggested that all the listeners have similar, relatively stable internal standards for 'normal' voice quality because of the every day experience they have with normal voices. Hence, consistency is observed when they rate normal or near normal voices. The internal standard for pathological voice may vary from a listener to listener depending upon their experience or exposure to it. This suggests that listeners need many years to develop a stable set of criteria for reliability in rating voice quality (Kreiman et al 1992). Also, in the present study high reliability of perceptual evaluation of voice using CAPE-V can be attributed to high sensitivity of the 100 mm VA scale incorporated in it. The findings of this study supports the findings of Kreiman et al (1993) who suggested that scaling systems that rely primarily on ordinal or equal-appearing interval scales may have limited reliability potential and proposed a visual analog scaling procedure. The finding of the present study supports this statement where high reliability in perceptual voice evaluation using CAPE-V, can be attributed to experience of judges contributing to the study. The present study adds to the few studies which have used CAPE-V to establish reliability of perceptual evaluation of voice.

The study does not support the findings of Wuyts et al (1999) who opined that though VA offers finer judgment of voice quality but with increased degree of freedom the inter rater agreement decreases considerably. They also reported that on VA scale a general trend was exhibited i.e. the raters tend to score the voice more to the middle of the 100 mm line. They suggested that VA scale has more variability in rating the voices than the ordinal scale.

II. Correlation between Ordinal and VA Scale

The dimensions on which both the scales of CAPE-V evaluate a voice sample are different. The ordinal scale has the classification as mild, mild- moderate, moderate, moderate- severe, severe where as VA scale has numbers assigned from 1-100, based on which judge has to provide description and also assign numbers for rating the voice on six parameters. In order to estimate the correlation between VA and ordinal scale, spearman's rank correlation coefficient was computed.

The perceptual scores provided by the judges were compared to find agreement between the two scales. Section II describes correlation across different tasks by estimating the relation between ordinal and VA scale. Table 5 illustrates the correlation coefficients between the two scales on phonation task for which spearman's correlation coefficient was computed and compared

Phonation / Parameter	Spearman's coefficient between both the scales	Significant/ Insignificant correlation
Roughness	r = 0.76***	Significant
Breathiness	r = 0.69**	Significant
Strain	r = 0.47*	Significant
Pitch	r = 0.72***	Significant
Loudness	r = 0.67**	Significant
Overall quality	r = 0.66**	Significant

Table 5: Correlation between ordinal and VA scale on phonation task

Here, '*' indicates $P < 0.05$; '**' indicates $P < 0.01$; '***' indicates $P < 0.001$

Table 5 reveals that there is significant correlation in all the parameters between VA and ordinal scale but comparatively higher correlation in roughness and pitch parameters (r = 0.76 & r = 0.72 respectively) is observed. Moderate correlation is been observed for parameters like breathiness (r = 0.69), loudness (r = 0.67) and for overall quality (r = 0.66). Low correlation is observed for strain parameter (r = 0.47) in the phonation task.

To evaluate the correlation between degrees on ordinal and VA Scale on oral - nasal sentence reading task spearman's correlation coefficient was computed for each of the parameter between both the scales. Table 6 illustrates the correlation coefficients for sentence reading task of the study.

Sentences / Parameter	Spearman's coefficient	Significant/ Insignificant correlation
Roughness	r = 0.59**	Significant
Breathiness	r = 0.73***	Significant
Strain	r = 0.69***	Significant
Pitch	r = 0.80***	Significant
Loudness	r = 0.64**	Significant
Overall quality	r = 0.72***	Significant

Table 6: Correlation between ordinal and VA scale on Sentences Reading task

Here, '*' indicates P < 0.05; '**' indicates P < 0.01; '***' indicates P < 0.001

The above table shows that there is a high correlation between the two scale in pitch (r = 0.80), breathiness (r = 0.73), overall quality (r = 0.72) and strain (r = 0.69) parameters in sentence reading/repetition task. However, moderate correlation was seen for loudness (r = 0.64) and roughness parameter (r = 0.59). It is apparent that loudness on both the task shows moderate correlation between the ratings on the two scales.

To evaluate the correlation between ordinal and VA scale on spontaneous speech task spearman's correlation coefficient was calculated for each of the parameter. Table 7 illustrates the correlation coefficients obtained from the computation.

Spont. speech / Parameter	Spearman's coefficient	Significant/ Insignificant correlation
Roughness	r = 0.75***	Significant
Breathiness	r = 0.95***	Significant
Strain	r = 0.71***	Significant
Pitch	r = 0.78***	Significant
Loudness	r = 0.48*	Significant
Overall quality	r = 0.82***	Significant

Table 7: Correlation between ordinal and VA scale on spontaneous speech task

Here, '*' indicates $P < 0.05$; '**' indicates $P < 0.01$; '***' indicates $P < 0.001$

Table 7 shows that there exists a significant correlation on all the parameters except for loudness. Highest correlation is observed in breathiness (r = 0.95) followed by overall quality (r = 0.82), pitch (r = 0.78), roughness (r = 0.75) and strain (r = 0.71). It is evident that, loudness parameter demonstrates lowest correlation of r = 0.48 compared to all the other parameters. This matches with the findings from the phonation and sentence reading task, where loudness demonstrates lower correlation between both the scales.

The present study was also aimed at finding the correlation between the two scales of CAPE-V i.e., the VA and ordinal scale. Thus, Spearman's rank correlation coefficient was computed across the three tasks. The results of the computation reveal that there is high correlation between the two scales amongst all the voice attributes of CAPE-V. However, some of the parameters like loudness consistently demonstrate lower correlation in comparison to parameters like roughness, breathiness, overall quality etc.

The present is study aimed to evaluate the correlation between the two scales (VA and ordinal) and also to examine the pattern of ratings given by the judges. Since the two scales are fundamentally different the statistical analysis is also different. Thus, direct comparison of the two scales could not be done. It was required to convert VA scale to an ordinal scale by pooling the scores in to bins in order to understand the correlation between the two scales (Wuyts et al, 1999). To do this VA rating was converted into '1-10' ordinal scale i.e. a VA value from 0 - 9 was given ordinal scale score of '1'and value from 10-19 was given a ordinal score of '2', value from 20 - 29 was given a score of '3' and like wise the ratings were converted from VA to ordinal scale.

Table 8 illustrates the scheme for conversion of VA score to 10 point ordinal scale. The converted VA values are called as modified VA values or mVA (which is actually ordinal). This method of conversion of scales for comparison of fundamentally two different types of scale was first reported by Wuyts et al, (1999). Yu et al (2002) also adapted similar method to discover the correlation between VA and ordinal scale. They reported higher correlation between the acoustic/aerodynamic values and perceptual evaluation using modified VA scale than ordinal scale. In addition, the ordinal ratings (degrees) i.e. mild, mild- moderate, moderate, moderate-severe, severe were given a numerical value of 1, 2, 3, 4 and 5 respectively.

The following table depicts the modified VA values on a 0 to 10 scale.

Original VA scale (in mm)	mVA (ordinal) scale
0-9	1
10-19	2
20-29	3
30-39	4
40-49	5
50-59	6
60-69	7
70-79	8
80-89	9
90-99	10

Table 8: Scheme for transforming VA score to 10 point modified VA scale

After converting the VA to mVA scale cross tabulations were acquired for both the scales. It was observed that, very few judges rate the voice quality as mild- moderate and moderate- severe i.e. ratings of 2 and 4 are being given much lesser than that of 1, 3 and 5 (degrees) on ordinal scale. Similar trend was observed on all the three types of voice samples being rated. Hence during cross tabulation the ‘2’ and ‘4’ (mild- mod and mod- severe degrees) have not been taken.

Table 9 shows the cross tabulation obtained for the task of phonation, sentences and spontaneous speech. Rating patterns for mild, moderate and severe degree are been depicted in the table.

	Phonation			Sentences			Spont. speech		
mVA	1 (Mild)	3 (Mod)	5 (Sev)	1 (Mild)	3 (Mod)	5 (Sev)	1 (Mild)	3 (Mod)	5 (Sev)
2	**8**	0	0	**3**	0	0	**5**	0	0
3	**15**	6	0	**28**	3	0	**7**	2	0
4	**14**	**12**	0	**10**	**12**	0	**16**	**10**	0
5	**6**	**10**	1	**5**	**14**	3	0	**17**	1
6	0	**11**	1	0	**11**	0	0	**11**	**2**
7	0	**5**	**2**	0	3	**3**	0	1	**4**
8	0	0	**4**	0	0	0	0	0	**3**
9	0	0	**3**	0	0	0	0	0	**2**
Total	43	44	11	46	43	6	46	43	6

Table 9: Cross tabulation across all samples

From Table 11 for phonation task it can be inferred that judges have given VA values from 10 to 49 corresponding to ‘mild’ ranking on the ordinal (degree) scale. Similarly, for ‘moderate’ degree the VA values ranged from 40 to 69. In the similar fashion conclusions can be drawn for ‘severe’ degree, which was marked in the VA range from 70 to 89.

However, an overlap can be observed between the entire three ordinal scales i.e., values from 30 to 49 (20 mm) indicate both the mild and moderate degree and 60 to 69 indicates both moderate and severe in the cross tabulation but grossly a range of VA values corresponding to each degree can be inferred for the task of phonation.

The table signifies the similar trend followed by judges in rating the voice sentences samples and it is observed that like in phonation VA values from 10 to 49 corresponds to ‘mild’ degree on ordinal scale. Similarly, VA values from 30 to 59 are being marked for ‘moderate’. An overlap exists in VA values, where VA values from 40 to 59 signifies both mild and moderate category on ordinal scale. Further, value greater than 69 on VA scale indicates severe degree on sentence reading task.

Also it can be clearly noticed similar rating patterns of judges was observed in phonation, sentences samples and spontaneous speech task. Hence the degree assigned to particular voice sample and the VA value given by judges, correlated well with each other. This signifies good correlation between the VA and ordinal scale.

Thus the above cross tabulations for phonation, sentences and spontaneous speech indicates that there is a predictable pattern in rating the voice quality on both the scales. On an average if a judge assigns VA value from 10 to 39 then he/she refers voice sample to have ‘mild’ degree. If he/she assigns VA value from 40 to 69 he refers the voice sample to be of ‘moderate’ degree. Finally, if he/she gives a VA value > 70 then it points to ‘severe’ degree.

The findings from the present study support the results of Yu et al (2002) who reported that the correlation between perceptual and objective voice judgment is better using modified visual analog scale (r = 0.88) than a conventional ordinal scale (r = 0.64). This can be attributed to the fact that distinguishing between normal and severely dysphonic voice is easy and independent of listener’s experience or level of skill (Kreiman et al 1992, Yu et al 2002). But the real challenge in perceptual analysis is recognition of intermediated grade of dysphonia (G1 and G2 and variations in degree of dysphonia). Hence when rating on ordinal scale, judgments even by skilled judges are subject to great variability probably because of lack of precise internal standard to distinguish intermediate grades of dysphonia (Yu et al, 2002). This also leads to a disadvantage that ordinal scale becomes too insensitive for small variations in voice quality (Wuyts et al, 1999) and an advantage that it limits inter and intra rater variability by providing a broad band for each level of severity (Yu et al, 2002). While, on the other hand it was observed that VA scale offers increased of freedom of judgment and thus finer judgment. But this may contribute to decreased inter rater agreement (Wuyts et al, 1999).

But when any rater uses the two scales simultaneously ‘cross over effect’ can be observed. When the two scales are used by a single clinician at the same time, the use of one scale probably impacts the use of the other. That is, the judges selected VA scale value and ordinal descriptor (degree) in such a way that it shows a relationship with each of the two. Given the strong reliability of both, it may be that the higher resolution of the VA system positively impacts the reliability of the ordinal scale. The simplicity of ordinal and the resolution of VA scale work together such that use of both the scales, in the clinical assessment of voice may be worth considering (Karnell et al, 2007). The above fact can be attributed to the high correlation between the two scales of CAPE-V in the present study.

Hence from the present study the correlation between VA and ordinal scale can be drawn as follows,

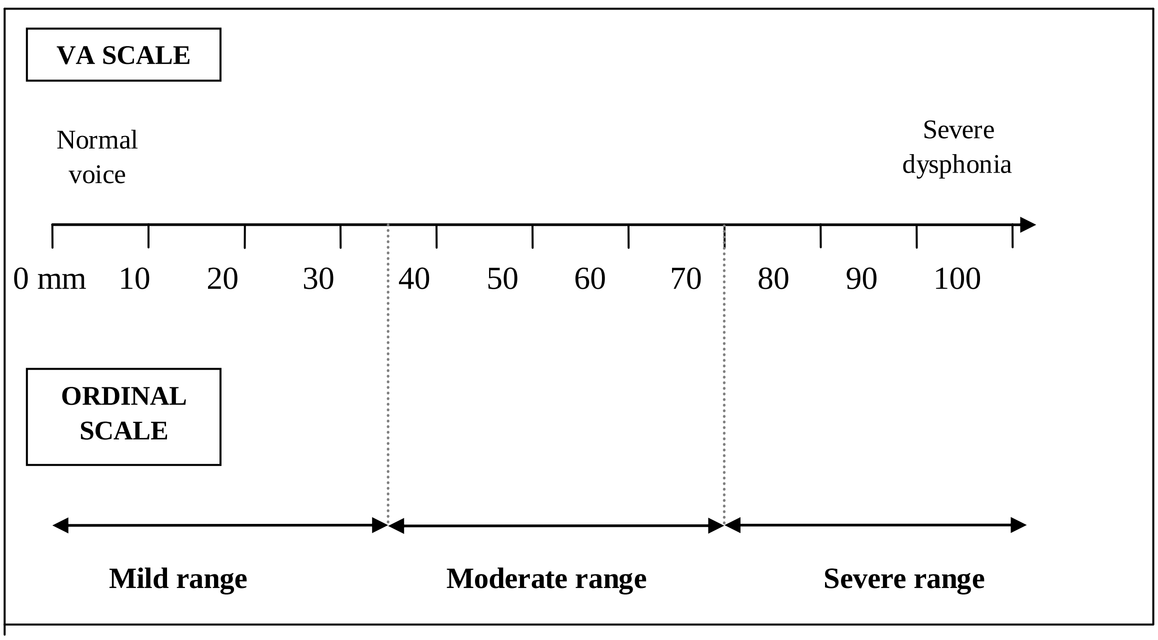

Figure 3: Visual representation of relationship between VA and ordinal scales

III. Comparison of Perceptual Task Across Voice Samples

The voice parameters were combined within phonation, sentences and spontaneous speech. Since the study was aimed to determine the suitable task for better perceptual evaluation of voice. In view of this, the perceptual attribute on all the dimensions were compared across the three tasks, spearman's correlation coefficient was computed in order to determine the task which shows highest correlation between the two scales of CAPE-V.

Table 10 shows the correlation coefficients obtained for each of the tasks of the study.

Task	Correlation coefficient
Phonation	r = 0.73***
Sentences	r = 0.76***
Spontaneous speech	r = 0.82***

Table 10: Overall correlation between the two scales on all the three tasks

Here '***' indicates $P < 0.001$

High degree of correlation was obtained for all the tasks. Since spontaneous speech task demonstrates highest correlation coefficient ($r = 0.82$) compared to sentences ($r = 0.76$) and phonation ($r = 0.73$), consequently this parameter can be considered to be most appropriate for perceptual evaluation of voice.

In view of the fact that the CAPE-V scale has different dimensions for classification of voice disorder, the comparison of VA and ordinal scale with six parameters was done using spearman's correlation coefficient. However to confirm the results the overall correlation on the three different tasks, the numerical values assigned by judges on VA were converted to ordinal ratings from 1 – 10 (mVA). The ordinal ratings i.e. degree were rated as mild = 1, mild- moderate = 2, moderate = 3, moderate-severe = 4 and severe = 5. On the acquired data again Spearman's correlation coefficient was administered.

The following table exhibits the values of spearman's coefficient for the three tasks.

Task	Correlation coefficient
Phonation	r = 0.70***
Sentences	r = 0.73***
Spontaneous speech	r = 0.81***

Table 11: Overall correlation between the two scales on all the three tasks (using mVA & ordinal scale)

Here '***' indicates P < 0.001

The above table suggest that overall correlation for spontaneous speech (r = 0.81) is better than that for oral- nasal sentences (r = 0.73) and phonation (r = 0.70) tasks. This authenticates the findings obtained based on the actual values in table 8. Also, the tasks follow a similar trend of correlation on both computations (table 10 & 11) i.e. spontaneous speech shows highest correlation followed by sentences and then phonation task. Hence, for perceptual evaluation the task of spontaneous speech elicits more reliability and correlation between the parameters than sentence reading or phonation tasks.

The following graphs represent the patterns of ratings by judges on the two scales of CAPE- V. X axis represents the ordinal scale where 1 refers to mild, 2 refers to mild moderate, 3 refers to moderate, 4 refers to moderate to severe and 5 refers to severe. Y axis refers to VA scale values from 0 to 100 on 100 mm VA scale.

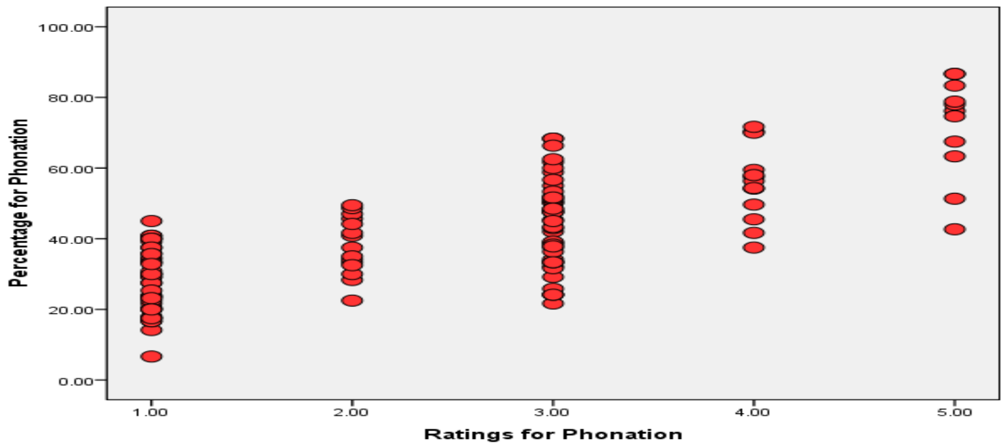

Figure 4: Graph representing the rating patterns on the phonation task

The above scatter plot was obtained for the phonation sample. From the figure it can be derived that as ratings or severity on ordinal scale enhances i.e. from mild to severe along the X axis the scatter plots show a clear rise in the VA scale on Y axis as well. A rise in the height of plots towards severe degree represents good correlation between the two scales in the phonation task.

The following figure was obtained for sentence task.

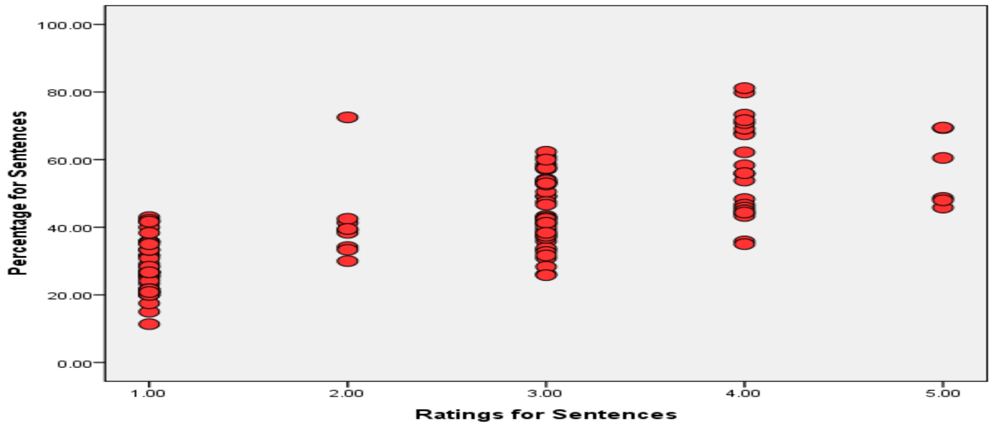

Figure 5: Graph representing the rating patterns on the sentences task.

The above scatter plot was obtained for the sentence (reading) sample. From the figure it can be derived that as ratings or severity on ordinal scale enhances i.e. from mild to severe along the X axis the scatter plots show a clear rise in the VA scale on Y axis as well. A rise in the height of plots towards severe degree represents good correlation between the two scales in the sentence reading task.

The following figure was obtained for spontaneous speech

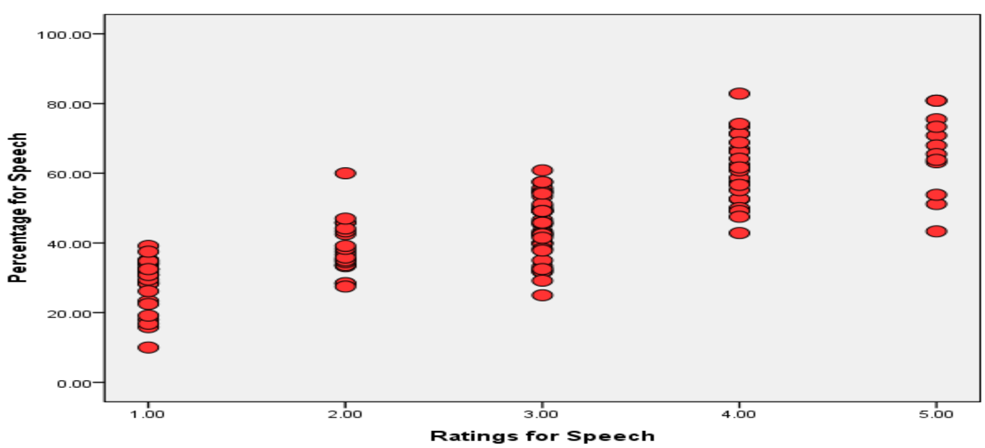

Figure 6: Graph representing the rating patterns on the spontaneous speech task.

The above scatter plot was obtained for the sentence (reading) sample. From the figure it can be derived that as ratings or severity on ordinal scale enhances i.e. from mild to severe along the X axis the scatter plots show a clear rise in the VA scale on Y axis as well. A rise in the height of plots towards severe degree represents good correlation between the two scales in the spontaneous speech task.

The conclusion obtained from the present study is in accordance with the studies done by Munoz et al (2002), Zraick et al (2005) and Wolfe and Cornell (1995) who support the use of connected speech for reliable perceptual evaluation. This can be attributed to fact that perceptual ratings of other vocal parameters, such as pitch and loudness vary as a function of the type of sample (Zraick et al, 2005).

Analysis of sustained vowels has always been given major importance in the perceptual (Aronson, 1980) as well as the acoustic evaluation (Greene, 1972) of voice disorders. However, many investigators have concluded that sustained vowels do not adequately represent continuous speech. According to Askenfelt and Hammarberg (1986), a sustained vowel tends to be representative of voice function status only in those cases where the dysphonia is due to severe laryngeal pathology, like laryngeal cancer or unilateral paralysis. Thus, for the majority of the subjects it is necessary to analyze running speech in order to obtain an adequate estimation of the voice status. Furthermore, according to Takahashi and Koiki (1975), the initial and the terminal parts of the voice may carry abundant information not contained in the steady-state vowel. Similarly, Hammarberg, Fritzell, Gauffin, Sundberg and Wedin (1980) stated "Changes in running speech such as vocal onset and termination, voice breaks, etc., are crucial to voice quality, and are not likely to appear in a single vowel sound."

Lieberman (1961) in a study of pitch perturbation reported that glottal irregularities that are introduced by the coarticulatory influences of connected speech can distort measures of vocal-fold function such as jitter and shimmer. This effect of coarticulation on glottal irregularities does not occur during vowel production. Hence, this effects the perceptual evaluation of voice too. Hirano (1981) reported that the laryngeal function measures are highly influenced by speech context because of co-articulatory influences. This hence effects both acoustic and perceptual assessment of voice. Potential differences in the severity or quality of phonation have been reported to exist between the two phonatory samples. During speech production, the quality of the laryngeal tone is subject to changing articulatory influences that do not operate during static vowel productions (Stevens, 1977).

The findings of Sapienza and Stathopoulos (1995) lead to the conclusion that fundamental frequency (f_o) is not the same during vowel prolongation as compared to contextual speech. Certain amplitude- and time-based measures are significantly different for vowel prolongation produced by speakers with disordered voices when compared to vowels within syllable and reading. This finding indicates that perceptual evaluation of voice highly depends on the sample being evaluated.

An important consideration for any evaluative measure of communicative disorders is its validity, that is, the degree to which the measure represents the speech or voice typically used for communicative purposes. Therefore, considering the widespread use of the sustained vowel in the evaluation of voice disorders, the degree to which it is representative of voice in connected speech should be of value to practicing clinicians (Wolfe et al, 1995). The authors did study to find the relationship between the vowels and connected speech on perceptual evaluation of voice and reported high correlation (r = 0.78) between the rated severity of sustained vowels and connected speech. But they also stated that sustained vowel sounds may not be an adequate clinical index to the severity of dysphonia as is indicated by continuous speech sample. This finding is in concordance with the results of present study which, too reports that the correlation coefficient of all the tasks (phonation, sentences, spontaneous speech) are high but is significantly high for spontaneous speech task.

However, other researchers like De Krom, (1994); Revis et al (1999) have reported that there is no effect of the type of sample, in reliability of perceptual voice evaluation and that the complete phonation sample is rated similarly as the connected speech by judges. De Krom, (1994) opined that stimulus type had virtually no effect on either intra-rater or inter-rater reliability. But when determined as a function of the overall degree of severity of a voice, the reliability of ratings for the breathiness and roughness parameters was slightly higher for vowel stimuli than for connected speech.

The present study adds to the available literature regarding the use of spontaneous speech as most appropriate speech sample for perceptual analysis of voice compared to phonation and reading.

FLOW CHART OF THE STEPS FOLLWED DURING THE STUDY

Voice samples (phonation, sentences & spontaneous speech) were recorded from 21 subjects with hoarseness of voice.

⇓

To avoid biasing these samples were then randomized and 21 sets were made. Each set had phonation, sentence and spontaneous speech sample of three different subjects. These sets were copied to CDs & were given to judges for rating the voice along with CAPE-V rating sheets (Appendix I).

⇓

After a period of one week, the rating sheets were collected from all the judges (Appendix IV) and the randomized data was organized back and each subject's ratings for the three tasks were grouped together.

⇩

Data was organized in a way that can be tabulated on SPSS (version - 16) software. Variables were made based on which task has been rated, on which parameter it has been rated, by which judge and on which scale (ordinal/ VA).

⇩

These variables were tabulated on horizontal axis and the total number of subjects on the vertical axis of SPSS data sheet I. Appropriate entries for each of the subjects was done.

⇩

To get the reliability of ratings between the judges on three tasks, on six different parameters and on three different scales Cronbach's alpha coefficient was computed.

⇩

To get coefficient of correlation between the ratings of different parameters on three different tasks by different judges, Spearman Correlation Coefficient was computed. For this Data sheet II was made by taking the maximally marked degree on the ordinal scale and average of numerical ratings, given by judges on each parameter of each task.

⇩

Again, to get the correlation between the two scales i.e. VA and Ordinal, the ratings on 100 mm VA were converted into modified VA (mVA) with values ranging from 1 – 10. On this data Spearman's correlation coefficient was applied.

The obtained statistical data was analyzed and results were formulated for evaluation of reliability of perceptual evaluation of voice (through cross tabulation/ graphs).

CHAPTER - V

SUMMARY AND CONCLUSION

The speaking voice conveys information about the speaking individual and the voice quality serves as a primary means by which speakers project their physical, psychological and social characteristics to the world. Voice can be evaluated objectively in many ways. But when objective acoustic measures alone are used to analyze vocal quality there appear to represent only a friction of the set of all the measures used by the human listener. Perception of patient's voice is the heart of evaluating and treating patients with voice disorders. Perceptual evaluation of voice is an integrated process of listening to and describing a particular voice. Rating voice quality perceptually is universally acknowledged as difficult task and one requires considerable experience in perceptual evaluation of voice. The perceptual importance of different aspects of voice depends on context, attention, a listener's background and the listening task (Kreiman & Garratt 2001).

Perceptual voice rating scales can be of different types, like - Categorical ratings, Equal Appearing Interval scales, Visual Analog (VA), Direct Magnitude Estimation (DME) and Paired Comparison. In literature, many types of perceptual scales are available for the judgment of the voice disorders. Some of the scales are - The Voice Profile (Wilson, 1970), The Vocal Profile Analysis Protocol (Laver, 1980), The GRABS Scale (Hirano, 1981), Voice Assessment Protocol for children and adults (Pindzola, 1987), Buffalo III Voice Profile (Wilson, 1987) and Consensus Auditory Perceptual Evaluation of Voice (ASHA, 2002). But the reliability of the perceptual data varies from study to study. CAPE-V is relatively a new scale for voice evaluation; hence only limited studies are available on its reliability and validity as a perceptual scale. The scale was developed as a tool for perceptual analysis of voice from a consensus meeting by speech language pathologists and invited experts in human perception held in Pittsburg (2002). The present study was aimed to study the reliability of perceptual evaluation of voice using CAPE- V rating scale.

The present study is an initial attempt to explore the reliability of CAPE-V on various parameters in Indian context. The study aimed at investigating the reliability of perceptual evaluations of voice disorders using CAPE-V scale for different tasks like phonation, sentences and spontaneous speech, in Indian context.

- Reliability across judges on different speech tasks (phonation, sentences and spontaneous speech) on categorical ratings (mild, moderate, severe) and on 100 mm Visual analog scale.
- Correlation between categorical naming (mild, moderate, and severe) and numerical value (on VAS) assigned across three tasks.
- Determination of the most suitable task or voice sample (phonation/ sentences/ spontaneous speech) for perceptual evaluation of voice using CAPE-V.

The participants considered for the present study were twenty one males (in age range of 25-45 years) diagnosed as having hoarse voice quality by a qualified speech language pathologist. The participants were native speaker of Kannada or English. The voice samples was collected by asking the participants to carry out three tasks i.e. Phonating /a/, /i/ and /u/; repetition/ reading of sentences (of respective language), and spontaneously speaking about their voice problem at their comfortable loudness level and in their native language. The tasks were recorded with an interval of 20 seconds between each of tasks. The samples were recorded on Cool Edit (Version II) software in a Compaq laptop using ‘hp’ microphone. To avoid biasing the samples were randomized.

Six speech and language pathologists, who have an experience of 3-4 years in diagnosis and management of voice disorders, were selected as judges. The analysis sheets along with the randomized voice samples were given to the judges for perceptual evaluation of voice. The complete evaluation of voice samples of twenty one participants was done in two sessions by each of the judge.

Obtained data was reorganized and was tabulated on to SPSS software for statistical analysis. Cronbach's alpha coefficient was computed to assess the reliability across judges. Spearman's correlation coefficient was computed to determine the correlation between the two scales i.e., VA and ordinal scale. Also, in order to determine the most suitable task for perceptual evaluation of voice spearman's rank correlation coefficient was computed.

The first aim of the study was to evaluate the reliability across judges on different speech tasks. Cronbach's alpha reliability coefficients were computed for both the scales of CAPE-V i.e., VA and ordinal scales across the tasks. High reliability is seen in all the parameters except for loudness, which demonstrates moderate reliability in most of the tasks. The reliability range was within acceptable limits of 0.05 (among a particular voice parameter in the two scales). This shows that the judges were reliable in giving ratings on VA and on ordinal scale. The results of the present study indicate that high reliability for perceptual evaluation of voice across tasks and judges was obtained using CAPE- V.

Zraick et al (2005) discussed high inter and intra judge reliability as a result of expertise of the listeners in identifying and describing dysphonia. Krieman et al 1992 reported that because of the expertise of the listeners, stable internal standard is formed against which they rate a particular voice sample. Thus, experienced listeners are more reliable in rating the voice than the less experienced ones. Good reliability in the present study can also be attributed to high sensitivity of the 100 mm visual analog scale incorporated in CAPE-V, besides an ordinal scale (Kreiman et al 1993; Karnell et al, 2006).

The second aim of the study was to investigate the correlation between ordinal and VA scale across task. Spearman's rank correlation coefficient was computed for each of the parameter between both the scales. The results of the computation revealed that there is high correlation between the two scales amongst all the voice attributes of the

CAPE – V. However, some of the parameters like loudness consistently demonstrate lower correlation in comparison to parameters like roughness, breathiness, overall quality etc.

Good correlation between the two scales which are fundamentally different can be attributed to the reason that when any rater uses the two scales simultaneously 'cross over effect' occurs. When the two scales are used by a single clinician at the same time, use of one scale probably impacts the use of the other. That is, the judges selected VA scale value and ordinal descriptor (degree) in such a way that it shows a relationship with each of the two. Given the strong reliability of both, it may be that the higher resolution of the VA system positively impacts the reliability of the ordinal scale. The simplicity of ordinal and the resolution of VA scale work together such that use of both the scales, in the clinical assessment of voice may be worth considering (Karnell et al, 2007).

The following figure is the visual representation of the relationship between VA and ordinal scale which can be concluded from the study.

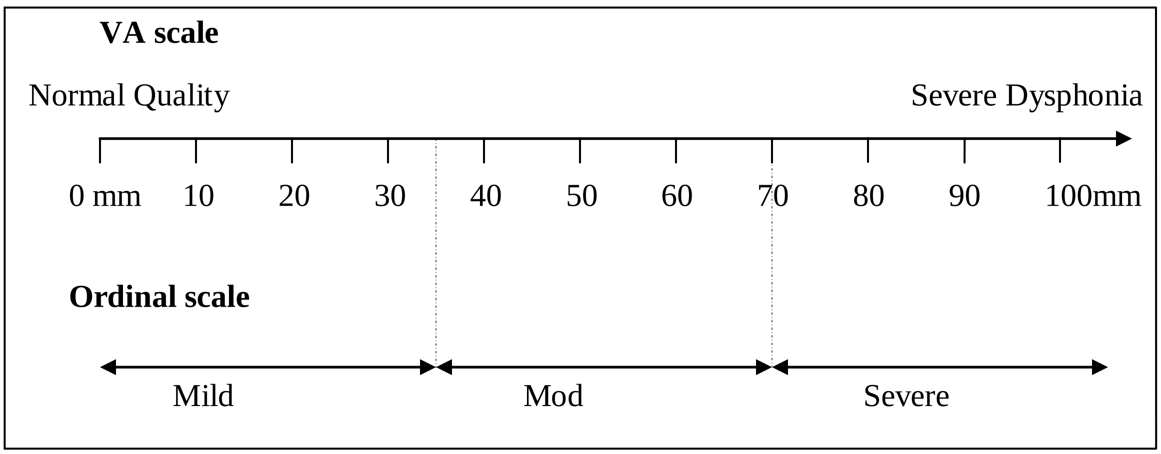

Figure 7: Visual representation of relationship between VA and ordinal scale

The third aim of the study was to determine the most appropriate task for perceptual analysis of voice using CAPE-V. Spearman's rank correlation coefficient was computed in order to determine overall correlation in all the three types of voice samples. It was observed that overall correlation for spontaneous speech (r = 0.81) is better than that for oral- nasal sentences (r = 0.73) and phonation (r = 0.70) tasks.

The conclusion obtained from the present study is in accordance with the studies done by various researchers like Wolfe et al (1995), Munoz et al (2002) and Zraick et al (2005) who support the use of connected speech for reliable perceptual evaluation. This can be attributed to fact that perceptual ratings of other vocal parameters, such as pitch and loudness vary as a function of the type of sample (Zraick et al, 2005).

Hammarberg, Fritzell, Gauffin, Sundberg and Wedin (1980) stated that changes in running speech such as vocal onset and termination, voice breaks, etc., are crucial to voice quality and are not likely to appear in a single vowel sound. Hence, it can be derived from the present study that spontaneous speech task better elicits the reliability in perceptual evaluation of voice.

Hence, following conclusions can be drawn from the present study

- CAPE-V scale which incorporates a Visual analog and an ordinal scale has good reliability.
- The use of the two scales together (VA and ordinal) facilitates reliability and correlation across parameters and judges.
- It can be concluded that, spontaneous speech sample (or connected speech sample) elicits more reliable perceptual evaluation of voice than sustained phonation of vowel and reading sample (sentences).

Implications of the Study

- The findings of the study are very helpful in evaluation of voice disorders using perceptual scale and can be used in combination with objective analysis for diagnosis as well as for evaluation of success rates after voice therapy.

- Perceptual evaluation of voice is a quick and reliable method of voice evaluation. Hence, its use is warranted for all the patients who have dysphonia.

Limitations of the Study

- Sample size taken for the study was small.
- Number of judges taken for study was less.
- Only single types of voice disorder i.e. patients only with hoarseness were included in the study.
- Only inter judge reliability has been gauged in the study. No evaluation for intra judge reliability has been done.

Directions for Future Research

- The present study considered only trained judges who have an experience of 3-4 years in diagnosis and management of voice disorders. The same method can be employed to determine the effect of experience or less experience on perceptual evaluation of voice.

- It may be interesting to examine the reliability of ratings of other groups of professionals also, such as general medical practitioners or otolaryngologists, who are often the first person called to make judgment about dysphonia and its significance in the presentation of signs and symptoms in the patients.

- Intra rater reliability can be assessed by giving the judges the same voice sample 2-3 times with an interval of some days/ week. This would be helpful in determining the internal standard of judges as stable or unstable.

REFERENCES

American Psychological Association. (1985). Standards of Psychological and Educational Testing.Washington, D.C, American Educational Research Association, America.

Anders, L., Hollien, H., Hurme, P., Soninen, A., Wendler, J. (1998). Perception of hoarseness by several classes of listeners. *Folia Phoniatrica Logopaedica*, 40, 91–100.

Anderson, V. A. (1961). The effective voice. *Training the Speaking Voice* (2^{nd} ed.), New York.

Kummer, A. W. (2008). *Cleft palate and craniofacial anomalies- Effects on speech and resonance*, (2nd ed). Thomson Delmar Learning, New York.

ANSI (1960). "Acoustical Terminology" ANSI S1.1-1960, American National Standard Institute, New York.

Aparna, P. S. (2000). Comparison of three different softwares for evaluation of voice disorders. *An Unpublished master's dissertation submitted to university of Mysore, Mysore.*

Aronson, A. (1980). *Clinical voice disorders,* New York, Thieme Stratton, 182.

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.

ASHA Special Interest Division 3, Voice and Voice Disorders, CAPE – V form and procedures. From http//:www.asha.org. Retrieved on April 2, 2009.

Askenfelt, A. G., & Hammarberg, B. (1986). Speech waveform perturbations analysis: A perceptual acoustical comparison of seven parameters. *Journal of Speech and Hearing Research,* 29 (1), 50-64.

Baken, R. J. (1987). *Clinical measurement of speech and voice.* Boston, College Hill.

Bassich, C. J., & Ludlow, C. L. (1986). The use of perceptual methods by new clinicians for assessing voice quality. *Journal of Speech Hearing Disorder*, 51, 125–133.

Baynes, R. A., (1966). An incidence study for chronic hoarseness among children. *Journal of Speech and Hearing Disorders,* 31 (2), 172 – 176.

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

Berg, B., & Eden, S., (2003). Perceptual evaluation of voice quality in three organic voice disorders – a comparison between Consensus auditory Perceptual Evaluation of Voice (CAPE-V) and Stockholm Voice Evalation Approach (SVEA) (in Swedish). *Unpublished Master's thesis, Karolinska Institute, Stockholm, Sweden*. From http//:www.asha.org. Retrieved on April 2, 2009.

Boone, D. R. (1971). *The Voice and voice therapy*. (2nd ed.), New Jercy, Prentice Hall, Inc.

Casper, M., Abramson, A. L., & Forman- Franco, B. (1981). Hoarseness in children: Summer camp study. *International Journal of Pediatric Otorhinolaryngology*, 3, 85-89. Cited in Sederholm, E., Mc Allister, A., Sundberg, J. & Dalkwist, J., (1992). Perceptual analysis of child hoarseness using continuous scales. *Speech transmission laboratory Quarterly Progress and status report*, Vol. 1, 99-113.

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

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

Crocker, L., Algina, J., (1986). *Introduction to classical and modern test theory*. Orlando, F. L., Holt, Rinehart & Winston Inc., 106. Cited in Bele, I. V. (2005). Reliability in Perceptual Analysis of Voice Quality. *Journal of Voice*, Vol. 19 (4), 555- 573.

Cullinan, W. L., Prather, E. M., & Williams, D.E. (1963). Comparisons of procedures for scaling severity of stuttering. *Journal of Speech and Hearing Research*, 6, 187 – 194.

Curtis, J. F. (1978). *Process and disorders of communication.* New York, Harper Row Publication.

Aronson, D. F., & Brown, A. J. (1975). *Motor Speech Disorders*. Philadelphia, PA, W.B. Saunders.

De Bodt, M. S., Wuyts, F. L., Van de Heyning, P. H., & Lambrechts, L. (1996). The perceptual evaluation of voice disorders. *Acta Oto Rhino Laryngologica*, 50, 283–291.

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

Divya, M. (1996). Analysis of voice before and after treatment. *Unpublished master's dissertation submitted to University of Mysore, Mysore.*

Eadie, T. L., Baylor, C. R. (2006). The effect of perceptual training on inexperienced listener's judgment of dysphonic voice. *Journal of Voice*, 20 (4), 527 – 544.

Emerick, L. L. & Hatten, J. T., (1974). *Diagnosis and evaluation on speech pathology*. Jercy Prentice Hall.

Fairbanks, G. (1960). *Voice and Articulation drillbook.* New York. Harper and Bros.

Fex, S. (1992). Perceptual evaluation. *Journal of Voice*, 6, 155- 158.

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

Gerratt, B. R., Kreiman, J., Barroso, N. A., & Berke, G. S. (1993). Comparing internal and external Standards in voice quality judgments. *Journal of Speech Hearing Research*, 36, 14–20.

Gerratt, B. R., Till, J. A., Rosenbeck, J. C., Wertz, R. T., & Boysen, A. E. (1991). Used and perceived value of perceptual and instrumental measures in dysarthic management. In Moore, C. A., Yorkston, K. M., & Beukelman D. R. (Eds.), *Dysarthia and Apraxia of Speech* (pp. 77- 93). Baltimore, MD, Brookes.

Green, M. C. L. (1964). *The Voice and its disorders.* Mitman Medical, London.

Greene, M. C. L. (1972). *Disorders of voice.* Bobbs – Merrill, New York.

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

Hammarberg, B., and Gauffin, J., and Hertegard, S. (1986). Irregularities in the Voice: A perceptual experiment using synthetic voices with subharmonics. Cited in Davis, P. J., & Fletcher, N. H. (Eds), *Vocal fold Physiology.* London, Singular Publishing group.

Hammarberg. B., Gauffin, J. (1995). Perceptual and acoustic characteristics of quality differences in pathological voices as related to physiological aspects. In:

Fujimura, O. & Hirano, M., (1995) eds. *Vocal fold physiology & voice quality control.* San Diego, CA: Singular Publishing Group, Inc.

Helmholtz, H. (1885). *On the sensations of tone.* Dover, New York.

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

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

Hirano, M. (1975). Phonosurgery: Basic and clinical investigations. *Otologia* (Fukuoka), 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, Springer, Wein.

Huang, D. Z., Manifie, F. D., Kasuya, H., & Lin, S. (1995b) Measure of vocal function during change in vocal effort level. *Journal of Voice,* Vol. 9 (4), 429 – 438.

Jaya Kumar (2004). Normative Scores in Nasometry in Kannada. *An Unpublished master's dissertation submitted to Mysore University, Mysore.*

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*, Vol. 21 (5), 576 – 590

Kelcner, L. N., Brehm, S. B., Weinrich, B., Middendorf, J., Dealarcon, A., Levin, L., Elluru, R. (2009) Perceptual evaluation of severe pediatric voice disorders: Rater reliability using the Consensus Auditory Perceptual Evaluation of Voice. *Journal of Voice,* Article in Press.

Kelmen, A. W., Gordan, M. T., Morton, F. M. and Simpson, I. C. (1981) Comparison of methods of assessing vocal function. *Folia Phoniatrica,* 33(1), 51-65.

Kitzing, P., Glottografisk frekvensindikering. Doctoral thesis. ENT-clinic. Lund University, Malmo¨, Sweden; 1979. Cited in Bele I.V., (2005) Reliability in perceptual analysis of voice quality. *Journal of Voice*, Vol. 19 (4), 555 – 573.

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

Kleven., T. (1995) Reliability as a pedagogical problem. In Norwegian, Reliabilitet som Pedagogisk Problem. Oslo, Norway, Department of Education, University of Oslo.

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

Kreiman, J., Gerratt, B. R. (2000). Sources of listener disagreement in voice quality assessment. *Journal of Acoustic Society of America,* 108 (4), 1867 - 1876.

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,* Vol.36, 21-40.

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

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

Kreiman, J., & Gerratt, B. R., (1998). Validity of rating scale measures of voice quality. Journal of acoustical society of America, Vol. 104(3), 1598 – 1608.

Ladefoged, P., Maddieson, I., & Jackson, M. (1988). Investigating phonation types in different languages. In O. Fujimora (Ed), Vocal fold physiology: *Voice production: Mechanisms and Functions,* New York, Raven Press, 297- 317.

Laver (1980). The phonetic description of voice quality. Cambridge U.K., Cambridge University Press.

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

Lively, M. A., & Emanuel, F. W., (1970). Spectral noise levels in roughness severity ratings for normal and stimulated rough vowels produced by adult females. *Journal of Speech and Hearing Research*, Vol.13 (3), 503 - 517. Cited in Sarita, K. (2000) Classification of Hoarse voice based on Acoustic Parameters, *An Unpublished masters dissertation submitted to Mysore University, Mysore*.

Ludlow, C. (1981). Research needs for the assessment of phonatory function. ASHA Reports, 11, 3 – 8. Rockville, MD: American Speech Languge Hearing Association.

McAllister, A. (1997). Acoustic Perceptual and physiological studies of ten-year-old children's voices. Department of Logopedics and Phoniatrics, Karolinska

Institute, Huddinge University Hospital and Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm, Sweden.

Michael, J. & Wendahl, R. (1971). Correlates of voice production. In Travis, L.E., (Ed.) *Handbook of Speech Pathology and Audiology*, (pp. 465- 480), New Jersy, Prentice - Hall, Inc. Englewood Clliffs..

Moll, K. L., (1964). Objective measures of nasality. *The Cleft Palate Journal*, Vol. 1, 371-374.

Moore, G. P., (1971). Voice disorders organically based (Chapter 21). In L.E.Travis (Ed.), *Handbook of Speech Pathology and Audiology*. , New Jercy: Prentice Hall, pp. 535 – 569.

Moore, G. P. (1975). Observation on the physiology of hoarseness – Proceeding of the fourth international congress of phonetic science. Helsinki: Finland (pp. 92 - 95). Cited in Childrens, D. G. and Lee, C. K. (1991). Vocal quality factors: Analysis, synthesis and perception. *Journal of Acoustical Society of America*, Vol. 90 (5), 2394 – 2410.

Moran, M. J., & Gilbert, H. R. (1984). Relation between voice profile ratings and aerodynamic and acoustic parameters. *Journal of Communication Disorders,* 17, 245 -260.

Munoz, J., Mendoza, E., Fresneda, M. D., Carballo, G., Ramirez, I. (2002). Perceptual analysis in different voice samples: agreement and reliability. *Percept Motor Skills*, 94, 1187-1195.

Revis, J., Giovanni, A., Wuyts, F., Triglia, J. M. (1999). Comparison of different voice samples for Perceptual Analysis. *Folia Phoniatricia Logopaed*, 51, 108–116.

Rossi, M., Pavlovic. C., & Espesser. R., (1990). Reducing contexts effects in the subjective evaluation of speech quality. Paper presented at the 120th Meeting of the Acoustical Society of America, San Diego, CA. Cited in Gerratt, B. R., Kreiman, J., Barroso, N. A., & Berke, G. S. (1993). Comparing internal and external Standards in voice quality judgments. *Journal of Speech Hearing Research*, 36, 14–20.

Sarita, K. (2000). Classification of Hoarse voice based on Acoustic Parameters, *An Unpublished masters dissertation submitted to Mysore University, Mysore.*

Sederholm E., Mc Allister, A., Sundberg, J. & Dalkwist, J., (1992). Perceptual analysis of child hoarseness using continuous scales. *Speech Transmission Laboratory Quarterly Progress and Status Report*, vol. 1, 99 -113.

Sederholm, E. (1996). Hoarseness in ten-year old children: Perceptual Characteristics, Prevalence and Etiology. Department of Logopedics and Phoniatrics, Karolinska Institute, Huddinge University Hospital and Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm, Sweden.

Shewell, C. (1998). The Effect of Perceptual Training on Ability to Use the Vocal Profile Analysis Scheme. *International Journal Language Communication Disorder.* Vol. 33, 322–326.

Shrivastav, R., Sapienza, C. (2004). Measurement of voice quality using rating scales: what do ratings tell us? Presented at The Voice Foundation's 33rd Annual Symposium: Care of the Professional Voice, Philadelphia, PA. Cited in Zraick, R. I., Wendel, K., & Olinde, L.S. (2005). The Effect of Speaking Task on

Perceptual Judgment of the Severity of Dysphonic Voice, *Journal of Voice*, Vol. 19 (4), 574–581.

Stevens, K. N. (1977). Physics of laryngeal behavior and larynx modes. *Phonetica*, 34.

Takahashi, H., & Koike, Y. (1975). Some perceptual dimensions and acoustic correlates of pathological voices. *Acta Otolaryngologica* (Suppl. 338), 2- 24.

Van Riper, C. & Irwin, J. V. (1978). *Voice and Articulation*. New Jersy: Prentice Hall, Inc.

Weismer, G., & Liss, J., (1991). Reductionism is a dead end in speech research: Perspective on a New Direction. In C. Moore, K. Yorkston & D. Beukelman (Eds.), Dysarthia and Apraxia of Speech: Perspective on Management (15 - 27) Baltimore: Brookes.

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.

Wendler, J., Doherty, E T., & Hollien, H. (1980). Voice Classification by means of long term speech spectra. *Folia Phoniatrica,* 32, 51- 60.

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

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*. Vol. 43, 697–705.

Wolfe, V., & Ratusnik, D., (1988). Acoustic and perceptual measurements of roughness influencing judgment of pitch. *Journal of Speech and Hearing Research*, 27, 2-6.

Wolfe, V. Cornell, R., Fitch, J. (1995). Sentence/ vowel correlation in the evaluation of dysphonia. *Journal of Voice*, Vol. 9 (3), 297 – 303.

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 GRABS scale for perceptual evaluation of dysphonia. *Journal of Voice*, Vol. 13 (4), 508-517.

Yanagihara, N. (1967). Significance of Harmonic Changes and Noise Component in Hoarseness, *Journal of Speech and Hearing Research*, Vol. 10 (3), 531 – 541.

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 Logopaedia, 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 acoustical society of America*, Vol. 71, pp.1544-1550.

Zraick, R. I., Wendel, K., & Olinde, L. S. (2005). The effect of speaking task on Perceptual Judgement of the Severity of Dysphonic Voice. *Journal of Voice*, Vol. 19 (4), 574–581.

APPENDIX I
CAPE – V RATING SHEET

Perceptual Judgment of Voice disorder   Voice Disorder: Sample No.___

Roughness	MI MO SE	C	I	-----/100
Breathiness	MI MO SE	C	I	-----/100
Strain	MI MO SE	C	I	-----/100
Pitch	Indicate the nature of abnormality----- MI MO SE	C	I	-----/100
Loudness	Indicate the nature of abnormality----- MI MO SE	C	I	-----/100

Over all severity	MI MO SE	C	I	-----/100

Phonation

Sentences

Roughness	MI MO SE	C	I	-----/100
Breathiness	MI MO SE	C	I	-----/100
Strain	MI MO SE	C	I	-----/100
Pitch	Indicate the nature of abnormality----- MI MO SE	C	I	-----/100
Loudness	Indicate the nature of abnormality----- MI MO SE	C	I	-----/100
Over all severity	MI MO SE	C	I	-----/100

Spontaneous speech

Roughness	MI MO SE	C	I	-----/100
Breathiness	MI MO SE	C	I	-----/100
Strain	MI MO SE	C	I	-----/100
Pitch	Indicate the nature of abnormality----- MI MO SE	C	I	-----/100
Loudness	Indicate the nature of abnormality----- MI MO SE	C	I	-----/100
Over all severity	MI MO SE	C	I	-----/100

Remarks :

APPENDIX II

Kannada Nasal sentences (Jaya Kumar, 2004)

1. ಮನು ಆನೆಯನ್ನು ನೋಡಿದ.
2. ನವೀನ ಮನೆಯಿಂದ ಬಂದನು.
3. ನಾನು ಆನೆಯನ್ನು ನೋಡಿದೆ.
4. ಮಂಗ ಮನೆಯ ಮೇಲಿದೆ.
5. ಮಾಮ ಮಂಡ್ಯದಿಂದ ಬಂದರು.
6. ಮೀನಾಳಿಗೆ ನೆಗಡಿ ಬಂದಿದೆ.
7. ನರಿ ನೆಲದಿಂದ ನೆಗೆಯುದು.
8. ಮಾಮನ ಮನೆ ಮಂಗಳೂರಿನಲ್ಲಿದೆ.

Kannada Oral sentences (Jaya Kumar, 2004)

1. ಕಾಗೆ ಕಾಲು ಕಪ್ಪು.
2. ಗೀತ ಬೇಗ ಹೋಗು.
3. ದನ ದಾರಿ ತಪ್ಪಿತು.
4. ಅಪ್ಪ ಪಟ ತಾ.
5. ಬಾಲು ತಬಲ ಬಾರಿಸು.
6. ಬೇಡ ಕಾಡಿಗೆ ಓಡಿದ.
7. ಸರಿತ ಕತ್ತರಿ ತಾ.
8. ಇದು ಹೊಸ ಬಟ್ಟೆ.

APENDIX III

ENGLISH SENTENCES

ORAL SENTENCES (CAPE- V, ASHA, 2002)

1. The blue spot is on the key again.
2. How hard did he hit him?
3. We were away a year ago.
4. We eat eggs every Easter.
5. Peter will keep at the peck.

NASAL SENTENCES (Kummer, 2008)

1. Mama made lemon jam.
2. Ten men came when Jane rang.
3. Ben can’t plan on a lengthy rain.
4. Amanda came from a bounding Maine.
5. Dan’s gang changed my mind.
6. My mama makes lemon muffins.

APPENDIX IV

CAPE- V Rating Scale

Perceptual Judgment of Voice disorder | Voice Disorder: Sample No. 19

Phonation KMJ (C-13) SM

Roughness	MI ~~MO~~ ✓ SE	C ✓	I	-----/100 40
Breathiness	MI ✓ MO SE	C ✓	I	-----/100 30
Strain	MI ~~MO~~ ✓ SE	C ✓	I	-----/100 40
Pitch	Indicate the nature of abnormality----- MI ✓ MO SE	C ✓	I	-----/100 30
Loudness	Indicate the nature of abnormality----- ~~MI~~ ✓ MO SE	C ✓	I	-----/100 20
Over all severity	MI ~~MO~~ ✓ SE	C ✓	I	-----/100 40

Sentences (Oral + Nasal) KMK (C-14)

Roughness	MI ~~MO~~ ✓ SE	C ✓	I	-----/100 40
Breathiness	~~MI~~ ✓ MO SE	C ✓	I	-----/100 20
Strain	MI ~~MO~~ ✓ SE	C ✓	I	-----/100 40
Pitch	Indicate the nature of abnormality----- MI MO ✓ SE	C ✓	I	-----/100 60
Loudness	Indicate the nature of abnormality----- MI ~~MO~~ ✓ SE	C ✓	I	-----/100 40
Over all severity	MI MO ✓ SE	C ✓	I	-----/100 60

Spontaneous Speech Sample KML (C-15) SM

Roughness	MI MO SE ✓	C ✓	I	-----/100 70
Breathiness	MI ✓ MO SE	C ✓	I	-----/100 30
Strain	MI MO SE ✓	C ✓	I	-----/100 80
Pitch	Indicate the nature of abnormality----- MI ✓ MO SE	C ✓	I	-----/100 30
Loudness	Indicate the nature of abnormality----- MI MO ✓ SE	C ✓	I	-----/100 60
Over all severity	MI MO ✓SE	C ✓	I	-----/100 70

Remarks: