4-2-2020 10:00 AM

Auditory- Perceptual and Pupillometric Evaluations of Dysphonic Voices

Mojgan Farahani, *The University of Western Ontario*

Supervisor: Doyle, Philip C., *The University of Western Ontario*
Co-Supervisor: Parsa, Vijay, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Health and Rehabilitation Sciences

Abstract

Background: This thesis reports the findings of three projects that included pupillometric and auditory-perceptual evaluation of three voice quality features (strain, roughness, and breathiness, respectively), and the simultaneous measurement of perceived listening effort.

Methods: In the first study, speech samples from individuals with adductor spasmodic dysphonia (AdSD) were perceptually evaluated by both naïve and experienced listeners on the features of vocal strain and listening effort. In the second project, speech samples of post-laryngectomy tracheoesophageal (TE) talkers were rated by two groups of naïve listeners on vocal roughness and listening effort; one group was provided with audio anchors, the other without. The final study focused on perceptual evaluation of breathiness and listening effort in talkers with vocal fold paralysis (VFP). The VFP speech samples were rated by two listener groups (with and without audio anchors). In all three studies, listeners' pupillary responses also were collected (EyeLink 1000) while listening to and perceptually rating voice stimuli.

Findings: Data obtained from the pupillary assessment, peak pupil dilation (PPD), may indicate a listener's cognitive load when perceptually evaluating disordered voices. Results revealed high correlations between each of the voice dimensions and listening effort. Also, various degrees of correlations were observed between perceptual ratings and PPD. In the first study of AdSD, high correlations were found between PPD and perceptual ratings for naïve listeners. A listener's previous exposure and training evoked different pupillary behavior when compared to naïve listeners. In the second study with TE speakers, moderate correlations were found between perceptual dimensions and PPD values of the with–anchor group; extra cognitive load was attributed to the inclusion of anchors. Anchors also improved interrater reliability for this listener group. Finally, in the third project with VFP, again a correlation was observed between perceptual ratings and PPD. The inclusion of anchor did not improve reliability over the no-anchor group.

Similar to the second study, PPD measures of the with-anchor group were impacted by the use of anchors.

Conclusions: Overall, our data offer valuable insights into auditory-perceptual evaluation of voice quality, the influence of listener experience, previous exposure to dysphonic voices, inclusion/exclusion of audio anchors, and voice features and the potential physiological or cognitive responses to dysphonic voices.

Keywords

Auditory- perceptual evaluation, voice quality, audio anchors, Adductor spasmodic dysphonia, Tracheoesophageal speech, Vocal fold paralysis, Listening effort, Pupillometry.

Summary for Lay Audience

The purpose of the experiments reported in this thesis was to examine how people evaluate the voices of individuals diagnosed with different voice disorders. Voice disorders may occur due to different reasons such as neurological problems that influence the muscles of the larynx (voice box) or due to laryngeal cancer. Depending on the type and underlying cause of the voice disorder, changes in the sound or "quality" of the voice may vary and will differ from that which is considered normal. Such disorders impact many aspects of an individual's life and many of them seek treatment. In order to assess the extent of the disorder at the time of diagnosis and to see how successful the treatment has been afterwards, voice quality is typically evaluated through various methods. One of the most commonly used measures is auditory-perceptual evaluation, a measure based on the judgments and impression of listeners. As a result, these studies recruited participants to listen to samples from speakers with three unique types of voice disorders. Listeners were asked to rate the voices of each speaker on one of three dimensions, how strained, rough, or breathy they sounded. The listeners also were asked to indicate how much "work" they thought they needed to listen to those samples by rating a feature termed "listener effort". In all three experiments, simultaneous data on each listener's variations in pupil dilation in response to these abnormal voice samples were gathered. This measure was obtained with the assumption that changes in pupil dilation, namely, what is termed peak pupil diameter could be used as an indicator of listening and cognitive effort. Results revealed that listeners may require more cognitive work and attention when listening to some disordered voices. This listening demand or cognitive load also may decrease and listeners may habituate to voices with an abnormal quality as they get more exposure to such voices.

Acknowledgments

I would like to express my deep gratitude to Professor Philip C. Doyle and Dr. Vijay Parsa, my supervisors, for their patient professional guidance and valuable support over that past couple of years. My PhD was a great opportunity to work with these two very knowledgeable, understanding, amiable and respectful supervisors.

I also wish to acknowledge the Voice Production and Perception Laboratory, for all the resources and support over the past five past years and also thank my friends and labmates.

My special thanks are also extended to the wonderful staff and friends at National Centre for Audiology, my second home for the past 5 years, for the very friendly atmosphere, their great support and also participation in the experiments.

My sincere thanks also go to Dr. Ingrid Johnsrude at the Brain and Mind Institute (Western University) for providing access to the EyeLink system and also the technical support there.

Finally, I wish to express my great appreciation to my parents and my brothers for their support and encouragement throughout my study.

Table of Contents

List of Figures (where applicable)

List of Appendices (where applicable)

Chapter 1

1. Introduction

Spoken language is transmitted through voice which is the sound produced by airflow through the larynx with periods of vocal folds vibration. This sound is then shaped as it moves into the vocal tract where it evolves into a unique acoustic form representing each individual. If there are any structural or functional abnormalities anywhere in the vocal tract during phonation or resonance, these alterations will result in output that will vary from the normal range of voice features (i.e., pitch, loudness, resonance, and overall voice quality). This will also be influenced by the speaker's age, gender, or geographic or language background, with particular variation noted for individuals who have experience or formal exposure to voice disorders (Boone, McFarlane, & Von Berg, 2005). Voice disorders influence the speaker's quality of life with changes that may impact personal, occupational, and social aspects of one's life. The influence of a voice disorder may be disabling even though the speaker is intelligible.

Upon being diagnosed with any type of voice disorder, more detailed assessments involve efforts to evaluate the extent of the disorder and potential treatment options. The process of voice evaluation is complex and multidimensional and often involves various assessment procedures. The success of treatment often depends on how much the speaker is noticed by listeners as being abnormal or based on the degree of difference from normal expectation (Eadie & Doyle, 2004). The primary focus of the thesis is on auditory-perceptual dimensions of voice and their relationship as key factors in voice quality evaluation. Within the sections to follow, the statement of the problem will be presented, followed by the objectives, hypotheses and the primary research questions.

1.1 Statement of the problem and rationale for the proposed study

Upon diagnosis of any voice disorder, as well as during or after treatment, various voice evaluation measures are used to assess the extent of the disorder and the success of treatment. These measures include various options such as auditory-perceptual methods, objective or acoustic and aerodynamic measures, and visualization techniques. Among all of these approaches, auditory-perceptual evaluation of dysphonic voices is of critical importance as it serves to describe the character and extent of the disorder and its potential for documenting rehabilitation and treatment success. Auditory-perceptual evaluation is still the most widely used assessment method for disordered voices in voice clinics and clinicians still rely on such methods in spite of access to sophisticated acoustic, aerodynamic and vocal fold imaging instrumentation and methods. However, several aspects of auditory-perceptual assessment must always be considered.

For example, listener reliability specific to their judgments of voice quality is very important with factors like a listener's experience, shifting internal standards, types of rating scales used, and the characteristics of the voice sample being evaluated potentially influencing the ratings and reliability of listener judgments. The key papers in the extensive literature of auditory-perceptual evaluation of voice quality recommend a framework for efficient control of these potentially confounding factors and their influence on reliability (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). The framework suggests that the use of anchors, using an appropriate scale which can appropriately account for both metathetic and prothetic voice features, as well as the influence of naïve and experienced listeners and the voice/speech stimuli are dependent on the purpose of any given study. In order to control for and reduce variability in voice quality ratings, listeners' idiosyncratic, unstable, internal standards should be replaced with anchors or referent voices for various voice qualities (Kreiman, Gerratt, Precoda, & Berke, 1992; Gerratt, Kreiman, Antonanza-Barroso, & Berke, 1993; Brinca et a;., 2015). The literature on auditory-perceptual evaluation of voice contains numerous studies on the influence of these factors ((Kreiman, Gerratt, Precoda, & Berke, 1992; Gerratt, Kreiman, Antonanza- Barroso & Berke, 1993; Brinca et al., 2015).

In addition to auditory-perceptual and objective approaches to voice evaluation, another dimension can be added to evaluation of voice quality. This dimension focuses on listeners and how they may be influenced by listening to and/or communicating with individuals who present with disordered voices, rather than on the signal (i.e., the voice) and the talker. Thus, how the listener responds to an abnormal voice is of concern. For example, might some voices elicit possible involuntary physiological changes in listeners upon processing disordered voice signals? In this regard, measures such as pupillometry which assesses changes in pupil response to stimuli may be of value. Pupillometry may also provide additional information relative to the auditory-perceptual process in that it may serve as an objective index of listening effort which may in turn provide an indication of cognitive load.

The literature on pupillometry contains many studies which indicate that a task evoked pupillary response is a reliable, indirect measure of cognitive processing load. Although pupil responses as a marker of cognitive load are not limited to visual stimuli; reactions have been evaluated during listening tasks with some type of auditory stimuli. However, to the best of our knowledge, no study to date has evaluated the potential relationship between the presentation dysphonic voice stimuli and pupillary reactions in listeners while perceptually evaluating abnormal voices. Through such evaluation, the subjective ratings of voice quality, and the degree of effort they put into the process of listening to that stimulus may be identified. Therefore, there is a gap in the literature regarding examining these three factors together: 1) pupil reactions while listening to dysphonic voices, 2) the auditory-perceptual rating of a range of voice disorders, and 3) concomitant ratings of perceived listener effort when presented with abnormal voice samples.

Given the importance of auditory-perceptual evaluation of voice quality in those with voice disorders, exploration of the aforementioned influential factors along with possible physiological reactions to dysphonic voices and the perceived listening effort ratings is a necessary addition to the literature. Examination of pupil dilation while asking listeners to perceptually rate dysphonic voices in addition to seeking their self-reporting of the degree of perceived listener effort can indicate the possible presence and extent of such reactions.

The aim of the series of studies to follow was to evaluate voice features using auditory-perceptual methods. More specifically, we evaluated the feature of “strain” in speakers with adductor spasmodic dysphonia (AdSD), “breathiness” in those with vocal fold paralysis (VFP), and finally, “roughness” in tracheoesophageal (TE) speakers through three experiments. Both experienced and naïve listeners, perceptual anchors and the use of continuous spoken stimuli (to represent running speech and its dynamic features) were used in the experiments. As recommended in the literature, all auditory-perceptual scaling by listeners was done using a visual analog scale (VAS). Further, and in order to measure the listener effort, listeners were also asked to indicate the degree of effort they put in rating each stimulus using a separate VAS scale. In addition, pupil responses as an index of cognitive load and listening effort while perceptually evaluating the stimuli (dysphonic voices). These data were simultaneously measured in an effort to investigate the potential processing load and changes in the pupil dilation measures of the listeners evoked by the voice quality in each group of experimental speakers. The outcome of these studies was anticipated to advance our understanding of the possible relationship between processing load and listening and listener effort during the auditory-perceptual rating of a range of dysphonic stimuli. To the best of our knowledge, no study to date has empirically evaluated measures of pupillometry in normal hearing individuals while being asked to listen to and rate dysphonic voices with the goal of assessing cognitive load and listening effort. The findings from these experiments may then offer greater understanding of the presence of listener effort while perceptually evaluating dysphonic voices.

1.2 Objectives

The purpose of this study is to perceptually evaluate voice features of strain, roughness and breathiness in three patient groups along with possible pupil reactions indicating listening effort and self-indication of listener effort via another VAS scale for each stimulus.

It is hypothesized that different listener groups (naïve, experienced) access a perceptual point-of-reference (judgments made with-anchor and no-anchor) and/or expend different levels of listening effort when listening to a variety of disordered stimuli. It is hypothesized that the nature of the voice disorder, the inclusion/exclusion of audio anchors and previous training, as well as the experience and background in voice evaluation by listeners would influence their pupillary responses in the context of the presentation of disordered voice samples. It was also presupposed that there would be a relationship between the objective and subjective measures of listening effort. Thus, this study was designed to explore the aforementioned hypotheses.

1.3 Primary Research Questions

Research question 1: Do normal hearing adult listeners expend more listening effort while listening and perceptually rating dysphonic voices in patients with spasmodic dysphonia, alaryngeal speakers (TE) and unilateral vocal fold paralysis?

Research question 2: Are the pupil responses in experienced and naïve listeners different in contact with dysphonic voices (AdSD)?

Research question 3: Does the inclusion or absence of anchors make any changes in listeners' pupil responses?

Research question 4: Is there any relationship between measures of objective listening effort and perceived listening effort?

Given the importance of auditory-perceptual evaluation of voice quality in assessment of voice disorders, the impacts of voice disorder, and the gap in the literature, the researchers conducted the experiments to answer the aforementioned questions. The following chapters include a review of the literature around auditory-perceptual and pupillometric evaluation, followed by the experimental details and results.

Chapter 2

2 Review of the related literature

Individuals diagnosed with any type of voice disorder may face challenges in everyday life. Many aspects of life such as occupational, personal, and social functioning may be affected once the normal production of voice is disrupted and the individual's voice and its perceptual features such as pitch, loudness, and/or quality fall outside of a normal range for the speaker's age and gender. Within sections to follow, a variety of issues related to voice disorders and the measurement of such abnormalities will be addressed. This includes normal and disordered voice, voice quality evaluation, measurement considerations, factors influencing reliability of voice quality measurement, pupillometry and listening effort.

2.1 Normal voice characteristics

Normal voice quality represents the acoustic product of a normal larynx which houses the vocal folds and its interaction with the vocal tract (Mathieson, 2000). According to Boone, McFarlane and Von Berg (2005) normal voice production has 5 characteristics: loudness, hygiene, pleasantness, flexibility, and speaker representation. These terms are defined as follows: loudness - heard and understood above environmental noises; hygienic - no trauma or laryngeal lesions to the vocal mechanism; pleasantness - pleasant to listen to and pleasing in vocal quality; flexibility - flexible and capable of expressing emotions; and finally, representation – appropriate for speaker age and gender. For instance, if the voice is produced inefficiently or with strain to the mechanism, sounds unpleasant, is abnormally loud or soft, or causes the listener to misjudge the age or gender of the speaker, then it is said to be defective (Boone et al., 2005). The negative changes in the voice noticed by the listener, typically referred to as dysphonia have been described with many names such as hoarseness, harshness, breathiness, etc. and there is debate over the definition of these terms. Therefore, the more general term dysphonic is often used to refer to voices that exhibit these changes as a result of any vocal dysfunction.

Characteristics perceived in a specific voice which results in it being identified as disordered and how much it deviates from normal voice are challenging issues (Ferrand, 2011). The difficulties lie in defining and describing what characterizes a “normal” voice; this is not an easy task and voice quality is a multidimensional entity determined by a wide range of features including health status, age, sex, physical stature, culture, personality and region (Ferrand, 2011). Hollien (2000) notes that the environment, situation, mood, or emotional state also impact on voice. The characteristics of voice is also determined through listener perception. Wilson (1987) defines ideal normal voice as the one with proper oral and nasal resonance balance, suitable loudness, and appropriate fundamental frequency for the speaker’s gender, age and physical size. If these characteristics are more towards the ideal end and are beyond the normal ranges, the voice is perceived as superior. Anderson (1942) defines a superior voice as one with a pleasant fundamental frequency, clear pure tone and clear diction, vibrant, and produced with ease and flexibility. Considering vocal output factors such as loudness, pitch and quality along with physical aspects of voice production like pain, strain, fatigue, discomfort, lifestyle, and the amount and type of daily use can yield a comprehensive description of vocal normality. As a result, if the voice quality is clear, production is effortless, without strain, free of pain and fatigue, and pitch and loudness are age, sex and situation appropriate, and the speaker is content with the use of their voice for emotional, social and vocational purposes, the voice can be labelled as normal (Ferrand, 2011). Accordingly, if any of the aforementioned features are compromised for any reason, a voice disorder may exist (Ferrand, 2011).

Many factors can cause voice disorders, ranging from structural, respiratory, neurological, psychological, or problems related to overuse or inefficient voice production, to physical injuries, systemic diseases, use of some medications, and lifestyle and it is quite common to find multiple causes for dysphonia (Ferrand, 2011). Considering all the variables that determine the normality of the voice, and also all those that can be potential causes of dysphonia, is necessary for professionals who provide services to individuals with voice disorders to pursue comprehensive voice evaluation

2.2 Nature of voice disorder

Multiple factors influence voice disorders including those which are anatomical, physiologic, neurogenic, psychological, as well as changes in lifestyle and medications (Ferrand 2011; Boone et al., 2005). For instance, hoarseness can have a physical, structurally-based cause such as a lesion on one or both vocal folds that hinders normal vibration or it can occur secondary to a neurogenic disorder which paralyses one fold, or, it might be due to a hyperfunctional disorder after voice abuse, heavy use, and/or misuse (Boone et al., 2005).

Some individuals suffer from muscle control deficiencies in respiration, phonation, resonance and/or articulation due to an injury to the peripheral or central nervous system. Patients in this category may vary from children with cerebral palsy who struggle with respiration and voice control to an adult who is challenged with a motor speech disorder due to a stroke (Boone et al., 2005; Darley, Aronson, & Brown, 1975). Unfortunately, most of the breathing, voice and resonance impairments due to neurological problems cannot be remedied, thus the aim of the therapy and intervention is to minimize the effects and improve the functions to normal level as much as possible (Boone et al., 2005). Sometimes the disorder does not have any of the above mentioned causes and is merely due to misuse of the vocal mechanism leading to “hyperfunctional” dysphonia (Hillman, et al, 1989). This includes faulty production of voice, or lack of laryngeal coordination. Individuals may speak with hard glottal attacks (which is the abrupt onset of voice) or speak excessively loud or with unsuitable pitch level, or abuse vocal folds (Boone et al., 2005). These are all examples of vocal hyperfunction, which is defined as employing excessive muscle force, and physical effort in respiration, and phonation. Its continuation over time may cause changes in the vocal fold tissue such as the thickening of their inner margins, or creation of nodules or polyps and leads to organic voice disorders. As Boone et al. (2005) state, changes in the vocal and oral cavity size and configuration brought about by muscle contractions and relaxations have functional effects that may influence the quality of the voice. This statement emphasizes that much of the voice quality is determined by the resonating chambers of the airway from the larynx up to the pharynx, oral and nasal cavities.

Other causes of voice disorders include lifestyle (alcohol, tobacco, and drug use, poor diet, exercise habits, vocational requirements and style of voice use), psychological problems (stresses, mental, and emotional problems), as the influence of some medications and respiratory challenges such as allergies and asthma (Ferrand, 2011).

Medications prescribed for acute or chronic health issues may have possible side effects, some of which influence vocal function. This is mostly seen in older adults as they may have the increased likelihood of taking multiple medications. For instance, Ferrand (2002) reported that 12 of the 14 older women in their study who were on medications such as Estrace (estrogen replacement), exhibited reduced phonatory stability. One of the most common side effects of more than 500 medications is drying out the mucosa of the oral cavity which is called xerostomia (dry mouth). Finally, hormonal medications may change the fluid content and structure of vocal folds. Hormones such as androgen which is used in breast cancer chemotherapy or endometriosis or postmenopausal sexual disorders might result in permanent F lowering in women and coarsening of the voice. These side effects were reported in studies on some of birth control pills in the 1960s and 1970s (Amir & Kishon-Rabin, 2004). Although, modern oral contraceptives which include lower doses of estrogen and progesterone with less androgenic derivatives are reported as having less negative effect on voice, jitter and shimmer values on sustained vowels may be lower than those who are not on pills (Amir, Biron- Shental, & Shabtai, 2006). There are reports that taking pills may in some cases improve vocal stability (Amir, Biron-Shental, Muchnik, & Kishon-Rabin, 2003).

2.2.1 Neurogenic disorders

Neurogenic disorders result from damage to the central nervous system (CNS) or the peripheral nervous system (PNS). In fact, one of the early manifestations of neurogenic diseases is a change in an individual’s speech or voice (Duffy, 1995). Congenital disorders or injuries to the peripheral or central nervous system may cause muscle control deficiencies which impact respiration, phonation, resonance and/or articulation; such disorders are not limited to any specific age group. One of the most common neurological disorders of voice is the result of a vocal fold paralysis.

2.2.1.1 Vocal fold paralysis

The vocal fold paralysis is categorized as a flaccid dysarthria which results from damage to brainstem, or a compromise of vagus (cranial nerve X) nerve, or its branches the recurrent laryngeal nerve (RLN) and superior laryngeal nerve (SLN). The etiology varies from lesions of the vagus, RLN and SLN, trauma, and neuritis (Kelchner, Stemple, Gerdeman, Le Borgne, & Adam, 1999).

If the damage to the cranial nerve is somewhere in the route between medulla to the larynx, the paralysis will involve either a partial, unilateral, bilateral, or complete loss. Unilateral vocal fold paralysis (UVFP), which is due to direct nerve trauma or disease of viral origin (idiopathic) in the recurrent laryngeal nerve (RLN) on one side, is the most common type of laryngeal paralysis (Case, 2002). The left RLN is more prone to surgical or traumatic injuries than the right RLN due to its longer path. It travels down the neck and loops around the aortic arch in the chest and then it travels up to the larynx (Boone et al., 2005). When such injuries occur and the RLN is compromised on one side, the performance of the laryngeal adductor muscles especially the lateral cricoarytenoid muscles are disrupted and the paralyzed fold rests in the paramedian position which is neither fully closed nor abducted (Boone et al., 2005). As a result, the voice of patients with UVFP is perceptually described as breathy, hoarse, and of limited pitch and loudness due to incomplete vocal fold closure. Also, such individuals have a very short phonation time and are unable to speak loudly (Bassich & Ludlow, 1986; Södersten & Lindestad, 1990). The auditory-perceptual feature that characterizes this group is “breathiness” due to leakage of air through the glottis (Hammarberg et al., 1980; Södersten & Lindestad, 1990).

2.2.1.2 Spasmodic dysphonia

Another type of neurogenic voice disorder is termed spasmodic dysphonia. This rather uncommon voice disorder can be classified as a form of focal laryngeal dystonia (Case, 2002; Yeung et al., 2015) and refers to muscle spasms in the vocal folds. Spasm most frequently impact the adductor muscles of the larynx, but abductory or mixed muscle spasms also may occur in subgroups of this disorder. In the adductory variety, or adductory spasmodic dysphonia (AdSD), the vocal fold muscles experience involuntary sudden movements or spasms which interfere with vocal folds vibration and voice production. Those exhibiting this disorder are described to be noticeably trying to push the outgoing airstream via a tightly closed larynx and may exhibit phonatory breaks and a strained, strangled voice quality (Yeung et al., 2015).

2.2.2 Cancer of larynx

Cancer can occur in different parts of larynx. The symptoms include feeling a lump in the throat, hoarseness or change in voice, shortness of breath, weight loss, and problems in swallowing. Patients receive treatments such as radiation, chemotherapy or surgery based on the location and size of the cancer. When surgery is necessary, part or all of the larynx is removed. When the entire larynx is surgically removed, it is called a total laryngectomy (Doyle, 1994).

Laryngectomy patients are left with three alternative modes of speech: esophageal speech (ES), tracheoesophageal speech (TE) or using an electrolarynx (EL). Esophageal speech involves injecting the air into the esophagus and then expelling it through the pharyngoesophageal (PE) segment where vibration for phonation occurs (Doyle, 1994). TE speech, which was introduced in 1980s (Singer & Blom, 1979), involves placing a voice prosthesis which is a one way valved appliance in the wall between the trachea and esophagus. Pulmonary air is directed into esophagus through trachea and the prosthesis and then is used to vibrate PE segment for TE phonation. Some laryngectomies use an electro- or artificial larynx which is a device placed on the neck against the external throat. The device has a vibrating diaphragm. Once the patient puts it against the throat and pushes a button, the vibrating diaphragm creates a sound source that moves into the vocal tract were speech sounds are then articulated.

The rehabilitative success of TE is generally higher due to fluency and intelligibility rate. The objective measures of TE mode such as intensity and durational measures and frequency are usually close to normal ranges, but when judging acceptability and naturalness, they are clearly recognizable as different from normal laryngeal speech (Robbins, Fisher, Blom & Singer, 1984). In terms of the differences between TE and ES, the F_0 in TE speakers is closer to laryngeal speakers and TE speakers can maintain faster speaking rate with less pause time (Robbins et al., 1984; Baggs & Pine, 1983). They also speak with more intensity than ES speakers. The vowel amplitude in TE speakers is also reported to be higher than ES speakers which is due to the fact that TE uses a pulmonary air supply which has a higher air pressure and flow compared to ES (Most, Tobin, & Mimran, 2000). In terms of auditory-perceptual evaluations, TE speech is generally reported to be more acceptable than ES speech, but not always more intelligible than ES (Most et al., 2000). Some other perceptual evaluation studies of ES and TE report that TE speakers are more intelligible than ES speakers (Doyle, Danhauer, & Reed, 1988; William & Watson, 1987), some report the opposite though (Trudeau, 1987; Ng, Kwok, & Chow, 1997). In terms of acceptability, some report more acceptability for TE (Williams & Watson, 1987), but others have reported no difference between the two modes (Sedory, Hamlet, & Connor, 1989). TE voice is, therefore, commonly reported to be generally more intelligible, fluent and acceptable than ES and more similar to laryngeal speech (D’Alatri, Bussu, Scarano, Paludetti, & Marchese, 2012).

2.3 Voice quality

Voice quality has interested experts for many years, however, there have always been different ideas about what the term covers. There have been discussions into whether voice quality is restricted to aspects derived from vocal fold activity or those from the supralaryngeal settings of articulators (Kent & Ball, 2000). Voice quality has also been referred to as the interaction between the acoustic stimuli and the listener and, thus, has been measured through auditory-perceptual methods (Sofranko & Prosek, 2012).

2.4 The voice evaluation

Voice evaluation traditionally includes acoustic and aerodynamic measures, laryngeal imaging, self-assessment and auditory-perceptual evaluations (Coelho, Brasolotto, Fernandes, De Souza Medved, Da Silva & Júnior, 2017). Evaluation of the voice should take place to determine the nature and scale of the disorder and can ultimately been used to assess the success of treatment. Voice evaluation can be done either instrumentally or non-instrumentally. Although voice professionals are capable of making judgments of dysphonia without any instruments through subjective, perceptual evaluation, the use of objective measures such as acoustics will add important elements in describing the problem (Boone et al., 2005). As Boone et al. (2005) mention, physical parameters such as intensity, frequency, and airflow rate are achieved through instrumental approaches, whereas non-instrumental auditory-perceptual achieved give insight into loudness, pitch, and quality.

2.5 Why auditory perceptual evaluation?

> “Even though instrumental methods may be more quantitative and objective, they cannot stand alone. Perceptual judgement is necessary for a voice quality to be identified. Perceptual and instrumental approaches are complimentary and the central task is to bring them together" (Kent & Ball, 2000, p. 1).

Auditory-perceptual evaluation is the most commonly used clinical assessment method for disordered voice quality and it is even considered by some to be the gold standard for documenting voice disorders (Oates, 2009). Several factors contribute to the popularity of the auditory-perceptual evaluation. Because voice is basically a perceptual phenomenon in response to an acoustic stimulus, perceptually evaluating it is the best approach. As Oates (2009) explains, the perceptual nature of the voice features have a shared reality among patients, clinicians, and other professionals which makes perceptual descriptions intuitively meaningful and interpretable. For instance, describing a specific voice quality as “breathy” is more easily understood than describing it through instrumental measures such as the rate of airflow or the harmonic-to-noise ratio.

Voice quality assessment is effective in terms of both cost and time required for the evaluation. Patients are usually comfortable in providing samples and not much technical information is required. In terms of instruments, a good microphone, a high quality audio recorder and good quality headphones are all that are needed. The evaluation itself can be done relatively quickly; further, intensive, sophisticated training is not required other than familiarizing listeners with the method of auditory-perceptual evaluation task and adequate descriptions of what is being assessed (Oates, 2009).

2.6 Voice descriptors

It is important to define the features underlying voice quality in order to obtain the most consistent ratings. Various terms have been used to describe different features of voice quality. Some of these features are subjective, but some have objective correlates. The most commonly used auditory-perceptual features for describing voice quality and its components are described briefly in the subsequent sections.

“Pitch”: the perceptual correlate of fundamental frequency; “breathiness”: the audible escape of air between poorly approximated edges of the glottis that fail to make optimum contact; “aperiodicity”: a lack of consistency between cycle-to-cycle waveforms during voicing; the more dissimilar each cycle is from the preceding and following ones, the more noise exists in the signal; “phonation break”: a temporary loss of voicing which may occur at any point in an utterance; “loudness”: perceptual correlate of vocal intensity (Kempster, Gerratt, Abbott, Barkmeier-Kraemer, & Hillman, 2009); “hoarseness”: includes features of both breathiness and harshness characterized by irregular vocal fold vibration and additive noise (Kreiman & Gerratt, 1998). Hoarseness is the most commonly identified voice quality disturbance and anything that interferes with optimum vocal fold adduction can cause variable degrees of hoarseness. In some instances, those with hoarseness may compensate by increasing vocal fold closure. They may also adopt an abrupt initiation of voicing (i.e., glottal attack). Other descriptors include: “overall severity”: global, integrated impression of voice deviance and relates to the general impression created by a voice quality (how normal or abnormal it sounds); “roughness”: perceived irregularity in the voicing source; “strain”: perception of excessive vocal effort during glottal closure); “naturalness”: conformity with listener's standards of rate, rhythm, intonation and stress pattern (Eadie, Doyle, Hansen, & Beaudin, 2008); “acceptability”: the acceptability of the voice to the listener regarding pitch, rate, understandability and voice quality (Eadie et al., 2008); and “listener comfort”: listeners’ feelings of what it would be like to communicate with a speaker in a social context (O'Brian, Packman, Onslow, Cream, O'Brian, & Bastock, 2003).

2.7 Measurement considerations

An important point relative to a possible source of voice evaluation error pertains to the choice of rating scales (Eadie & Doyle, 2002). Many studies have been done on the suitability of scales (Kreiman & Gerratt, 2000). Interval or ordinal scales are not that suitable for evaluating voice quality, especially if the voice sample being rated is multidimensional and complex. In such situations, listeners have to focus and listen selectively for a specific aspect of voice such as breathiness or roughness, and then evaluate the extent to which the feature is present in a given voice (Kreiman, Gerratt, & Berke, 1994). Not surprisingly, listeners often have difficulty isolating single perceptual dimensions form complex stimuli (Kreiman & Gerratt, 2000).

An ideal auditory-perceptual method will reliably differentiate a normal voice from one that is disordered; it also is capable of tracking changes in patients’ voices across time, correlates with pathophysiology and objective measures, and is clearly established in terms of type of scale to be used and if appropriate, the anchors to be used, as well as the amount of user training required (Kempster et al., 2009).

2.7.1 Prothetic vs. metathetic continua

Before choosing a specific perceptual scale, one must determine if the dimension under study is a prothetic or metathetic dimension; this consideration addresses differences between quantity and quality, magnitude or kind, or size or sort (Stevens, 1975). Stevens (1975) defines a prothetic continuum as additive and quantitative in nature and recommends using direct magnitude estimation (DME) scales. DME is to be used with some perceptual dimensions because a prothetic continuum cannot be subdivided into equal intervals. In contrast, a metathetic continuum is defined as a substitutive, qualitative continuum which can be scaled with either DME or equal appearing interval

(EAI) scale; an example of this type of scale would be “pitch”. "Metathetic, positional, qualitative continuum seem to concern what, and where as opposed to how much" (Stevens, 1975, p. 13). Stevens (1975) also indicates that metathetic continua include smaller and less orderly categories of perceptual variables than do those that are prothetic.

Selecting and defining individual scales remain a critical aspect of scale development and use, to specify what is being measured, to justify why those aspects of voice and not others are of interest and to clarify the relationship among different scales. "Individual scales are usually validated by appeals to intuition, consensual validity, and face validity or by reference to their association with purported acoustic, aerodynamic and or physiological correlates" (Kreiman & Gerratt, 1998, p: 76).

2.7.2 Direct magnitude estimation (DME)

With DME scales, raters make perceptual evaluations relative to a standard or modulus. This modulus is usually given an arbitrary value (e.g., 100) and the judges are asked to perceptually evaluate a specific dimension of voice quality secondary to dysphonia. For example, features such as breathiness or roughness within a sample are assessed relative to the modulus and are then given a numeric value that is less than or greater than the modulus (Eadie & Doyle, 2002).

2.7.3 Visual Analogue Scale (VAS)

A visual analog scale (VAS) is usually an undifferentiated line scale (100-mm) with anchors/endpoints labeled. VAS may be used to evaluate any individual voice feature. The left end of the scale usually reflects normality for features such as severity or loudness or reflects no presence of the feature under assessment, for example, breathiness or strain. The right end section of the scale however, represents the judgement of the listeners of the most extreme presence of the given feature. Listeners rate samples within that range of judgment by bisecting the scale at a point that they believe best represents the voice feature under evaluation.

2.7.4 Equal Appearing Interval (EAI)

As Eadie and Doyle (2002) discuss, in perceptual evaluations with EAI scales, ratings are provided based on a predetermined, fixed interval scale. The scale usually has 7 points plus or minus two points (i.e., 5 point or 9 point). The perceptual distance, weight or magnitude between scale points are assumed to be equal. The left end point, which is 1, usually implies normality and the right end point, for instance 5, 7, or 9, reflects the severity extreme. Some differences between EAI and DME scales include the fixed end points in EAI and the use of whole number ratings.

2.7.5 Paired comparison

The other type of perceptual task to measure perceived voice quality is the forced-choice, paired comparison procedure. With this approach, listeners are presented with two stimuli and then asked to compare the two on the extent of the difference or similarity on some dimension (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). In doing so, the judge is forced to select one sample from the pair that best represents the feature under assessment.

2.7.6 Reaction time to process stimuli

One way of studying the impact of a disordered voice on the listener is by measuring reaction times (RTs). RTs represent an index of cognitive workload which is placed on the listener during the task (Evitts et al., 2016). When performing any task, the listener's cognitive system undergoes an amount of mental demand which is called the cognitive workload (Evitts et al., 2016). The additional cognitive workload is attributed to the increased time required by the listener to extract basic acoustic-phonetic data from the dysphonic speech due to the altered nature of that and the task requested from the listener (Evitts et al., 2016).

2.7.7 Instruments in common use

2.7.7.1 GRBAS

This approach, which was developed by Hirano in 1981, is widely used and represents five dimensions of phonation which are: grade (G), roughness (R), breathiness (B), asthenia (A), and strain (S). The GRBAS procedure uses a 4-point Likert-type scale in which 0 represents normal and 3 represents extreme for each of the 5 parameters. It is a relatively fast method of auditory-perceptual evaluation and has been widely used as an auditory perceptual assessment tool (Zraick, Kempster, Connor, Thibeault, Klaben, Bursac, & Glaze, 2011).

However, there is no published standardized protocol to be followed in English for GRBAS (Kempster et al., 2009). Also, only an ordinal judgement on a four point scale (normal, mild, moderate & severe) are available for GRBAS which clearly limits the application of this scale in research design (Kempster et al., 2009). GRBAS is an ordinal scale that does not allow parametric statistical analysis, a problem that has been identified as one of its limitations (Zraick et al., 2011). There are also many other criticisms with this instrument regarding the influence of task order, type and amount of listener training, and variability of listening samples on the reliability and validity of voice quality judgements (Kreiman et al., 1993). Additionally, because of the restricted nature of the GRBAS scaling method (i.e., 4 points), concerns about accurately assessing the reliability of ratings within and across listeners is of concern. These concerns led to the development of a newer tool for voice perceptual measurement which uses continuous scaling.

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

The Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) measure was developed in 2000 under Special Interest Division 3 (Voice & Voice Disorders) of the American Speech-Language Hearing Association (ASHA). It was developed with the hope of creating a standardized, valid and reliable clinical and research tool for the perceptual measurement of voice quality. It is standardized due to the consistency in its administration and scoring procedure (Kempster et al., 2009). It is a continuous VAS which evaluates six parameters: overall severity, roughness, breathiness, strain, pitch, and loudness. Authors of the CAPE-V agreed on a set of principles: “dimensions should reflect a set of clinically meaningful perceptual voice parameters, procedures should be obtained easily and quickly, should be applicable to a broad range of vocal pathologies and clinical settings, ratings should be demonstrated to optimize reliability within and across clinicians through later validation studies and anchors must be considered for training and future use" (Kempster et al., 2009, p. 126).

2.7.7.2.1 CAPE-V tasks

Assessment tasks using the CAPE-V are based on samples of sustained vowels, six sentences with different phonetic contexts, and finally, conversing in response to interview questions (e.g., "tell me about your voice problem"). Vowels provide information without articulatory influences. Each of the six sentences has been developed for a specific reason: 1) "The blue spot is on the key again" to examine the coarticulatory effect of three vowels (/a, i, u/), 2) "How hard did he hit him?" to evaluate soft glottal attacks and voiceless to voiced transitions, 3) "We were away a year ago" having all voiced phonemes, it provides a context to assess possible voiced stoppages/spasms and speakers' ability to "link" (i.e., maintain voicing) from one word to another, 4) "We eat eggs every Easter" has several vowel-initiated words and may provoke hard glottal attacks and provides the opportunity to assess whether these occur, 5) "My mama makes lemon jam" includes numerous nasal consonants, providing the context to assess hyponasality and possible stimulability for resonant voice therapy, 6) "Peter will keep at the peak" having no nasal consonants, this sentence provides the context for evaluating intraoral pressure and possible hypernasality or nasal air emission (Kempster et al., 2009).

The six voice quality features selected for consistent appraisal in CAPE-V are labeled and defined as follows: "Overall severity": global, integrated impression of voice deviance, "Roughness": perceived irregularity in the voicing source, "Breathiness": audible air escape in the voice, "Strain": perception of excessive vocal effort (hyperfunction), "Pitch": perceptual correlate of fundamental frequency, "Loudness": perceptual correlate of sound intensity (Kempster et al., 2009).

The reason and rationale for including these six auditory-perceptual features is that both clinicians and researchers find them meaningful and they have appeared in literature for decades (Kempster et al., 2009). Kempster et al. (2009) also mentions that the feature “hoarseness” is excluded since it is perceived by many as a combination of “roughness” and “breathiness”. The CAPE-V also includes two unlabeled scales which allow for documentation of other salient perceptual features (i.e., spasm, tremor, degree of nasality, falsetto, intermittent aphonia or glottal fry).

One advantage of CAPE-V is that it may offer more sensitivity to small differences in each voice quality dimension (Kempster et al., 2009). Its other advantages are the defined elicitation protocol and the use of a consistent conversational speech probe, and inclusion of phonologically diverse speech contexts (Zraick et al., 2011). In addition, both prothetic and metathetic continua can be assessed (Zraick et al., 2011).

Zraick et al. (2011) compared CAPE-V and GRBAS to evaluate interrater and intra-rater reliability of experienced voice clinicians' judgements using the two tools. They also tried to establish the empirical validity of CAPE-V by examining the relationship between the two. 21 experienced raters (16 female & 5 male) were selected based on specific criteria regarding the type of training and experience level. They evaluated the 22 normal and 37 disordered voices using both CAPE-V and GRBAS.

Their results showed that the interrater and intra-rater reliability coefficients for CAPE-V are a bit higher than those for GRBAS. In terms of empirical validity, there is strong correlation between both scales which means the CAPE-V is a valid tool that measures similar constructs of voice quality. "Strain" was the least perceptually salient dimension and "asthenia" had the highest intra-rater reliability value in their study. Zraick et al. (2011) had the largest number of experienced raters to date. Having more experienced raters introduces variability into the evaluation (Kreiman & Gerratt, 2000), as these raters use various strategies to make their perceptual evaluations and their assessment is constantly fine-tuned along the way (Kreiman & Gerratt, 2000). Their raters were all voice specialists, had diverse background and training which reflects inconsistencies compared to a cohesive group. The evaluations were also made during two different sessions to eradicate fatigue and possible order effects. However they did not age and gender match the normal and dysphonic voice samples that were assessed and only 11 of the stimuli were repeated for the intra- rater reliability assessment and their raters were all experienced so naïve raters' performance with CAPE-V is not included.

Nemr et al. (2012) also investigated if applying both GRBAS and CAPE-V to the same stimuli at different times would yield the same reliability and consensus. The reproduction of the six sentences, sustained vowels /i/ and /a/, and the spontaneous speech in response to "tell me about your voice" by 60 subjects were analyzed using CAPE-V and GRBAS by three expert SLP raters. They report similarities in dysphonia distribution values and a high degree of correlation between the two tools. The judges indicate that GRBAS is faster to apply and CAPE-V is more sensitive, especially for identifying small differences. CAPE-V is also more complete as it allows evaluation of additional parameters. One interesting point about this study is that they used the Portuguese-translated version of CAPE-V and the findings are similar to the original English version, a finding which supports the fact that CAPE-V may be used in different populations with different languages.

2.8 Reliability vs. agreement

Reliability and agreement are not the same in statistical terms. "Listeners are in agreement to the extent that they make exactly the same judgements about the voices rated" and "Ratings are reliable when the relationship of one rated voice to another is constant and voice ratings are parallel or correlated" (Kreiman et al., 1993). When listeners are in agreement, they assign identical meanings to each point on the scale point and they have the same idea of what defines, for example, extreme breathiness. Their definition of normal and extreme are the same and the rater's idea of the distance between intervening points on the scale are the same (Kreiman et al., 1993). When listeners judge voices in a parallel fashion, ratings are reliable. This does not imply that the scale values have the same meaning for the listeners. However, good agreement does not necessarily guarantee good reliability.

If the ratings range is restricted, for instance due to not much of a variation with regard to the quality being rated or due to avoidance of end points on an EAI scale, reliability may be low but the judges might be in good agreement. High intra- rater agreement is required from raters, however, when it comes to between rater evaluation, reliability is critical (Kreiman et al., 1993). When comparing raters, that may be done in pairs or by overall coherence of the entire listener group. Most often, Pearson's r is calculated across all possible pairs of raters and is reported as a single averaged value. This across-all-pairs evaluation reveals those raters who disagree with the majority and also the extent of their disagreement.

2.8.1 Factors influencing ratings & reliability

2.8.1.1Stimuli

2.8.1.1.1 Different types of stimuli: Sustained vowel, Sentences, Conversation

One of the issues in the literature on auditory-perceptual evaluation is which type of stimulus yields the best intra-rater and interrater reliability. Each stimuli type looks at a specific aspect in speakers' voice/speech. The differences between the voice tasks create perceived differences in the type and degree of judged severity. In fact, type and severity of voice quality also differs between stimuli, for example, sustained vowel and continuous speech (Barsties & Maryn, 2017). The advantage of continuous speech is that it resembles everyday conversation and represents the dynamic features of voice in daily speech such as vocal fluctuations during voicing onset and termination, and differences in frequency and amplitude (Gerratt, Kreiman, & Garellek, 2016). However the quality ratings are more variable because of non-vocal phenomena such as phonetic context, or prosodic fluctuations are present (Barsties & Maryn, 2017).

On the other hand, sustained vowels are free from such phonetic variability but are not a good representative of everyday conversation and voice use patterns (Barsties & Maryn, 2017). Sustained vowels are time invariant, not under the effects of phonetic context, stress, intonation and speaking rate, are easy to elicit and evaluate, and they are not influenced by dialect (Gerratt et al., 2016). Sustained vowels also are held at relative constant subglottal, glottal and supraglottal pressure levels. Voice onsets, voice terminations, vocal pauses, voiceless phonemes, phonetic context, prosodic fluctuations in F_0 and intensity and speech rate may cause temporal and spectra variations. As such, some studies report a higher reliability for oral reading and connected speech (Bele, 2005; De Krom, 1994). Brinca, Batista, Tavares, Pinto and Araujo (2015) also have reported high levels of agreement and reliability for oral reading and connected speech stimuli.

2.8.1.1.2 Length of the stimuli (in seconds, number of syllables, number of words)

Barsties and Maryn (2017) investigated the effect of different stimuli length on the degree of severity in voice quality judgements. They had three different stimulus lengths: 17, 35 and 93 syllables. Reliability results for ratings of severity were significantly different between the 17 and 35 syllable samples, but not between the 35 and 93 syllable samples. They suggest that speech material can be reduced in length as a possible option. The authors conclude that the 17 syllable length is not sufficient "because significant differences were found in extended length of continuous speech" (Barsties & Maryn, 2017). Their results showed that shorter length samples are judged to be less severe than longer ones. In fact, the longer the stimuli, the more chances of phonetic variability to appear (Barsties & Maryn, 2017). In addition to the above points, perceptual context might also create changes in ratings. For instance, if judges are asked to evaluate moderately rough voices after listening to and rating several mildly rough voices, the moderately rough voices are assessed to be more severe due to the shift in the raters' internal standards (Coelho et al., 2017).

2.8.1.2 Speaker factors

2.8.1.2.1 Speaker gender

As a means for communication, human voice carries information about the speaker which is used by listeners to make some estimates about speakers' personality and physical characteristics, which at times may not be precise (Amir & Levine-Yundof, 2013). Speech and communication disorders influence the judgement by listeners about the speaker. People with such features are usually evaluated negatively, as less intelligent, less educated and capable, emotionally unstable and more aggressive and stressed (Amir

& Levine-Yundof, 2013). Doyle (1994) notes that because of the stigma associated with the perceptual characteristics of alaryngeal voice and its deviation from society's standards of normality, such speakers have the risk of being socially penalized.

From a listener's perspective, gender identification is the result of comparing a speaker's voice to a form of internal referent or prototype stored in their mind for each gender. Although both men and women suffer the negative impacts of voice disorders, there is specific concern for females in social context. That is, they may be more likely to be penalized compared with men because of the Western society's more inflexible and higher standards of femininity (Newell, 2007). In fact, according to Haeberle (1981) the standards of society requires women to be feminine and not just female.

With respect to the aforementioned attitudes toward people with disordered voice and speech, gender considerations are of importance. Amir and Levine-Yundof (2013) examined listeners' attitude toward dysphonic people with an attempt to evaluate effects of gender and age on the attitudes. Their study used multiple male and female speakers. Their 26 male and 48 female naïve listeners grouped in two groups of younger (age ≤ 40) and older (age>41) judges, evaluated 3 male and 3 female dysphonic patients and 6 matching non-dysphonic individuals who read the Hebrew passage "Thousand Island". They used a semantic differential questionnaire which included 12 seven-point rating scales (positive-negative, healthy-ill, etc.) or contrastive adjectives representing three underlying attitude factors (evaluation, potency, activity). The results of factor analysis indicate that dysphonic speakers were rated lower on "Evaluation" and "Potency" factors and higher on "Activity". Being rated higher than non-dysphonic speakers on the activity factor which includes personality traits (tense, aggressive, etc.) means that dysphonic patients were perceived to have more negative personality traits. The authors discuss that although in Israel 16% report to have, and 34% reported that they have had voice problems, and while raters are familiar with voice issues or had these issues themselves, they still held a solid negative attitude towards those with voice disorders. One interesting result of this study is that in non-dysphonic speakers, women are rated more positive than men on a variety of their scales such as healthy, successful, sexy and calm, but when it comes to rating dysphonic women, this inclination toward females is reduced and they are rated lower and are penalized more than men and are more severely affected by listeners' judgement. In fact, dysphonic voices are usually characterized by low pitch and it affects females more than males because it is deviation from the stereotype; this may be the reason why women seek voice therapy more often than men (Amir & Levine-Yundof, 2013).

Eadie et al. (2008) investigated gender identification and influences of gender identification on listener judgement. Their subjects were tracheoesophageal (TE) speakers. They made an effort to collect information on listeners' preference, listeners' performance on determining gender of patients and also their judgements of the multidimensional features of “acceptability” and “naturalness” (speech rate) of TE speakers with and without knowing their gender. They report the hierarchical ranking of speakers based on the preference score which does not show any preference for speaker gender in this study. They also report correct identification of all the 6 male but only 2 of their 6 female and they relate the perfect male identification to a set of collective attributes emerging from frequency amplitude, temporal domain or the combination of them. However, two correctly identified females had the F_0 within the range for females and combination of rate and F_0, respectively, which might have assisted in correct identification. Eadie et al. (2008) also report that females were judged to be less acceptable and less natural and males were evaluated to be more natural when gender was known. These results identify that gender is critical in determining a speaker's acceptability as their female speakers were penalized once gender was revealed. It was also suggested by Eadie et al. (2008) that when listeners evaluate male and female TE speakers, they likely compare them with an internal referent formed by their experience with normal laryngeal speakers. Since female TE speakers may deviate more from laryngeal speech than their male counterparts, they may be judged to be more severe and less natural and acceptable.

2.8.1.2.2 Speaker age

Voices change as the result of aging process. These age-related anatomic, physiologic changes are normal and the specific characteristics of aging help distinguish normal from pathological features of the voice. As a person ages, the thoracic cage undergoes structural changes which leads to chest wall compliance reductions. If the rib cage is calcified due to osteoporosis, this may lead to a reduction in thoracic vertebrae height, kyphosis and hunching of the back and this process leads to thoracic stiffness. As a result, gradual reduction in thoracic cage expansion capability is experienced during inspiration which disturbs the diaphragm effective contraction (Sharma & Goodwin, 2006). Sharma and Goodwin (2006) also mention that respiratory system or the chest wall and lung compliance undergoes changes with aging. Chest wall compliance is a change in volume relative to change in pressure. The elastic load during inspiration is determined by thoracic compliance and the expiration rate and force is influenced by lung compliance (Sharma & Goodwin, 2006).

Some of these features are gender variant. For instance, decreases in the strength of respiratory muscles is observed more often in men than women (Sharma & Goodwin, 2006), but vocal fold edema is mainly observed in older women due to hormonal changes during menopause (Linville, 2000).

In terms of acoustic alterations as people age, Linville (2000) reported changes in speaking fundamental frequency (SF_0), maximum phonational frequency range (MPFR), stability of SF_0 and amplitude, and jitter and shimmer. Changes in SF_0 occurs in both men and women throughout their lives, though less prominently for women. In men, it lowers from early adulthood into middle ages and then it increase as they reach older ages. The drop is reported to be related to normal vocal use and the increase is due to muscle atrophy and changes in vocal fold tissue stiffness (Linville, 2000). The SF_0 pattern is quite stable after 20 years of age in women until around age 50 when menopause may exist at which point a drop in SF_0 is experienced because of hormonal changes causing vocal fold edema (Linville, 2000).

Hormonal changes in middle-aged women which leads to increases in vocal fold mass enables women to produce lower frequencies than young and aging women. However, when it comes to higher frequencies, females face limitations due to vocal fold mass changes, weakness of the muscles and calcification and ossification of cartilages of larynx (Linville, 2000). The voices of older adults is also reported to display increased vocal fold vibration instability, vocal tremor, pitch breaks, harshness and hoarseness due to anatomical and physiological changes. In addition, the mean jitter and shimmer values are higher for older compared to younger adults (Linville, 2000).

Anatomical changes due to aging can bring about changes in the voice. Age induced thyroarytenoid weakening may cause imperfect closure from the vocal processes to the anterior commissure and introducing a posterior chink (Linville, 2000). It is not clear exactly why laryngeal muscles atrophy, but it is reported that long term voice use might weaken the muscles (Linville, 2000). However, there are contrary reports regarding this issue. Given the anatomical changes, older women are expected to have high incidence of posterior chink, however, it is young women who display a high incidence. Young females may physiologically choose not to achieve complete vocal fold closure for functional reasons with the purpose of introducing a bit of breathiness into their voice quality (Linville, 2000). Some of the other anatomical age related changes include losing teeth, temporomandibular movement restrictions, facial skeletal growth, tongue musculature atrophy or hypertrophy, pharynx musculature atrophy, larynx lowering due to ligament stretching and atrophy of neck strap muscles and vocal tract lengthening in women (Linville, 2000).

The glottal gap deficiencies in older adults have consequences. Some men increase the adductory forces to compensate for the gap, but they end up being perceived as having strained voice. The glottal closure failure also leads to shorter syllables per breath group which is confirmed by the fact that many older individuals require more breath pauses than young people while talking (Linville, 2000). The rate of speech in aging individuals is also influenced by neuromuscular slowing and alterations in the respiratory system (Linville, 2000). Loud phonation also is difficult to achieve due to loss of elasticity recoil of lung tissue.

These age related processes bring about changes in voice which are considered normal and influence the voice quality of every individual and the perception of the listeners of the speakers. In fact, acoustic properties form listeners' judgement of speakers' personality, emotional state, cognitive ability and physical characteristics and judgement of speech quality carries information on various characteristics of the talkers such as social pathological condition and social characteristics like age and gender. In addition, such perceptions shape the interactions between speakers in a conversation. For instance, if one of the speakers in a conversation is cognitively impaired or hard of hearing, the other speaker may adjust by using exaggerated intonation and a slower speech rate (Goy & Pichora-Fuller, 2016). Knowing normal age related changes of voice is important in contextualizing data gathered from individuals who experience pathological conditions that affect their voice.

As discussed earlier, voices age and they display specific characteristics. Aging females have a lower F_0 compared to younger females and they also have a longer maximum phonation time (MPT) compared to the younger females and older males who talk at a slower rate and have more shimmer in their speech (Goy & Pichora-Fuller, 2016). When judging both age groups, even if there are not big differences between older and younger speakers acoustically, the older speakers are judged as having reduced vocal quality and less precise articulation, less normal, less powerful, and are less engaged, and present with more negative personality stereotypes than younger speakers. Such biases were specifically observed when listeners are informed of speaker age (Goy & Pichora-Fuller, 2016).

2.8.1.3 Listeners (Judges) factors

In order for perceptual evaluation of dysphonic voices to be meaningful and interpretable, raters or judges listen to them and indicate their ratings of particular features. The question which is then raised is who is the best listener or judge to perceptually evaluate voice and what characteristics should they assess? Many studies have used different types of raters from "experienced judges" to "naïve listeners" (Doyle, Swift, & Haaf, 1989; Eadie & Doyle, 2002; Evitts et al., 2016; Helou, Solomon, Henry, Coppit, Howard, & Stojadinovic, 2010; Sofranko & Prosek, 2012; Zraick et al., 2011). Graduate and undergraduate students in speech pathology and otolaryngology specialists have been most commonly used for evaluations. There is no evidence regarding the optimum number of judges needed for evaluations; Kreiman et al. (1993) report that in the 57 papers reviewed, the number of listeners varied from 1 to 461, but many (60%) used only between 3 and 12 (Carding, 2000).

2.8.1.3.1 Listeners' experience with the voices

When rating voices, listeners compare the stimulus to an internal standard. As Kreiman et al. (1993) state, these internal standards represent average or typical samples for different quality levels according to the experience of listeners. These experiences shape the choices of listeners and where along the severity continuum they place their internal standard. Sofranko and Prosek (2012) mention that an experienced listener is someone who has worked in the field of voice for more than a third of their career and over a 3 year period. The amount of detail in internal representation of voices features and the severity level in memory is different for every listener. For normal and near normal voices all listeners have extensive and almost equal experience because of every day contacts. It is the assessment of intermediate voices which is controversial among judges (Coelho et al., 2017). Experienced listeners have different internal standards due to various levels of exposure to pathological voices (Kreiman et al., 1993). They appear to use a flexible strategy to determine prominent perceptual characteristics and make ongoing adjustments and fine tune their decisions (Coelho et al., 2017). This makes them use different rating strategies (Coelho et al., 2017). On the other hand naïve listeners do not have significant formal exposure to pathological voices and, therefore, may lack specific standards to judge voice quality. For naïve listeners the judgment is mainly based on standards for normal voices. These internal standards are unstable and variable.

Level and type of experience is reported to have an impact on quality judgements of synthesized voice (Sofranko-Kisenwether, & Prosek, 2014). Those investigators evaluated 6 groups of 10 listeners each who had various levels and types of experience. Raters used the CAPE-V to evaluate the synthesized vowel /a/. They report that of all rater groups, listeners with a singing background rated the stimuli as sounding more severely than those with speech pathology backgrounds. For that reason, Sofranko-Kisenwether and Prosek (2016) indicate that experience with dysphonic voices might desensitize perception of the degree of dysphonia.

The effects of level and type of experience on response time when judging synthesized voice quality was also studied by Sofranko et al. (2016). Their 60 experienced and naïve listeners evaluated the synthesized stimuli. The stimuli were systematically altered on the components of jitter, shimmer, noise to harmonics ratio (NHR) for parameters of overall severity, breathiness, roughness, pitch and strain. They reported a significant effect for group and type of stimulus on response time. Their results showed a longer response time for experienced raters. They also took a longer time assessing stimuli that were altered on two acoustical components, jitter and shimmer, compared to NHR. This can be an evidence that judging stimuli with multidimensional nature takes more time. They conclude that naïve listeners who do not have much information may make "snap" judgements compared to those who rely on their varied experience and information, thus, potentially listening for additional components in the signal.

Helou et al. (2010) also looked at the effect of experience on the way judges perceptually evaluate and rate voice quality. Their 10 experienced and 10 naïve listeners rated 21 post-thyroidectomy voices using CAPE-V. Their results show that ratings by the experienced raters are less severe than the naïve group. These authors attributed this to the fact that naïve judges have limited exposure to disordered voices and so they may rate the voices as maximally severe, a process that is not the same for experienced raters. Prior exposure of experienced listeners to more severe voices had probably lead them to rating the stimuli as less severe. According to Helou et al. (2010) expert judges use a variety of strategies for perceptual evaluation such as feature assessment and attention to idiosyncratic characteristics of voices, and skills developed through experience. Further, experienced raters were also shown to demonstrate higher levels of interrater reliability.

Sofranko and Prosek (2012) report significant differences across their three groups of listeners (speech language pathologists, singing teachers, and inexperienced listeners) who were asked to classify voice samples as breathy, rough or normal. The SLPs demonstrated substantial interrater agreement (0.67), the singing teachers with a moderate level of agreement (0.53), and inexperienced listeners with fair interrater agreement (0.24). As the results indicate, inexperienced listeners had the lowest interrater agreement results.

Suhail, Kazi, and Jagade (2016) differentiate between experience and professional backgrounds for listeners. They report that a professional background has a larger influence on perceptual evaluation than experience. They also note that trained expert raters use the range of the perceptual scales better which indicates that these raters differentiate more between various perceptual aspects of voice quality. It is recommended that for research purposes where the perceptual evaluations are used as a standard and other measures will be compared to them, experienced raters should be used (Suhail et al., 2016).

2.8.1.3.2 Individual perceptual habits and biases and overall sensitivity to the judged quality

Listeners' (dis)agreement with one another in ratings of voice quality is one of the sources of unreliability. When asked to make auditory-perceptual evaluations of voice quality, listeners usually compare the stimuli to be rated with their internal standards. These internal standards are formed over years and with experience and exposure to voices. They are, however, unstable and change constantly; this is why auditory anchors are used to minimize the context related variability and as a result increase reliability of judgments (Helou et al., 2010; Kreiman, Gerratt, & Precoda, 1990).

Voiers (1964) notes that "there are independent auditory perceptual channels available to every typical listener for information to the identity of a speaker". He also mentions that there are some extra-stimulus factors in listeners' perceptual responses to voices. Listeners' biases which are the constant errors and the constant errors of interaction are listener idiosyncrasies" manifest only for specific combinations of speaker and listener" or interactions between listeners and voice samples (Voiers, 1964). Kreiman, Gerratt, Precoda and Berke (1992) also report that different listeners, whether naïve or experienced who judge the same voice, evaluate different cues and acoustic parameters. However, listeners even within the same group are not the same in how they use vocal features for evaluating different voice quality. In fact, listeners deviate from an average perceptual strategy with respect to relative importance given to these perceptual characteristics (Kreiman et al., 1992).

Kreiman et al. (1993) discuss, listeners' internal standards for normal and near normal voices are relatively similar. This similarity is due to extensive and almost equal exposure to normal voices through everyday life and normal speakers. The internal standards for pathological voices though, differ from listener to listener. Naïve listeners who do not have internal standards for pathological voices, apparently rate dysphonic speakers according to internal standards that are more appropriate for normal voices (Kreiman et al., 1993). The sensitivity and the internal standards of the listeners interacts with the scale resolution and mismatches create variability in the results. A multidimensional quality (e.g., breathiness or roughness) must be evaluated on a multidimensional scale. If not, listeners selectively focus on one dimension or another and reliability is decreased (Kreiman et al., 1993). Therefore, variability in voice quality evaluation can be reduced by providing a constant set of perceptual referents which replace the idiosyncratic unstable internal standards for various voice qualities (Kreiman et al., 1993).

2.8.1.3.3 Listener age

Investigating the differences between younger and older listeners' judgement of perceptual voice quality is also of interest. In choosing listeners, caution should be exercised regarding their characteristics and inclusion criteria. Older and younger people have different social experiences and internal standards, hearing abilities, and have different expectations and reaction to young and old voices. In addition, age related alterations in the auditory system may influence their perception of speech (Goy & Pichora-Fuller, 2016). Goy and Pichora- Fuller (2016) investigated the effects of listener and speaker age on speech and voice quality perceptions and report that both speaker groups (younger adults: mean age of 19.0 and older adults: mean age of 71.3) were perceived similarly on most features except age. Both listener groups (younger adults, mean age of 19.06 and older adults, mean age of 74.1) rated younger voices as more pleasant and less rough compared to those of older speakers. Younger listeners were more exact at guessing age, but older listeners were more exact at gender identification than younger listeners. They conclude that age of listeners influences some of the talker characteristics' evaluations.

However, Amir and Levine-Yundof (2013) reported no influence of listeners' age on their judgements of dysphonic voices based on 10 of 12 perceptual scales. The only two scales rated differently by older and younger listeners were “healthy-ill” and “positive-negative”, indicating that older raters were more tolerant of dysphonic samples. This suggests that listeners have consistent attitudes towards dysphonia regardless of their age (Amir & Levine-Yundof, 2013). Among the features judged by listeners, guessing the age of speaker is reported to be influenced by the speaking fundamental frequency (SF_0). The accuracy rate of judging age from phonated vowels is much higher compared to that based on whispered vowels by younger listeners. Elderly listeners require both voicing and resonance for such judgements (Linville & Korabic, 1986).

Older listeners use a large variety of age related information to evaluate speakers than younger raters. Goy and Pichora-Fuller (2016) have reported that older listeners rely more on speech than voice information for identifying age. In fact, older and younger listeners are reported to be different in choosing which auditory cues they rely on for judging age and gender and all of this can be due to differences in social experiences and age-related changes in hearing. Regarding the age factor, it is better not to reveal speakers' age in order not to create any biases in the perceptual evaluations. Although age of the listener is not that influential on their perceptual evaluations, researchers are also advised to pick judges of various age groups depending on the purpose of the study. Also, since hearing ability also goes through changes with aging, hearing abilities should be taken into consideration. Different age groups also have different experiences, internal standards and expectations which can impact their judgements.

2.8.1.3.4 Listener gender

Differences between male and female and gender expectations and their influence on voice quality may provide information both regarding the individuals with voice disorders and also the way male and female listeners evaluate voices. Amir and Levine-Yundof (2013) studied listeners' attitude towards people with dysphonia and reported that in terms of the influence of listener gender on judgements of dysphonic voice, no specific differences between male and female listeners are observed. They arranged the responses according to listener's age group (younger/older) and gender (male/female) and voice (dysphonic/non-dysphonic) and repeated measures *ANOVA* revealed no significant influence for listener gender and age.

2.8.1.3.5 Other factors

2.8.1.3.5.1 Hearing impairment

The use of listeners or judges with hearing loss to perceptually evaluate voice quality has not been widely studied. Hearing impairment acts like a filter that may impact how the listener receives and perceives certain features depending on the type and degree of their impairment. Hearing loss delays or impairs development of speech perception and hinders the process of decoding utterances, especially unfamiliar ones, which is a real obstacle in speech perception (Pittman, Vincent, & Carter, 2009). Certain acoustic elements are inaudible due to the hearing loss and, therefore, such patients are not capable of perceptual evaluation (Pittman et al., 2009).

Hearing loss at any age or degree can create communication problems for the patients and influences their speech perception, therefore; such patients may not be reliable judges of voice quality. Briefly, relative to some consequences of hearing loss, it can be noted that speech consists of time-varying acoustic cues and hearing loss due to, for instance, aging adversely influences the ability to process such temporal cues. Speech perception is also dependent on multiple spectral, temporal and intensity cues. Hearing loss can also result in inability to process voice onset time (VOT) potentially resulting in problems distinguishing voiced and voiceless sounds (Trembly, Piskosz, & Souza, 2003). With sensorineural hearing loss, the use of hearing aid or the amplification does not fully compensate for the loss. Harkrider, Plyler and Hedrick (2009) also report that hearing loss influences identification and neural response patterns of stop-consonant + vowel stimuli (CVs). Hearing loss also influences identification of F2 formant transitions, leading to decreased audibility and distortion of the sample being judged.

2.8.1.3.5.2 Language (Bilingualism)

Shifting languages, from native language (L1) to a foreign or second language (L2) may influence the individuals’ voice production. The effects can vary from influences on the voice source, perceptual voice quality, and a change in mean fundamental frequency and pitch range to other challenges such as mental stress and vocal fatigue due to increased phonatory and articulatory effort (Jarvinen, Laukkanen, & Geneid, 2017). According to Jarvinen et al. (2017) speaking L2 may be more loading than L1. There is mechanical load on vocal folds which increases with intensity, fundamental frequency (F_0) and degree of adduction or phonatory type. Degree of adduction and phonatory type refer to the same concept, for instance breathy voice involves low adduction and pressed or strained voice corresponds to high adduction (Jarvinen et al., 2017). Jarvinen et al. (2017) investigated if the perceived phonation is more pressed when speaking L2 than L1. They also studied if voice source features are different in L1 and L2 and if there is more L2 vocal fatigue in L2 than L1.

Based on a questionnaire which asked about vocal fatigue in L2, Jarvinen et al. (2017) selected 12 subjects who responded “yes” and 12 who reported “no” vocal fatigue when switching to L2 (each having 6 male and 6 female). They also had equal number of native and second language speakers of English/Finnish. They report obvious perceptual differences between speaking L1 and L2. L2 speech had poorer voice quality and was more stressed and strenuous (Jarvinen et al., 2017). They also suggest that the perceptual evaluation of some characteristics depend on the language background of listeners. For instance, asthenia, roughness and strain evaluations are influenced by language background, but breathiness is not (Jarvinen et al., 2017). They report correlation of acoustic features with perceptual evaluations. There is a decrease in normalized amplitude quotient and closing quotient which represents pressed and strenuous L2 voice. Decreased amplitude quotient also shows raised pitch and pressed phonation increases vocal loading which results in vocal fatigue. However, pressed phonation alone does not always point to vocal loading because voicing time, F_0, and intensity are influential factors too. Finally, experience and proficiency in L2 plays a crucial role as lack of experience enhances psycho-physiological stress and mental effort and this can cause muscle tension and increased F_0 and pressed speech and it causes feeling of vocal overloading (Jarvinen et al., 2017).

2.8.2 Increasing reliability

2.8.2.1 Use of descriptive anchors

Factors like the individual's memory and the acoustic context in which the voice is evaluated influence idiosyncratic unstable internal standards (Brinca et al., 2015). As a result, external anchors are recommended to minimize the effects of these internal standards and studies show these external standards do in fact improve reliability and agreement among listeners even the ones with diverse backgrounds (Gerratt et al., 1993). Barsties et al. (2017) report that anchors do improve the reliability of ratings and suggest using them for better and more reliable ratings of voice quality. Similarly, Gerratt et al. (1993) studied the use of external anchors improving the reliability; they gathered ratings via 2 tools: a 5-point EAI scale, 1 representing normal and 5 representing severe roughness and a 5-point scale with an external anchor/example for each point on the scale. Listeners provided ratings with a one week gap in between and rated each stimuli twice per session without knowledge. Gerratt et al., (1993) reported significantly more reliable ratings when using the explicitly anchored EAI. They also reported that ratings with the unanchored EAI drifted significantly within the same listening session. These data confirm previous findings on stability of internal standards for normal and extremes as “the internal standards for normal and extreme qualities are well-developed and stable” (Gerratt et al., 1993).

2.8.2.2 Synthetic stimuli

Any voice which is produced by a rule-based text-to-speech system (TTS) is termed synthetic speech. TTS converts an input of string of text characteristics into an output speech waveform. Synthetic speech is different from natural speech, but being able to synthesize pathological voices can help with the development of a tool for assessment of voice quality. Speech synthesizers which are capable of modeling a wide range of voice qualities have many applications (Bangayan, Long, Alwan, Kreiman, & Gerratt, 1997).

Speech synthesis involves classifying voice qualities and synthesizing those qualities to demonstrate that perceptual voice classes do exist (Kreiman & Gerratt, 2000). In fact, voices are grouped based on perceptual criteria and are then investigated to see what synthesis strategies and parameters listeners used to model and reproduce those phonation types (Kreiman & Gerratt, 2000). The other method which can be used is the method of adjustment task. In this method, listeners are presented with natural stimuli and then asked to change the parameters of a synthetic stimulus until it matches. The agreement increases by the mean ratings for the voice and the existing challenges in the way of evaluating the acoustic features are reduced (Kreiman, Gerratt, & Ito, 2007).

Bangayan et al. (1997) studied pathological voices via analysis-by-synthesis using the Klatt synthesizer (Klatt, 1980). Ten expert listeners listened to 24 stimuli pairs of a natural voice and a synthetic copy of the vowel /a/ and judged how much they matched using a 7-point scale (1 being the perfect match). They included 3 pairs of natural/natural stimuli and repeated the whole set of 17 pairs twice randomly to raters. They reported raters could synthesize the less severe voices, as well as male voices better than female voices. Using the version of the Klatt synthesizer used at that time, voices with notable frequency and amplitude perturbation were more difficult to synthesize. Expert listeners found half of the synthesized voices well-matched the natural ones and tokens included parameters such as rough, breathy, bifurcated, and strained-rough. Analysis-by-synthesis is a way of improving reliability as it gives the raters the opportunity and time to move adjust different parameters and constructs which leads to improved agreement among raters.

2.8.2.3 Training

In addition to providing anchors or external standards, training listeners can increase reliability and agreement. Brinca et al. (2015) propose two levels of training: orientation and extensive training which included providing anchor stimuli, a few practice trials, and definitions of the scale terms (grade, roughness, breathiness on the GRBAS scale). Training was provided by a coach who was a speech and language therapist with more than 15 years of experience. Definitions were provided for the rating parameters under study (GRB); 10 samples which were mild to moderate in severity were provided. Since there is no upper limit for severity as per Brinca et al. (2015), they did not provide anchors for severity level. Judges could listen to the stimuli as many times as they wanted and discussed their ratings with the others afterwards. Their extensive training which took place 1 month later included definition presentation and anchor familiarization. They were then provided with 10 anchors for each stimulus and made their evaluations and classified samples based on severity (Brinca et al., 2015). The extensive training and use of anchors improved the interrater and intra-rater agreement and reliability. The highest interrater reliability was obtained using the oral reading stimuli, particularly for the phonation dimensions of grade and breathiness. Breathiness was the easiest and roughness was the most difficult to rate.

Barsties et al. (2017) also investigated the influences of training and visual feedback on rating voice quality of the features grade, roughness and breathiness by naïve listeners who underwent two, two-hour training sessions. Training involved stimulus-response feedback using spectrograms as visual support for the auditory perceptual judgement of quality of voice. The raters listened to the stimuli after training and made their ratings. Training involved providing sustained phonation, continuous speech and concatenated speech of normal and slightly, moderately and severely disordered quality. They report a training effect in the improvement of rating reliability of roughness within their no feedback and the auditory perceptual feedback group, and for breathiness within visual plus auditory feedback group. Based on their data, it was suggested that the use of visual and auditory anchors for rating are of value, but also recommend longer training sessions.

Coelho et al. (2017) emphasized the importance of perceptual training and calibration for achieving a reliable and valid resources as they increase agreement and consistency. However, it is not clear which type and what amount of training is required for the optimum result, but there is no doubt about the positive influence. Coelho et al. provided extensive training for raters who perceptually evaluated voice quality of those who had a cochlear-implant and normal hearing individuals. Three experienced SLP raters took part in three, four-hour sessions; in each session they were trained for the 3 stimuli types: sustained vowels, connected speech and conversational speech. Training also involved listening to normal hearing and adults with cochlear implants, a discussion of all parameters to be evaluated, efforts to reach agreement on the presence, absence, and severity of each feature and their definitions.

2.8.3 Physiological reactions to audio stimuli

In addition to the perceptual evaluations of voice quality and the objective measures of voice evaluation, another dimension can be added to the voice quality evaluation and that dimension is assessing possible physiological responses to dysphonic voices. This trend, investigates possible involuntary physiological changes in listeners when presented with and processing voice signals. In fact, there are some physiological indicators to mirror various internal responses and reactions which can be used as an indicator of listening and cognitive effort (Pichora-Fuller et al., 2016).

2.8.3.1Listening effort

Pichora-Fuller and Kramer (2016) defined listening effort as "the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task that involves listening" (p: 10s). Pichora-Fuller et al. (2016) also proposed a new "Framework for Understanding Effortful Listening" (FUEL). According to this framework, motivation arousal, cognitive capacity, and task demand modulate listening effort independently. Listening effort can be evaluated subjectively through self-report and questionnaires and objectively through physiological measures (Pichora-Fuller et al., 2016). Effortful listening is something beyond the challenges faced by listeners in everyday conversations such as audibility. According to Pichora-Fuller et al. (2016) there are two main groups of physiological measures for assessing listening effort.

The first category involves those dealing with brain activity: magnetic encephalography (MEG), evoked-response potential (ERPs), alpha power in electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). All provide information regarding timing and the precise localization of brain activity. The second category they present refers to measures of autonomic nervous system which involves both sympathetic and parasympathetic responses. More specifically, pupillary changes, hormonal shifts, skin conductance and cardiac responses can be used for autonomic response measurement (Pichora-Fuller et al., 2016). By examining pupil reactions, or pupillometry, the underlying mental effort required while listening can be evaluated (Kramer et al., 2013). In fact, the behavior of autonomous nervous system and how much parasympathetic and sympathetic nervous systems intercede with iris muscles while listening are indicated in pupillometry (Kramer et al., 2013). Using this technique, pupil diameter is constantly recorded using infrared eye tracking technology and pupil size measurement is synchronized in time with the presentation of the stimuli (Kramer et al., 2013).

2.8.3.2 Pupillometry

Pupillometry refers to evaluating the fluctuations in the size of the eye pupil which has been used in experimental psychology to evaluate memory processes, task performance dynamics, fluctuations in autonomic arousal and alertness, and attention studies (McGarrigle, Dawes, Stewart, Kuchinsky, & Munro, 2017). It has been used in various studies since the mid-20 century (Hess & Polt, 1960; Hess & Polt, 1964). The main measure generated from pupillometry is the peak pupil dilation (PPD) which is defined as the maximum dilation of pupil within the time interval between onset and offset of the stimuli (Wendt, Hietkamp, & Lunner, 2017). PPD is calculated relative to the baseline pupil dilation. Perhaps the first thing that comes to mind about changes in the pupil size is its reflexive response to light. Studies report that brightness illusions and imagining a dark room yield differential pupil response (Laeng & Endestad, 2012; Laeng & Sulutvedt, 2014). In addition, pupil reflux is apparently under considerable cognitive control reacting to multiple emotional and cognitive process and states (Einhauser, 2017). PPD also indicates cognitive load, and for instance, its size increases while solving a math problem and the increase enhances with the difficulty level (Hess & Polt, 1964). Hess and Polt (1960) also report that pupil diameter changes sensitively signal mental state; an increase in pupil size occurs while viewing interesting or emotionally toned stimuli. Pleasure dilates and displeasure constricts the pupils. In terms of emotions, high audience anxiety creates larger dilation compared to low (Simpson & Molloy, 1971). Kahneman and Beatty (1966) note that the peak pupil diameter serves as a measure of the amount of material under active processing. Mental activities such as solving arithmetic problems are reported to dilate the pupil and in short term memory tasks the pupils dilate when the listener is trying to listen and they constrict as the reporting phase begins (Kahneman & Beatty, 1966). Therefore, decision making process signifies larger dilation if the decision is to be signaled (Simpson & Hale, 1969). However, pupil reactions are not limited to visual stimuli only as audio stimuli can trigger pupil reactions too. For instance, pupil dilation is reported to be influenced by speech intelligibility - the less speech intelligibility, the more pupil dilation (Zekveld & Kramer, 2014).

The underlying physiological change in pupil size is referred to as the locus coeruleus norepinephrine (LC-NE) system. Changes in pupil size are reported to covary with changes in the blood oxygen level-dependent (BOLD) response in the locus coeruleus. As mentioned earlier, this technique can be used to evaluate various mental or emotional processes, one of those is listening effort by individuals when listening to various audio stimuli.

To date, it has been reported that pupil dilation responses are sensitive to syntactic complexity, speech intelligibility, type of masking noise, and divided attention (Wang et al., 2018). Kramer et al. (2013) report task-evoked pupillary responses to be a reliable, albeit indirect measure of cognitive processing load. They also report it to be reflective of task demands and stimulus features associated with language processing tasks. In terms of the type of tasks assessed while monitoring pupil responses, different studies have used various tasks and conditions. In 1966, Kahneman and Beatty designed their experiment with four blocks of seven trials. Listeners heard: i) strings of digits and asked to recall them immediately, ii) a string of four high frequency monosyllabic nouns and asked to recall immediately, and iii) a string of four digits presented for transformation. Five pictures were taken of the listeners' eyes and the stimuli were presented on the sixth click of the camera and four additional pictures were taken while subjects were reporting. Results confirmed the relationship between pupillary response and task difficulty while performing a mental task. Kahneman and Beatty (1966) reported no dilation during the listening phase, but more during the word recall and transformation task which means allocating cognitive resources and making an effort to recall was accompanied by more pupil dilation.

Wang et al. (2018) also investigated the relationship between self-reported daily-life fatigue, hearing status, and pupil dilation during a speech perception in noise task. They reported a negative correlation between fatigue and peak pupil dilation. Higher levels of fatigue were associated with smaller PPDs. Some studies also assessed pupillary fluctuations in hearing impaired and normal hearing individuals with findings of significantly smaller pupil peak dilation in hearing impaired individuals when compared to normal participants when confronted with a challenging listening condition (Wang et al., 2018).

In another study, Kramer et al. (2013) investigated the extent of pupil response evoked by various tasks and information complexity. Specifically, they designed different conditions to observe the extent of the pupillary response evoked by each condition. The conditions involved tasks of various linguistic and auditory complexity including background noise alone- answer prompt, no response (listeners just listened, no answer was required); background noise alone- prompt- response (listeners were required to verbally answer "yes" after each prompt); noise- in- background- noise detection (listeners answered "yes" if they detected the target noise burst and 'no' if they didn't.); words- in background- noise identification for evaluating (hearing the background noise and then repeating the word presented or saying 'not understood') (Kramer et al., 2013). They included normal hearing subjects from three different locations and languages in their experiments. All of their experiments consisted of "anticipation to verbal responding to a prompt signal, auditory detection, and the identification of meaningful words" (p. 426). Based on data gathered, Kramer et al. reported the largest peak dilation in the words-in- noise identification task, where listeners had to understand what word was presented and repeat it.

McGarrigle et al. (2017) investigated physiological arousal in normal hearing young adults during a sustained listening task. They used response time and pupillometry as markers of listening related fatigue through a speech picture verification task. The listening conditions which included short passages were both easy and difficult and had contrasting signal-to- noise ratios (+ 15 dB SNR for the easy and -8 dB SNR for the difficult passages). Listeners were presented with the speech passages each containing three sentences (45–50 words long and between 13–18 seconds in duration). Two lists of speech-picture pairs were then developed and presented to listeners. If the object depicted in the picture was mentioned in the passage, they were requested to say ‘yes’ and if not, then ‘no’. Researchers also collected participants’ self-report of effort and a fatigue scale after each block and pupil size was measured throughout the session. The self-report contained the question of “How hard did you have to work to accomplish your level of performance?” on an incremental Likert scale. Results indicated a steeper linear decrease in pupil size in the more challenging SNR condition in their second half of the trials (hard passages).

Wendt et al., (2017) used hearing impaired listeners (pure-tone average from 500 to 4000 Hz ranging from 34 to 70 dB HL) to perform a speech recognition task. Their aim was to discover the effect of noise reduction (NR) and intelligibility level on processing effort using pupillometry. The stimuli (Danish sentences from the Hearing in Noise Test (HINT) were presented in a four-talker babble. The files were presented through a loudspeaker and participants heard them through hearing aids with a no noise reduction scheme and once with a noise reduction scheme applied. Results indicated that processing effort and recognition were influenced by intelligibility level (L50 versus L95) and NR scheme (no NR versus NR). They also reported an increase in PPD which indicates more processing effort in the L50 condition. Thus, effort increases with the reduced intelligibility. The NR scheme also brought about less effort indicated by smaller PPD. In their second experiment, they found that the processing effort depends on the type of the NR scheme and they used two hearing aids with different NR schemes. Listening effort and associated fatigue is an important challenge for hearing impaired individuals. It is, therefore, reasonable to note that task-evoked pupil dilation is a combination of attention, arousal, engagement, effort and anxiety and not a unitary concept of effort (Nunnally, Knott, Duchnowski, & Parker, 1967; Pichora-Fuller et al., 2016).

2.8.3.3 Subjective listening effort

In addition to the objective listening effort, listener burden or subjective listening effort is a unique perceptual construct. It is defined by modifying the definition by Whitehill and Wong (2006) as the amount of work needed [by a listener] to listen to the speaker. When investigating this dimension which is reported to be experienced differently by individuals, attention is drawn away from the signal and toward the listener; ultimately, it forces listeners to think about their reactions to the speech they are presented (Nagle & Eadie, 2012). When examining listener effort, listeners are asked to report the perceived cognitive resources which are allocated to speech processing (Beukelman, Childes, Carrell, Funk, Ball, & Pattee, 2011). In fact, such data, also referred to as perceived listener effort (PLE), is achieved through subjective ratings and reports from listeners with no familiarity with disordered voices (Nagle & Eadie, 2012). Asking listeners to judge the amount of effort required to listen to a speech sample can reveal paralinguistic parameters beyond dimensions such as intelligibility (Nagle & Eadie, 2012).

Subjective listening effort is reported to be affected by factors such as familiarity with specific speaker groups or speech type such as disordered or accented speech (Nagle & Eadie, 2012). Some studies evaluate this construct through qualitative measures such as eliciting statements like ‘it was hard to listen to this sentence; I got distracted by the way the speech sounded; or I had to completely attend to the sentence to understand it’ (Klasner & Yorkston, 2005). Zekveld and Kramer (2014) used pupillometry as the objective measure of cognitive load and an EAI scale for PLE; in doing so, they reported that their listeners “gave up trying” to perceive the stimuli more often in the medium- and low intelligibility conditions. Some studies have also used a VAS for obtaining ratings of PLE (Mackersie & Cones (2011). They report a high association between PLE and physiological values (skin conductance) at high effort levels.

Subjective listening effort is also reported not to be influenced by age as the older listeners can be less accurate and take longer time to react, but do not indicate a higher perceived effort (Larsby, Hallgren, Lyxell & Arlinger, 2006). In terms of intelligibility, a strong negative correlation between ratings of perceived or subjective listening effort and intelligibility are reported for most of the speakers in a study by Whitehill and Wong (2006). However, some of their speakers who were highly intelligible were moderately rated on PLE which indicates that ratings of PLE provides unique information beyond intelligibility. As Koelewjn, Zekveld, Festen and Kramer (2012) report, two listeners may receive the same intelligibility score but experience different levels of listening effort. Identifying words correctly does not mean all aspects of speech perception are covered.

Speech perception involves any aspects such as understanding and analyzing a talker's intention (Snedeker & Trueswell, 2004; Tanenhaus, Spivey, Eberhard, & Sedivy, 1995), recognizing prosodic emphasis (Dahan, Tanenhaus, & Chambers, 2002), identifying a speaker (Best, Streeter, Roverud, Mason, & Kidd, 2017), deciding if the speech makes sense (Best, et al., 2017), foreseeing what information comes next (Altmann & Kamide, 1999; Tavano & Scharinger, 2015), and translating what was heard into another language (Hyona, Tommola, & Alaja, 1995). For that reason, all are necessary for successful speech communication and mere identification of words correctly does not guarantee success. In conclusion, listening and perceived listener effort are unique constructs and a decrease in the processing speed and performance does not necessarily lead to the perception of "working hard" by listeners (Gosselin & Gagne, 2011; Nagle & Eadie, 2012; Zekveld et al., 2011).

2.8.3.4 Utilizing pupillometry

Given the growing popularity of applying pupillometry in different fields of research, specific attention needs to be paid to logistics of the experiment, timing and data cleaning (removing blinks) and analysis. Winn, Wendt, Koelewijn, and Kuchinsky (2018) suggest a variety of guidelines for pupillometry experiments which includes the best practices and advice for using pupillometry as an indicator of listening effort.

Pupil responses are reported to dilate and contract between 3 mm and 7 mm due to a variety of reasons (Laeng, Sirous, & Gredeback, 2012) with the biggest changes being due to illumination. Changes due to cognitive tasks are smaller, 0.1 to 0.5mm, based on the task and conditions of an experiment. Any small physical movements during a given task can create dramatic changes in pupil responses that are not due to the task (Winn et al., 2018). Therefore, all the other sources of dilation and constriction must be managed and controlled in order to assure that the resulting changes are due to the task (Winn et al., 2018).

In terms of task selection, there should be a balance between stimuli ease and effort demand. Stimuli that are too easy require too little effort and a very difficult one makes the effort useless and may then cause the participant to lose interest. The stimuli must be of value to the participants as well. Boredom must be avoided by providing breaks during the experiment as boredom interferes with eliciting reliable pupillary values.

Determining how many participants and how many stimuli to use is also crucial in securing reliable results. Although the number of trials in any experiment depends on the effect size of interest and power of the analytical approach, Winn et al., (2018) recommend that a minimum of 16 to 18 good pupil recordings should be used. For a sentence perception task, 20 to 25 trials are recommended. Sufficient data should be recorded as some data may be missed due to contamination or mistracking. However, if the task is very difficult, as few as 10 trials have been reported to be sufficient (Winn et al., 2018).

In all pupillometry studies, participants who are taking any of the medications known to interfere with pupil reactions must be excluded. A list of medications known to cause pupil dilation is fairly extensive, however, the most common are listed in Appendix A. For those medications which only cause pupillary changes in overdose the word is indicated in brackets.

2.9 Conclusions

Auditory-perceptual evaluation of voice quality is an important step in the diagnosis and treatment of voice disorders. The literature includes many studies which highlight the importance and efficiency of these measures. Speech clinicians’ value and use such measures and gauge objective measures with them. In order for the perceptual measures to be meaningful and, reliable though, a comprehensive theoretical framework must be followed for selecting the task and stimuli, measurement scales and tools, listeners and the standards against which ratings are made.

Auditory-perceptual measures may be used independently or along with a wide range of acoustic and objective measures. This study aims at using such gold standards along with pupillometry to objectively evaluate cognitive load and listening efforts while listening to various dysphonic voices.

Chapter 3

3 Adductor Spasmodic Dysphonia

Adductor spasmodic dysphonia (AdSD) is categorized as a focal laryngeal dystonia which influences motor control during voice and speech production. AdSD is characterized by intermittent hyper-adductions of the vocal folds creating voice and/or pitch breaks (Nash & Ludlow, 1996; Eadie et al., 2007). Those presenting with AdSD are reported to exhibit a variable range of adductory spasms which ultimately result in a strained strangled voice quality that may be varied and intermittent (Nash & Ludlow, 1996). Like any other voice disorder, the identification of AdSD often begins with evaluation of the problem using auditory-perceptual evaluations methods. The present project specifically focuses on auditory-perceptual evaluation of “vocal strain” in a sample of speakers presenting with AdSD. The primary objective of this study sought to evaluate how much physiological or listening effort normal hearing listeners expend while listening to voice stimuli produced by those with AdSD.

3.1 Method

3.1.1 Participants

The participants of the project consisted of two listener groups: naïve and experienced. Both groups were recruited based on the following inclusion and exclusion criteria. For the naïve group, these criteria were being a native English speaker, being between the ages of 18 to 35 (availability and inclusion of university students as listeners), having no prior exposure to or training in voice disorders (formal coursework or clinical experience), having no previous experience with auditory- perceptual research, or a personal history of any speech, voice, language, or hearing difficulties. Also, if the potential participant reported having had an upper respiratory infection within the week prior to the experiment, they were not able to participate until that problem had been resolved for 14 days. In addition, participants were asked if they were taking any of the medications listed in Appendix A and if so, they were excluded from participation as such medications are pharmaceutically reported to be influential on pupil reactions which may have an influence on the results of the study (Appendix H, REB # 112674).

For the experienced listener group, inclusion criteria also included being a native English speaker and being between the ages of 25 and 60 years of age. The experienced listeners' age range was selected based on the literature and to include clinicians/speech pathologist with minimum two years of experience and more and also availability. Since hearing problems are more prevalent after the age of 60, that age was the cut off. Also, if they were a speech-language pathologist or a voice researcher, a minimum of two years' experience in the field or having education related to and/or clinical training and exposure to voice disorders and/or having direct experience in the formal evaluation of voice disorders was necessary.

The naïve listener group was comprised of 20 adults (11 males, 9 females; age range = 18-29 years; mean: 22.75 years). The recruited number of participants was based on a power analysis calculated using GPower, with an effect size of 0.4. The experienced group included 3 female clinicians, age range 41 to 56 years of age (mean = 49 years).

Each listener participated in a single listening session which required approximately 45 minutes (i.e., 10 to 15 minutes for task instruction, instrumentation adjustment and calibration, 7 to 10 minutes for the experimental protocol, and 7 to 9 minutes for the retest procedure). Task instructions included the oral and written explanation of the auditory-perceptual dimensions under study (i.e., strain and listening effort). Definitions were provided to listeners and a written description of these dimensions was provided to listeners for reference purposes during the experiment. During the experimental task, participants sat in a softly lighted room. The light was consistent throughout the room to prevent reflexive dilation in reaction to changing luminance on the retina. This is reported to be the best method for the collection of the pupillary response and data gathered are reported to be more reliable in a reduced light experimental setting as opposed to dark settings (Winn et al., 2018).

3.1.2 Auditory Stimuli

Stimuli used in the current study included speech samples from 23 talkers (6 males, 17 females). All speakers had been diagnosed with AdSD with these samples obtained from an archive of the Voice Production & Perception Laboratory at the University of Western

Ontario. All talkers had been diagnosed to have AdSD by a board-certified laryngologist. All voice samples were gathered via digital recording obtained in either a quiet clinical setting free from ambient noise or within a sound-treated environment. All voice samples included sustained vowels, a standard reading (The Rainbow Passage; Fairbanks, 1960), a short monologue (approximately 60 seconds), and a variety of single word stimuli along. In some instances, standard sentences from the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) were also gathered. Recording of speech stimuli occurred after informed consent was obtained, along with demographic information for each speaker. A directional microphone (Shure PG-81) attached to a desktop microphone stand was used in all recording. During the collection of each sample, a microphone-to-mouth distance of 15 centimeters was maintained and was checked prior to each task. Each participant speaker's sample of voice/speech stimuli was recorded onto a laptop computer at a sampling rate of 44.1 kHz using the Multidimensional Voice Profile (SonaSpeech II, Kay Pentax, Lincoln Park, NJ). Volume input levels were adjusted for each speaker at the beginning of each session and were monitored during the recordings using SonaSpeech II to avoid any under- or over-driving of the input signal. Once the reading passage was collected, the second sentence ("The rainbow is a division of white light into many beautiful colors.") was extracted for use in the current study. The stimuli were also equalized for their Root Mean Square (RMS), so that they were presented at approximately same intensity during playback.

The experimental procedure for each listening trial was as follows. Each trial began with the spoken cue "Please listen to the following stimulus"; this preparatory stimulus was spoken by a normal speaking adult. This cue lasted three seconds and indicated the impending onset of the upcoming experimental stimulus. Upon cue presentation, one of the 23 sentences from the AdSD talkers was presented. One second after sentence offset, the spoken sentence "Please indicate your ratings after the beep" instructed participants to begin rating strain and listening effort using a computer-based slider procedure that represented a visual analog scale.

3.1.3 Assessment of strain and listening effort

After the presentation of each sentence, each listener rated their perceived judgment of strain and listening effort using two separate 100 mm long electronic sliders (see Figure 3-1) that represented a visual analog scale that had 100 intervals (i.e., 1 through 100). The end points of the slider for the feature of ‘strain’ was marked “mild” toward the left side of the scale and “profound” toward the right side. The end points of the slider for ‘listening effort’ indicated “none” on the left and “extreme” on the right. Listeners could manual move the slider handle and mark the scale at any point along the continuum that they thought best indicated the degree of strain and also their own listening effort that was represented for that stimulus sentence.

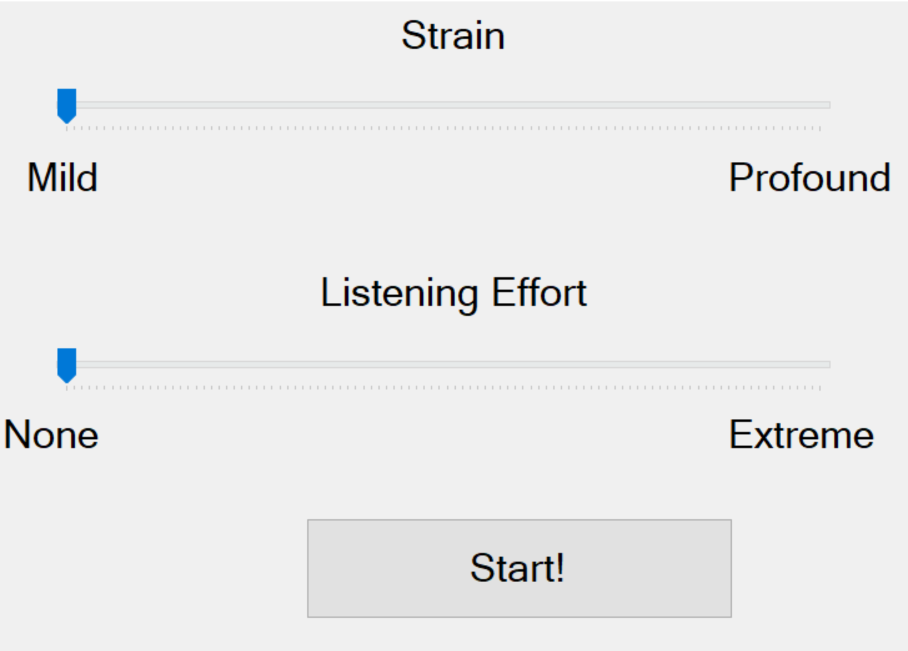


Figure 3-1 Appearance of slider rating scales for "strain" and "listening effort".

3.1.4 Pupillometry data recording and data analysis

Pupil dilation during the presentation of AdSD voice stimuli for each participant was recorded continuously using an EyeLink 1000 (SR Research, Ottawa, Canada) eye tracker (Figure 3-2 & 3-3). Participants were seated comfortably on a stationary chair at the instrumental tower mount; the participant’s chin was positioned on a chin rest and their forehead placed against a rest while they visualized the monitor ahead of them. The device collected the pupil responses of the right eye at a sampling rate of 1000 Hz.

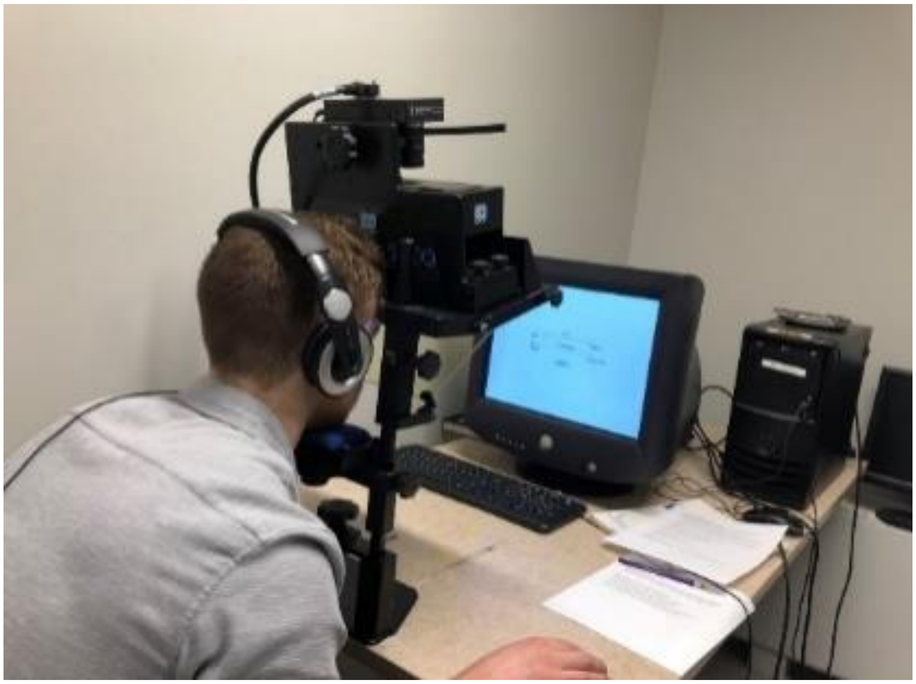


Figure 3-2 EyeLink 1000 set-up

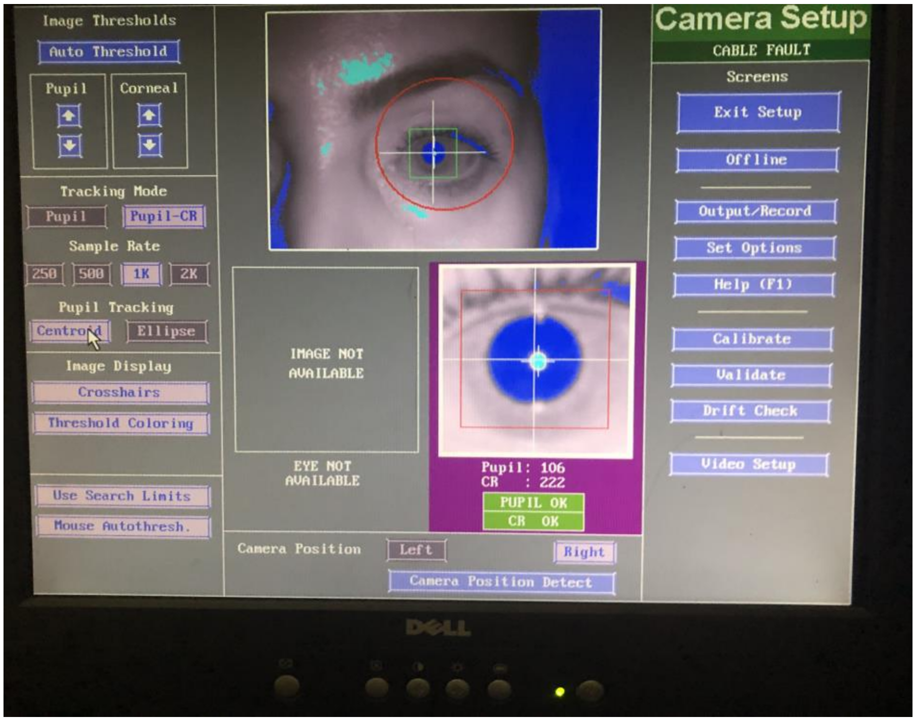


Figure 3-3 Pupil image on EyeLink 1000 monitor

3.1.5 Procedure

On the day of the experiment, each listener was individually familiarized with the experimental tasks they would perform. They were trained about the voice dimension of “strain” and “listening effort” and all were provided written definitions. Strain was defined as the “perception of excessive vocal effort” and listening effort was defined as “the amount of work required while auditing the speaker samples”. The height and general positioning of the eye tracker were adjusted for each listener to provide the best and most direct view of the pupils. Listeners were instructed not to move their head or body or to look down or away from the monitor at any point during the experiment. During the task, they were asked to keep looking at the center of the monitor and were requested to avoid blinks as much as possible, or at least try not to blink excessively when listening to the stimulus. Listeners were asked to wear headphones (Sennheiser, HD 205) and prior to the start of the experimental task, each listener was asked to test the output volume before beginning the experiment. Once the optimum position was reached and the listeners were ready to proceed, calibration of the visual gaze and then validation of the measure was performed.

At the beginning of the experimental task, listeners were asked to maintain visual focus on a fixation circle on the screen and to follow it when requested in order to calibrate the system. Upon obtaining satisfactory results in calibration and subsequent validation, the rating experiment was initiated by the experimenter. Speaker stimuli were randomized and presented to listeners one by one in a randomized order. After listening to each stimulus followed by the beep, the listener used the first computer slider to indicate their ratings of the voice feature of strain which ranged between mild and profound and the second slider to indicate their judgement of listening effort. Once they were done with both ratings, they clicked the “next” box to hear the following stimulus. All AdSD voice stimuli were randomized and presented twice in a test and re-test condition for the evaluation of intra-rater reliability. After the test phase of the experiment, each listener was given a 10-minute break to rest and then the re-test phase of the experiment was undertaken. Once all stimuli and reliability samples were rated, a message appeared on the screen indicating the end of the test.

Given the fact that speaker tracks were randomized for presentation and presented at different time stamps during the experiment, all stimuli were normalized first so that the starting point of each sentence was at zero (0) second. Raw pupil data which were recorded throughout the experiment by the eye tracker had to be processed in several steps before formal final analysis and visualization. Given the nature of the experiment, eye blinks, or changes due to factors other than the experimental task are potential confounds and, thus, needed to be identified. Quick blinks (<125 milliseconds) were identified, removed, and interpolated (interpolation began roughly 50 msec before the blink and end at least 150 msec after the blink) without changing the overall pattern of the tracking sequence. However, some tracks still (13%) had to be discarded due to response dropouts, too many variations and long blinks; this process was required in an effort to eliminate the risk of data distortion.

In the current study, we focused on peak pupil dilation (PPD) as a dependent measure secondary to the presentation of the AdSD samples. For each trial, the peak in pupil dilation was determined as the maximum dilation during the presentation time of the speech sample relative to the time immediately before the stimulus. It is also called baseline subtracted absolute pupil size. The last second of the three second prompt (Figure 3-6) served as the baseline against which the stimulus was compared to determine the PPD or the highest point in the track.

3.2 Results (naïve group)

Once all listeners had completed the experimental task, their ratings for the voice dimension of strain and listener effort were first analyzed for reliability.

3.2.1 Auditory-perceptual data analyses

Intra-rater reliability was obtained for each listener by correlating the rating results of test-retest for both strain and listening effort. The correlation range for strain ranged from 0.56 to 0.96 and from 0.58 and 0.90 for effort, indicating a moderate-to-high correlation. Interrater reliability was calculated through Cronbach’s alpha in SPSS (Version 24, Armonk, NY) for each of the two rated features. The Cronbach’s alpha was 0.98 for strain and 0.97 for effort. The interrater reliability outcome confirms very high reliability among listeners for the rating task.

Two sets of strain rating scores that could range between 1 and 100 and two sets of ratings for effort, again which could range between 1 and 100, were generated for each speaker sample by each listener. Once all 20 listeners completed the experiment, ratings per feature were averaged across trials (test, retest) to achieve a single strain value and a single effort value per talker for each listener and then averaged across all listeners to achieve a single strain and a single effort score for each speaker along with the standard error of the mean (Figure 3-4) with additional information presented in Appendix B. These data were then plotted against each other to represent the correlation between the two measures (Figure 3-5). As can be seen from Figure 3-4, Talkers 8 and 10 were rated as having the least perceived levels of strain and effort and Talkers 1 and 18 were rated as demonstrating the greatest degree of strain and effort.

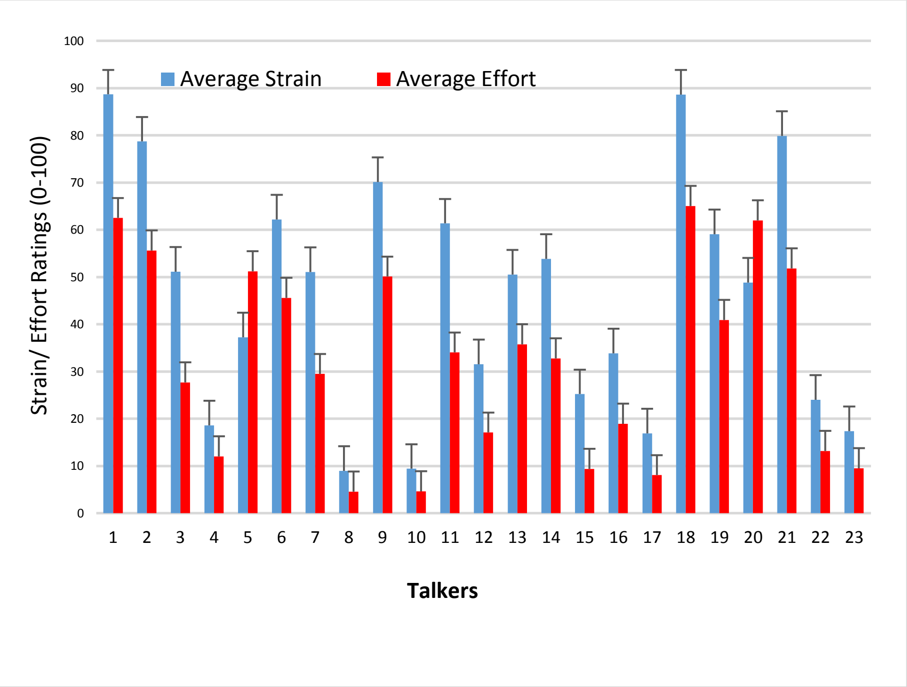


Figure 3-4 Average strain (blue) and effort (red) ratings for each talker along with the standard error of the mean.

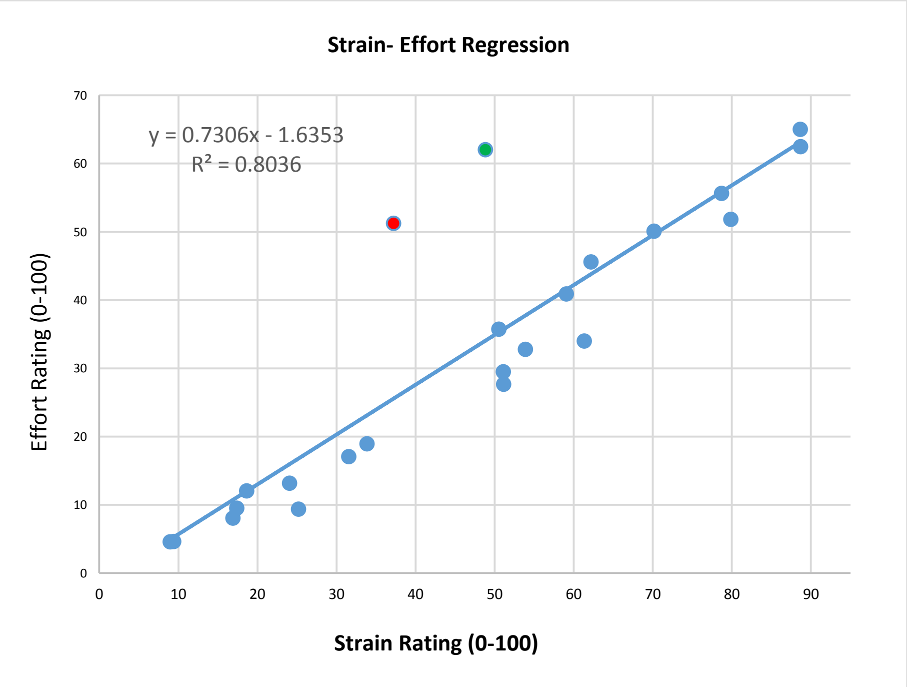


Figure 3-5 Regression between strain and effort.

A repeated measures ANOVA was conducted to statistically compare strain and effort, talkers, and the potential interaction between the auditory-perceptual features of strain, and effort and talkers. The a priori significance level was set at 0.05 for all statistical tests. Significant effects were found for the auditory-perceptual features (strain and effort) (F1, 19 = 37.13, $p <0.05$, η_P^2 = 0.662); talkers (F (22, 418) =72.08, $p < 0.05$, η_P^2= 0.791); and revealed an interaction between features and talkers (F (22, 418) = 12.88, $p < 0.05$, η_P^2= 0.404). In addition, post-hoc comparisons using a Bonferroni correction revealed a feature-talker interaction indicating that Talkers 5 and 20 were perceived to be significantly different from the others (Figure 3-5). Unlike the rest of talkers, and as depicted in Figure 3-5, Talkers 5 (red dot) and 20 (green dot) were rated higher on the feature of effort when compared to strain, indicating that a talker can be not highly strained but still demanding high listening effort from listeners.

In order to examine the relationship between strain and effort ratings, the correlation coefficient was calculated. The results indicated that a strong correlation (r = 0.89) existed between rating of these two auditory-perceptual features (Figure 3-5).

3.2.1.1 Pupillometry

The recorded time stamps for all stimuli were first normalized so that the starting point of each sentence was at 0 second. Raw pupil data which were recorded throughout the experiment by the eye tracker had to be processed in several steps before formal final analysis and visualization and they were cleaned explained earlier.

In addition, the pupil responses were plotted to provide a better understanding of what the pupil tracking looked like in terms of listener reactions during presentation of the talker speech stimuli. Figure 3-6 shows a sample of pupil reactions for Talker 1 and depicts how each track appeared after the pupil responses were averaged across all listeners along with their corresponding waveform. As indicated in the figure, the PPD typically falls between approximately 5000 msec and 6500 msec following the initiation of the voice stimuli.

In order to illustrate how talkers who induced high and low PPD results appear relative to each other, the pupil tracks of 4 talkers, two representing the highest auditory-perceptual ratings of strain and effort (Talkers 1 and 18) and two with the lowest (Talkers 8 and 10) are displayed in Figure 3-7, respectively. This indicates that the higher the perceptual ratings, the higher the PPD indicating cognitive load and the lower the perceptual ratings, the lower the PPD.

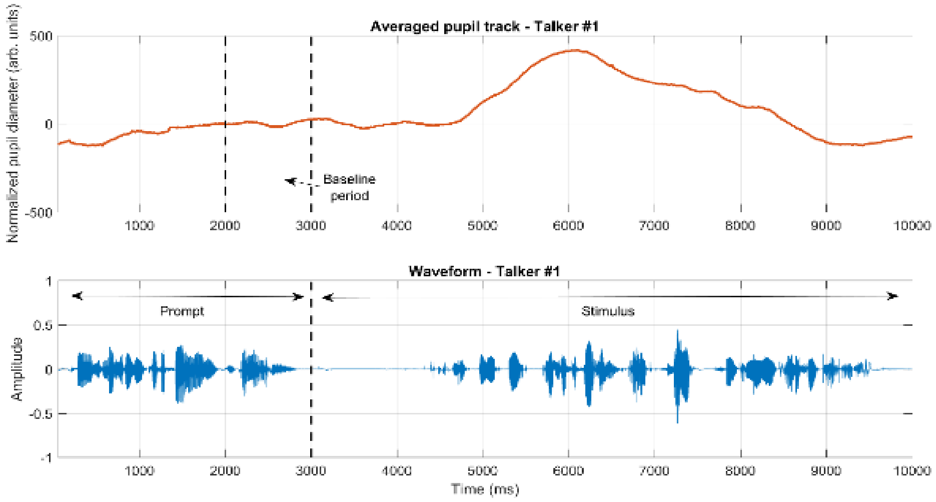


Figure 3-6 Pupil response track for Talker 1 averaged across all listeners along with its waveform.

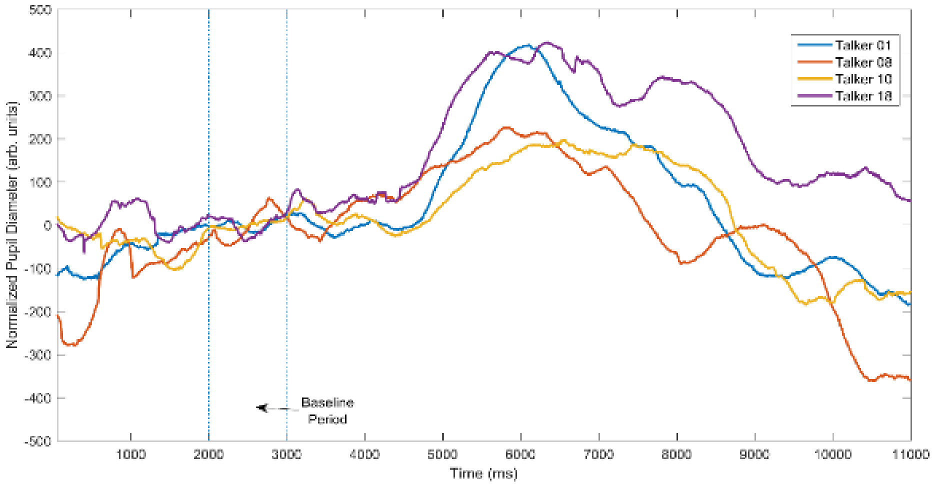


Figure 3-7 Pupil response tracks for the two perceptually high and two perceptually low strain/ effort talkers (averaged across all listeners).

3.2.1.2 PPD

Once the PPD results were extracted, these data were averaged across listeners and one set of values were obtained for each talker for the features of strain and listening effort. The average PPDs are displayed in Figure 3-8. The two talkers provoking the highest PPD are represented in purple and the two talkers creating low PPD responses are shown in red.

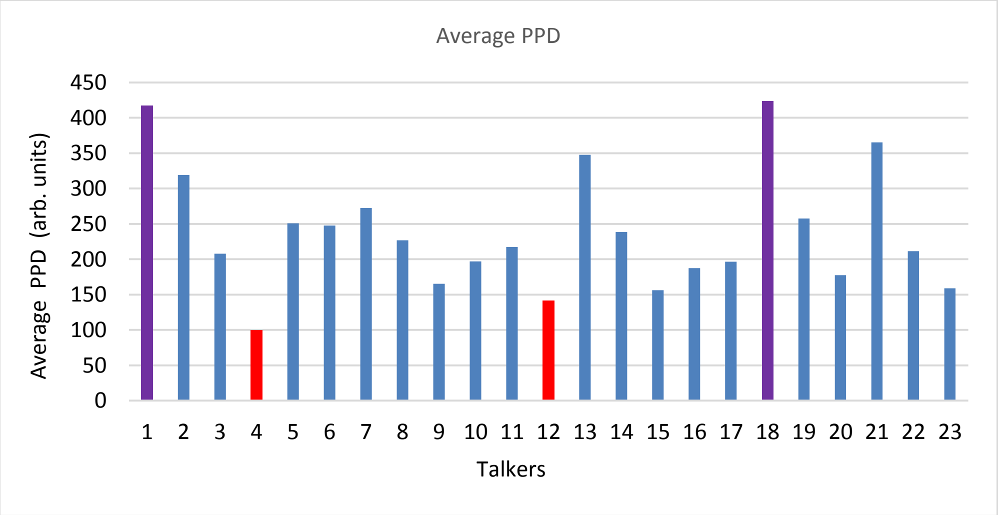


Figure 3-8 Average PPD (across all listeners) for each talker.

3.2.2 Relationship between PPD, strain and effort

The correlation between PPD and strain, and PPD and effort was calculated, and this analysis indicated statistically significant values of r = 0.73 and r = 0.66, respectively. A linear regression was calculated to predict PPDs based on strain (Figure 3-9) ratings. A significant function R^2 of 0.53 was found. Furthermore, predicting PPD for effort (Figure 3-10) values also revealed a significant R^2 of 0.43.

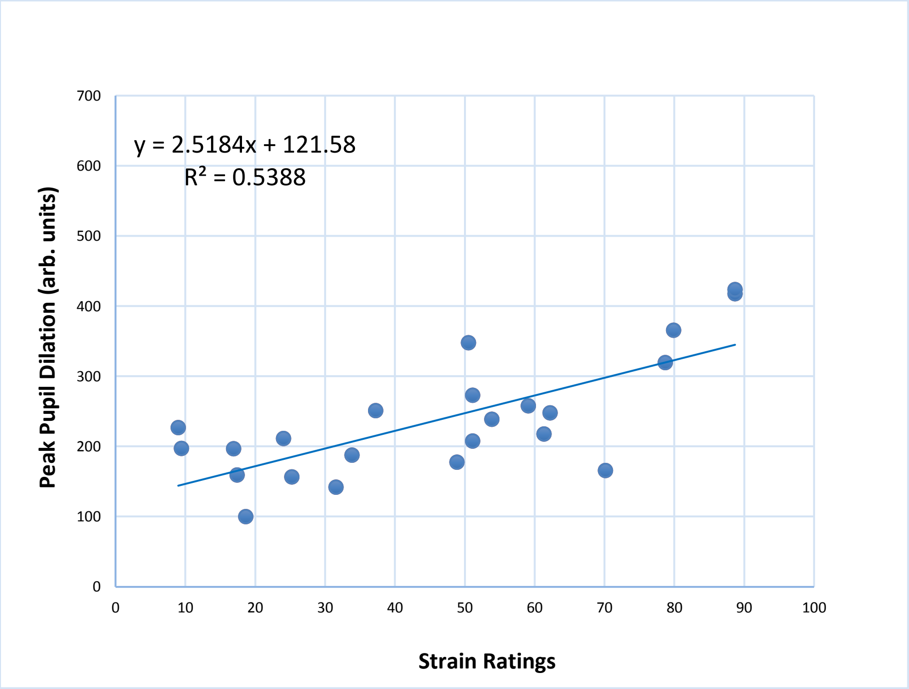


Figure 3-9 Simple linear regression plots for predicting PPD based on perceptual ratings of strain.

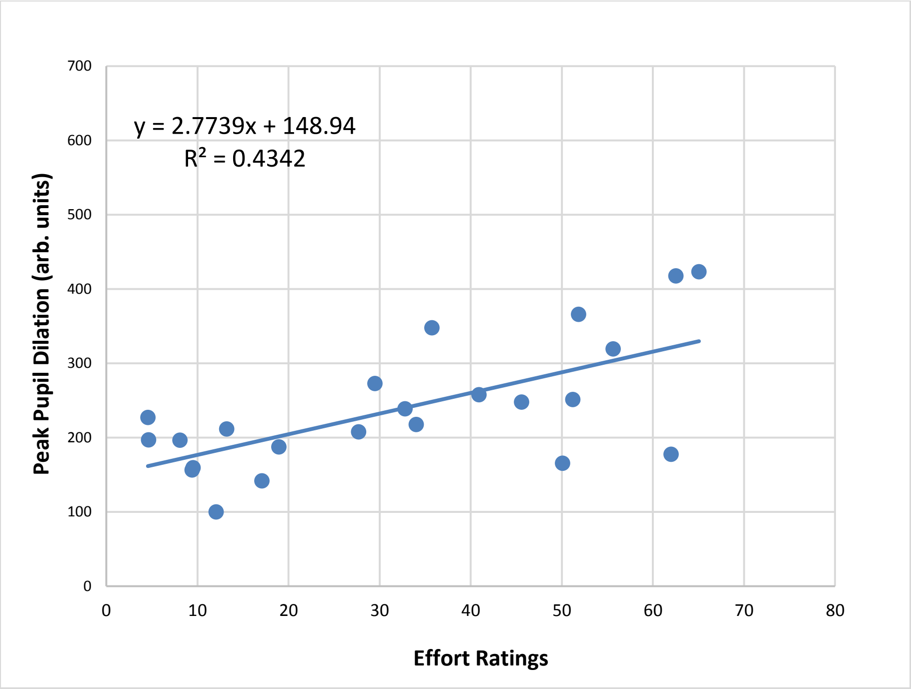


Figure 3-10 Simple linear regression plots for predicting PPD based on perceptual ratings of effort.

3.3 Results (Experienced group)

3.3.1 Auditory-perceptual data analyses

Intra-rater reliability was calculated for each experienced listeners, by correlating the rating results of test-retest for both strain and listening effort. The correlation range for strain ranged from 0.72 to 0.86 and from 0.71 to 0.78 for effort, indicating moderate-to-high correlations. Interrater reliability was calculated for each of the two rated features through SPSS (Version 24, Armonk, NY). The Cronbach's alpha value was 0.86 for strain and 0.83 for effort. The interrater reliability outcome confirms high reliability among experienced listeners for the rating task. For each of the 23 talkers, 2 sets of strain rating scores that could range between 1 and 100 and 2 sets of ratings for effort, again ranging between 1 and 100 each, were generated based on each listener's test and retest auditory-perceptual ratings. Ratings per single feature were then averaged across trials (test, retest) to achieve a single strain value and a single effort value per talker for each listener (Appendix C). Once all 3 experienced listeners completed the experiment, the strain and effort rating values were averaged across all listeners to achieve a single value for each dimension and standard error of the mean was also calculated (Figure 3-11).

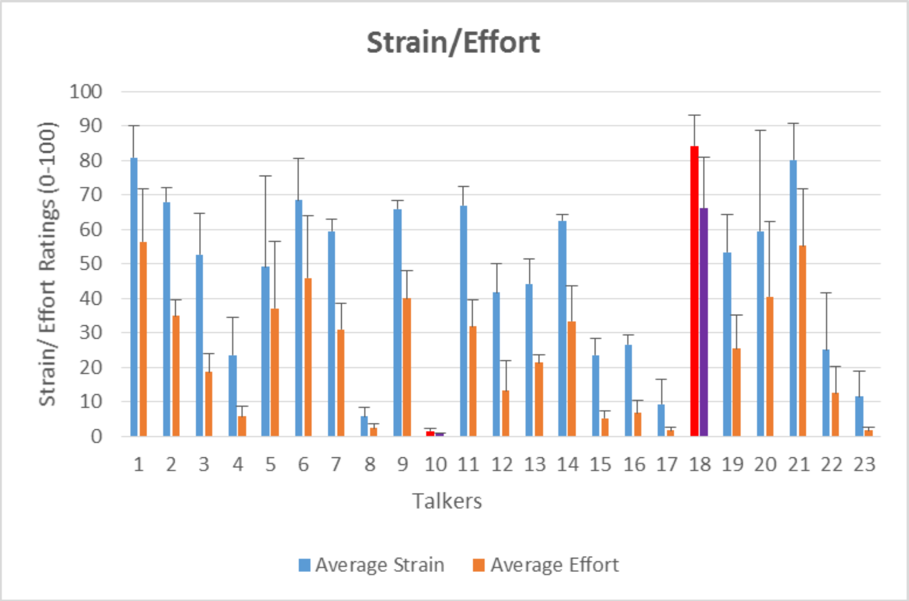


Figure 3-11 Average strain/effort ratings per talkers; standard error of the mean; highest/lowest rated talkers color distinguished.

The correlation coefficient was also calculated in order to examine the relationship between strain and effort ratings in the experienced group. The results indicated that a very strong correlation (r = 0.94) existed between rating of these two auditory-perceptual features (Figure 3-12).

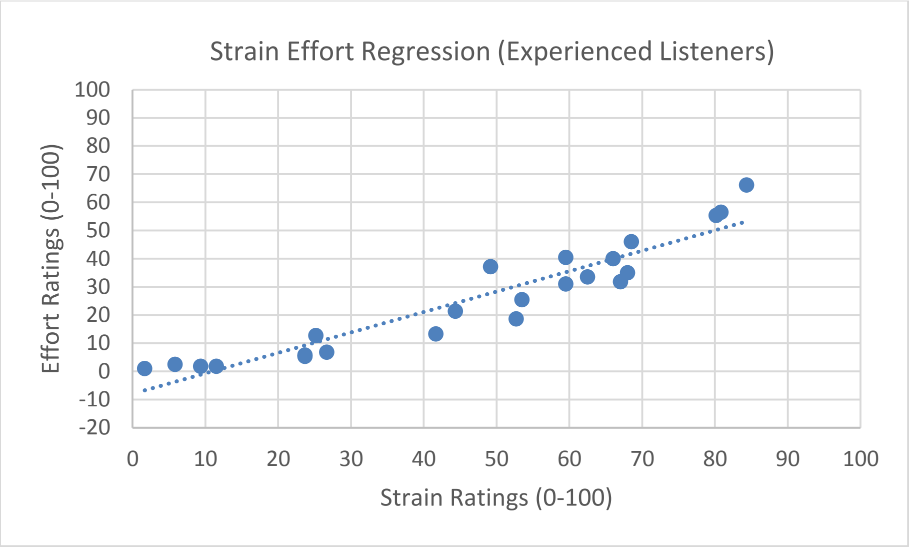


Figure 3-12 Regression between strain and effort.

3.3.2 Pupillometry

Pupil data were recorded throughout the experiment via the eye tracker. In order to start analyzing pupil data, the raw data had to be processed in several steps before formal final analysis and visualization and the same process for cleaning tracks which was followed for the naïve listeners were also used for experienced listeners' raw pupillary data. About 4.3% of tracks in this group were discarded due to dropouts, too many variations, or long blinks. Like the naïve group, our interest was on PPD values. For this group too, the same PPD determining procedure as the naïve group was followed. Figure 3-13 displays all pupil tracks averaged across listeners along with the baseline and PPD.

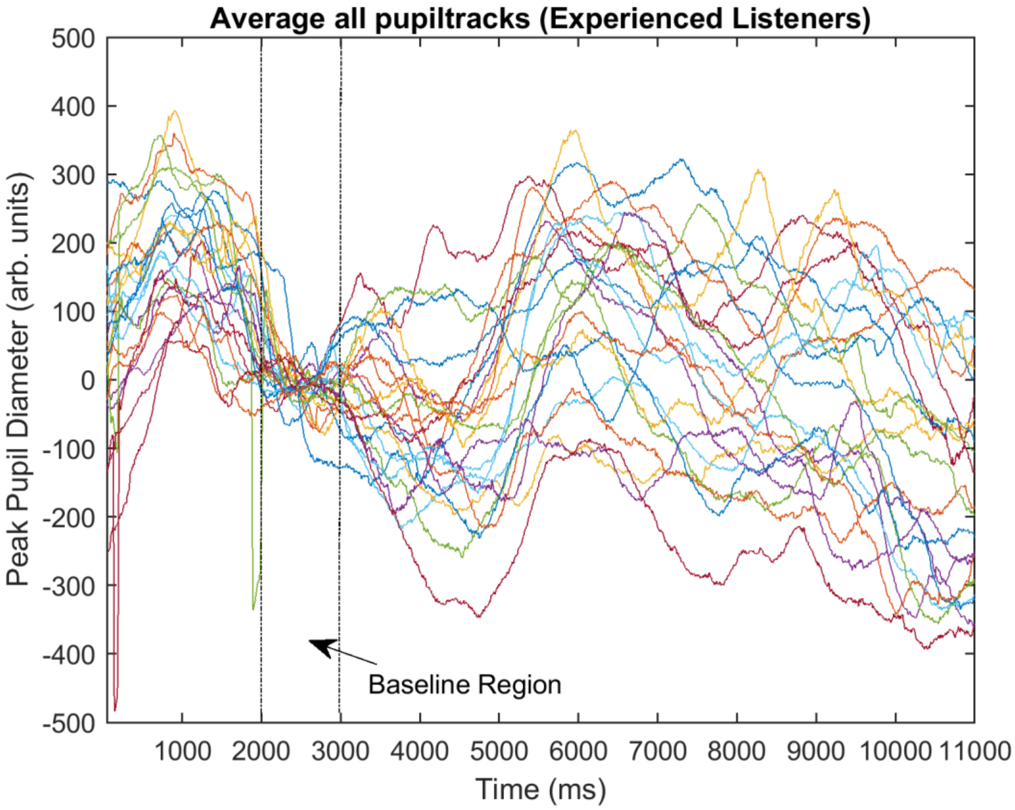


Figure 3-13 Pupil tracks averaged across all experienced listeners (n=3).

Also, the correlation between the PPD and strain and PPD and effort were calculated and displayed in Figures 3-14 and 3-15, respectively.

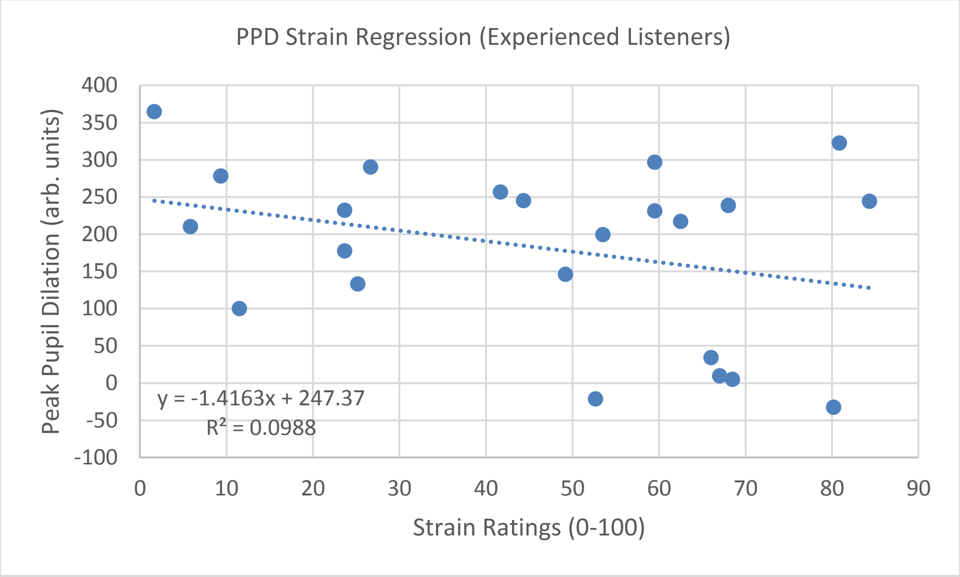


Figure 3-14 Strain PPD Regression (Experienced Listeners).

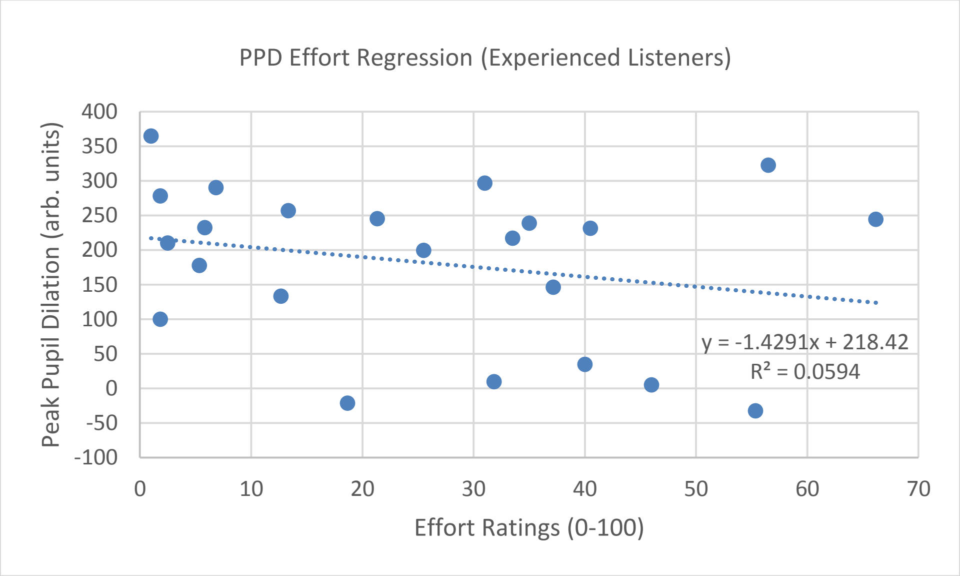


Figure 3-15 Effort PPD Regression (Experienced Listeners).

In order to explore if talkers who were judged to exhibit high and low perceptual results appear relative to each other, the pupil tracks of 4 talkers, two representing the highest auditory-perceptual ratings of strain and effort (Talkers 18 and 1) and two with the lowest on effort (Talkers 10 and 17) are displayed in Figure 3-16, respectively.

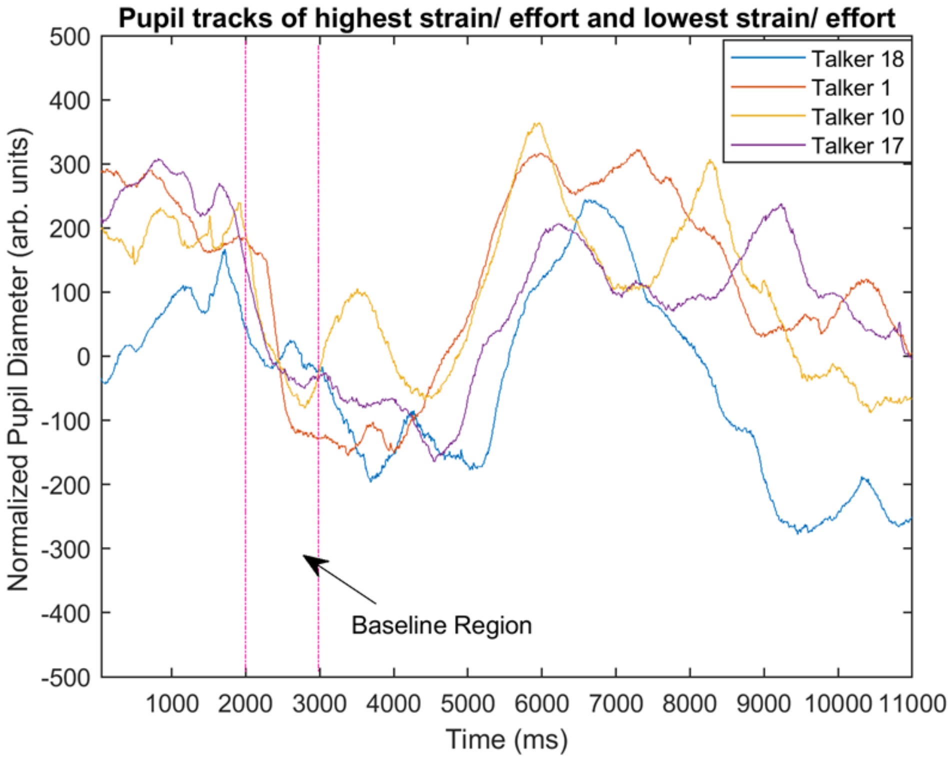


Figure 3-16 Pupil tracks of all listeners (averaged across all) listening to the 2 talkers with the highest strain/effort (18 and 1) and the 2 with the lowest effort ratings (10 and 17).

3.4 Discussion

3.4.1 Naïve Listeners (Auditory-Perceptual Evaluation)

This study was designed to examine pupillary reactions evoked by dysphonic voices of speakers with AdSD in listeners who were presented with these speech samples. This involved auditory-perceptual ratings of both strain and effort by normal-hearing listeners. The AdSD voice samples selected for use varied widely in severity in order to evaluate potentially differential responses to the stimuli by listeners. The objective of this investigation sought to assess whether voice samples characterized by increased levels of strain would also be found to correspond to the perception of listening effort.

The results of the auditory-perceptual evaluation of strain revealed that talkers demonstrated various degrees of strain and were rated to require increase listening effort. This finding is consistent with previous studies which report that the voice quality of AdSD speakers is perceptually judged as being significantly strained and effortful (Eadie et al., 2017; Isetti, Xuereb, & Eadie, 2014; Cannito et al., 1997). This was confirmed by data from the current study where some of the voice stimuli were rated as less strained (e.g., 8, Talkers 4, 8, 10 and 15) compared to others who were consistently judged as highly strained (e.g., Talkers 1, 2, 9, 18 and 21). Also, when ratings were averaged across listeners the data demonstrate that the higher the ratings for strain, the more listening effort was expended; this finding is clearly indicated by subjective ratings of this perceptual feature. Of substantial importance is the fact that relative to listener ratings, our data showed that interrater reliability was quite high for both the feature of strain (0.98) and effort (0.97) and the intra-rater correlation was moderate-to-high for both strain (0.56 to 0.96) and effort (0.58 and 0.97).

The results indicated that Talkers 8 and 10 were rated the lowest in terms of strain and also judged to require the lowest degree of listener effort; in contrast, Talkers 1 and 18 were judged as exhibiting the most strained voices and evaluated as requiring the most listening effort. To our knowledge, no study to date has evaluated perceived listening effort in the context of speakers with AdSD and our results confirm increased listener effort is required as speaker severity increases.

As indicated, out of 23 talker stimuli, 21 samples were judged to have higher strain ratings than listening effort, a finding that was not unexpected. Interestingly, results revealed that listeners rated stimuli from Speakers 5 and 20 to have higher ratings for effort than for strain a finding that was consistent across all listeners. These results are consistent with previous findings suggesting that the challenges faced by listeners are beyond those related to audibility (Pichora-Fuller et al., 2016) or intelligibility (Whitehill & Wong, 2006). Such perceptual challenges increase when more cognitive effort is expended to channel attention and concentration in order to achieve a listening goal; this is particularly important when the quality of an auditory signal is distanced from optimal (Pichora-Fuller et al., 2016). The auditory-perceptual ratings for these two talkers (5 and

20) also confirm that listeners were in fact rating the target stimuli as requested. In other words, although the stimuli were not perceived to be highly strained, they still deviated from normal which subsequently required increased listener effort, an observation that was uniformly indicated by their ratings.

3.4.2 Pupillometry

The other aim of this study sought to evaluate whether the pupillary response (i.e., PPD) and its sensitivity to various stimuli retains its value as a measure of cognitive load and listening effort. That is, the relationship between the current AdSD samples and listening effort was examined using pupillometry. To our knowledge, this is the first study to empirically evaluate pupil responses and the amount of effort expended while listening to disordered voices. The goal herein was to explore the variability in processing effort indicated by PPD. Pupil size is reported to be impacted by cognitive load and more specifically, language processing tasks such as hearing and reading words (Brown et al., 1999) or sentences (Hyona et al., 1995). The present aim was to determine whether a more strained voice sample would be associated with an increased PPD with respect to baseline. If confirmed, then such increased PPD would be assumed to reflect increases in processing load or listening effort and the amount of cognitive resources utilized by a listener in a speech reception task (Wendt, Dao, & Hjortkjær, 2016).

Processing demand is reported to be imposed by either stimulus factors such as linguistic complexity or noise (Pichora-Fuller et al., 2016; Wendt, Hietkamp, & Lunner, 2017), or as addressed in our study, the quality of the voice sample being assessed. Additionally, it is possible that listener factors such as the capacity of working memory or hearing impairment will influence both perceptual ratings and PPD. Thus, consideration of both speaker and listener factors is essential as they are reported to influence processing demands (Pichora-Fuller & Singh, 2006; Rabbitt, 1968). Our results also revealed a strong, positive correlation between strain and PPD (0.73), as well as for effort and PPD (0.65) when averaged across all listeners. Figure 3-17 shows the pupil tracks along with the PPDs and baselines for all 23 speaker samples.

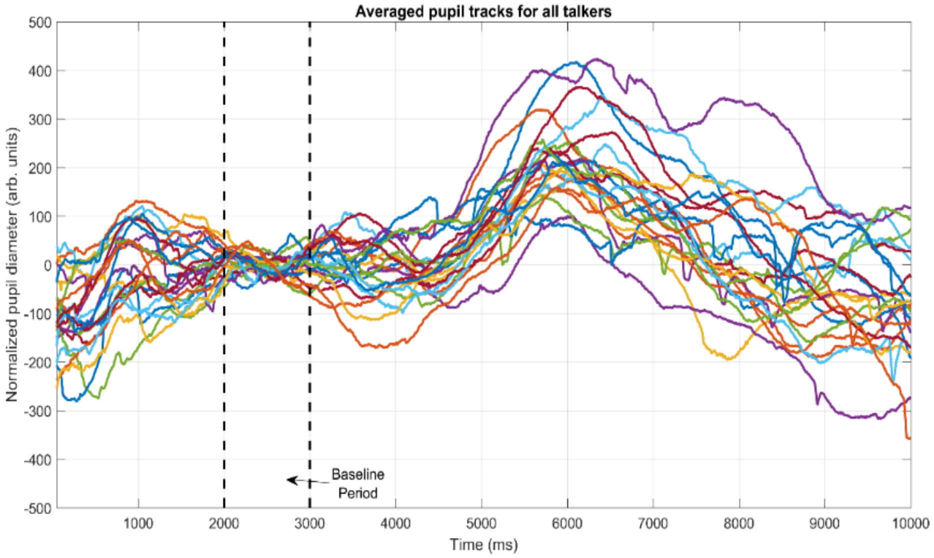


Figure 3-17 Average pupil tracks for all 23 talkers.

Since pupillometry is often reported as an indicator of cognitive load, the pattern indicated by our pupillary response tracks (Figure 3-17) appears to be consistent with past suggestions (Xie & Salvendy, 2000) for measuring cognitive load (i.e., "the load imposed on working memory by the cognitive processes that learning materials evoke and can be measured at different levels" (Antonenko, Paas, Grabner, & van Gog, 2010, p. 426). Various cognitive loads on memory have been defined and distinguished as: "instantaneous load (dynamics of cognitive load which fluctuates every moment as of the onset to the offset of the carrying out a task or tasks), peak load (the maximum point of instantaneous load during a task), average load (the mean intensity of the load), overall load (the experienced load based on the working procedure), and the accumulated load (total amount of load experienced during the task and falls below the peak load" (Xie & Salvendy, 2000, pp. 88-89).

In regard to the present data, variations in pupil response from the onset of the voice stimulus to its offset can indicate instantaneous load; the peak load matches the PPD and the accumulated load is the total effort required throughout the task of listening to each sample. In fact, pupil fluctuations may allow tracking the smallest variations in brain activity associated with cognitive load through direct evaluation of specific time instances such as the peak.

Based on the above information, the PPD data from our listeners' were analyzed further and interestingly, the peak of pupil responses was observed to occur at almost the same time periods during the presentation of stimuli. More specifically, the time to peak latency ranged between 5000 to 6560 ms which included the first 3 seconds of the prompt. Latency is reported to vary from individual to individual and between eyes depending on the intensity of the stimulus (Bergamin & Kardon, 2002); however, it is negatively correlated with factors such as signal-to-noise ratio (SNR) and decreased SNR increases the time to peak latency (Zekveld, Kramer, & Festen, 2010). Our participant listeners were consistent in terms of their time to peak latency ranges and all PPD ranges observed were between 5000 to 6500 msec. For this reason, we concluded that the PPD had been evoked in response to the unique quality of the voice/speech stimuli. A closer assessment of the pupillary data revealed that two talkers (1 and 18) who were perceptually rated as having the highest perceived levels of strain and increased listener effort, also demonstrated the highest PPDs. Moreover, talkers rated lower on strain and effort, demonstrated smaller PPDs.

The study data were collected in both test and retest conditions from each listener and the same stimuli were presented randomly in both scenarios. Then, the differences between pupil responses and PPD in both test and retest were examined to see if the similar responses were observed. For Talker 1 who was one of the talkers with the highest level of strain and for whom listeners exhibited a high PPD value, his data were examined more closely in various test-retest presentation orders for those listeners who had these two stimuli with the most distance in the order of presentation in test and re-test. These presentation orders included position order 7th in the initial test exposure and in position order 22nd in the re-test exposure (Figure 3-18) and for presentation order 21 (test) and 11 (re-test) (see Figure 3-19). In all these instances, the first presentation was always followed by a greater PPD when compared to the second presentation of the same stimulus.

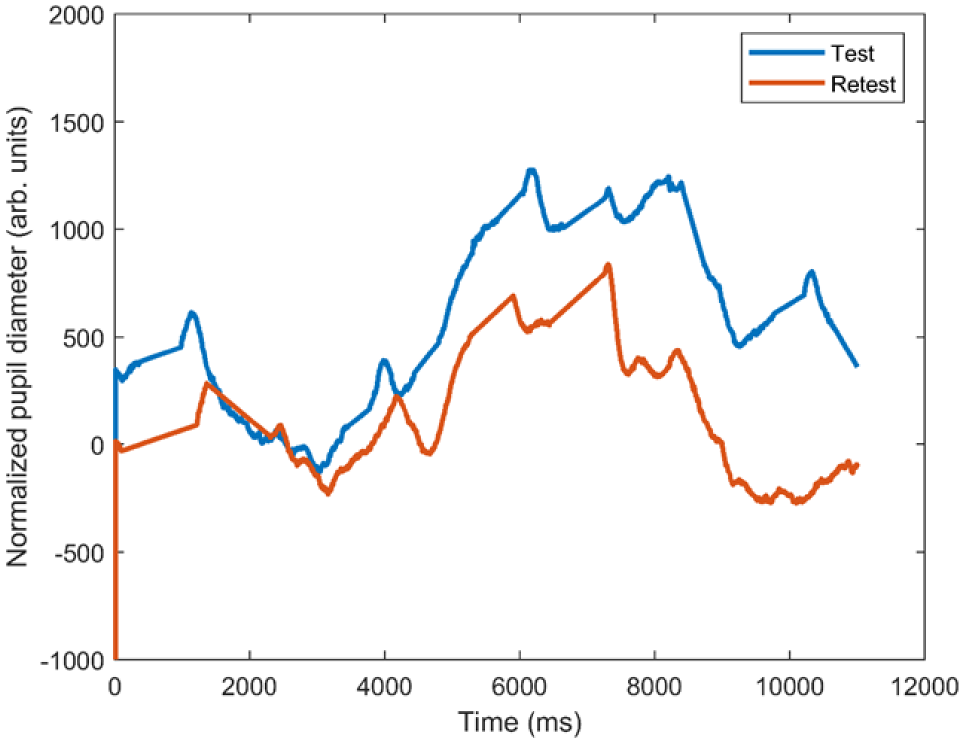


Figure 3-18 Talker 1, presentation order 7th and 22nd.

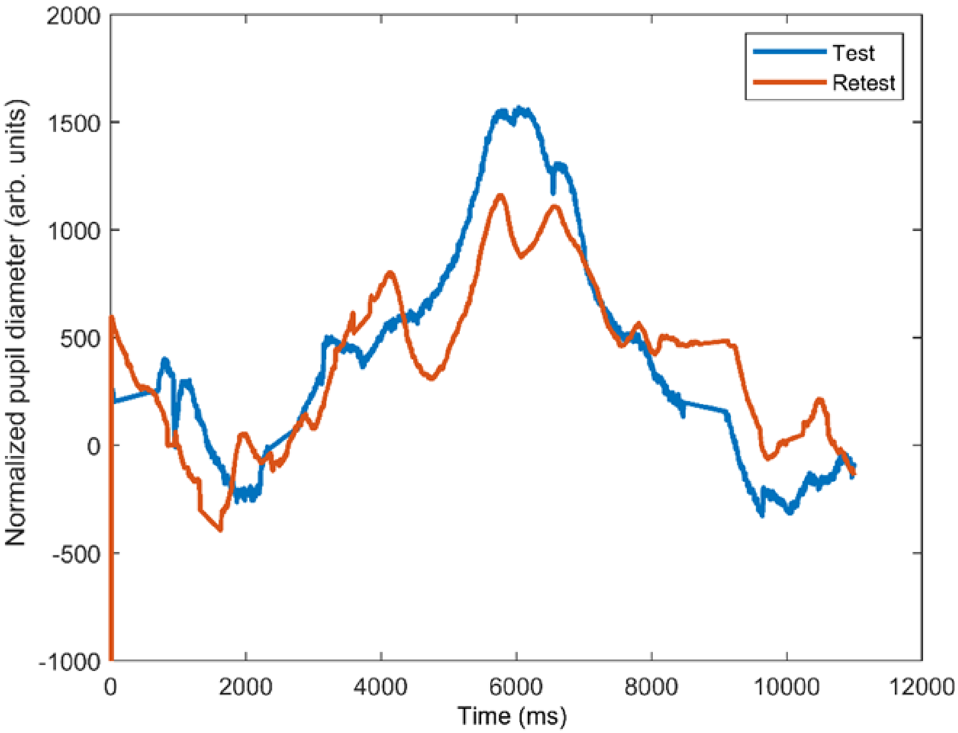


Figure 3-19 Talker 1, presentation order 21st and 11th.

The decrease in PPD values may be due to the fact that listeners have, at least to some extent, already habituated to the stimulus and a cognitive schemata is formed. Therefore, the load on working memory is reduced as a schemata is handled as a single element of information (Antonenco et al, 2010). According to cognitive load theory (CLT), working memory is limited in capacity. Accordingly, the time for holding and processing information via working memory (Miller, 1956) is restricted in comparison to long-term memory which is of virtually unlimited capacity (Sweller, Van Merrienboer, & Paas, 1998). Working memory also is reported to be able to host 7±2 information elements (Miller, 1956; Cowan, 2001) and this reduces when information needs to not only be stored and remembered, but also processed (Antonenco et al., 2010). In contrast, information already learned or experienced, appears to be stored in long-term memory in the form of cognitive schemata. Accordingly, this type of information is handled as a single information element and the cognitive processing load is decreased. As a result, cognitive load imposed by a given task is lowered if prior knowledge and expertise exists

(Antonenco et al., 2010). Cognitive schemata can even become automated, requiring little processing load if the task or aspects of a task are repeated and practiced (Schiffrin & Schneider, 1977), a process which in turn may lead to freeing working memory resources (Antonenco et al., 2010). Given the current experiment, the repetition and exposure to the voice stimuli evaluated seems to have provided listeners with a prior exposure that potentially may have led to lower cognitive load during the retest session.

The results of the present study are consistent with previous studies which indicate more listening effort is required in challenging and adverse language processing conditions (Hällgren, Larsby, Lyxell & Arlinger, 2005; Koelewijn, Zekveld, Festen, & Kramer, 2012; Wendt, Dau, & Hjorkjær, 2016). Further, the current data support prior assumptions that cognitive load contributes to the effort demand experienced during challenging listening conditions (Rabbitt 1968, Zekveld et al., 2010). There are also consistencies with previous studies that report influences on pupil dilation in language processing tasks such as reading or with the auditory presentation of words and sentences (Brown et al., 1999; Hyona et al., 1995). The present findings are consistent with those reported by Kramer, Kapteyn, Festen, and Kuik, (1997) and Zekveld et al. (2010). Both of these studies examined listener effort through pupillometry and reported larger mean PPD for their normal hearing listeners in low intelligibility compared to high intelligibility conditions, ascribing larger mental effort to such challenging language conditions. When viewed collectively, our data on AdSD samples support the notion that when confronted with stimuli characterized by an abnormal vocal quality, listeners demonstrate a physiologic response that corresponds to their auditory-perceptual assessments. These findings provide valuable insights into the demands of effective verbal communication in general, and the challenges that may occur in the presence of disordered speech or an abnormal vocal quality specifically.

3.4.3 Discussion (Experienced Listeners)

This phase of the study evaluated the extent of pupillary responses elicited in listeners with prior exposure to voice sample from those with AdSD, as well as other voice disordered speaker samples. The normal hearing listeners who were all professional clinicians and had ample experience with disordered voices and various patient groups with dysphonia, were asked to indicate their perceived judgement of the degree of strain in audio stimuli and their listening effort. The samples were of various degrees of severity in an attempt to elicit a wide of range of ratings and reactions to the stimuli and correspondence to the perceived assessment of listening effort. As indicated for the naïve group, the talkers had various degrees of strain which was indicated in strain and effort ratings of this group. The intra-rater reliability in the experienced group, although still high for both strain (0.72-0.86) and effort (0.71-0.78), and the inter-rater reliability (strain: 0.86; effort: 0.83) were lower than the novice group which is similar to the interrater reliability results reported by Eadie et al. (2007) with a similar talker group.

The way listener experience (with dysphonia) influences the ratings is not totally clear and some report increased sensitivity of the disorder and some indicate that naïve listeners rate disordered voices more severely than experienced ones as the exposure may reduce sensitivity (Kreiman, et al., 1990; Kreiman, et al., 1992; Damrose, Goldman, Groessl, & Orloff, 2004; Laczi, Sussman, Stathopoulos & Huber, 2005; Eadie et al., 2007). Generally, in our study, experienced listeners rated the talkers lower on perceived listening effort (less effort demanding) and 17 talkers out of the 23 had lower ratings from this group compared to the judgements naïve listeners had provided for them on effort. More specifically, average rating for effort by the experienced group was lower for Talker 1 (56.5) which was rated by both groups as demonstrating the highest effort compared to naïve ratings (62.52). Experienced listeners also gave lower values to low effort demanding talkers: Talker 10 (1) and Talker 8 (2.5) compared to naïve ratings for the same talkers (Talker 10= 4.62 & Talker 8= 4.57) indicating that those talkers demanded less listening effort from experienced listeners. This differs from the report by Eadie et al. (2007) that showed no significant differences in ratings between the two listener groups. They do report a strong trend for naïve raters to perceive AdSD talkers as being more effortful, an observation that corroborates our current results. This higher effort rating by naïve listeners may be the result of internalizing self-perceived effort associated with voice production (Brandt, Ruder, & Shipp, 1969).

In terms of strain, there were no differences between the two groups' ratings. The experienced group performed slightly differently, rating 12 out of 23 talkers lower on strain compared to the values on this dimension given by naïve listeners. The auditory-perceptual ratings for Talkers 1 and 18 which were rated by both groups to exhibit the highest levels of strain, are slightly lower when compared to the experienced group (M = 84.33 vs. 80.83, respectively) when compared to the average ratings from the naïve group for the same talkers (65.05 vs 62.52). The same pattern is observed for Talkers 8 and 10 who were rated on average as having low degrees of strain by both groups (5.83 vs. 8.95 & 1.66 vs. 9.41). As the ratings indicate, naïve listeners gave them higher values on strain than the experienced group. Based on these data the level of a listener's experience and exposure seems to have reduced the impact of voice disorders on the experienced group's ratings as has been previously reported in the literature (Laczi et al., 2005). Of course, it also must be noted that the experienced group was fairly small (3 participants) compared to the naïve group (n = 20). For that reason, data obtained from a larger sample of experienced judges may help clarify the potential impact of exposure more fully.

3.4.4 Pupillometry

Similar to the naïve group, pupillary data from the experienced listeners were also analyzed to evaluate physiological reactions (i.e., PPD) and the potential function of PPD as a measure of cognitive load and listening effort. In other words, the aim was to assess the variability in processing effort as indicated by PPD and to see if experienced listeners would also go through more or less listening effort when processing strained and listening effort demanding voices. Listener factors like hearing impairment, capacity of working memory (Pichora-Fuller & Singh, 2006; Rabbitt, 1968), attention control (Unsworth & Robinson, 2018) are reported to be influential and we were curious how experience and training with disordered voices would impact PPD values. The results revealed a negative correlation (-0.31) between PPD and strain ratings and negative correlation (-0.24) between strain and perceived listening effort ratings. One of the high strained rated talkers (Talkers 1) had evoked high PPD values in experienced listeners but this pattern was not true for all and we did observe some highly strained voices with low PPD values. Also, Talker 10 who was rated very low on strain and listening effort had created a high PPD value in experienced listeners. It must be noted that we only had three experienced listeners in this project and this small number could have impacted our results. Examining the pupil tracks of the highest and lowest rated talkers on effort (Figure 3-16) revealed that the PPD patterns of the highest perceptually rated Talkers 1 and 18, and the two lowest rated Talkers 10 and 17, were similar. This could mean that our experienced listeners were allocating similar attention and concentration on processing Talkers stimuli independent from perceptual ratings. It is reported that experienced listeners may take more time judging voice samples of because they have more voice information to process and rely on for judgement compared to naïve listeners who make a snap evaluation (Sofranko-Kisenwether & Prosek, 2015). In this study the length of rating time was not controlled, however, PPD was observed as an indicator of cognitive load and effort to process stimuli and results show that our experienced listeners had higher pre-baseline PPD activity compared to naïve listener and they also displayed uniform pupillary responses for all talkers. This is due to their engagement in the rating task and the effortful allocation of attention from before the onset of task and when they are instructed and prompted on what is to be presented. Due to the nature of training and experience of such listeners and their profession which requires them to listen to many disordered voices and diagnose and judge them by accessing their mental resources, their PPD activity suggests continuous attention allocation throughout the task independent of the extent voice disorder dimension. In fact, such PPD responses as an indication of maintaining items active in working memory and continuous allocation of attention to tasks have been reported in the literature (Unsworth & Robinson, 2018) and our experienced listeners' result is in line with that.

3.4.5 Conclusions

This study addressed auditory-perceptual evaluation of two features related to abnormal voice quality, namely strain and listener effort, in relation to pupil responses. The present data offer important observations and provide valuable insights into how naïve listeners rate aspects of voice quality, in this case strain, and their simultaneous evaluation of the work required to audit these samples or what is termed listening effort. First, listeners consistently assigned greater listening effort (i.e., more demand placed on the listener) to voice samples that were judged to exhibit greater perceived levels of strain. Second, because listening effort may include multiple perceptual factors, a disordered voice might be rated lower on strain (e.g., Talkers 5 and 20), but higher on listening effort due to the overall, composite quality of the voice. Given the nature of voice quality deviation in those diagnosed with AdSD, this suggestion is not unwarranted. Third, like previous studies, intelligible voices are rated as demanding variously increased degrees of listening effort which confirms the fact that listening effort goes beyond simply understanding what is being said and processing its linguistic content. However, it should be noted that none of the speaker samples used in the present study were characterized by reduced intelligibility; rather, the samples were altered by changes in the consistency and flow of speech production consistent with the cardinal characteristics of AdSD. Fourth, the stimuli which were subjectively rated by listeners as being more were also generally observed to provoke an increase in PPD; this finding suggests a potential relationship of the listening task to aspects of cognitive load and subsequently, listening effort. It is, however, important to acknowledge that this cognitive load also was observed to decrease with exposure and habituation over the course of the experiment.

Our listeners also rated speaker stimuli based on their individual internal standards. While excellent reliability was documented in our study, it would be valuable to determine if adding perceptual anchors to the scale might influence the ratings and concurrent PPD values. Finally, the temporal gap between test-retest was relatively short (10-15 minutes). Future studies might seek to assess longer gaps between test-retest to identify whether the exposure to the stimuli would fade away and PPD would be altered within the context of an increased break. It is possible that the duration of a break, or the distance in time away from a request for judgments might have a differential influence on one’s physiologic response to abnormal vocal stimuli.

Chapter 4

4 Auditory-perceptual and pupillometric evaluation of vocal roughness in tracheoesophageal speech

The focus of the second project was on perceptual and pupillometric evaluation of vocal roughness in total laryngectomies who used tracheoesophageal (TE) speech as the alternative speech production mode. As indicated before, TE is clearly distinguishable from laryngeal speech as it is aperiodic and low in pitch. In addition to the gold standard of perceptual evaluation of the TE speech, listeners' cognitive and physiological responses to abnormalities of TE voice was measured via pupillometry as an indicator of listening and cognitive effort.

4.1 Methodology

4.1.1 Participants

A total of twenty adults (8 males, 12 females; age range = 18-32 years; mean: 24.65 years) participated in the current study. It must be noted here that some of these participants had already taken part in the first experiment (Chapter 3). They were divided into two groups: with-anchor (4 male, 6 female; mean age =24.7) and no-anchor (4 male, 6 females; mean age = 24.6). The allocation to each group was random and based on their participation order: every other listener was assigned to the with-anchor/no-anchor group. The recruited number of participants was based on a power analysis calculated using GPower with an effect size of 0.4. Participants were all native English speakers and self-reported as normal hearing. They did not have any formal education or training in voice disorders such as clinical or course work. Also, if they had an upper respiratory infection within the week prior to the experiment, they were asked to come back once the problem had been resolved for 14 days. Participants were also scrutinized in terms of medications they were on. If they were on any of the listed medications (Appendix A), they were not be able to participate in the study as such medications might interfere with the results. Each participant completed the study in a single 45-minute session. Each session involved 10 to 15 minutes for task instruction, instrumentation adjustment and calibration, 7 to 10 minutes for the experimental protocol, and 7 to 9 minutes for the retest procedure. They were also given a 10 minute break between the test and retest. The experiment session began with task instruction which included explanation the aim of the study and the features under assessment. Participants were also provided with the written explanation of what the definition of the dimensions (roughness and listening effort) were in case they wanted to refer to during the experiment (Appendix H, REB # 112674). The experiment was conducted in a softly lighted room and the light was consistent throughout the room to prevent reflexive dilation in reaction to changing luminance on the retina (Winn et al., 2018).

4.1.2 Auditory Stimuli

The audio stimuli of this study came from an archive of the Voice Production & Perception Laboratory at the University of Western Ontario. The samples were read by 20 talkers who relied on TE speech production. All voice samples were gathered via digital recording obtained in either a quiet clinical setting free from ambient noise or within a sound-treated environment. All voice samples included sustained vowels, a standard reading (The Rainbow Passage; Fairbanks, 1960), a short monologue (approximately 60 seconds), and a variety of single word stimuli along. In some instances, standard sentences from the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) were also gathered. Recording of speech stimuli occurred after informed consent was obtained, along with demographic information for each speaker. A directional microphone (Shure PG-81) attached to a desktop microphone stand was used in all recording. During the collection of each sample, a microphone-to-mouth distance of 15 centimeters was maintained and was checked prior to each task. Each participant speaker's sample of voice/speech stimuli was recorded onto a laptop computer at a sampling rate of 44.1 kH using the Multidimensional Voice Profile (SonaSpeech II, Kay Pentax, Lincoln Park, NJ). Volume input levels were adjusted for each speaker at the beginning of each session and were monitored during the recordings using SonaSpeech II to avoid any under- or over-driving of the input signal. The second sentence of the Rainbow Passage ("The rainbow is a division of white light into many beautiful colors") was then extracted to serve as the stimulus of this study. The stimuli were also equalized for their Root Mean Square (RMS), so that they were presented at approximately the same intensity during playback. Once the experiment started, listeners were presented each stimulus in a random order. All stimuli began with the oral cue “Please listen to the following stimulus” to prompt the initiation of the upcoming stimulus. This oral cue had been read by a male normal native English speaker and lasted 3 seconds. Listeners were then instructed through another oral cue, at the end of stimulus presentation, to indicate their ratings of roughness and listening effort through the computerized visual analog sliders. The spoken cue “Please indicate your ratings after the beep” was played after the offset of the stimulus and lasted for three seconds as well. For the no-anchor group, the spoken cues and the stimuli were played automatically and once listeners indicated their ratings, they clicked “Next” to hear the next stimulus. For the with-anchor group, they were instructed to listen to the mild anchor and severe anchors before listening to each experimental stimulus sample. These anchors were selected from the TE archive by an experienced speech pathologist and clinician and were not part of the stimuli which were being rated. The left end point anchor represented a TE talker with a mild rough voice and the right endpoint anchor indicated a TE speaker with a severely rough voice.

4.1.3 Assessment of roughness and listening effort

Once each sentence was presented, participants indicated their judgement of perceived roughness and listening effort using two separate 100 mm long electronic sliders which represented a VAS (Figure 4-1) with 100 intervals (i.e., 1 through 100). The end points on the “roughness” VAS were marked as “mild” (left end) and “severe” (right end). The second slider, which was used to collect listeners’ ratings of listening effort, had end points marked as “none” (left end) and “extreme” (right end). Listeners were instructed to move the slider handles between the end points to mark their ratings of roughness and their own listening efforts which would be between 0 and 100.

While participants were listening to the stimuli and were rating them, their pupil responses were recorded using the EyeLink 1000 (SR Research, Ottawa, Canada) eye tracker (Figures 3-2 & 3-3). This EyeLink 1000 consisted of the instrument tower mount. Listeners were asked to sit on a stationary chair at this tower mount with their chin on a chin rest and their forehead placed against a rest head. The EyeLink 1000 collected their pupil activity at a sampling rate of 1000 Hz. Participants could see the sliders on the monitor ahead of them and listened to the stimuli through headphones (Sennheiser, HD 205).

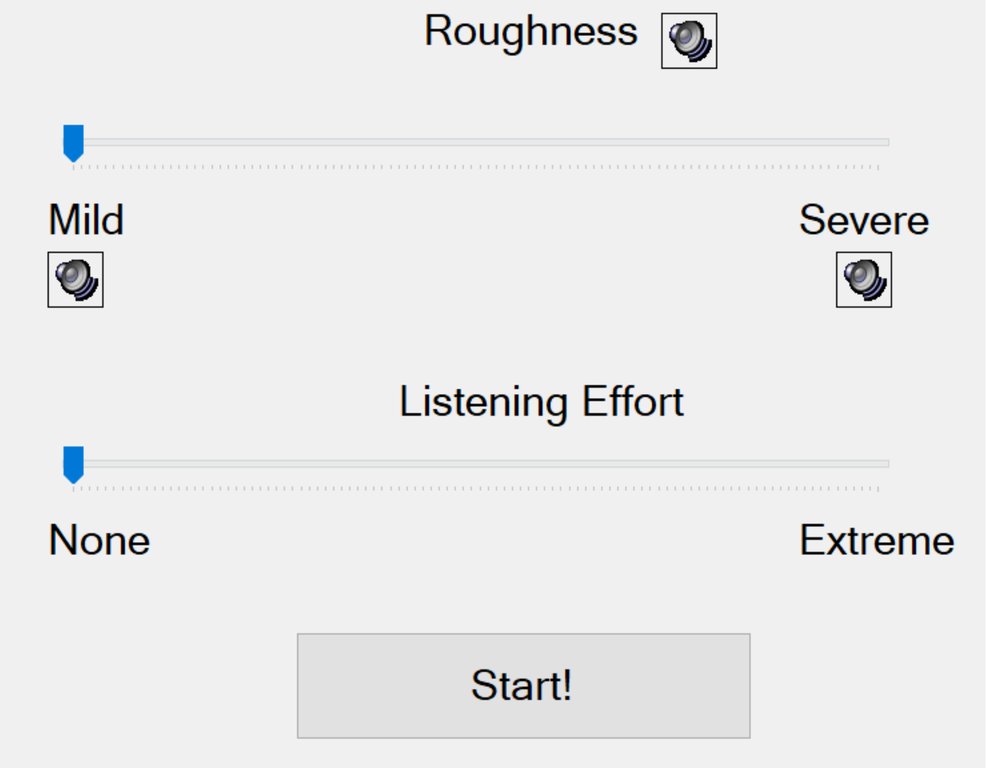


Figure 4-1 Appearance of slider rating scales for "Roughness" and "listening effort".

Procedure

At the beginning of the experiment session, participants were familiarized with the nature of the tasks they had to complete. First, they received brief explanation on the definition of “roughness” dimension and “listening effort”. Roughness can be defined as irregular fluctuation in vocal fold vibration which influences the vocal pitch period and in vocal amplitude, variation of which in a combined manner indicates the non-harmonic and low-pitched noise components of the voice (Imaizumi, 1986). Roughness is a long standing auditory-perceptual feature and is uniformly found in a majority of voice disorders. In this study, the CAPE-V definition of roughness (“perception of irregularities in the voicing source) was used (Kempster, Gerratt, Abbott, Barkmeier-Kraemer, Hillman, 2009)”. Listening effort was defined as “the amount of work required while auditing the speaker samples”. For the with-anchor group, they were instructed on how to rate the stimuli with respect to the audio anchors provided. Listeners in this group were asked to first play the mild anchor and then the severe anchor before playing each stimulus to refresh the standards against which they were rating the stimuli. The written definitions and the picture of the experiment scales, which included the order of playing the anchors, were also provided in case listeners wanted to refer to them while doing the experiments. Also, for the with-anchor group, the order of playing the anchors and the stimulus were numbered on the printed picture to prevent any confusions. For all participants, their positioning and height at the EyeLink 1000 was adjusted before the beginning of the experiment to guarantee best direct view of the right eye pupil. In addition, all participants were asked to keep their head and body stable and still throughout the test and avoid looking down or looking away from the monitor. They were also instructed to keep looking at the center of the monitor and were requested not to blink as much as possible or at least while listening to the stimulus.

Once listeners were seated at the optimum position and had headphones (Sennheiser, HD 205) on, calibration of the visual gaze and then validation of the measure was performed. This required listeners to maintain visual focus on a fixation circle on the screen and to follow it in order to calibrate the system. The rating experiment was initiated by the experimenter once satisfactory individual listener calibration and validation were achieved. The study stimuli were presented to listeners in a randomized fashion and they moved on to rating them after hearing the beep which followed the verbal cue at the end of each stimulus. Once both roughness and listening effort ratings were put in through the two sliders, they clicked “Next” button to move on to the next stimulus. The with-anchor group had to listen to the mild and severe anchor before each stimulus and then moved to the rating phase. The stimuli were randomized and presented twice as test and re-test for intra-rater reliability evaluation, with a 10 minute break between the sessions. Once all stimuli in the test and re-test phases were rated, a message appeared on the screen indicating the end of each trial.

Once the experiment was completed, all pupil tracks were normalized first so that the starting point of each track was at zero (0) second. This normalization process was due to the fact that that speech stimuli were randomly presented at different time stamps during the experiment. In addition, the raw pupil data which was gathered through the EyeLink 1000 during the experiment had to be processed in several phases before starting the actual analysis. Due to the nature of the experiment, there were inevitable blinks or dropouts which had to be identified. Once quick blinks (<125 milliseconds) were identified through examining each individual track, they were interpolated (interpolation began roughly 50 ms before the blink and end at least 150 ms after the blink) without changing the overall pattern of the tracking sequence. If the blinks or dropouts were too long, they were removed as interpolation might have led to manipulation of pattern. About 0.5% of tracks (n = 20) in the with-anchor group and 2.25% (n = 9) in the no-anchor group were had to be removed to eliminate risk of data distortion.

One of the foci of this experiment was on the peak pupil dilation (PPD) as a dependent measure. The peak in pupil dilation was determined as the maximum dilation during the presentation time of the speech sample relative to the mean dilation in the baseline period before the stimulus for each stimulus per listener. It was the last second of the three second prompt (Figure 4-7) which served as the baseline against which the stimulus was compared to determine the PPD or the highest point in the track.

4.2 Results

After listeners in both groups (with-anchor and no-anchor) completed the experiments, their perceptual ratings for the voice dimension of roughness and perceived listening effort were analyzed for reliability and other measures.

4.2.1 Auditory-Perceptual Evaluations (With-Anchor Listener Group)

Intra-rater reliability was calculated by correlating each listener’s test-retest ratings for both roughness and listening effort. The range of correlation for roughness was 0.44 to 0.84 and 0.27 to 0.85 for listening effort, indicating a moderate to high correlation for roughness. For listening effort, 7 out of 10 listeners had intra-rater correlations above 0.52, which indicates a moderate to high correlation for the majority of the listeners in this group. Interrater reliability was obtained through Cronbach’s alpha in SPSS (Version 24, Armonk, NY) for each of the two features rated perceptually. The interrater value was

0.96 for roughness and 0.95 for listening effort which confirms a very high reliability among listeners for the tasks.

The perceptual ratings generated for both features ranged between 0 and 100 (two sets per listener). After averaging them across (Appendix D), all talkers (roughness & effort rates) were plotted along with their standard error of the mean (Figure 4-2). Talker 15 was the highest rated on roughness and effort and Talker 6 was rated as the lowest on both features. All the talkers were generally rated higher on roughness than listening effort by listeners in this group. The range of perceptual ratings was 16.95 to 91.06 for roughness and 10.45 to 72.05 for listening effort in this group.

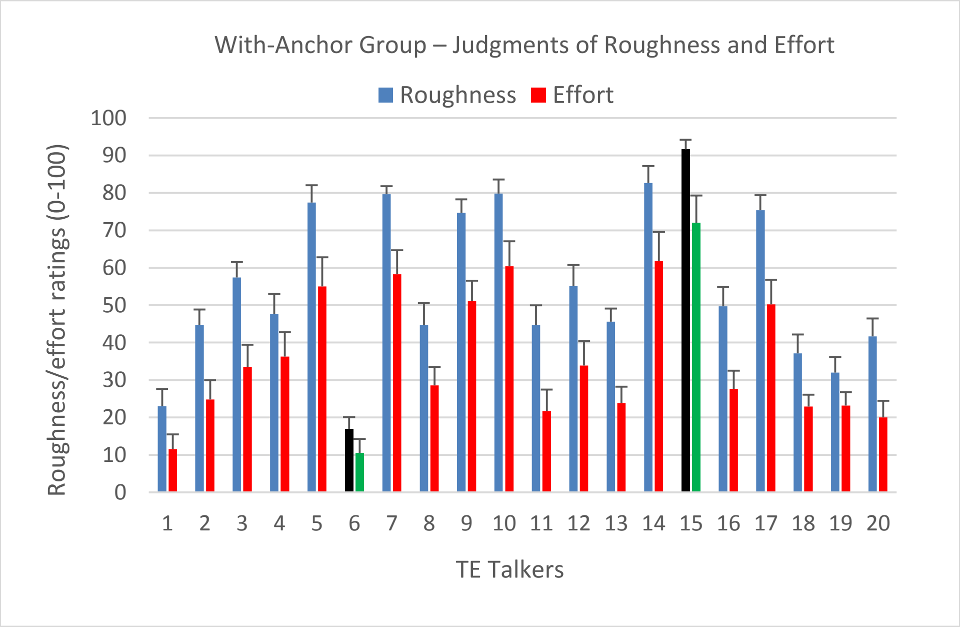


Figure 4-2 Roughness and effort ratings per TE talkers; highest and lowest rated talkers color distinguished (Black: roughness; green: listening effort).

The data were also plotted to represent the correlation between the two measures (Figure 4-3). Results indicate a very high correlation between the ratings of the two features.

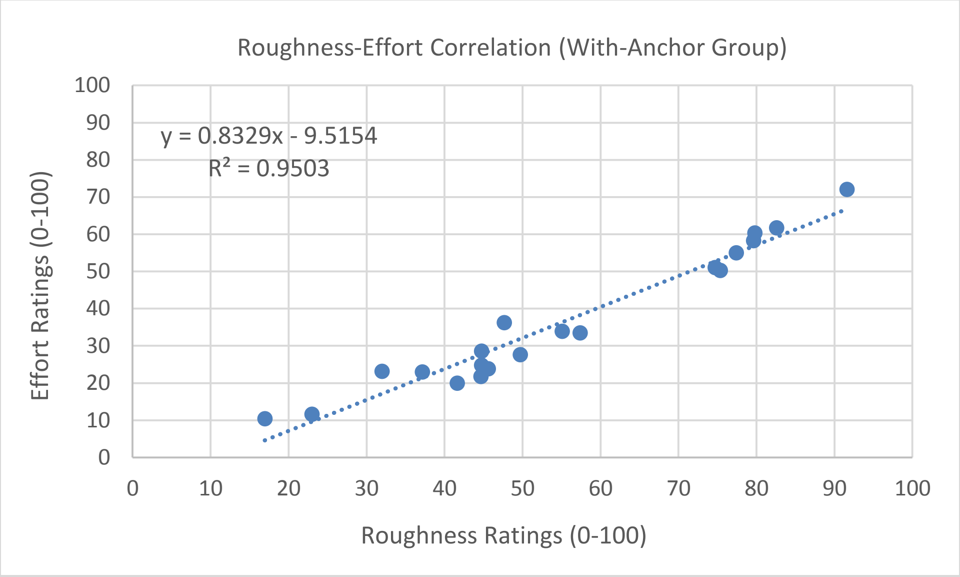


Figure 4-3 Regression between roughness and effort ratings by the with-anchor group.

4.2.2 Auditory-Perceptual Evaluations (No-Anchor Listener Group)

Intra-rater reliability was calculated for the no-anchor listener group in the same manner as the with-anchor group data. The intra-rater correlation range for roughness was 0.37 to 0.72 and for the subjective listening effort the range was 0.09 to 0.81. The interrater reliability was also calculated through Cronbach’s alpha in SPSS (Version 24, Armonk, NY) for the two perceptually rated features. The value for roughness was 0.93 and it was 0.90 for listening effort. The inter-rater reliability values obtained for this group is also very high although they are lower than the inter-rater reliability values of the with-anchor group.

The two sets of ratings were used to generate the plot of all talkers (Figure 4-4) with more information presented in Appendix E. Talkers 15 and 1 (Figure 4-4) were rated by this group as highest and lowest on roughness and the listening effort, respectively and are color distinguished on the graph. The range of ratings for roughness was 28.6 to 83.1 and the range for listening effort was 18.75 to 64.65.

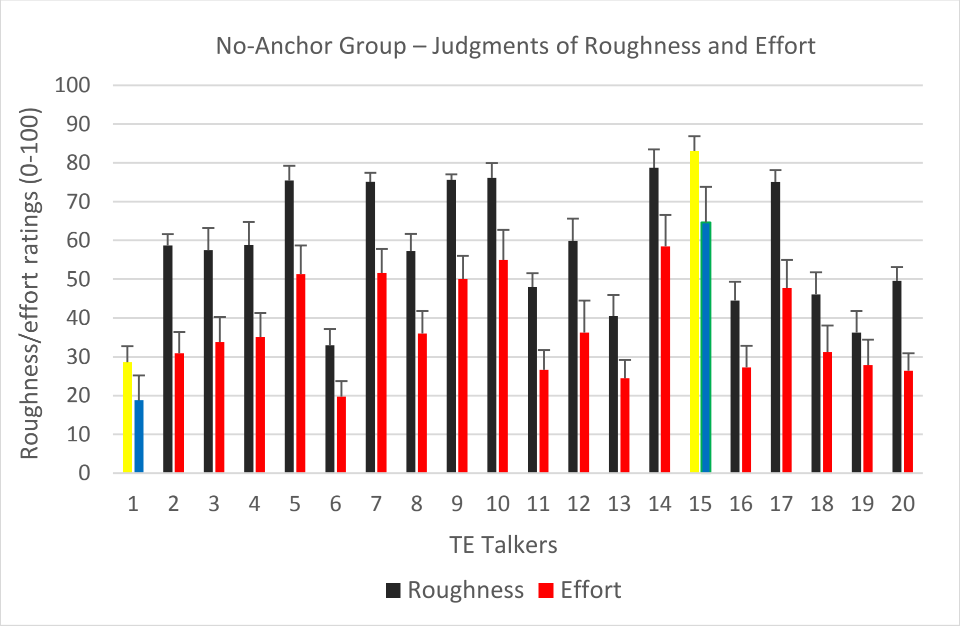


Figure 4-4 Roughness and Effort ratings per TE talkers; highest and lowest rated talkers color distinguished (Yellow: roughness; blue: listening effort).

The data were also plotted to represent the correlation between the two measures (Figure 4-5) and a very high correlation was achieved between the two features.

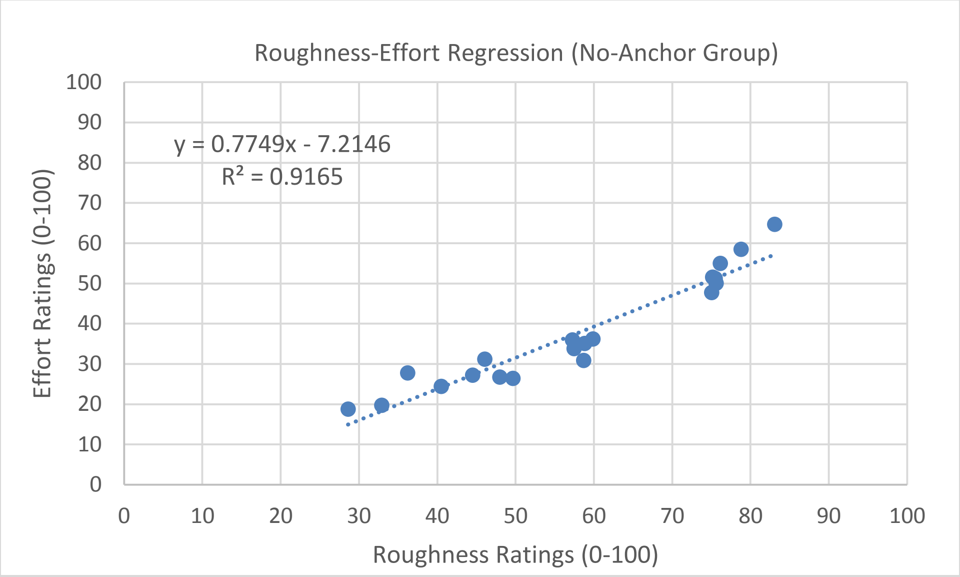


Figure 4-5 Regression between roughness and effort ratings by the No-anchor group.

4.2.2.1Statistical analysis

A repeated measures *ANOVA* was conducted to examine the effect of features (roughness and listening effort) and groups (with-anchor and no-anchor) and the interaction between them. The *a priori* significance level was set at 0.05 for all statistical tests. Significant effects were found for features (roughness and effort) (F 1, 18 = 49.36, p < 0.05, $\eta_P^2 =$ 0.733). No significant effect was found for group (With-anchor, No- anchor), (F (1, 18) = 0.24, p > 0.05, η_P^2= 0.013). In addition, no interaction was found between features and groups (F (1, 18) = 0.07, p > 0.05, η_P^2 = 0.004).

4.2.3 Pupillometry results (With-Anchor Listener Group)

In order to compare the pupil data, the recorded time stamps of all stimuli were first normalized so that the starting point of each sentence was at zero (0) second. Before final analysis and visualization, the raw pupil data were processed and cleaned as discussed earlier by selecting and removing distorted tracks (Figure 4-6). Pupil responses were then plotted to examine what the pupil tracks looked like in terms of listener reactions during presentation of the talker speech stimuli for listeners. Figure 4-7 shows the pupil tracks of all listeners in this group averaged across all with PPD and baseline regions marked.

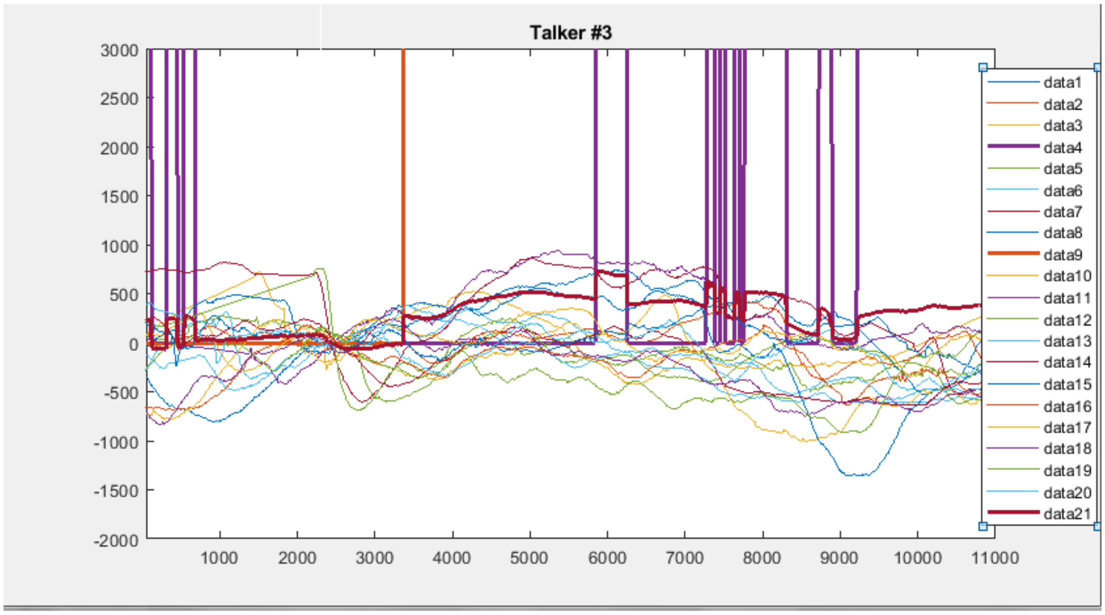


Figure 4-6 Sample pupil tracks with distorted listener tracks (bold; purple & orange) to be removed.

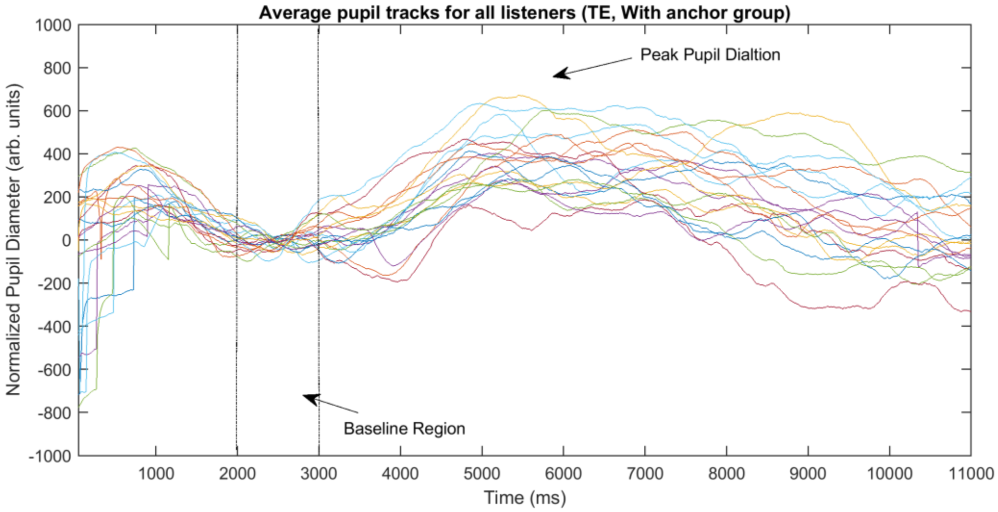


Figure 4-7 Pupil tracks averaged across all (TE, with-anchor group).

The extracted PPD values were examined to see if they correlated with the perceptual data. The correlation between PPD and roughness was 0.58 (Figure 4-8). The correlation between PPD and perceived listening effort was 0.64, as can be gauged from the scatter plot in Figure 4-9.

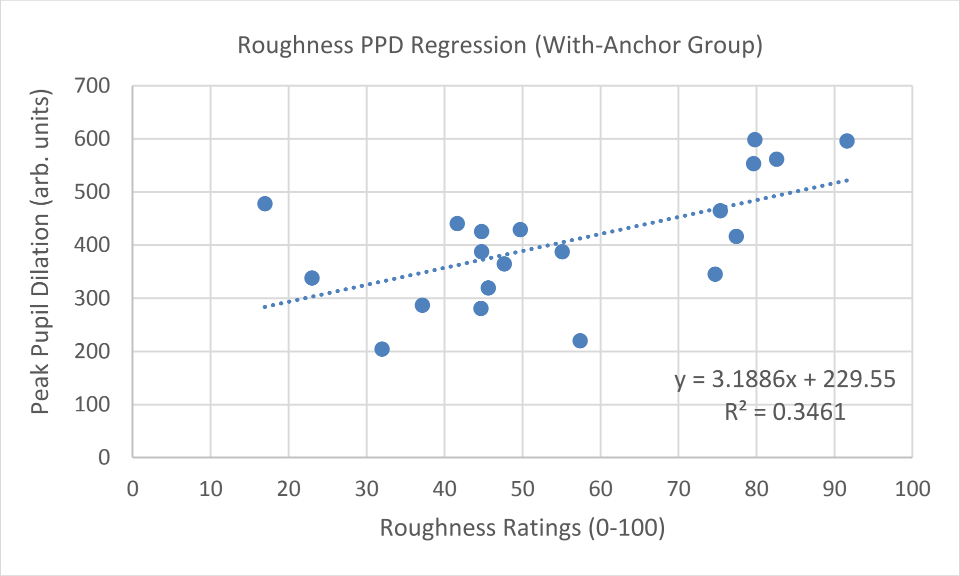


Figure 4-8 Roughness PPD Regression (TE, with-anchor group).

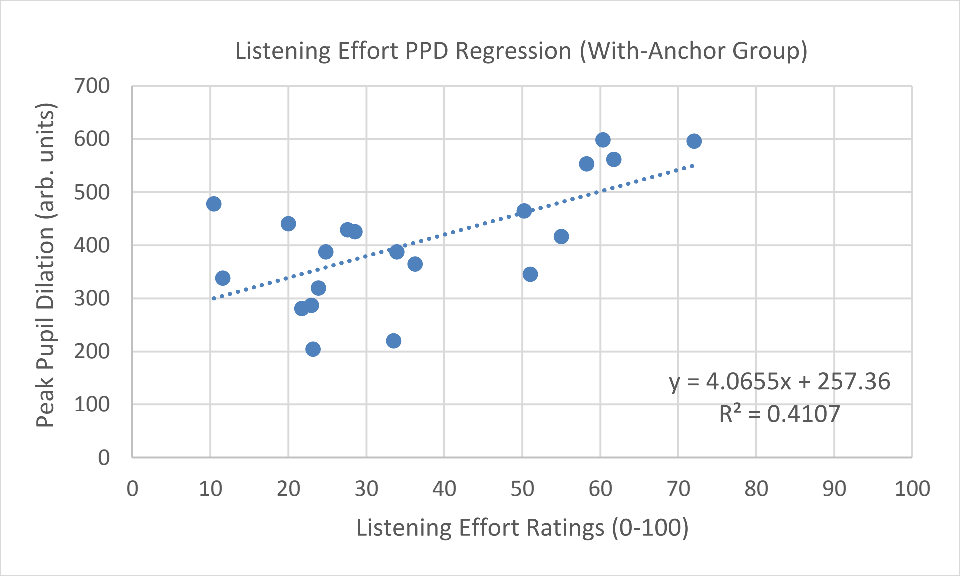


Figure 4-9 Listening Effort PPD Regression (TE, With-Anchor Group).

4.2.4 Pupillometry results (No-Anchor Listener Group)

Similar to the with-anchor group, pupil data were first normalized and then processed to be analyzed and visualized. Averaged all listener pupil tracks were plotted along with PPD and baseline regions (Figure 4-10).

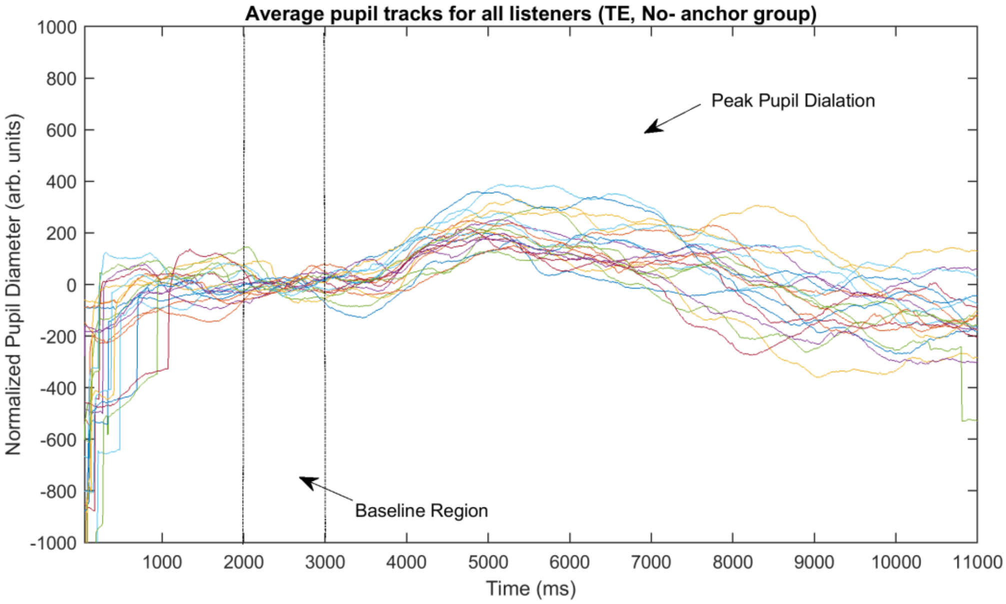


Figure 4-10 TE (No-anchor group) pupil tracks averaged across all listeners.

As the next step, the extracted PPD values were examined for correlation with the perceptual data. The correlation between PPD and roughness was 0.14 (Figure 4-11). The correlation between PPD and perceived listening effort was 0.22 which is plotted in Figure 4-12.

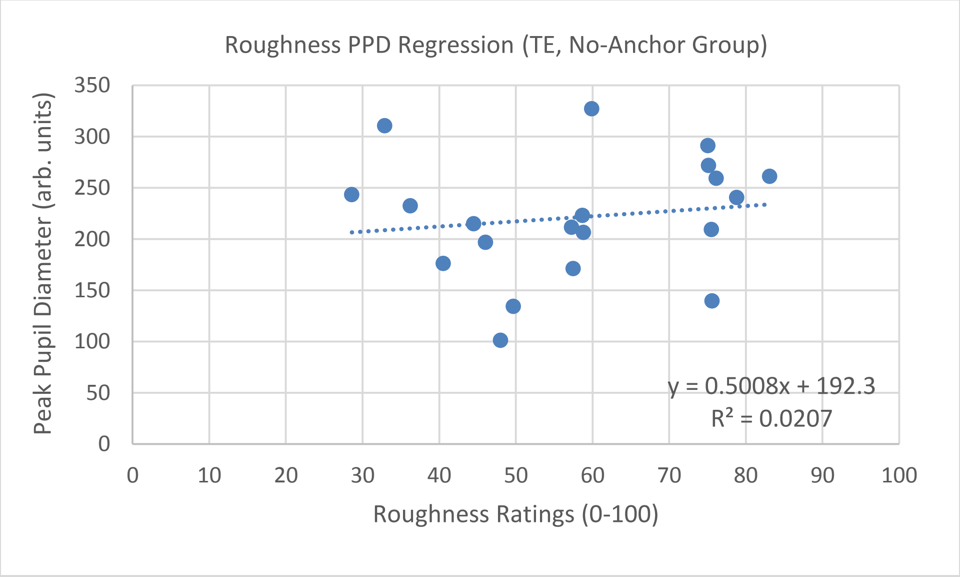


Figure 4-11 Roughness PPD Regression Plot (TE, No-Anchor Group).

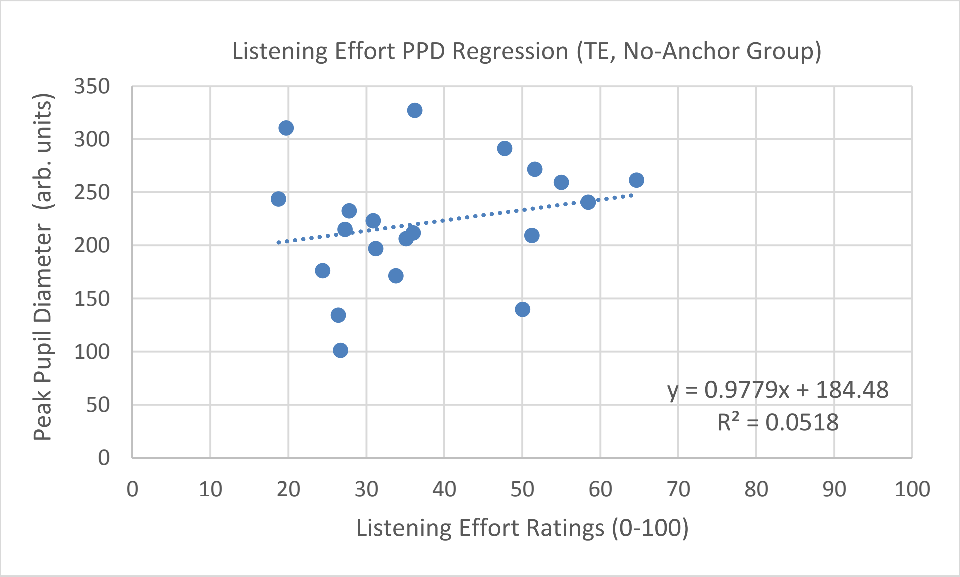


Figure 4-12 Listening Effort PPD Regression Plot (TE, No-Anchor Group).

4.3 Discussion

The objectives of this study were to perceptually assess vocal roughness and perceived listening effort, and also to evaluate potential pupillary reactions evoked by dysphonic voices of tracheoesophageal speakers in listeners. The study also involved auditory perceptual assessment of roughness with and without the help of audio anchors in two listener groups along with subjective/perceived listening effort. The inclusion of audio anchors in one of the two listener groups sought to assess how reliability was impacted by the inclusion and exclusion of audio anchors. The voice/speech samples selected for this study varied widely in roughness to account for potentially different reactions to the stimuli by listeners. Researchers in this study sought to evaluate if highly rough voices correspond to high levels of perceived listening effort by listeners and their pupillary responses to the stimuli.

The auditory-perceptual evaluation of roughness revealed that talkers demonstrated different degrees of roughness and were judged to require increased listening effort by both listener groups for stimuli with greater roughness ratings. Results showed both with-anchor and no-anchor group rated the stimuli reliably, although both the intra-rater ranges (0.44-0.84 vs. 0.37-0.72) and inter-rater reliability (0.96 vs. 0.93) for roughness were higher for the with-anchor group. This finding is in line with multiple past studies which report higher reliability with the use of anchors as the external voice standards, which are explicit and constant compared to unstable varied internal standards and control the context in which quality ratings are made and, therefore; increase the reliability (Kreiman, Gerrat, Precoda, & Berke, 1992; Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993; Brinca et al., 2015). Roughness may be a difficult dimension to rate for some listeners as the internal standard may not be clear or it may be influenced by other features such as breathiness (Brinca et al., 2015). In the with-anchor group, perceptual ratings of roughness and listening effort were highly correlated (0.97). Similar high correlation was obtained in the no-anchor group between roughness and perceived listening effort (0.95).

Repeated measures *ANOVA* also revealed the effect of the features (roughness and listening effort) on the perceptual ratings but neither a significant difference was found between the with-anchor and no-anchor groups in perceptual ratings nor an interaction between the features and group.

The other goal of the study was to evaluate the physiological reactions to rough voices through pupillometry. Pupillimotery, specifically PPD, has been used as a measure of cognitive load and listening effort due to language processing such as sentence processing (Hyona et al., 1995) and hearing and reading words (Brown et al., 1999). Moreover, it has been used to examine processing demand and cognitive load in studies on intelligibility and noise reduction schemes (Wendt, Hietkamp, & Lunner, 2017). Various studies on the listening effort indicate that listeners experience difficulties in daily life beyond sound audibility and speech intelligibility (Pichora-Fuller, et al., 2016; Johnsrude & Rodd, 2016) meaning that voice can be audible and intelligible but still demanding on listening effort.

In the with-anchor group, moderate roughness and PPD (0.58), and listening effort and PPD (0.64) correlations were observed indicating that listeners experienced cognitive load while processing rough voices. For instance, Talkers 14, 15, and 17 which were perceptually rated high (Figure 4-4) on roughness and effort, also evoked very high PPD values from listeners in this group. Such correlation pattern was not observed for all talkers with low perceptual roughness/effort ratings though. Only few of the talkers with low perceptual ratings had evoked lower PPD values (Talkers 11 & 19).

In the no-anchor group, the correlation between PPD and roughness was low (0.14) but in this group, PPD correlated with listening effort slightly better (0.22). Talkers with lower perceptual effort ratings had also lower PPD values (Talkers 11, 13, 20) and talkers with moderate perceptual effort ratings, had evoked moderate PPDs.

Upon closer examination of the results, it was noticed that the PPD values obtained from with-anchor group were collectively higher than the values from no-anchor group (Figure 4-13). Also the averaged PPD tracks of all talkers from both groups seemed slightly different especially with respect to pre-baseline activity (Figures 4-7 & 4-10).

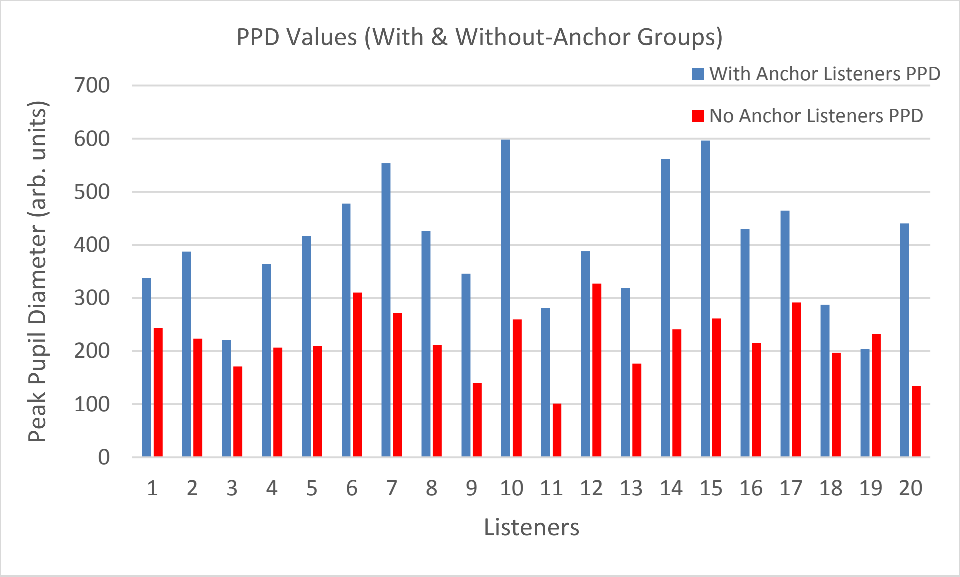


Figure 4-13 PPD values averaged across both groups for each TE talker.

Given the fact that listeners were all naïve, met all the inclusion criteria and were assigned to each group randomly, further investigation was carried out to explore the possible explanation of such differences. It should be noted that this project was the second one on the relation between pupillometry and voice disorders, and some of our listeners had already participated in the first project (Chapter 3) a few months prior to the present study. As a result, listeners in both groups (with-anchor & no-anchor) were examined in terms of their participation in the first project and the possible influence of exposure or lack of exposure to the experiment setting.

In the with-anchor, three out of ten had already participated in the previous project and therefore, seven of our listeners were first time participants. The PPDs of these two sub groups (within the anchor group) were extracted and a paired-samples t-test was conducted to compare the PPDs of first time participants and PPDs of repeated participant. There was a significant difference between the PPDs for the first time listeners (M = 463.97, SD = 132.57) and repeated listeners (M = 304.65, SD = 130.06) condition; t (19) = 4.82, p = .000 in the with- anchor group.

In the no-anchor group, five out of ten listeners were first time participants. The same procedure was done for this group as well. PPDs of first time and repeated listeners in the no-anchor group were extracted and compared via a paired-samples t-test. There was not a significant difference in PPDs of first time participants (M = 258.16, SD = 107.76) and repeated participants (M = 210.71, SD = 77.79) conditions; t (19) = 1.87, p = .077, however; the mean values were higher for the first time participants and the p value was close to significance. In addition, the correlation was calculated between the PPD and roughness (-0.11) and PPD and listening effort (-0.12) in the first time participants and between the PPD and roughness (0.21) and PPD and listening effort (0.28) in the repeated listeners.

The exposure to the experiment setting and disordered voices through participation in the first project seemed to make a difference in the with-anchor group but not in no-anchor group. Given the fact that half of participants in the no- anchor group were also first time and half repeated participants and since their mean analysis had not revealed the influence of previous exposure, the PPD responses were examined more closely. Comparing the average all pupil tracks of both groups (Figures 4-7 & 4-10) revealed that the with-anchor listeners were also showing collective higher pre-baseline pupillary activity. As the graph indicates, for this group the range of pupillary responses were between -750 and 450 with more concentration around the region of –200 and 400. For the no- anchor group, however; the range was between -1000 and 100, with the concentration in the region of -200 and 100. Thus, visual inspection of the graphs confirmed lower and smoother pupil activity for the no- anchor group even before the task begins. Given the fact that phasic pupillary responses are indicators of active maintenance of information in working memory, utilization of its capacity, and even allocating attention to the items in it during a delay period causes pupils to dilate (Kahneman, 1973; Just & Carpenter, 1993; Unsworth & Robinson, 2018), the behavior of our with-anchor group seems in line with the previous reports in the literature (Kursawe & Zimmer, 2015; Unsworth & Robinson, 2018). While doing the experiment and having their pupil behavior recorded, our with-anchor listeners were required to listen to the mild anchor (same sentence as the stimulus read by a mildly rough TE speaker) and the then severe anchor (same sentence as the stimulus read by a severely rough TE speaker) and then to the stimulus itself which had a prompt at the beginning and at the end and were then asked to rate the stimulus with regard to the two anchors provided. Each anchor was approximately 6 seconds and the stimulus was almost the same length plus a 3 second prompt at the beginning and a three second prompt and a one second beep at the end. As a result, each listener in the with-anchor group was in a delay period of maintaining both anchors and the stimulus in their working memory until they indicated their ratings through the two sliders. Their active maintenance of all the anchors and stimulus and attention allocation justified the increased PPD and pre-baseline activity.

4.4 Conclusions

This study addressed the auditory-perceptual evaluations of two dimensions related to voice disorder, namely vocal roughness and perceived listening effort of tracheoesophageal speakers. The obtained data highlights important findings and insights into how naïve listeners judge aspects of voice quality such as roughness and also their cognitive and processing load imposed by such auditory samples with and without the help of external standard. Listeners in both groups consistently assigned more subjective effort to highly rough voices. Similar to previous studies (Koelewjn et al., 2012; Nagle & Eadie, 2012; Whitehill & Wong, 2006), intelligible voices are still rated as demanding effort from the normal hearing individuals highlighting the fact that listening effort goes beyond simply understanding every single word and processing the linguistic content. Also, like previous studies (Gerratt et al., 1993; Barsties et al. 2017), replacing unstable varied internal standards with stable external anchors increases the reliability but imposes more cognitive load on the listeners. Due to the nature of the vocal dimension under study, in this case roughness, adding external anchors can help listeners have a better understanding of the feature they are rating. Pupil behavior seems to be influenced by the degree of roughness and inclusion of anchors. As listeners are required to maintain anchors and compare them with the upcoming stimulus, the cognitive load is impacted and increased. Delay period is also observed to increase the PPD.

Chapter 5

5 Auditory-perceptual and pupillometric evaluation of breathiness in talkers with vocal fold paralysis

Vocal fold paralysis (VFP) is an example of a voice disorder of neurogenic origin and it is categorized as a flaccid dysarthria (Darley, Aronson, & Brown, 1975), VFP is a result of damage to either the central or peripheral nervous system; however, peripheral losses are the most common. Damage leading to VFP can be to the vagus nerve (cranial nerve X), the brainstem, or the recurrent laryngeal nerve or superior laryngeal nerve or their branches. Depending on where the damage to the cranial nerve occurs, individuals may suffer from partial or unilateral VFP (Crumley, 1994).

Unilateral vocal fold paralysis may be due to trauma or of idiopathic (viral) nature. The recurrent laryngeal nerve is the most commonly observed type of laryngeal paralysis (Case, 2002). Individuals can experience flaccidity of the vocal fold which results in an incapacity to adduct/abduct the vocal folds, resulting in various degrees of dysphonia, aphonia, and-or excessive aspiration while phonating (Crumley, 1994). The acute effects of VFP are immediate flaccidity of the vocal fold with a direct impact on voice production. Individuals with VFP are usually described as perceptually sounding breathy, exhibit reductions in vocal intensity, and are not capable of sustaining phonation for long durations (Ferrer, Haderlein, Maryn, De Bodt, & Nöth, 2018).

The purpose of this study was to perceptually evaluate speech samples from talkers with unilateral vocal fold paralysis on the voice dimension of breathiness by two listener groups of normal hearing listeners. Also, the reliability ratings were examined for the impact of inclusion and exclusion of audio anchors to the listener groups. Furthermore, listeners' physiological/cognitive responses to exposure to such disordered voices were assessed through pupillometry.

5.1 Methodology

Twenty adults (5 males, 15 females; age range = 19- 33 years; mean = 24.4 years) served as participant listeners in the current project. Participants were divided into two groups: with–anchor (2 males, 8 females; age range = 19-33, mean age = 24) and the no- anchor group (3 males, 7 females; age range = 20-31, mean age= 24.8). Participants were randomly assigned to each group based on the order that they participated in the study with every other participant assigned to either group. Determination of the total number of participants required for this experiment was obtained through a power analysis calculated by GPower, with an effect size of 0.4. All participants self-reported as being normal hearing and were also native English speakers. In addition, they did not report any formal training and/or education as in the area of voice or voice disorders. If a participant reported an upper respiratory infection within the week prior to the experiment, they were asked to postpone their participation for at least 14 days or until the problem had resolved. As part of the exclusion criteria, a list of medications was provided to all potential participants (Appendix A) if a potential participant indicated that they were taking one or more of the medications listed, they were excluded from participation in the study (Appendix H, REB # 112674).

5.1.1 Auditory Stimuli

Digitally recorded speech samples of 20 talkers with VFP from an archive of the Voice Production & Perception Laboratory at the University of Western Ontario served as the stimuli of the project. Identical to all other studies, all voice samples were gathered via digital recording obtained in either a quiet clinical setting free from ambient noise or within a sound-treated environment. All voice samples included sustained vowels, a standard reading (The Rainbow Passage; Fairbanks, 1960), a short monologue (approximately 60 seconds), and a variety of single word stimuli along. In some instances, standard sentences from the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) were also gathered. Recording of speech stimuli occurred after informed consent was obtained, along with demographic information for each speaker. A directional microphone (Shure PG-81) attached to a desktop microphone stand was used in all recording. During the collection of each sample, a microphone-to-mouth distance of 15 centimeters was maintained and was checked prior to each task. Each participant speaker's sample of voice/speech stimuli was recorded onto a laptop computer at a sampling rate of 44.1 kH using the Multidimensional Voice Profile (SonaSpeech II, Kay Pentax, Lincoln Park, NJ). Volume input levels were adjusted for each speaker at the beginning of each session and were monitored during the recordings using SonaSpeech II to avoid any under- or over-driving of the input signal. The second sentence of the Rainbow Passage ("The rainbow is a division of white light into many beautiful colors.") was then extracted from the recording and used as the study stimuli. The stimuli were also equalized for their Root Mean Square (RMS), so that they were presented at approximately same intensity during playback. All speech stimuli were presented to the listeners in a randomized order during the experiment.

All stimuli presentations began with the oral cue "Please listen to the following stimulus" which prompted the listener for the initiation of the upcoming stimulus. This cue lasted 3 seconds and was read by a normal English speaking adult male. After each stimulus was played, listeners were instructed through another verbal cue to indicate their scaled ratings of breathiness and listening effort using two computerized visual analog slider scales. For the no-anchor group, once the experiment began the oral cues and the stimuli were played randomly and automatically. Once participants completed their ratings and clicked "Next", the rest of the stimuli were played until the experiment was completed. For the with-anchor group, the experimental protocol was slightly different. That is, stimuli were not automatically played; listeners were required to click on speaker icons for breathiness and the anchors to listen to them. After indicating their rating, they clicked "Next" to move forward. These anchors (mild and severe) were selected by an experienced voice scientist as exemplar anchors from the VFP archival samples and represented one talker with a mild and one with a severely breathy voice. The mild anchor icon was placed at the left endpoint of the scale and the severe anchor icon was placed at the right endpoint of the scale.

5.1.2 Assessment of breathiness and perceived listening effort

Auditory-perceptual judgments of breathiness and ratings of listening effort were collected using two separate 100 mm long electronic sliders. These sliders represented a

VAS (Figure 5-1) with 100 intervals (i.e., 1 through 100) and listeners could move the slider bar between the end points to mark their ratings of the dimensions under study. For the bottom slider which was used for collecting ratings of listening effort, the end point was labelled as “none” and the right as “extreme”. The experiment protocol was quite similar for both groups, except for the inclusion of audio anchors for one group. They were instructed to rate with reference to those anchors. Participants were instructed on the order of playing and listening to the anchors and stimuli to prevent any confusions.

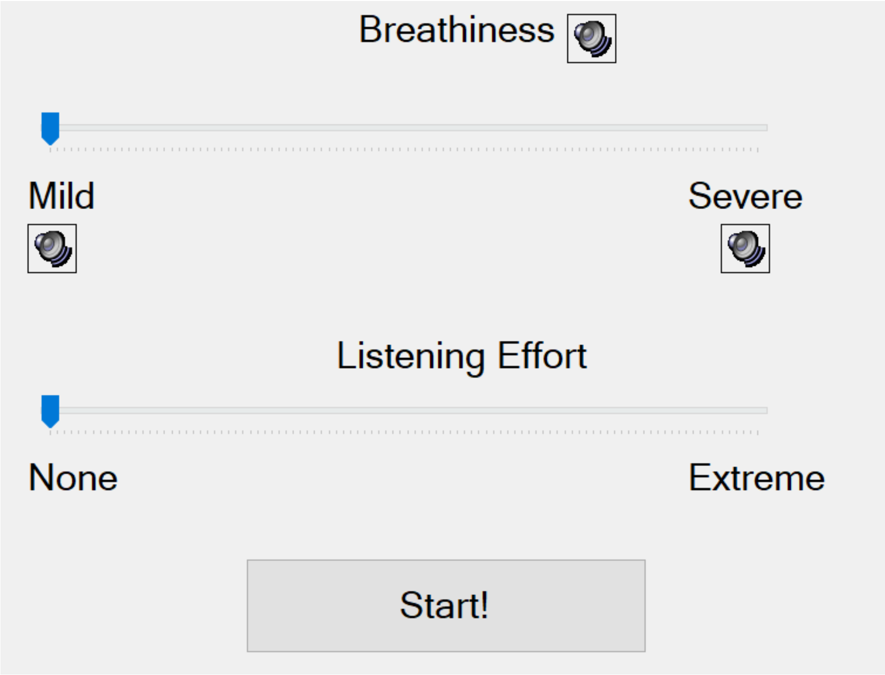


Figure 5-1 Appearance of slider rating scales for "Breathiness" and "listening effort".

While participants in either group were listening to and rating the stimuli, their pupil responses were recorded using the EyeLink 1000 (SR Research, Ottawa, Canada) eye tracker (Figures 3-2 and 3-3). Before beginning the experiments, listeners were asked to sit on a stationary chair at the tower mount with their chin positioned on a chin rest and their forehead placed against a head rest. The EyeLink 1000 collected pupil activity at a sampling rate of 1000 Hz. Participants could visualize the sliders on the monitor ahead of them while simultaneously listening to the speech stimuli through headphones (Sennheiser, HD 250).

5.1.3 Procedure

At the beginning of the experimental session, participants were familiarized with the nature of the tasks and the objectives of the study. They were provided with the oral and written definitions and a brief explanation of the dimensions under evaluation: breathiness and listening effort. Breathiness was defined as the "audible air escape in the voice" (Kempster, Gerratt, Abbott, Barkmeier-Kraemer, & Hillman, 2009). Listening effort was defined as "the amount of work required while auditing the speaker samples".

For the with-anchor group, they were instructed on listening and rating the stimuli relative to the anchors. Listeners were also instructed on the function of the EyeLink1000 and were positioning at the device to ensure that the best direct view of the right eye was obtained. All listeners were instructed to keep their head and body stable during the experiment and to avoid looking away from the monitor. Finally, they were asked to continue gazing at the center of the monitor and avoid blinking as much as possible while listening to the stimulus.

Once the optimum position was obtained at the tower mount and listeners had headphones on, their visual gaze was calibrated and then validated. This process involved maintaining visual focus on a fixation circle on the screen. Once satisfactory calibration and validation were achieved, the experimenter initiated the listening procedure and participants rated the randomly presented stimuli after hearing the verbal cue and the beep. Listeners were prompted via a message on the screen once all the stimuli were rated at which point they were given a 10 minute break before the retest phase began. After the listening experiment was completed, the pupil tracks were normalized and all had a start point at 0 second. The tracks had to be normalized because the stimuli were presented randomly to each listener during the test and retest phases. In addition, before the actual analysis of pupil tracks, they had to be processed and quick blinks (<125 msec) or dropouts had to be interpolated or removed. Interpolation of quick blinks began approximately 50 msec before the blink and at least 150 msec after a blink. Long blinks or dropouts were removed as interpolation would change the overall pattern of tracking sequence. In this project, about 4.75 % (n = 19) of the tracks were removed in the with-anchor and 12.25% (n = 49) were removed from the no-anchor group to eliminate the risk of data distortion.

One of the foci of this experiment was on the peak pupil dilation (PPD) as a dependent measure. The peak in pupil dilation was determined per listener as the maximum dilation during the presentation of the speech sample relative to the mean dilation in the baseline period of every stimulus. The baseline period was the last second of the three second prompt (Figure 5-6) against which the stimulus was compared to determine the PPD or the highest point in the track.

5.2 Results

5.2.1 Auditory-Perceptual Evaluations With-anchor group

Each listener's test-retest intra-rater reliability for both breathiness and listening effort were calculated by correlating the rating values. The intra-rater correlation for breathiness ranged from 0.47 to 0.90, indicating a moderate-to-high relationship between the measures. The correlation for listening effort ranged from -0.7 to +0.88. For listening effort, 6 out of 10 listeners in this group had intra-rater correlations above 0.49 which indicates a moderate-to-high reliability for this listener subset.

Interrater reliability was calculated through Cronbach's alpha in SPSS (Version 24, Armonk, NY) for each of the two features rated perceptually. The interrater reliability was determined to be 0.94 for breathiness and 0.88 for listening effort which confirms a very high reliability among listeners for these two tasks.

The auditory-perceptual ratings for both features ranged between 0 and 100 (two sets per listener). After averaging each talker's value across all listeners (Appendix F), all talkers (breathiness and listening effort) were plotted along with the standard error of the mean (Figure 5-2). Talker 12 was rated the highest on breathiness followed by talker 19. Talker

19 was rated the highest on listening effort. Talker 18 is rated as the lowest on both breathiness and listening effort. The data were also plotted to represent the correlation between the two measures (Figure 5-3). Results indicate a very high correlation (0.93) between the ratings of the two features.

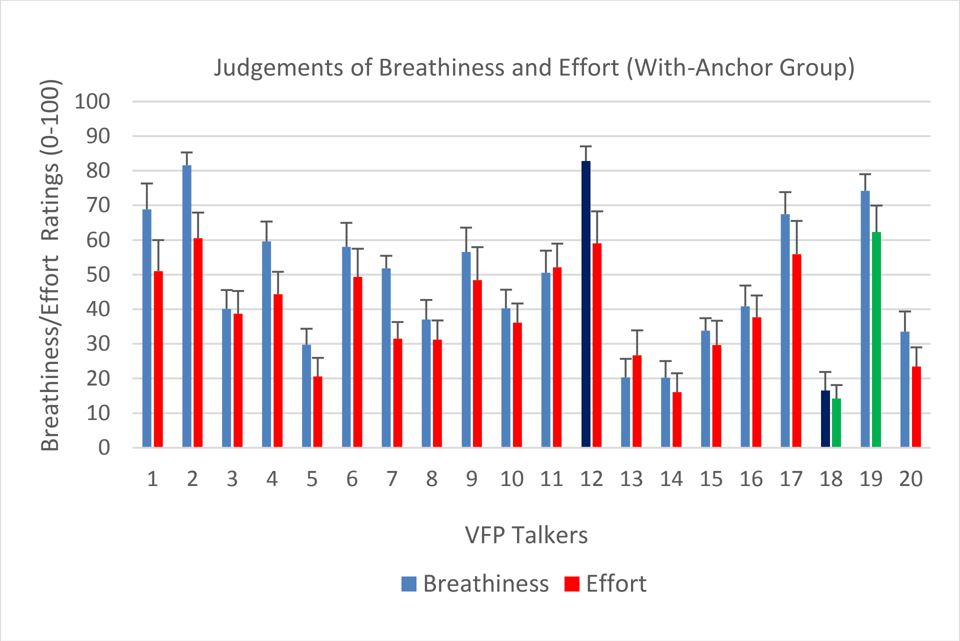


Figure 5-2 Breathiness and effort ratings per VFP talkers; highest and lowest rated talkers color distinguished (Breathiness: black; effort: green).

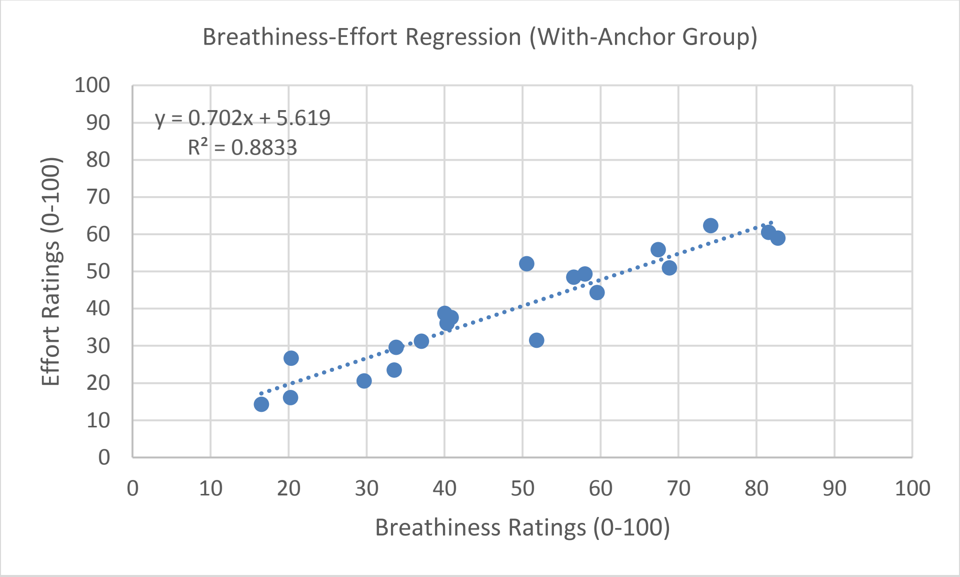


Figure 5-3 Regression between breathiness and effort ratings by the with-Anchor group.

5.2.2 Auditory-Perceptual Evaluations (No-Anchor Listener Group)

The intra-rater reliability of the no anchor group also was calculated by correlating the test-retest ratings of breathiness and effort. The intra-rater value for breathiness ranged from 0.55 to 0.92 and for effort from 0.25 to 0.89.

Interrater reliability was calculated through Cronbach’s alpha in SPSS (Version 24, Armonk, NY) for each of the two auditory-perceptually features. The interrater value was 0.96 for breathiness and 0.94 for listening effort which confirms a very high degree of reliability among listeners for the two tasks.

The auditory-perceptual averaged ratings for breathiness and effort (Appendix G), were also plotted to display how each talker was rated for each feature (Figure 5-4). Similar to the with-anchor group, Talker 12 was rated as the highest on breathiness and Talker 19 as the highest on effort. The listening effort value was also high for Talker 12 and the breathiness rating was high for Talker 19. Talker 18 is the lowest rated for both breathiness and listening effort.

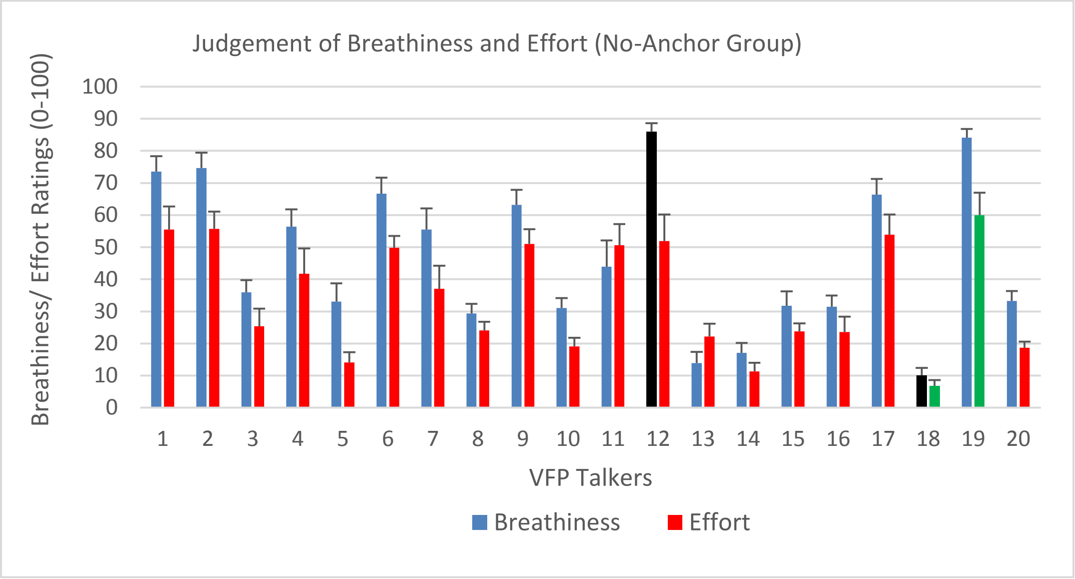


Figure 5-4 Breathiness and effort ratings per VFP talkers; highest and lowest rated talkers color distinguished (Breathiness: black; effort: green).

Similar to the other group of listeners, the breathiness and effort ratings were plotted to display the correlation between the two measures; (Figure 5-5) this analysis indicated a very high correlation between the listener ratings for these two dimensions

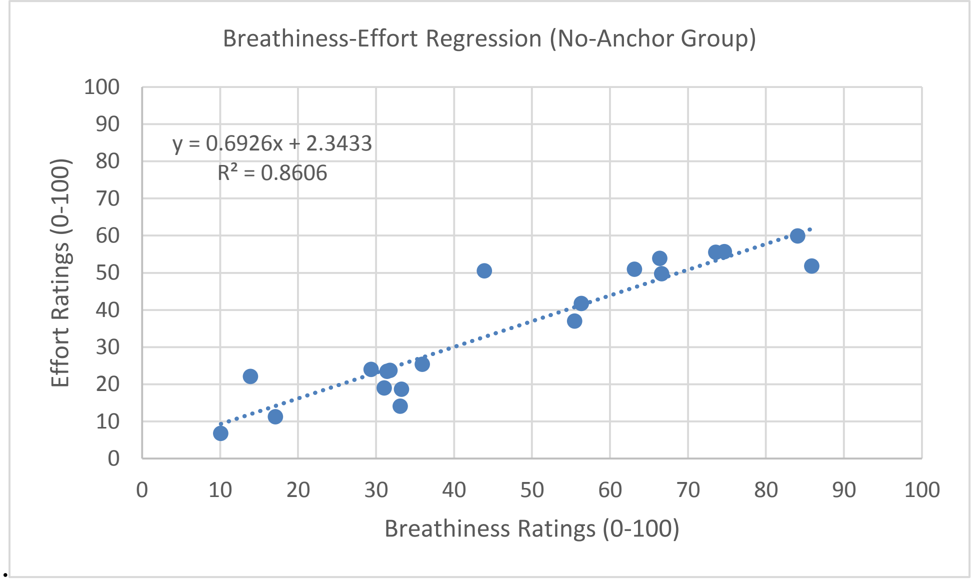


Figure 5-5 Regression between breathiness and listening effort.

5.2.3 Statistical analysis

A repeated measures *ANOVA* was conducted to compare the main effects of features (breathiness and listening effort) and the groups (with-anchor no-anchor) and the interactions between them. The a priori significance level was set at 0.05 for all statistical test. Significant effect was found for features (breathiness and listening effort) (*F* (1,18) = 21.10, p< 0.005, η_p^2= 0.540). Also, no group effects (with-anchor and no- anchor) were found (*F* (1,18) = 0.540, p> 0.05, η_p^2= 0.029). The results also revealed an interaction between the features and talkers (*F* (19,342) = 7.97, p< 0.05, η_p^2= 0.307).

5.2.4 Pupillometry results (With-anchor group)

After normalizing the preparing the pupil tracks for analyses by removing the blinks and dropout (Figure 4-6), tracks were plotted for visualization. Figure 5-6 displays the pupil activity of listeners in this group (averaged across all) while listening to each of the stimuli with the baseline region and PPDs identified.

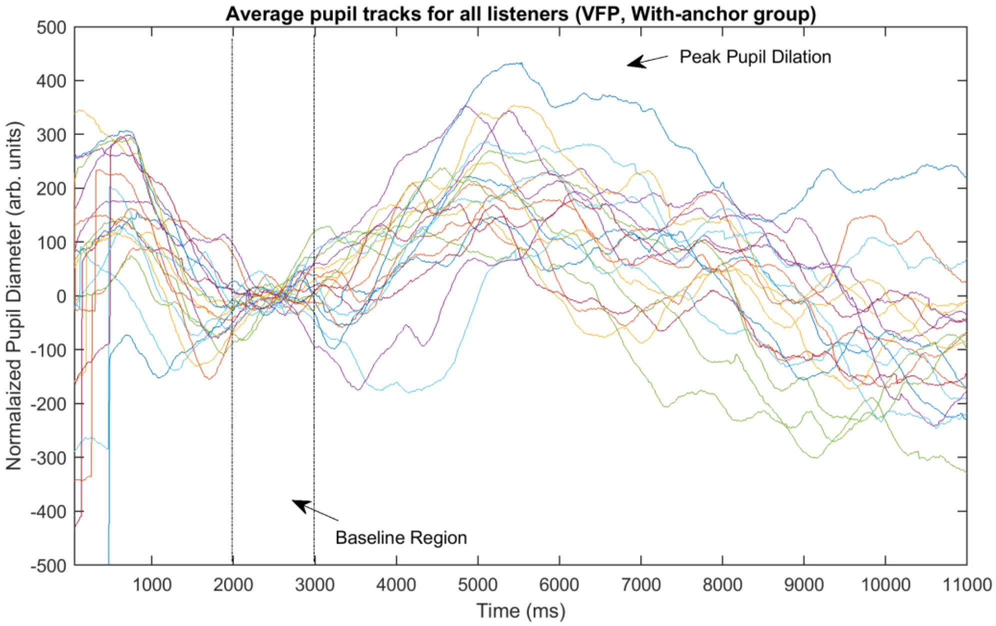


Figure 5-6 Pupil tracks averaged across all listeners (With-anchor group).

The extracted PPD values were examined to determine their correlation with the auditory-perceptual data. The correlation between PPD and breathiness ratings was 0.24 (Figure 5-7) and between PPD and listening effort (Figure 5-8) was 0.25. The obtained correlation results were plotted in the following figures.

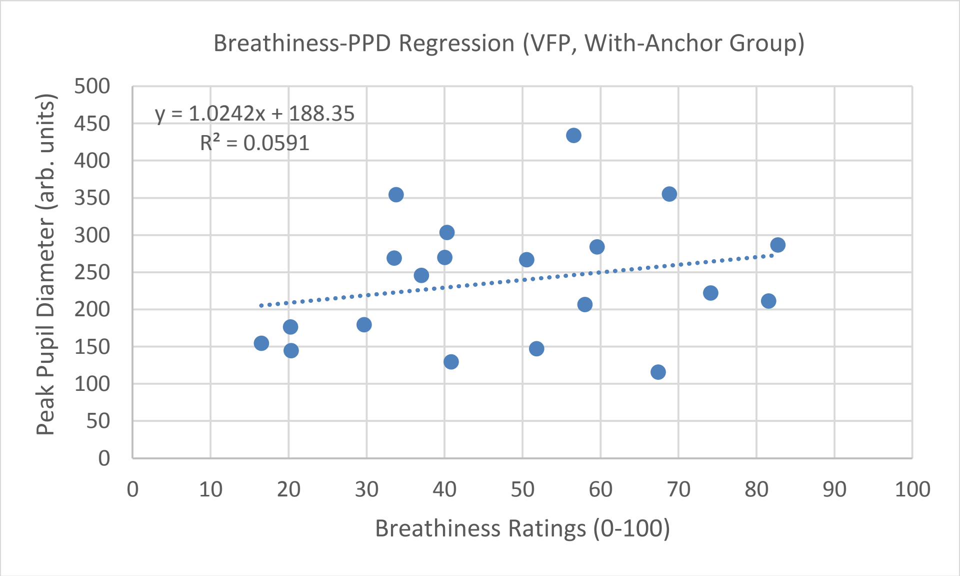


Figure 5-7 Breathiness-PPD Regression (VFP, With-Anchor group).

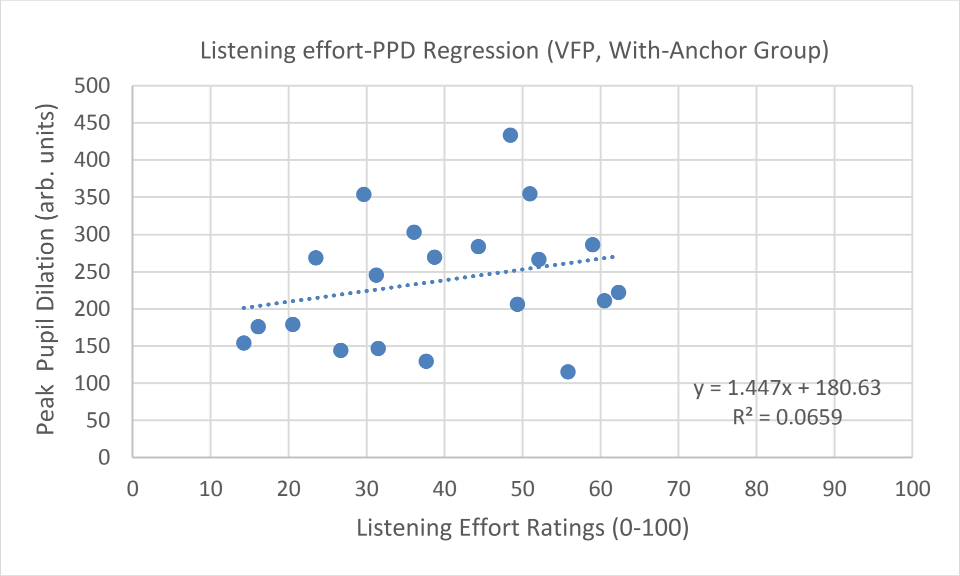


Figure 5-8 Listening effort-PPD Regression (VFP, With-Anchor group).

5.2.5 Pupillometry results (No-Anchor group)

The same procedure was followed to analyze and visualize the pupil data from the no-anchor listener group. The average across all listeners' pupil tracks are presented in Figure 5-9 with the baseline region and the PPD identified. Once the tracks were examined and cleaned from blinks and dropouts, PPDs were extracted and correlated with auditory-perceptual ratings of breathiness (Figure 5-10) and listening effort (Figure 5-11). The correlation between ratings of breathiness and PPD was 0.25. The listening effort and PPD was 0.13 for this group.

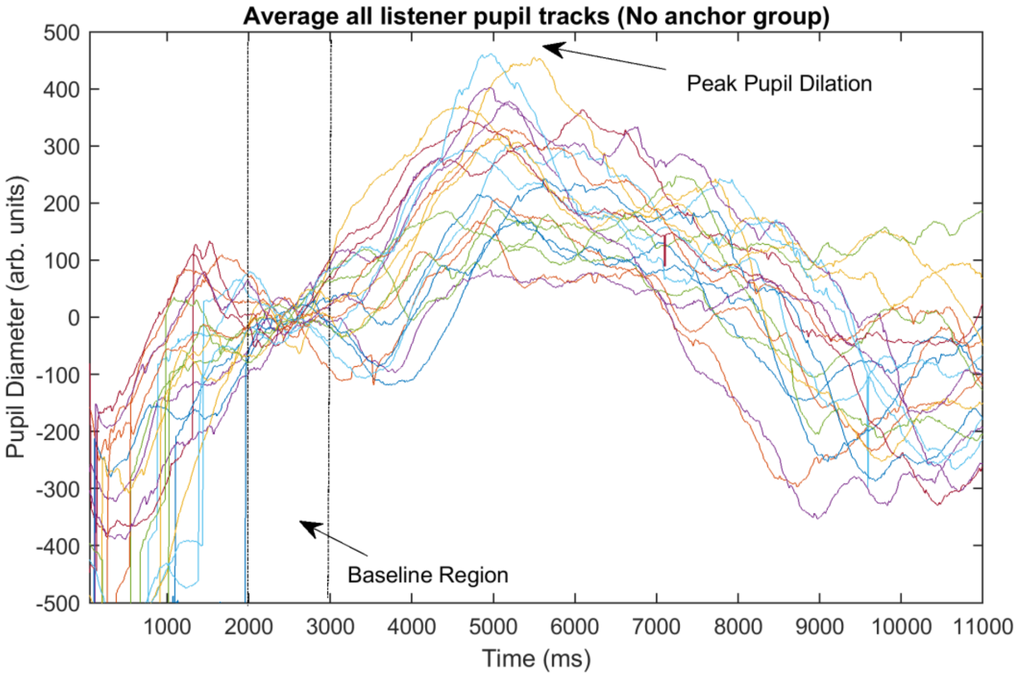


Figure 5-9 Pupil tracks averaged across all listeners (No-Anchor group).

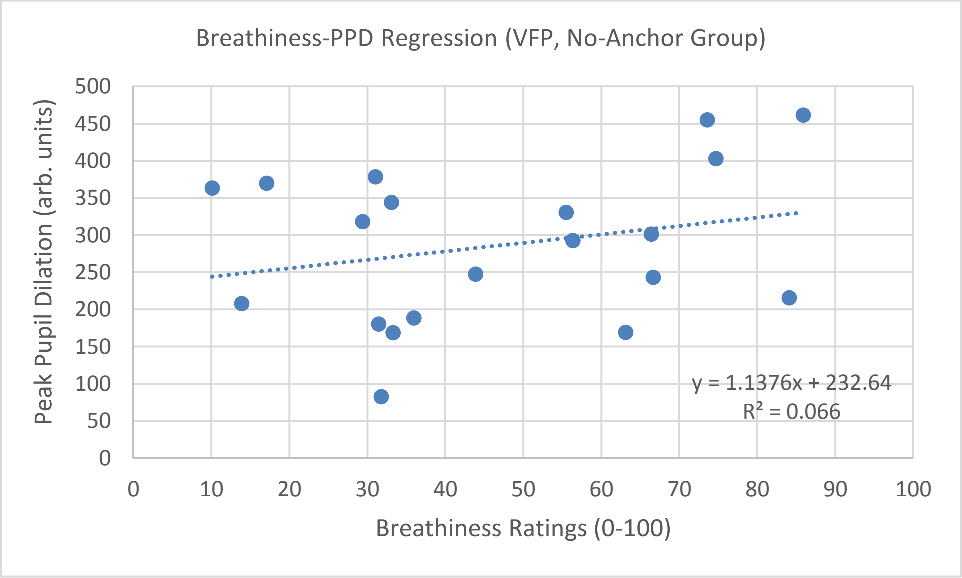


Figure 5-10 Breathiness-PPD Regression (VFP, No-Anchor group).

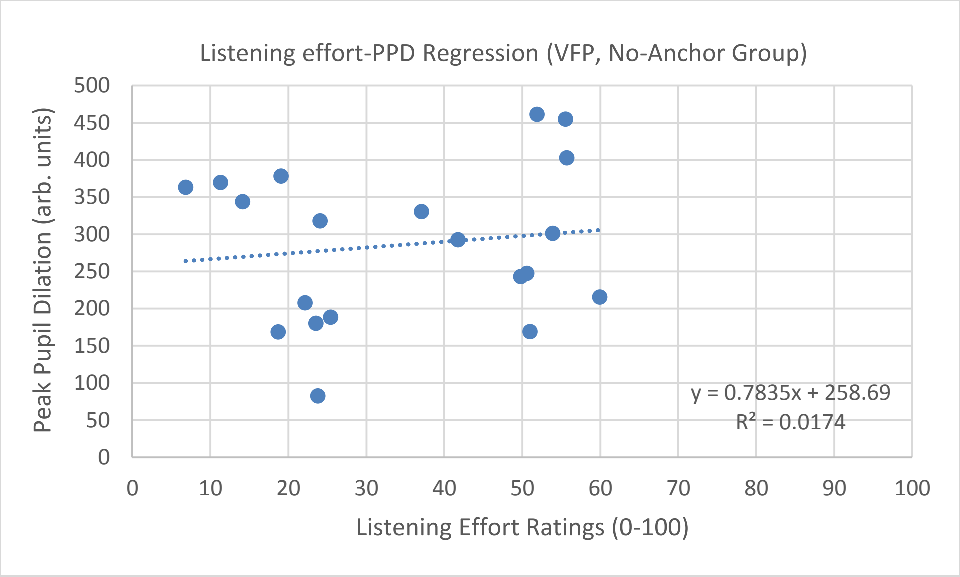


Figure 5-11 Listening effort-PPD Regression (VFP, With-Anchor group).

5.3 Discussion

5.3.1 Auditory-perceptual ratings

The objective of this study was to examine auditory-perceptual evaluations of breathiness and listening effort in relationship to pupillary responses evoked by experimental speech samples of talkers with VFP. In addition, the perceptual evaluation examined the potential influence of the use of audio anchors during the scaling procedure.

Similar to the study on TE (Chapter 4), the inclusion of these external anchors in one of the two listener groups sought to assess how reliability was affected by their exclusion or inclusion on ratings generated. In order to access a wide range of potential physiological responses, the speech samples selected varied widely in the degree of breathiness. In fact, we sought to examine if talkers who were rated as having highly breathy voices were also judged as highly demanding specific to listening effort, as well as whether their pupillary responses corresponded to such auditory-perceptual ratings.

The perceptual evaluations indicate that talkers demonstrated various degrees of breathiness. Both groups rated the samples reliably. Also, the correlation between breathiness and effort was very strong in both groups (i.e., With-anchor group: 0.93; No-anchor group: 0.92).

In the with-anchor group, the ranges of perceptual ratings were slightly higher (perceived as being more breathy) for minimum breathiness and the listening effort range. Breathiness ratings (averaged across all listeners) ranged between 16.5 to 82.75 and listening effort 14.25 to 62.3. In the no-anchor group, breathiness ranged from 10.1 to 85.9 and listening effort ranged from 6.82 to 59.95. With the exception of Talkers 11 and 13, both listener groups rated all other talkers, higher on breathiness than effort. These two talkers (11 and 13) were rated slightly higher on listening effort than breathiness, but the difference was small and they were closely aligned. Talker 11 was a male talker who was very breathy and had a slower speech rate. Although we did not investigate the influence of gender on either listeners or talkers, there seems to be a wide range of variation in the perception of breathiness within and between genders (Hillenbrand, Cleveland & Ercikson, 1994). It should be noted that Talker 11 was rated more on effort than breathiness. Further examination of this talker revealed the presence of vocal roughness in their voice. In addition, the rating could be due to his gender as males are being penalized for the presence of this feature compared to women who may be less penalized for having breathy voices (Hillenbrand et al., 1994). These results were also indicated in the interaction analysis results of the repeated measures *ANOVA*. As indicated in the results, an interaction was revealed between features and talkers, meaning that not all talkers were rated higher on breathiness.

Results showed that both with-anchor and no-anchor listener groups rated the stimuli reliably with slightly higher intra-rater correlation values for the no-anchor group (breathiness intra-rater correlation ranges of 0.47- 0.9 vs. 0.55- 0.92 and listening effort ranges of -0.71- 0.88 vs. 0.25- 0.89, respectively). The same pattern was observed in the inter-rater reliability with the values for breathiness (0.94 vs. 0.96) and listening effort (0.88 vs. 0.94) being higher the no-anchor listener group. Although the reliability values are quite strong for both groups, the use of audio anchors did not appear to improve the values over the reliability of the no- anchor group. This finding is different from the higher reliability for with-anchor group presented in Chapter 4, as well as when compared to previous studies in the literature (Brinca et al., 2015; Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993; Kreiman, Gerrat, Precoda, & Berke, 1992). Generally, improved reliability is reported with the use of anchors which are explicit and constant compared to unstable varied internal standards and control the context in which quality ratings are made (Kreiman, Gerrat, Precoda, & Berke, 1992; Gerratt, Kreiman, Antonanzas-Barroso, & Berke, 1993; Brinca et al., 2015).

The repeated measures *ANOVA* also revealed the effect of features (breathiness and listening effort) on the auditory-perceptual evaluation ratings, but no effect was found for the group (with-anchor and no-anchor) meaning that both groups rated the features similarly.

5.3.2 Pupillometry

The other objective of the study evaluated whether any physiological reactions to breathy voices existed. One measure achieved through pupillometry is the peak pupil dilation in response to the cognitive demands of a task (Beatty, 1982). Examining the pupillary data revealed similar correlations between PPD and breathiness (with-anchor group = 0.24 vs. 0.25 for no-anchor group), but a better correlation between PPD and listening effort for the with- anchor group (0.25 vs. 0.13).

The with-anchor group had lower PPD values (Figure 5-12) for most of the stimuli (16 out of 20). This could be due to the fact that 8 out of 10 participants in this group had already taken part in the previous studies. In fact, 5 out of theses 8 listeners had participated in both the ADSD (Chapter 3) and TE (Chapter 4) experiments and 3 of those 8 had done the TE study. The time interval between the first two studies for these repeated participants was approximately 3 months, whereas that between the second and the third project ranged between 3 to 6 weeks. However, in the no-anchor group, 7 out of 10 listeners were first time participants. Given that some people may experience varied workloads for the same task at various stages, and that workload can reduce through learning or training (Xie & Salvendi, 2000; Vidulich & Pandit, 1986), the lower PPD values in this group may be attributed to their previous exposures in this project.

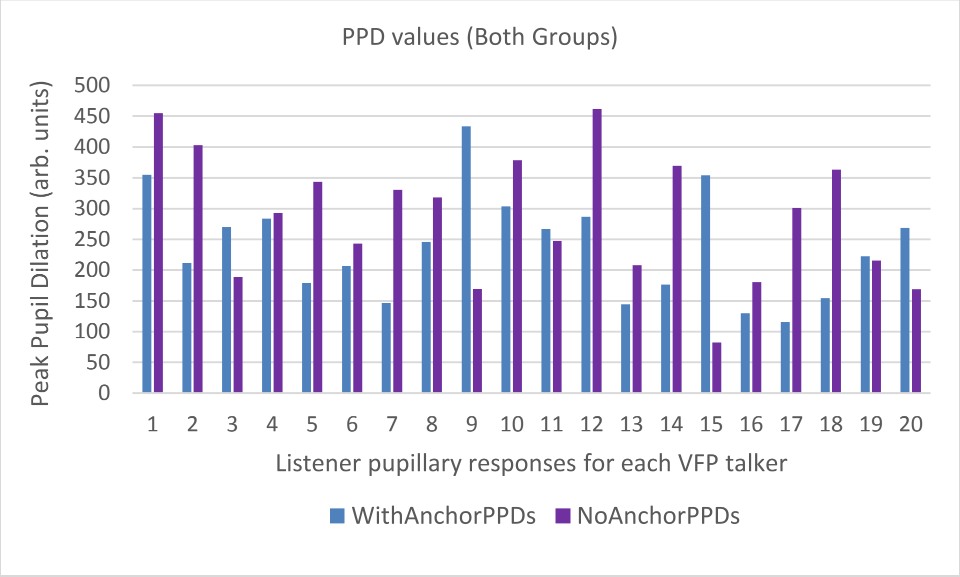


Figure 5-12 PPD values of both groups in response to the stimuli.

Talker 12 who was perceptually rated the highest on breathiness by both groups, also evoked the highest PPD in the no-anchor group, but only had a moderate PPD value in the with-anchor group. Talker 1 had who had been perceptually rated high on breathiness and listening effort, had evoked high PPD values in both groups; however, this pattern was not observed for all perceptually (breathiness and effort) high rated talkers. The lowest perceptually rated talker (Talker 18), evoked different PPD responses (low in With-anchor group, pretty high in no-anchor group). Therefore, not all the highly breathy/listening effort demanding samples evoked high PPDs.

In terms of the pre-baseline pupillary activity, examining tracks of both groups revealed a similar pattern as that reported for the TE study (Chapter 4). The with-anchor group showed relatively high pre-baseline pupil activity. As it can be seen in Figure 5-6, the range of the pre-baseline activity for the with-anchor group is between -400 and 350 msec with the concentration in the region between 0 and 300 msec. In the no-anchor group (5-9), this range is between -400 and 100 msec indicating a lower and smoother pupillary activity in this group. The phasic pupil reactions indicate the active maintenance of information in the working memory, in this case anchors. Listeners were also asked to compare the forthcoming stimuli to the anchors. In fact, listeners were forced to both keep the anchors and also allocate attention to them while they are waiting for the stimuli to be played and therefore more pupil activity is observed during this delay period (Kahneman, 1973; Just & Carpenter, 1993; Kursawe & Zimmer, 2015; Unsworth & Robinson, 2018).

The length of time listeners were supposed to hold anchors in their working memory and the delay period to wait for the stimulus is imposing cognitive load. Each anchor was approximately 6 seconds long. The stimulus was almost the same length plus a 3 second prompt at the beginning and a 3 second prompt and a one second beep at the end. As a result, each listener in the with-anchor group was in a delay period of maintaining both anchors and the stimulus in their working memory until they indicated their ratings through the two sliders. Their active maintenance of all the anchors and stimulus and attention allocation justified the increased pre-baseline activity. These generally low average PPDs may be due to the fact that listeners in this group (8 of 10) had already been habituated to listening to the dyphonic voices, potentially due in part to their participation in previous experiments. In this case, it is possible that a cognitive schema had already been formed and as a result, the cognitive load is reduced due to previous knowledge and expertise (Antonenco et al., 2010).

5.4 Conclusions

This study addressed auditory-perceptual evaluations of breathiness and listening effort by listeners in two groups, those exposed to anchors and those who were not. Listeners generally assigned greater listening effort (demand placed on the listener) to voice samples that were rated as exhibiting more breathiness. Listening effort includes multiple perceptual factors and as a result, a dysphonic voice (Talkers 11 and 13) may be rated lower on breathiness but higher on listening effort because of the composite quality of the voice.

Based on these data, the processing demand and cognitive load are generally reflected in the PPD values. This may be more pronounced if the listener is maintaining items in the working memory where attention allocation and a delay period are involved for the listening task. This load seems to be decreasing if exposure, learning, and training are provided as cognitive schemata are being formed. The cognitive load in response to listening to the VFP stimuli were initiated from the onset of the stimuli in both groups (Figures 5-6 & 5-9) due to the perception of breathiness from the onset of the speech stimuli.

Chapter 6

6 General Conclusions

The three studies reported in this thesis addressed auditory-perceptual evaluation of four specific features related to abnormal voice quality. This series of experiments were addressed with the objective of identifying how various voice dimensions (e.g., strain, roughness, breathiness), listening effort, and pupil responses varied between naïve and experienced listeners. The other research question addressed was: how would the inclusion/exclusion of audio anchors potentially impact the auditory-perceptual ratings in association with pupil responses of our participant listeners? Lastly, the relationship between subjective measures of voice quality (auditory-perceptual) and objective measure (pupillary), as well as perceived listening effort was examined. In order to answer the questions posed, three research projects were designed, each focusing on a particular voice disorder and a specific voice feature: AdSD- vocal strain, TE- vocal roughness and VFP- breathiness. In all three studies listening effort was consistently measured and pupillary responses were also evaluated in addition to the previously noted vocal dimensions. A summary of the finding from each of these three experiments will be outlined in the subsequent sections.

6.1 Experiment 1 (Adductor Spasmodic Dysphonia)

This study involved listener ratings of perceived degrees of vocal strain and listening effort by normal-hearing listeners; both naïve and experienced listeners were evaluated in this study. High correlations were found between strain and listening effort in both listener groups. These data suggest that higher auditory-perceptual ratings on the perceived strain (more strained) generally indicates higher values on perceived listening effort. This finding indicates samples which were rated as really strained, were also evaluated as demanding a lot of listening effort. High correlations were found between peak pupil dilation values (PPDs) and perceptual ratings in the naïve listener group indicating that listeners expended cognitive resources while auditing speech samples. However, no such results were found in the experienced group. Previous exposure seems to lead to listeners' habituation to dysphonic voices. Listener groups PPDs revealed other patterns which can be attributed to the strategies used by them when listening to disordered voices.

6.2 Experiment 2 (Tracheoesophageal Speech)

In the second experiment, vocal roughness was evaluated for voice samples obtained from TE talkers. Also, listening effort was assessed by listeners with and without the use of audio-anchors along with pupillary responses. In the with-anchor group, high correlations were observed between roughness and perceived listening effort ratings. When assessing the reliability of listener ratings, higher reliability was achieved for the with-anchor group which confirms the potentially positive influence of anchors on improving reliability. Further, moderately high correlations were observed between PPDs and auditory-perceptual ratings in the with-anchor listener group; this finding indicates a potentially increased cognitive load, as well as reported listening effort experienced by those in the with-anchor group while listening to the TE voice samples. The TE study also revealed additional information regarding the use of anchors. Since listeners had to maintain anchors in their working memory prior to the onset of every stimulus, higher pupillary activity was observed in the pre-baseline period. For listeners in the no-anchor group, their pupillary responses were slightly lower indicating. The correlation between their PPDs and auditory-perceptual ratings were not found to be as strong as the with-anchor group and exclusion of anchors seemed to have reduced the potential cognitive loads in these listeners. In addition, five out of ten participants in this group (no-anchor) had participated in the first experiment and had previous exposure to the dysphonic voices and habituation may have reduced the average PPD of the group. Their pre-baseline pupillary activity was also smaller compared to the with-anchor group.

6.3 Experiment 3 (Vocal Fold Paralysis)

The focus of the third study was on auditory-perceptual evaluation of breathiness in individuals with vocal fold paralysis and listening effort and objective pupillary assessments. Similar to Experiment 2, the potential influence of audio anchors on ratings was examined. Both groups’ ratings were highly reliable, but the no-anchor group had slightly higher reliability values. Correlations between breathiness and effort were found to be very high in both groups (with-anchor group: 0.93; no-anchor group: 0.92). The correlation between PPDs and breathiness ratings were similar in both groups indicating that an increase in PPD did vary together with an increase in listeners' perceptions breathiness. However, the correlation between PPD and listening effort was observed to be higher in with-anchor group than the no-anchor listeners meaning that the PPD was higher for the talkers with more listening effort ratings. Similar to Experiment 2, the PPD behavior of with-anchor group seems to be influenced by the inclusion of anchors.

6.4 Discussion

Because one of the objectives of the study was to evaluate experienced listeners' pupillary responses and listening effort, three experienced listeners participated in the Experiment 1. The next two projects, Experiments 2 and 3 were conducted only with naïve listeners who were divided into two groups per study, those who provided their rating with-anchors and those without any anchors. For all three experiments, data were collected through two computerized sliders representing VAS and EyeLink1000. The computerized sliders were consistent across the three experiments except with the vocal dimensions (strain, roughness, breathiness), which were unique per each experiment.

Data from all three experiments offer significant observations and valuable insights into how naïve and experienced listeners (experiment 1) judge various aspects of voice quality. These findings are also enhanced when evaluated in the context of the listeners' simultaneous ratings of listening effort. Thus, not only was an auditory-perceptual feature assessed, but listeners were requested to make judgment of "how much work" was required for them to process the audio samples. Our results also provide insights into how listeners cognitively and physiologically respond while doing a listening task. Voice disorders, although intelligible, are found to be imposing listening effort and cognitive load on listeners. This load may vary depending on the vocal feature and prior experience of the listeners.

Our data represents two categories of measurement: perceptual data and pupillary/physiological or objective data obtained through pupillometry. The paragraphs below briefly summarize the similarities and differences observed in all three research projects with consideration of both auditory-perceptual and pupillometric data.

Speech stimuli for all three projects varied extensively in the degree of voice dimensions. A wide range of ratings were generated for the stimuli in all three experiments. In terms of intra-rater reliability, the ratings of the AdSD samples by the experienced listeners had the highest range for auditory-perceptual feature of strain (0.72 to 0.86) and the TE no-anchor group had the lowest intra-rater range (0.37 to 0.72) for roughness. The experienced group in the AdSD study also demonstrated the narrowest intra-rater reliability range (0.71 to 0.78) for listening effort, whereas the range from judgments with-anchor group of the VFP samples was the widest (-0.7 to 0.88). All participants in the 6 groups (two groups per experiment) rated the voice feature and listening effort consistently and correlations between the voice feature and listening effort was very high. The highest correlations between the voice feature and listening effort value (0.97) was obtained between roughness and listening effort for the TE speaker samples judge by listeners in the with-anchor group, with the lowest value (0.89) obtained from the naïve listeners between strain and listening effort in AdSD experiment.

Interrater reliabilities from all experiments were quite good with the highest value (0.98) observed for the voice dimension of strain obtained for judgments of the AdSD by the naïve listener group. The highest interrater reliability for listening effort (0.97) was also obtained from this same group. In contrast, the lowest interrater values for the voice feature of strain (0.86) and listening effort (0.83) were provided for the AdSD speaker group by experienced listeners.

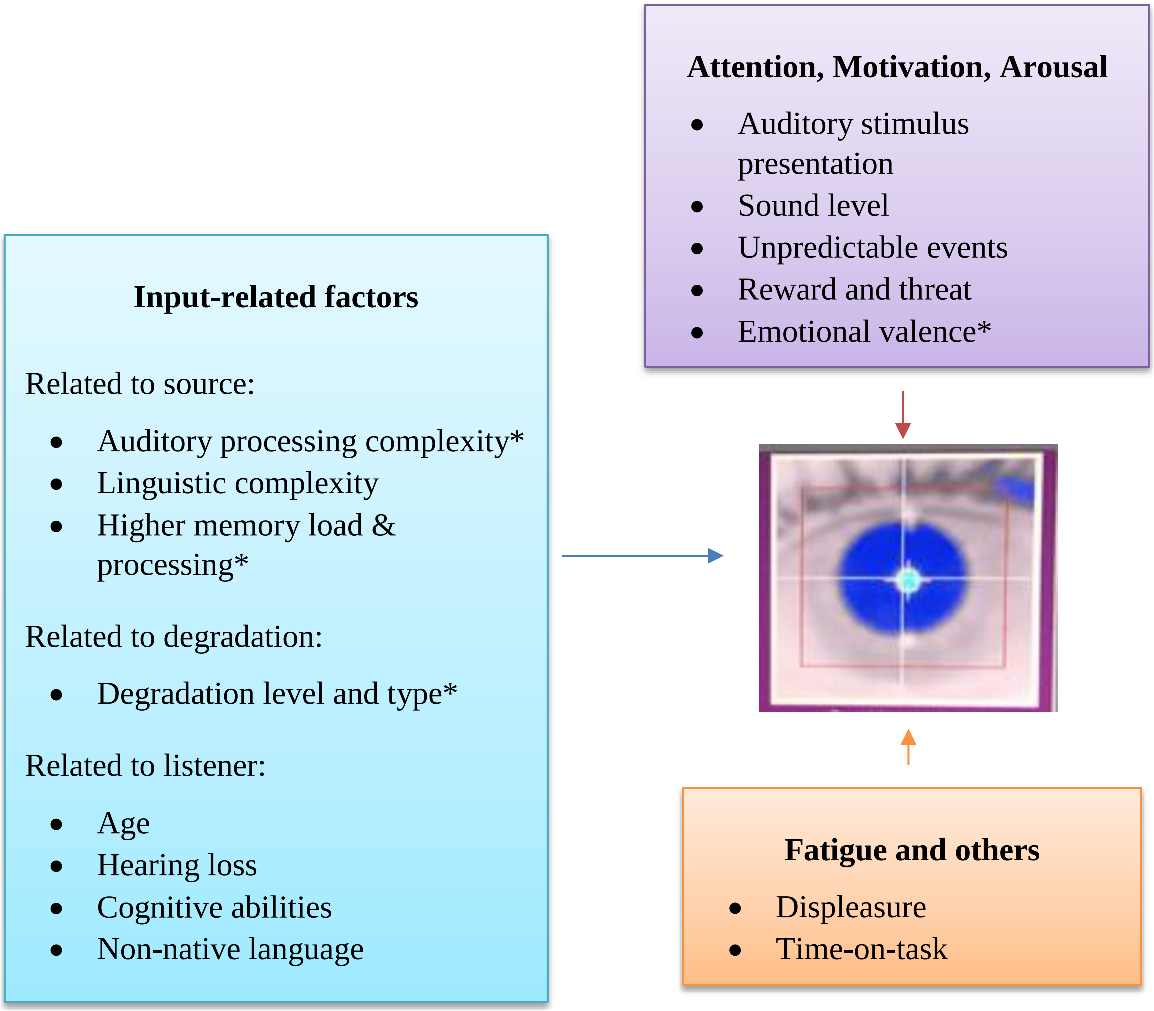
Regarding the use of anchors, ratings were generally rated consistently with less fluctuations in the with-anchor groups. Talkers with higher breathiness ratings were also generally rated higher on listening effort. However, the reliability was slightly lower in the VFP with-anchor group for both features (0.94 for breathiness, 0.88 for effort) compared to the no-anchor group (0.96 for breathiness, 0.89 for effort).

Statistical analysis in all three studies (auditory-perceptual data) showed significant effects for perceptual features (strain, roughness, breathiness, listening effort). There was no group effect for TE and VFP (with-anchor and no-anchor) and differences between the group ratings were not statistically significant. This indicates that they all rated the features similarly. In terms of any interactions between features and talkers, none were found in the TE study (Experiment 2) meaning that all talkers received ratings in a similar manner for both roughness and listening effort. However, there was an interaction found between features and talkers in both the AdSD and VFP studies (Experiments 2 and 3) which revealed that two talkers in each study were rated more on listening effort when compare to that of the voice dimension (i.e. strain, roughness, breathiness). Such interactions confirm the fact that listening effort may involve multiple perceptual factors. A disordered voice may be rated as less breathy or less rough for example but high on listening effort due to the overall composite quality of the voice.

6.4.1 Pupillometry

The other objective of this study was directed towards the examination of pupillary reactions in response to a listener's exposure to disordered voices. The pupillary results did appear to vary across studies, listener groups, and voice dimensions under evaluation. While the discussion in individual chapters focused more on the interpretation of pupil dilation data from a cognitive load perspective, a more holistic discussion of other potential contributing factors is presented here.

As described in Section 2.8.3.2, task-evoked pupil dilation is a combination of attention, arousal, engagement, effort and anxiety and not a unitary concept of effort (Nunnally, Knott, Duchnowski, & Parker, 1967; Pichora-Fuller et al., 2016). Zekveld et al. (2018) reviewed the current state of knowledge on pupil dilation response to auditory stimulation and identified the plausible factors contributing to pupil dilation during an auditory behavioral experiment, which are summarized in Figure 6-1. Some or all of these elements may be present in a pupillometry study at varying degrees.



Input-related factors

Related to source:

- Auditory processing complexity*
- Linguistic complexity
- Higher memory load & processing*

Related to degradation:

- Degradation level and type*

Related to listener:

- Age
- Hearing loss
- Cognitive abilities
- Non-native language

Attention, Motivation, Arousal

- Auditory stimulus presentation
- Sound level
- Unpredictable events
- Reward and threat
- Emotional valence*

Fatigue and others

- Displeasure
- Time-on-task

Figure 6-1 Factors influencing pupil responses during auditory processing, derived from Figure 3 in Zekveld et al. (2018). Those factors that potentially played a role in our pupil data are marked with a *.

A subset of these potential influential factors can be discounted for the present set of auditory experiments. For example, all participants in our studies were native English speakers, and reported normal hearing and cognitive functioning. Furthermore, the naïve participants were young adults (age range 18–33 years, across all experiments). The experienced listeners in the AdSD study were older (age range 41–56 years), and while there is some evidence that baseline pupil size may be smaller for older listeners, there is no concrete evidence that the baseline normalized PPD reduces with age (see Zekveld et al. (2018) review). As such, all listener-related factors from Figure 6-1 can be ignored as potential contributors to the pupil data collected in our experiments.

Among the other input-related factors, degradation level and type are highly relevant to our studies. The stimuli in the three experiments reported here were all degraded as they were recorded from individuals presenting with different voice disorders (degradation type) and various degrees of vocal feature severity pertaining to those disorders (degradation level). With respect to source-related factors, linguistic complexity was not a factor, as all stimuli were recordings of the second sentence of the Rainbow Passage. Auditory processing complexity and memory load/processing factors are relevant and are discussed separately later with respect to each study.

Under the attention and arousal parameter group, auditory stimulus presentation is not a factor as it was cued, predictable (i.e. listeners knew that they would hear disordered voice samples), and not sudden. Similarly, sound level is not a contributing factor as all stimuli were RMS-equalized prior to their presentation at a comfortable listening level. There were no reward, penalty, or threat parameters within the experimental paradigms. The remaining factor in this group is emotional valence, which represents the attractiveness (positive affect) or aversiveness (negative affect) to an auditory stimulus (Francis & Love, 2020). Evidence exists for increased pupil dilation when listening to auditory stimuli with negative affective connotations (Francis & Love, 2020; Zekveld et al. 2018). As such, emotional valence may be a contributing factor to our pupil data, especially for those naïve listeners who are exposed to abnormal voice samples for the first time and perceived them to be aversive.

As shown in Figure 6-1, pupil responses may also be mediated by fatigue due to the listening task itself or time spent on completing the listening task (Beatty, 1982; Kahneman & Beatty, 1967). While the potential effect of these factors cannot be completely ignored, our experiments were relatively shorter in duration (~ 45 minutes per session), which included a ten-minute break between the test and retest sessions. Listeners who participated in more than one session (i.e. for more than one experiment), did so after a break of at least a week. It is therefore unlikely that fatigue and time-on-task significantly influenced the pupillary reactions in our studies.

The potential contribution of the aforementioned relevant factors can now be discussed in further detail for each study. In the AdSD study, there was a strong correlation identified between PPD and strain (0.73) and between PPD and listening effort (0.65) in the naïve group indicating that the more strained a voice sample, the higher the PPD value. This indicates that the auditory processing complexity and degradation level factors are potentially contributing to the increased pupil dilation observed with this group. However, the experienced listeners seemed not to be as greatly impacted by the disordered voices as did those in the naïve group. The PPD values of the experienced listener group did not appear to be following a similar pattern as naïve listener group. In fact, the correlation between strain and PPD and listening effort and PPD were very poor for the experienced group. The pupil tracks of the perceptually best and worst talkers were examined relative to pupillometry and no difference was identified between them. This, perhaps, points to the role of the emotional valence factor. Experienced listeners are used to listening to disordered voices and are therefore emotionally neutral to their presentation. On the other hand, the strained/strangled voice quality perhaps led to aversiveness and increased pupil dilation in the naïve listener group. It is also pertinent to highlight the changes in the pupil tracks between test and retest sessions, as shown in Figures 3.18 and 3.19. The peak pupil dilation, albeit pronounced, is reduced in magnitude on the second presentation of the Talker #1 stimulus. This perhaps insinuates that repeated exposure may reduce the negative emotional valence and these results are in line with previous studies that report habituation due to repetition and exposure (Dahlman, Sjors, Lindstrom, Ledin, & Falkmer, 2009; Damsma & Van Rijn, 2017; Marois et al., 2018).

In addition, the experienced listener group in the AdSD study demonstrated high pre-baseline pupillary activity. Their pupillary reactions were all homogenously high for the pre-baseline duration of the audio stimuli (Figure 3.13). This pupillary response pattern from the experienced group may be attributed to the fact that experienced listeners (often clinicians) may follow different strategies for rating voices (Kreiman & Gerratt, 2000).

Because their professional duties require them to allocate attention, focus and expend cognitive resources for assessment and diagnosis, their cognitive preparedness and allocation of resources may be illustrated by their variation in pupillary reactions pre-baseline. These findings allude to the role of the "memory load & processing" factor in impacting the pupil dilation during an auditory task.

In the TE study (Experiment 2), moderate correlations were found for the with-anchor group between PPD and roughness (0.58) and for PPD and listening effort (0. 64). Such a correlation between PPD and auditory-perceptual ratings was not as good for listeners in the no-anchor group. This could potentially be due to the higher proportion (50%) of listeners in the no-anchor group participated in the AdSD study than with-anchor group (30%). Due to their prior exposure to disordered voices, it is plausible that half of the no-anchor group participants may have tempered pupil response arising from the emotional valence component. One interesting finding related to the TE experiment though, was the higher pre-baseline activity and generally higher pupillary and PPD responses in the with-anchor group. Such results were due to the fact that their working memory capacity was occupied by holding the two anchors and comparison of each stimulus to those anchors held in the working memory. The attention allocation and the necessary cognitive load is being indicated through their pupillary tracks (Unsworth, & Robison, 2018). Also, it is of value to note that the increase in the pupil diameter values began shortly after the onset of the stimulus. This observation would correspond to the audio sample as roughness may have been perceived shortly after initiation of the stimuli and, interestingly, this finding differs from the pupillary behavior generated by naïve listeners in the AdSD study.

In the VFP study (Experiment 3) similar non-significant correlations between perceptual ratings and PPD were obtained for both listener groups (i.e., with-anchor and no-anchor), even though a majority of the naïve listeners in the no-anchor group were first-time participants in our studies. These results indicate that the degree and level of degradation associated with the VFP stimuli were not influential in evoking pupillary response. Furthermore, it is plausible that the emotional valence of the VFP stimuli is neutral, due to the "noisiness" of the voice stimuli. Interestingly, the tracks from listeners in the with-anchor group revealed higher pre-baseline activity. This was similar to what was observed in the TE with-anchor group, indicating that again talkers are actively maintaining the anchors in their working memory waiting for the onset of stimuli. The PPD values were lower in the with-anchor group, a finding that is believed to be a result of the fact that the majority of listeners in that group had participated in the previous studies. Thus, it is possible that exposure to the prior dysphonic voices may have resulted in the perceived decrease in their cognitive load specific to listening effort due to already developed cognitive schema. In addition, the fact that the majority in the no-anchor group were first time participants justified their higher PPDs because the information was new to them and there was no previous exposure and habituation.

In summary, our results were mainly pertinent to the behavioral ratings and pupillometric measurements of effort when listeners audited intelligible disordered voice samples. The following general conclusions may be drawn based on the observations across all three studies: (a) pupil dilation was dependent on stimulus degradation level and type, which in turn were related to the voice disorder type and degree of severity; (b) emotional response to the underlying voice abnormality may have been a contributing factor, with the strained/strangled voice quality having a greater impact than breathiness; and (c) increased memory load was observed when naïve listeners were instructed to base their ratings on pre-defined anchors and or when experienced listeners rated disordered voice samples.

6.5 Limitations of the present studies

While the present data offer valuable insights on various aspects of auditory-perceptual evaluation of voice quality, there are some limitations which deserve mention. Talker or listener gender was not controlled in these studies, so we are not certain if any of the results potentially may have been influenced by gender. Although some evidence was found regarding the possible influence of previous participation, the temporal gap between test re-test was relatively short (10 minute). Interestingly, the data acquired from some of our participants, namely, those who wore contact lenses while doing the experiment and/or wore mascara, had more blinks and dropouts in their pupillary response tracks compared to other listeners. Also, physiological differences regarding the shape of the eye due to race in few of our participants was causing missing pupil data.

6.6 Clinical Implications

This series of projects focused on potential relationships between auditory-perceptual judgments of voice disorder and physiological responses of the listener via pupillometry. Examining pupillary reactions in response to a variety of stimuli has long been used as a measure of an individual's response to such stimuli (Kahneman & Beatty, 1966; Kramer et al., 2013; Zekveld & Kramer, 2014; Zekveld et al., 2018). In terms of its application in the area of clinical voice disorders, however; no direct clinical implications of pupillometry can be suggested at this time. Although pupillary responses were tracked in relation to auditory-perceptual ratings in the three studies reported herein, the responses were noted to be variable. However, in this series of studies, pupillary evaluation was paired with auditory-perceptual methods which currently serve as the gold standard in voice assessment (Kreiman et al, 1992). In addition to the subjective evaluations obtained from this widely used and standard method of voice assessment, the current pupillometry data may also demonstrate physiologic reactions in a listener in response to abnormal voice qualities. This potential relationship has been demonstrated specific to abnormal voice quality that characterizes three specific voice disorders – adductory spasmodic dysphonia, vocal fold paralysis, and tracheoesophageal speech. Again, at the current time, we do not see any direct application of this method of physiologic measurement in the context of voice disorders. However, several additional considerations can be made based on the data gathered in this series of projects.

First, although a clear and distinct capacity to track pupillary responses was possible, variability in these responses were individualized. The presence of such physiologic reactions to vocal abnormalities regardless of the underlying dimension assessed (i.e., strain, roughness, and breathiness) may suggest larger, communication considerations. That is, listeners do appear to have an involuntary, physiologic response to abnormal vocal stimuli. This observation would appear to raise questions about the influence of a voice disorder on the communication interactions between a speaker and their listener (Eadie & Doyle, 2004). Yet, this response and its impact apparently varies by feature and disorder type. In fact, the degree of voice abnormality, the inherent changes in one's voice quality, and the severity of the disorder may enhance the impact on listeners. The data from this study may, therefore, signify the importance of assessing participation considerations in that if listeners are not comfortable with speaker, they may avoid interacting with them. More directly, if one presents with a voice disorder, this may challenge effective communication with negative impact on both the speaker and the listener. This rather negative influence on the communication dyad clearly raises the issue of better counselling for individuals who present with voice disorders; such counselling would encourage clinicians to educate their patients so that they are aware that their disorder may create larger difficulties in communication, and make listeners uncomfortable. This type of education, at least to some extent, may then serve to reduce the communication demand in the dyad. Under these circumstances, clinical efforts to document the perceived disability experienced by the speaker regardless of voice disorder type may provide a valuable index of the true impact of the disorder on communication (Eadie & Doyle, 2004). Finally, data from this study also highlights the importance of gathering an array of information from those with disorders how they perceive listeners to respond to their abnormal voices. Clinical education and ongoing counselling may then serve to provide the speaker with enhanced understanding of how listeners may respond to their voice quality, as well as serving to guide the best level of patient care to those presenting with dysphonia.

6.7 Directions for future research

The present data provide a strong foundation for future work on pupillary response and the auditory-perceptual evaluation and description of voice disorders. Future studies can examine potential gender variations in terms of pupillary reactions to disordered voices. It would be interesting to recruit older listeners and investigate their pupillary responses to disordered voices as well. Future studies also may seek to assess longer gaps between test and re-test to examine whether the exposure to stimuli would fade away and PPD would be altered with increased break. Our speech stimuli were all intelligible. Future studies may investigate dysphonic voices with various degrees of intelligibility.

References

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247-264.

Amir, O., Biron-Shental, T., & Shabtai, E. (2006). Birth control pills and nonprofessional voice: Acoustic analyses. *Journal of Speech, Language, and Hearing Research.*

Amir, O., Biron-Shental, T., Muchnik, C., & Kishon-Rabin, L. (2003). Do oral contraceptives improve vocal quality? Limited trial on low-dose formulations. *Obstetrics & Gynecology, 101*(4), 773-777.

Amir, O., & Kishon-Rabin, L. (2004). Association between birth control pills and voice quality. The *laryngoscope, 114*(6), 1021-1026.

Amir, O., & Levine-Yundof, R. (2013). Listeners' attitude toward people with dysphonia. *Journal of Voice, 27*(4), 524-e1.

Anderson, V. A. (1942). Training the speaking voice. Oxford University Press.

Antonenko, P., Paas, F., Grabner, R., & Van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review, 22*(4), 425-438.

Baggs, T. W., & Pine, S. J. (1983). Acoustic characteristics: Tracheoesophageal speech. *Journal of Communication Disorders, 16*(4), 299-307.

Bangayan, P., Long, C., Alwan, A. A., Kreiman, J., & Gerratt, B. R. (1997). Analysis by synthesis of pathological voices using the Klatt synthesizer. *Speech Communication, 22(*4), 343-368.

Barsties, B., Beers, M., Ten Cate, L., Van Ballegooijen, K., Braam, L., De Groot, M., & Maryn, Y. (2017). The effect of visual feedback and training in auditory-perceptual judgment of voice quality. *Logopedics Phoniatrics Vocology, 42*(1), 1-8.

Barsties, B., & Maryn, Y. (2017). The influence of voice sample length in the auditory-perceptual judgment of overall voice quality. *Journal of Voice, 31*(2), 202-210.

Bassich, C. J., & Ludlow, C. L. (1986). The use of perceptual methods by new clinicians for assessing voice quality. *Journal of Speech and Hearing Disorders, 51*(2), 125-133.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin, 91*(2), 276.

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

Bergamin, O., & Kardon, R. H. (2002). Greater pupillary escape differentiates central from peripheral visual field loss. *Ophthalmology, 109*(4), 771-780.

Best, V., Roverud, E., Streeter, T., Mason, C. R., & Kidd Jr, G. (2017). The benefit of a visually guided beamformer in a dynamic speech task. *Trends in hearing,* 21, 2331216517722304.

Beukelman, D. R., Childes, J., Carrell, T., Funk, T., Ball, L. J., & Pattee, G. L. (2011). Perceived attention allocation of listeners who transcribe the speech of speakers with amyotrophic lateral sclerosis. *Speech Communication, 53*(6), 801-806.

Boone, D. R., McFarlane, S. C., Von Berg, S. L. (2005). The voice and voice therapy. Allyn and Bacon.

Brandt, J. F., Ruder, K. F., & Shipp Jr, T. (1969). Vocal loudness and effort in continuous speech. *The Journal of the Acoustical Society of America, 46*(6B), 1543-1548.

Brinca, L., Batista, A. P., Tavares, A. I., Pinto, P. N., & Araújo, L. (2015). The effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli. *Journal of Voice, 29*(6), 776-e7.

Brown, G. G., Kindermann, S. S., Siegle, G. J., Granholm, E., Wong, E. C., & Buxton, R. B. (1999). Brain activation and pupil response during covert performance of the Stroop Color Word task. *Journal of the International Neuropsychological Society, 5*(4), 308-319.

Cannito, M. P., Burch, A. R., Watts, C., Rappold, P. W., Hood, S. B., & Sherrard, K. (1997). Disfluency in spasmodic dysphonia: A multivariate analysis. *Journal of Speech, Language, and Hearing Research, 40*(3), 627-641.

Carding, P. (2000). Evaluating voice therapy: Measuring the effectiveness of treatment. John Wiley & Sons Incorporated.

Case, J. L. (2002). Clinical management of voice disorders. (4th ed.). Austin, TX: Pro-Ed.

Coelho, A. C., Brasolotto, A. G., Fernandes, A. C. N., de Souza Medved, D. M., da Silva, E. M., & Júnior, F. B. (2017). Auditory-Perceptual Evaluation of Voice Quality of Cochlear-implanted and Normal-hearing Individuals: A Reliability Study. *Journal of Voice, 31*(6), 774-e1.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences, 24*(1), 87-114.

Crumley, R. L. (1994). Unilateral recurrent laryngeal nerve paralysis. *Journal of voice, 8*(1), 79-83.

D'Alatri, L., Bussu, F., Scarano, E., Paludetti, G., & Marchese, M. R. (2012). Objective and subjective assessment of tracheoesophageal prosthesis voice outcome. *Journal of Voice, 26*(5), 607-613.

Damrose, J. F., Goldman, S. N., Groessl, E. J., & Orloff, L. A. (2004). The impact of long-term botulinum toxin injections on symptom severity in patients with spasmodic dysphonia. *Journal of Voice, 18*(3), 415-422.

Damsma, A., & van Rijn, H. (2017). Pupillary response indexes the metrical hierarchy of unattended rhythmic violations. *Brain and Cognition, 111,* 95-103.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language, 47*(2), 292-314.

Dahlman, J., Sjörs, A., Lindström, J., Ledin, T., & Falkmer, T. (2009). Performance and autonomic responses during motion sickness. *Human factors, 51*(1), 56-66.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975). Motor speech disorders. Saunders.

De Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech, Language, and Hearing Research, 37*(5), 985-1000.

Doyle, P. C. (1994). Foundations of voice and speech rehabilitation following laryngeal cancer. Singular Pub Group.

Doyle, P.C., Danhauver, J.L., & Reed, C.G. (1988). Listeners' perception of consonants produced by esophagael and tracheoesophageal talkers. *Journal of Speech and Hearing Disorders, 53*, 400–407.

Doyle, P. C., Swift, E. R., & Haaf, R. G. (1989). Effects of listener sophistication on judgments of tracheoesophageal talker intelligibility. *Journal of Communication Disorders, 22*(2), 105-113.

Duffy, J. R. (1995). Motor speech disorders: Substrates, differential diagnosis, and management. St. Louis, MO: Mosby-Year Book.)^(Eds.):'Book Motor speech disorders: Substrates, differential diagnosis, and management. St. Louis, MO: Mosby-Year Book' (Inc, 2005, edn.).

Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (TE) speakers. *Journal of Speech, Language, and Hearing Research, 45*(6), 1088-1096.

Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *The Journal of the Acoustical Society of America, 112*(6), 3014-3021.

Eadie, T. L., & Doyle, P. C. (2004). Auditory-perceptual scaling and quality of life in tracheoesophageal speakers. *The Laryngoscope, 114*(4), 753-759.

Eadie, T. L., Doyle, P. C., Hansen, K., & Beaudin, P. G. (2008). Influence of speaker gender on listener judgments of tracheoesophageal speech. *Journal of Voice, 22*(1), 43-57.

Eadie, T. L., Nicolici, C., Baylor, C., Almand, K., Waugh, P., & Maronian, N. (2007). Effect of experience on judgments of adductor spasmodic dysphonia. *Annals of Otology, Rhinology & Laryngology, 116*(9), 695-701.

Eadie, T. L., Rajabzadeh, R., Isetti, D. D., Nevdahl, M. T., & Baylor, C. R. (2017). The effect of information and severity on perception of speakers with adductor spasmodic dysphonia. *American journal of speech-language pathology, 26*(2), 327-341.

Einhäuser, W. (2017). The pupil as marker of cognitive processes. In Computational and Cognitive Neuroscience of Vision (pp. 141-169). Springer, Singapore.

Evitts, P. M., Starmer, H., Teets, K., Montgomery, C., Calhoun, L., Schulze, A., & Adams, L. (2016). The Impact of Dysphonic Voices on Healthy Listeners: Listener Reaction Times, Speech Intelligibility, and Listener Comprehension. *American journal of speech-language pathology, 25*(4), 561-575.

Fairbanks, G. (1960). The rainbow passage. Voice and articulation drillbook, 2.

Ferrand, C. T. (2011). Voice disorders: Scope of theory and practice. Pearson Higher Ed.

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

Ferrer, C. A., Haderlein, T., Maryn, Y., De Bodt, M. S., & Nöth, E. (2018). Collinearity and sample coverage issues in the objective measurement of vocal quality: The case of roughness and breathiness. *Journal of Speech, Language, and Hearing Research, 61*(1), 1-24.

Francis, A. L., & Love, J. (2020). Listening effort: Are we measuring cognition or affect, or both?. *Wiley Interdisciplinary Reviews: Cognitive Science, 11*(1), e1514.

Gerratt, B. R., Kreiman, J., Antonanzas-Barroso, N., & Berke, G. S. (1993). Comparing internal and external standards in voice quality judgments. *Journal of Speech, Language, and Hearing Research, 36*(1), 14-20.

Gerratt, B. R., Kreiman, J., & Garellek, M. (2016). Comparing measures of voice quality from sustained phonation and continuous speech. *Journal of Speech, Language, and Hearing Research, 59*(5), 994-1001.

Gosselin, P. A., & Gagné, J. P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research*.

Goy, H., Kathleen Pichora-Fuller, M., & Van Lieshout, P. (2016). Effects of age on speech and voice quality ratings a. *The Journal of the Acoustical Society of America, 139*(4), 1648-1659.

Haeberle, E. J. (1981). The sex atlas. New York, NY: Continuum Publishing Company.

Hällgren, M., Larsby, B., Lyxell, B., & Arlinger, S. (2005). Speech understanding in quiet and noise, with and without hearing aids: Comprensión del lenguaje en silencio y con ruido, con y sin auxiliares auditivos. *International Journal of Audiology, 44*(10), 574-583.

Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta oto-laryngologica, 90*(1-6), 441-451.

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

Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M., & Vaughan, C. (1989). Objective assessment of vocal hyperfunction: An experimental framework and initial results. *Journal of Speech, Language, and Hearing Research, 32*(2), 373-392.

Harkrider, A. W., Plyler, P. N., & Hedrick, M. S. (2009). Effects of hearing loss and spectral shaping on identification and neural response patterns of stop-consonant stimuli in young adults. *Ear and hearing, 30*(1), 31-42.

Helou, L. B., Solomon, N. P., Henry, L. R., Coppit, G. L., Howard, R. S., & Stojadinovic, A. (2010). The role of listener experience on Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ratings of postthyroidectomy voice. *American Journal of Speech-Language Pathology, 19*(3), 248-258.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science, 132*(3423), 349-350.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143*(3611), 1190-1192.

Hirano, M. (1981). Clinical examination of voice. Disorders of human communication, 5, 1-99.

Hollien, H. (2000). The concept of ideal voice quality. In: Kent, R. D., & Ball, M. J. (Eds.), Voice quality measurement (pp. 13-24). San Diego, CA: Singular.

Hyönä, J., Tommola, J., & Alaja, A. M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology, 48*(3), 598-612.

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

Isetti, D., Xuereb, L., & Eadie, T. L. (2014). Inferring speaker attributes in adductor spasmodic dysphonia: Ratings from unfamiliar listeners. *American Journal of Speech-Language Pathology, 23*(2), 134-145.

Järvinen, K., Laukkanen, A. M., & Geneid, A. (2017). Voice Quality in Native and Foreign Languages Investigated by Inverse Filtering and Perceptual Analyses. *Journal of Voice, 31*(2), 261-e25.

Johnsrude, I. S., & Rodd, J. M. (2016). Factors that increase processing demands when listening to speech. In Neurobiology of Language (pp. 491-502). Academic Press.

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 47*(2), 310.

Kahneman, D. (1973). Attention and effort (Vol. 1063). Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154*(3756), 1583-1585.

Kelchner, L. N., Stemple, J. C., Gerdeman, B., Le Borgne, W., & Adam, S. (1999). Etiology, pathophysiology, treatment choices, and voice results for unilateral adductor vocal fold paralysis: a 3-year retrospective. *Journal of Voice, 13*(4), 592-601.

Kempster, G. B., Gerratt, B. R., Abbott, K. V., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology, 18*(2), 124-132.

Kent, R. D., & Ball, M. J. (Eds.). (2000). Voice quality measurement. San Diego, CA: Singular.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America, 67*(3), 971-995.

Klasner, E. R., & Yorkston, K. M. (2005). Speech intelligibility in ALS and HD dysarthria: The everyday listener's perspective. *Journal of Medical Speech-Language Pathology, 13*(2), 127-140.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing, 33*(2), 291-300.

Kramer, S. E., Kapteyn, T. S., Festen, J. M., & Kuik, D. J. (1997). Assessing aspects of auditory handicap by means of pupil dilatation. *Audiology, 36*(3), 155-164.

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

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

Kreiman, J., Gerratt, B. R., & Berke, G. S. (1994). The multidimensional nature of pathologic vocal quality. *The Journal of the Acoustical Society of America, 96*(3), 1291-1302.

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

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech, Language, and Hearing Research, 36*(1), 21-40.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech, Language, and Hearing Research, 35*(3), 512-520.

Kreiman, J., Gerratt, B. R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech, Language, and Hearing Research, 33*(1), 103-115.

Kramer, S. E., Lorens, A., Coninx, F., Zekveld, A. A., Piotrowska, A., & Skarzynski, H. (2013). Processing load during listening: The influence of task characteristics on the pupil response. *Language and cognitive processes, 28*(4), 426-442.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing, 33(*2), 291-300.

Kursawe, M. A., & Zimmer, H. D. (2015). Costs of storing colour and complex shape in visual working memory: Insights from pupil size and slow waves. *Acta psychologica, 158*, 67-77.

Laczi, E., Sussman, J. E., Stathopoulos, E. T., & Huber, J. (2005). Perceptual evaluation of hypernasality compared to HONC measures: The role of experience. *The Cleft palate-craniofacial journal, 42*(2), 202-211.

Laeng, B., & Endestad, T. (2012). Bright illusions reduce the eye's pupil. *Proceedings of the National Academy of Sciences, 109*(6), 2162-2167.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious?. *Perspectives on psychological science, 7*(1), 18-27.

Laeng, B., & Sulutvedt, U. (2014). The eye pupil adjusts to imaginary light. *Psychological science, 25*(1), 188-197.

Larsby, B., Hällgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: effects of different noise backgrounds in normal-hearing and hearing-impaired subjects Desempeño cognitivo y percepción del esfuerzo en tareas de procesamiento del lenguaje: Efectos de las diferentes condiciones de fondo en sujetos normales e hipoacúsicos. *International Journal of Audiology, 44*(3), 131-143.

Linville, S. E. (2000). The Aging voice. In: Kent, R. D., & Ball, M. J. (Eds.), Voice quality measurement (pp. 359-376). San Diego, CA: Singular.

Linville, S. E., & Korabic, E. W. (1986). Elderly listeners' estimates of vocal age in adult females. *The Journal of the Acoustical Society of America, 80*(2), 692-694.

Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology, 22*(2), 113-122.

McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2017). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology, 54* (2), 193-203.

Marois, A., Labonté, K., Parent, M., & Vachon, F. (2018). Eyes have ears: Indexing the orienting response to sound using pupillometry. *International Journal of Psychophysiology, 123*, 152-162

Mathieson, L. (2000). Normal- disordered continuum. In: Kent, R. D., & Ball, M. J. (Eds.), Voice quality measurement (pp. 3-12). San Diego, CA: Singular. McDowell, I., & Newell, C. (1987). Measuring health: A guide to rating scales and questionnaires. Oxford: Oxford University Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review, 63*(2), 81.

Most, T., Tobin, Y., & Mimran, R. C. (2000). Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *Journal of communication disorders, 33*(2), 165-181.

Nagle, K. F., & Eadie, T. L. (2012). Listener effort for highly intelligible tracheoesophageal speech. *Journal of Communication Disorders, 45*(3), 235-245.

Nash, E. A., & Ludlow, C. L. (1996). Laryngeal muscle activity during speech breaks in adductor spasmodic dysphonia. *The Laryngoscope, 106*(4), 484-489.

Nemr, K., Simoes-Zenari, M., Cordeiro, G. F., Tsuji, D., Ogawa, A. I., Ubrig, M. T., & Menezes, M. H. M. (2012). GRBAS and Cape-V scales: high reliability and consensus when applied at different times. *Journal of Voice, 26*(6), 812-e17.

Newell, C. J. S. (2007). Gender perception for tracheoesophageal speech: A comparative evaluation of paired comparison and visual analog scaling paradigm. (Master's thesis). Western University, London, Ontario.

Ng, M.L., Kwok, C.L., & Chow, S.F. (1997). Speech performance of adult Cantonese-speaking laryngectomees using different types of alaryngeal phonation. *Journal of Voice, 11,* 338–344.

Nunnally, J. C., Knott, P. D., Duchnowski, A., & Parker, R. (1967). Pupillary response as a general measure of activation. *Perception & psychophysics, 2*(4), 149-155.

Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality. *Folia Phoniatrica et Logopaedica, 61*(1), 49-56.

O'Brian, S., Packman, A., Onslow, M., Cream, A., O'Brian, N., & Bastock, K. (2003). Is listener comfort a viable construct in stuttering research?. *Journal of Speech, Language, and Hearing Research, 46*(2), 503-509.

Pichora-Fuller, M. K., & Kramer, S. E. (2016). Eriksholm workshop on hearing impairment and cognitive energy.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., ... & Naylor, G. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing, 37,* 5S-27S.

Pichora-Fuller, M. K., & Singh, G. (2006). Effects of age on auditory and cognitive processing: implications for hearing aid fitting and audiologic rehabilitation. *Trends in amplification, 10*(1), 29-59.

Pittman, A., Vincent, K., & Carter, L. (2009). Immediate and long-term effects of hearing loss on the speech perception of children. *The Journal of the Acoustical Society of America, 126*(3), 1477-1485.

Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly journal of experimental psychology, 20*(3), 241-248.

Robbins, J., Fisher, H. B., Blom, E. C., & Singer, M. I. (1984). A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. *Journal of Speech and Hearing disorders, 49*(2), 202-210.

Sedory, S.E., Hamlet, S.L., & Connor, N.P. (1989). Comparisons of perceptual and acoustic characteristics of tracheoesophageal and excellent esophageal speech. *Journal of Speech and Hearing Disorders, 54,* 209–214.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological review*, *84*(2), 127.

Sharma, G., & Goodwin, J. (2006). Effect of aging on respiratory system physiology and immunology. *Clinical interventions in aging, 1*(3), 253.

Simpson, H. M., & Hale, S. M. (1969). Pupillary changes during a decision-making task. *Perceptual and Motor Skills, 29*(2), 495-498.

Simpson, H. M., & Molloy, F. M. (1971). Effects of audience anxiety on pupil size. *Psychophysiology, 8*(4), 491-496.

Singer, M. I., & Blom, E. D. (1979, May). Tracheoesophageal puncture: a surgical prosthetic method for postlaryngectomy speech restoration. In unpublished paper presented to the Third International Symposium on Plastic-Reconstructive Surgery of the Head and Neck, New Orleans (Vol. 4).

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive psychology, 49*(3), 238-299.

Södersten, M., & Lindestad, P. Å. (1990). Glottal closure and perceived breathiness during phonation in normally speaking subjects. *Journal of Speech, Language, and Hearing Research, 33*(3), 601-611.

Sofranko, J., & Prosek, R. A. (2012). The effect of experience on classification of voice quality. *Journal of Voice, 26*(3), 299-303.

Sofranko Kisenwether, J., & Prosek, R. A. (2014). The effect of levels and types of experience on judgment of synthesized voice quality. *Journal of Voice, 28*(1), 24-35.

Sofranko Kisenwether, J., & Prosek, R. A. (2016). The Effect of Experience on Response Time When Judging Synthesized Voice Quality. *Journal of Voice, 30*(4), 394-397.

Stevens, S. (1975). *Psychophysics: Introduction to its Perceptual, Neural and Social Prospects.* (Wiley, New York).

Suhail, I. S., Kazi, R. A., & Jagade, M. (2016). Perceptual evaluation of tracheoesophageal speech: Is it a reliable tool? *Indian journal of cancer*, *53*(1), 127.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational psychology review*, *10*(3), 251-296.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632-1634.

Tavano, A., & Scharinger, M. (2015). Prediction in speech and language processing. Cortex; *a journal devoted to the study of the nervous system and behavior,* 68, 1.

Tremblay, K. L., Piskosz, M., & Souza, P. (2003). Effects of age and age-related hearing loss on the neural representation of speech cues. *Clinical Neurophysiology, 114*(7), 1332-1343.

Trudeau, M. D. (1987). A comparison of the speech acceptability of good and excellent esophageal and tracheoesophageal speakers. *Journal of communication disorders, 20*(1), 41-49.

Unsworth, N., & Robison, M. K. (2018). Tracking working memory maintenance with pupillometry. *Attention, Perception, & Psychophysics, 80*(2), 461-484.

Vidulich, M. A., & Pandit, P. (1986, September). Training and subjective workload in a category search task. In *Proceedings of the Human Factors Society Annual Meeting* (Vol. 30, No. 11, pp. 1133-1136). Sage CA: Los Angeles, CA: SAGE Publications.

Voiers, W. D. (1964). Perceptual bases of speaker identity. *The Journal of the Acoustical Society of America, 36*(6), 1065-1073.

Wang, Y., Naylor, G., Kramer, S. E., Zekveld, A. A., Wendt, D., Ohlenforst, B., & Lunner, T. (2018). Relations between self-reported daily-life fatigue, hearing status, and pupil dilation during a speech perception in noise task. *Ear and hearing, 39*(3), 573-582.

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in psychology, 7,* 345.

Wendt, D., Hietkamp, R. K., & Lunner, T. (2017). Impact of noise and noise reduction on processing effort: A pupillometry study. *Ear and hearing, 38*(6), 690-700.

Whitehill, T. L., & Wong, C. C. Y. (2006). Contributing factors to listener effort for dysarthric speech. *Journal of Medical Speech-Language Pathology, 14*(4), 335-342.

Wilson, D. K. (1987). Children's voice problems. Voice Problems of children, 3rd ed. Philadelphia: Williams & Wilkins, 1-15.

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in hearing, 22*, 2331216518800869.

Xie, B., & Salvendy, G. (2000). Prediction of mental workload in single and multiple tasks environments. *International journal of cognitive ergonomics, 4*(3), 213-242.

Yeung, J. C., Fung, K., Davis, E., Rai, S. K., Day, A., Dzioba, A., ... & Doyle, P. C. (2015). Longitudinal variations of laryngeal overpressure and voice-related quality of life in spasmodic dysphonia. *The Laryngoscope, 125*(3), 661-666.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and hearing, 31*(4), 480-490.

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology, 51*(3), 277-284.

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., & Glaze, L. E. (2011). Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V). *American Journal of Speech-Language Pathology, 20*(1), 14-22.

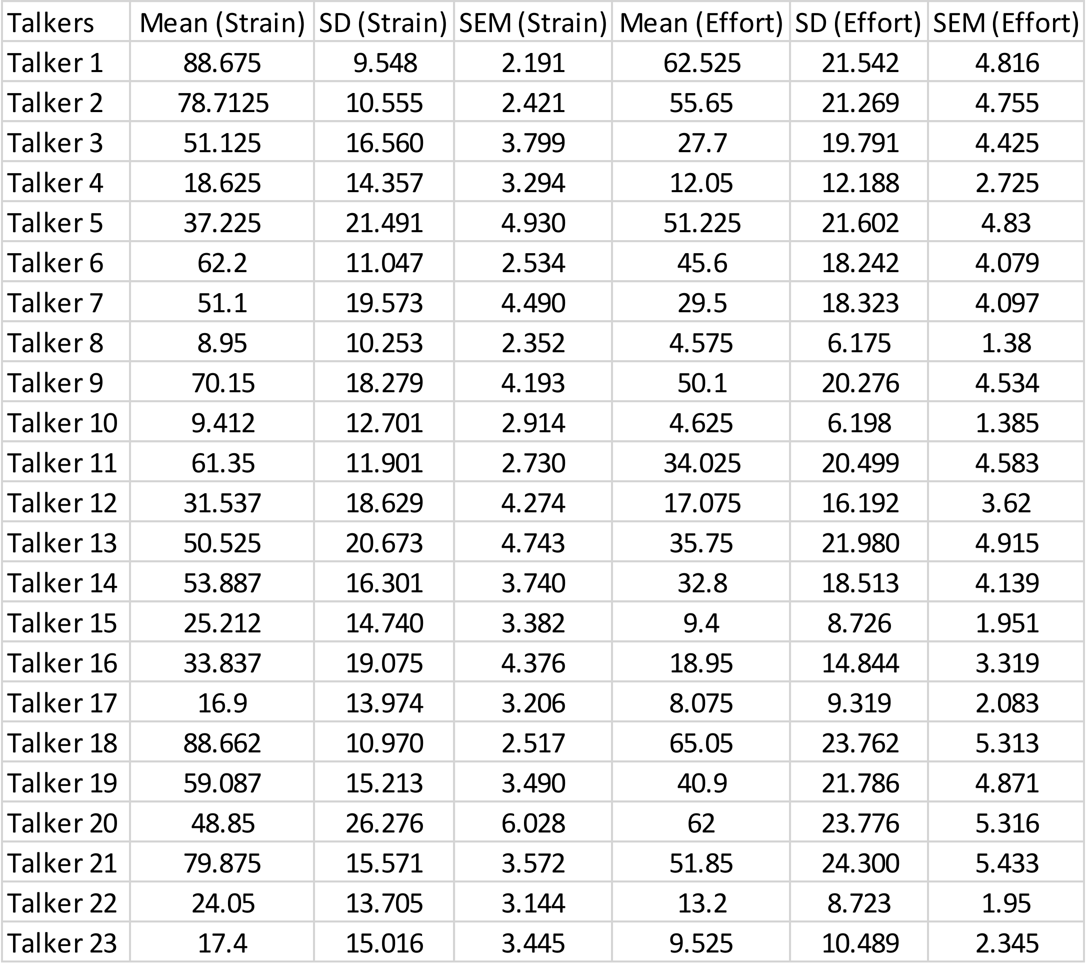
Appendices

Appendix A: Medications Causing Pupil Dilation

Generally, opioids cause pupil dilation, however, withdrawal syndrome and a severe overdose of opioids may cause pupil constriction.

- Dopaminergic agent: Levodopa, Levodopa-carbidopa, Levodopa- benserazide
- Anticholinergic: Ipratropium bromide, Tiotropiu bromide, Scopolamine, Benztropine, Atropine (overdose), Oxybutynin (overdose), Solifenacin (overdose)
- SSRI (in overdose): Citalopram, Escitalopram, Fluoxetine, Fluvoxamine, Paroxetine, Sertraline
- Antihistamine: Cetirizine, Doxylamine (overdose)
- Aminoglycoside antibacterial: amikacin, Gentamicin, Tobramycin
- Alpha-adrenergic agonist: Midodrine, phenylephrine
- Neuromuscular Paralytic Agent: OnabotulinumtoxinA
- CNS stimulant: Methylphenidate (overdose)
- SNRI antidepressant: Desvenlafaxine
- MAOI: Tranylcypromine
- Sympathomimetic: Pseudoephedrine (overdose)
- Opioid antagonist: Naloxegol, Methadone

Appendix B: ADSD, Naïve Listeners (Strain, Listening Effort), Descriptive Statistics Table



Talkers	Mean (Strain)	SD (Strain)	SEM (Strain)	Mean (Effort)	SD (Effort)	SEM (Effort)
Talker 1	88.675	9.548	2.191	62.525	21.542	4.816
Talker 2	78.7125	10.555	2.421	55.65	21.269	4.755
Talker 3	51.125	16.560	3.799	27.7	19.791	4.425
Talker 4	18.625	14.357	3.294	12.05	12.188	2.725
Talker 5	37.225	21.491	4.930	51.225	21.602	4.83
Talker 6	62.2	11.047	2.534	45.6	18.242	4.079
Talker 7	51.1	19.573	4.490	29.5	18.323	4.097
Talker 8	8.95	10.253	2.352	4.575	6.175	1.38
Talker 9	70.15	18.279	4.193	50.1	20.276	4.534
Talker 10	9.412	12.701	2.914	4.625	6.198	1.385
Talker 11	61.35	11.901	2.730	34.025	20.499	4.583
Talker 12	31.537	18.629	4.274	17.075	16.192	3.62
Talker 13	50.525	20.673	4.743	35.75	21.980	4.915
Talker 14	53.887	16.301	3.740	32.8	18.513	4.139
Talker 15	25.212	14.740	3.382	9.4	8.726	1.951
Talker 16	33.837	19.075	4.376	18.95	14.844	3.319
Talker 17	16.9	13.974	3.206	8.075	9.319	2.083
Talker 18	88.662	10.970	2.517	65.05	23.762	5.313
Talker 19	59.087	15.213	3.490	40.9	21.786	4.871
Talker 20	48.85	26.276	6.028	62	23.776	5.316
Talker 21	79.875	15.571	3.572	51.85	24.300	5.433
Talker 22	24.05	13.705	3.144	13.2	8.723	1.95
Talker 23	17.4	15.016	3.445	9.525	10.489	2.345

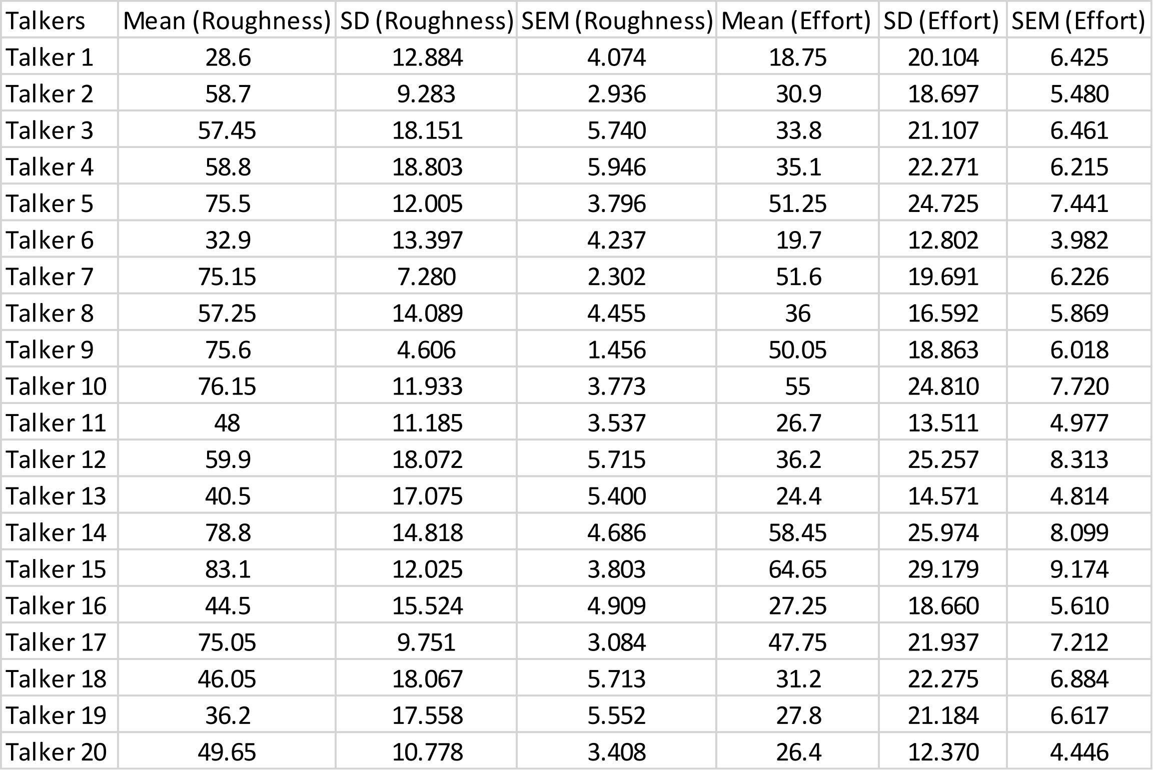
Appendix C: ADSD, Experienced Listeners (Strain, Listening Effort), Descriptive Statistics Table

Talkers	Mean (Strain)	SD (Strain)	SEM (Strain)	Mean (Effort)	SD (Effort)	SEM (Effort)
Talker 1	80.833	16.158	9.329	56.500	26.557	15.332
Talker 2	68.000	7.089	4.093	35.000	7.937	4.583
Talker 3	52.667	20.763	11.987	18.667	9.465	5.465
Talker 4	23.667	18.711	10.803	5.833	5.008	2.892
Talker 5	49.167	45.941	26.524	37.167	33.828	19.531
Talker 6	68.500	20.839	12.031	46.000	31.301	18.072
Talker 7	59.500	6.062	3.500	31.000	13.229	7.638
Talker 8	5.833	4.752	2.744	2.500	2.179	1.258
Talker 9	66.000	3.969	2.291	40.000	14.292	8.251
Talker 10	1.667	1.155	0.667	1.000	0.000	0.000
Talker 11	67.000	9.260	5.346	31.833	13.769	7.949
Talker 12	41.667	14.835	8.565	13.333	14.978	8.647
Talker 13	44.333	12.251	7.073	21.333	4.072	2.351
Talker 14	62.500	3.122	1.803	33.500	17.414	10.054
Talker 15	23.667	8.129	4.693	5.333	3.753	2.167
Talker 16	26.667	4.646	2.682	6.833	6.292	3.632
Talker 17	9.333	12.741	7.356	1.833	1.443	0.833
Talker 18	84.333	15.003	8.662	66.167	25.663	14.816
Talker 19	53.500	18.993	10.966	25.500	16.889	9.751
Talker 20	59.500	50.767	29.310	40.500	37.951	21.911
Talker 21	80.167	18.237	10.529	55.333	28.537	16.476
Talker 22	25.167	28.829	16.644	12.667	12.965	7.485
Talker 23	11.500	12.971	7.489	1.833	1.443	0.833

Appendix D: TE, With-Anchor Group (Roughness, Listening Effort), Descriptive Statistics Table

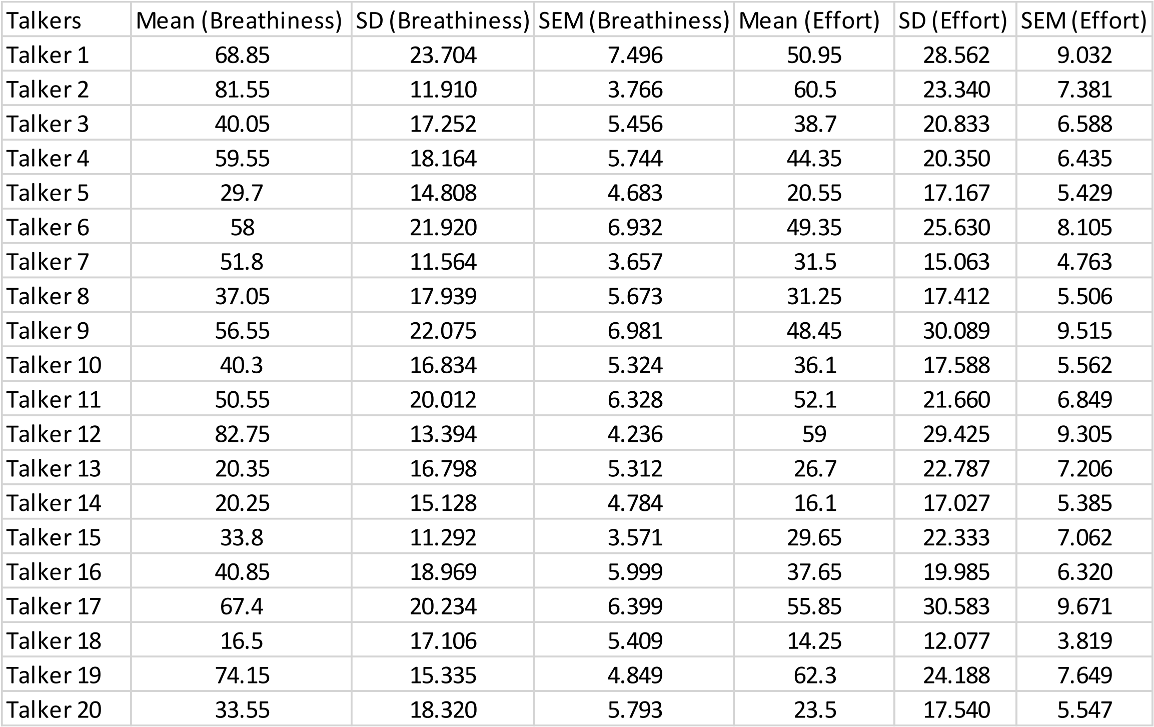
Talkers	Mean (Roughness)	SD (Roughness)	SEM (Roughness)	Mean (Effort)	SD (Effort)	SEM (Effort)
Talker 1	23	14.587	4.613	11.6	12.211	3.861
Talker 2	44.75	12.828	4.057	24.8	16.340	5.167
Talker 3	57.4	13.057	4.129	33.5	18.852	5.961
Talker 4	47.65	17.090	5.404	36.25	20.500	6.483
Talker 5	77.4	14.712	4.652	55	24.667	7.800
Talker 6	16.95	9.861	3.118	10.45	12.273	3.881
Talker 7	79.65	6.638	2.099	58.25	20.278	6.412
Talker 8	44.75	18.270	5.777	28.55	15.902	5.029
Talker 9	74.7	11.216	3.547	51.05	17.238	5.451
Talker 10	79.8	11.975	3.787	60.35	21.181	6.698
Talker 11	44.65	16.877	5.337	21.75	18.137	5.735
Talker 12	55.1	17.890	5.657	33.9	20.471	6.473
Talker 13	45.6	11.177	3.535	23.85	13.852	4.381
Talker 14	82.6	14.294	4.520	61.75	24.722	7.818
Talker 15	91.6	8.103	2.562	72.05	22.929	7.251
Talker 16	49.7	16.224	5.131	27.6	15.427	4.878
Talker 17	75.35	12.680	4.010	50.25	20.615	6.519
Talker 18	37.15	15.979	5.053	22.95	9.982	3.157
Talker 19	32	13.331	4.216	23.15	11.426	3.613
Talker 20	41.65	15.091	4.772	20	14.085	4.454

Appendix E: TE, No-Anchor Group (Roughness, Listening Effort), Descriptive Statistics Table



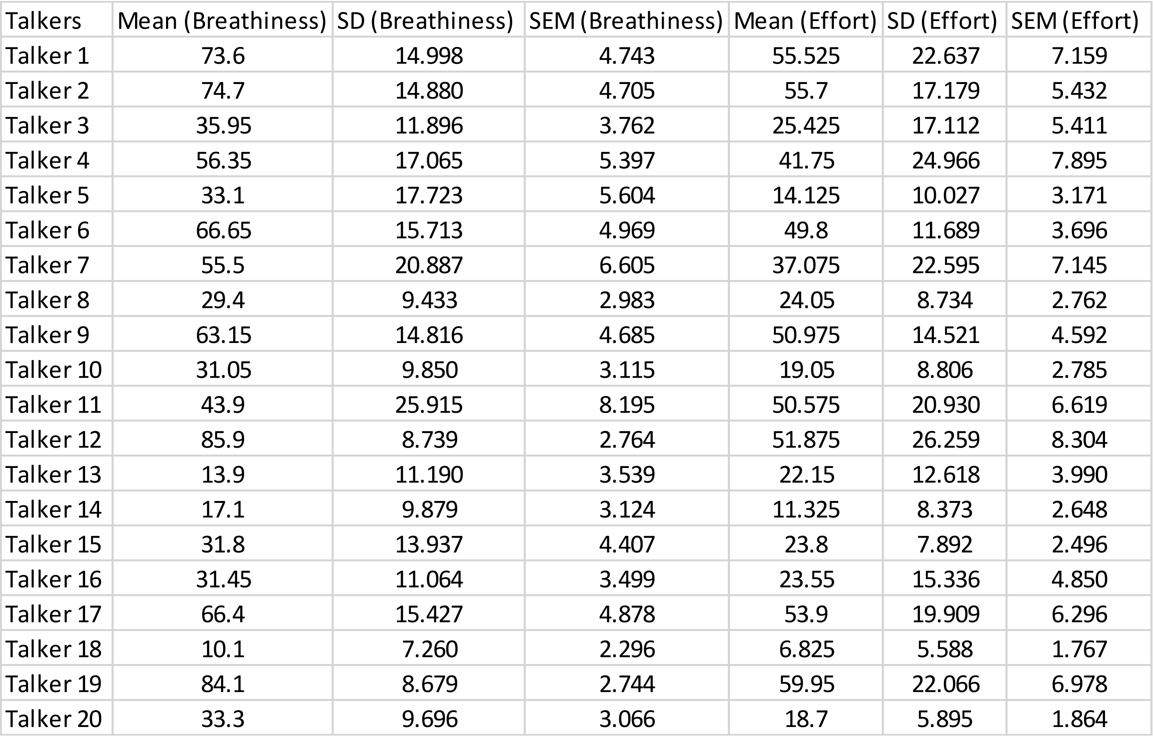
Talkers	Mean (Roughness)	SD (Roughness)	SEM (Roughness)	Mean (Effort)	SD (Effort)	SEM (Effort)
Talker 1	28.6	12.884	4.074	18.75	20.104	6.425
Talker 2	58.7	9.283	2.936	30.9	18.697	5.480
Talker 3	57.45	18.151	5.740	33.8	21.107	6.461
Talker 4	58.8	18.803	5.946	35.1	22.271	6.215
Talker 5	75.5	12.005	3.796	51.25	24.725	7.441
Talker 6	32.9	13.397	4.237	19.7	12.802	3.982
Talker 7	75.15	7.280	2.302	51.6	19.691	6.226
Talker 8	57.25	14.089	4.455	36	16.592	5.869
Talker 9	75.6	4.606	1.456	50.05	18.863	6.018
Talker 10	76.15	11.933	3.773	55	24.810	7.720
Talker 11	48	11.185	3.537	26.7	13.511	4.977
Talker 12	59.9	18.072	5.715	36.2	25.257	8.313
Talker 13	40.5	17.075	5.400	24.4	14.571	4.814
Talker 14	78.8	14.818	4.686	58.45	25.974	8.099
Talker 15	83.1	12.025	3.803	64.65	29.179	9.174
Talker 16	44.5	15.524	4.909	27.25	18.660	5.610
Talker 17	75.05	9.751	3.084	47.75	21.937	7.212
Talker 18	46.05	18.067	5.713	31.2	22.275	6.884
Talker 19	36.2	17.558	5.552	27.8	21.184	6.617
Talker 20	49.65	10.778	3.408	26.4	12.370	4.446

Appendix F: VFP, With-Anchor Group (Breathiness, Listening Effort), Descriptive Statistics Table



Talkers	Mean (Breathiness)	SD (Breathiness)	SEM (Breathiness)	Mean (Effort)	SD (Effort)	SEM (Effort)
Talker 1	68.85	23.704	7.496	50.95	28.562	9.032
Talker 2	81.55	11.910	3.766	60.5	23.340	7.381
Talker 3	40.05	17.252	5.456	38.7	20.833	6.588
Talker 4	59.55	18.164	5.744	44.35	20.350	6.435
Talker 5	29.7	14.808	4.683	20.55	17.167	5.429
Talker 6	58	21.920	6.932	49.35	25.630	8.105
Talker 7	51.8	11.564	3.657	31.5	15.063	4.763
Talker 8	37.05	17.939	5.673	31.25	17.412	5.506
Talker 9	56.55	22.075	6.981	48.45	30.089	9.515
Talker 10	40.3	16.834	5.324	36.1	17.588	5.562
Talker 11	50.55	20.012	6.328	52.1	21.660	6.849
Talker 12	82.75	13.394	4.236	59	29.425	9.305
Talker 13	20.35	16.798	5.312	26.7	22.787	7.206
Talker 14	20.25	15.128	4.784	16.1	17.027	5.385
Talker 15	33.8	11.292	3.571	29.65	22.333	7.062
Talker 16	40.85	18.969	5.999	37.65	19.985	6.320
Talker 17	67.4	20.234	6.399	55.85	30.583	9.671
Talker 18	16.5	17.106	5.409	14.25	12.077	3.819
Talker 19	74.15	15.335	4.849	62.3	24.188	7.649
Talker 20	33.55	18.320	5.793	23.5	17.540	5.547

Appendix G: VFP, No-Anchor Group (Breathiness, Listening Effort), Descriptive Statistics Table



Talkers	Mean (Breathiness)	SD (Breathiness)	SEM (Breathiness)	Mean (Effort)	SD (Effort)	SEM (Effort)
Talker 1	73.6	14.998	4.743	55.525	22.637	7.159
Talker 2	74.7	14.880	4.705	55.7	17.179	5.432
Talker 3	35.95	11.896	3.762	25.425	17.112	5.411
Talker 4	56.35	17.065	5.397	41.75	24.966	7.895
Talker 5	33.1	17.723	5.604	14.125	10.027	3.171
Talker 6	66.65	15.713	4.969	49.8	11.689	3.696
Talker 7	55.5	20.887	6.605	37.075	22.595	7.145
Talker 8	29.4	9.433	2.983	24.05	8.734	2.762
Talker 9	63.15	14.816	4.685	50.975	14.521	4.592
Talker 10	31.05	9.850	3.115	19.05	8.806	2.785
Talker 11	43.9	25.915	8.195	50.575	20.930	6.619
Talker 12	85.9	8.739	2.764	51.875	26.259	8.304
Talker 13	13.9	11.190	3.539	22.15	12.618	3.990
Talker 14	17.1	9.879	3.124	11.325	8.373	2.648
Talker 15	31.8	13.937	4.407	23.8	7.892	2.496
Talker 16	31.45	11.064	3.499	23.55	15.336	4.850
Talker 17	66.4	15.427	4.878	53.9	19.909	6.296
Talker 18	10.1	7.260	2.296	6.825	5.588	1.767
Talker 19	84.1	8.679	2.744	59.95	22.066	6.978
Talker 20	33.3	9.696	3.066	18.7	5.895	1.864

Appendix H: Ethics Approval Letter

Date: 20 February 2019

To: Dr. Vijay Parsa

Project ID: 112674

Study Title: Auditory-perceptual evaluation of dysphonic voices: A pupillometry study

Application Type: NMREB Initial Application

Review Type: Delegated

Full Board Reporting Date: March 1 2019

Date Approval Issued: 20/Feb/2019

REB Approval Expiry Date: 20/Feb/2020

Dear Dr. Vijay Parsa

The Western University Non-Medical Research Ethics Board (NMREB) has reviewed and approved the WREM application form for the above mentioned study, as of the date noted above. NMREB approval for this study remains valid until the expiry date noted above, conditional to timely submission and acceptance of NMREB Continuing Ethics Review.

This research study is to be conducted by the investigator noted above. All other required institutional approvals must also be obtained prior to the conduct of the study.

Documents Approved:

Document Name	Document Type	Document Date	Document Version
Letter of Information Experienced Resubmission Clean	Written Consent/Assent	13/Feb/2019	2
Letter of Information Naive Resubmission Clean	Written Consent/Assent	13/Feb/2019	3
Medications Causing Pupil Dilation	Written Consent/Assent	27/Nov/2018	
Recruitment Script Experienced Clean	Recruitment Materials	05/Feb/2019	1
Recruitment Script Resubmission Clean	Oral Script	22/Jan/2019	2
Updated Poster	Recruitment Materials	22/Jan/2019	
Visual Analog Rating Scale	Other Data Collection Instruments		

No deviations from, or changes to the protocol should be initiated without prior written approval from the NMREB, except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involves only administrative or logistical aspects of the trial.

The Western University NMREB operates in compliance with the Tri-Council Policy Statement Ethical Conduct for Research Involving Humans (TCPS2), the Ontario Personal Health Information Protection Act (PHIPA, 2004), and the applicable laws

and regulations of Ontario. Members of the NMREB who are named as Investigators in research studies do not participate in discussions related to, nor vote on such studies when they are presented to the REB. The NMREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000941.
Please do not hesitate to contact us if you have any questions.

Sincerely,

Kelly Patterson, Research Ethics Officer on behalf of Dr. Randal Graham, NMREB Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).

Curriculum Vitae

Name:	Mojgan Farahani
Post-secondary Education and Degrees:	University of Azad, Tehran South Branch Tehran, Iran, B.A. (Translation Studies)
	The University of Azad, Tehran Central Branch Tehran, Iran M.A. (Teaching English as a Foreign Language)
	The University of Western Ontario London, Ontario, Canada M.A. (Linguistics)
	The University of Western Ontario London, Ontario, Canada Ph.D. (Speech & Language Science)
Related Work Experience	Teaching Assistant The University of Western Ontario (MA & PhD)