August 2013

Cepstral- and Spectral-Based Acoustic Measures of Normal Voices

Rachel Garrett
University of Wisconsin-Milwaukee

CEPSTRAL- AND SPECTRAL- BASED ACOUSTIC MEASURES OF NORMAL VOICES

by

Rachel K. M. Garrett

A Thesis Submitted in

Partial Fulfillment of the

Requirements of the Degree of

Master of Science

in Communication Sciences and Disorders

at

The University of Wisconsin – Milwaukee

August 2013

ABSTRACT

CEPSTRAL- AND SPECTRAL- BASED ACOUSTIC MEASURES OF NORMAL VOICES
by

Rachel K. M. Garrett

The University of Wisconsin-Milwaukee, 2013
Under the Supervision of Dr. Marylou Pausewang Gelfer

A review of recent literature suggested that cepstral- and spectral-based acoustic measures showed good potential as objective measures of dysphonia for clinical application. However, the small numbers of normal subjects in previous research and wide age ranges prevent a good estimation of the performance of normal speakers of various ages on these measures. Therefore, the purpose of this study was to provide normative data for Long-Term Average spectral- and cepstral-based measures for both men and women in two different age groups to aid clinicians with assessing and treating voice disorders. Sixty participants consisting of fifteen males and fifteen females, ages 20-30 years, and fifteen males and fifteen females, ages 40-50 years contributed speech samples to be analyzed in this study. Speakers were asked to sustain the vowels /a/ and /i/, read out loud four CAPE-V stimulus sentences, and the 2nd and 3rd sentence of the Rainbow Passage. Dependent variables were Cepstral Peak Prominence (CPP), Low-to-High Spectral Ratio (L/H spectral ratio), and Cepstral Peak Prominence Fundamental Frequency (CPP F_0) for both vowels and connected speech. Male voice quality (CPP and L/H spectral ratio) was better in vowels /a/ and /i/, but female voice quality was better (CPP values) for connected speech. Age did not affect voice quality for vowels /a/ and /i/;

however, it did affect it for connected speech. Younger speakers had better voice quality (CPP) than older speakers. In general, for both vowels and connected speech, younger women had markedly higher CPP F_0 values than older women, while older men had slightly higher CPP F_0 values compared to younger men. It was concluded that separate normative data should be applied clinically for all four age/gender groups. The maximum limit of the ADSV extraction range for male participants should be changed from 300 Hz to 200 Hz for connected speech readings to obtain accurate CPP F_0 measures. Furthermore, due to limited research, data should be analyzed both with and without vocalic detection until it becomes clear which one is more valid. Further research is recommended to improve both the procedures and reference data available for voice quality.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Cepstral- and Spectral- Based Acoustic Measures of Normal Voices.

Introduction

One essential aspect of interpersonal communication is voice quality. The term "voice quality' is generally used to refer to the perceptual characteristics of a voice arising from the phonatory actions of the laryngeal system (Kent & Ball, 2000). There are many reasons why voice quality is important, both to listeners and speakers. For example, the quality of one's voice affects how a person is perceived. Listeners make judgments about an individual's health and personality based on how the voice sounds (Boone, 1991). For this reason, different types of individuals seek certain voice qualities. Professional singers want a smooth and confident voice, while lawyers aspire to a powerful, loud voice that resonates well.

The vocabulary used to describe voices allows professionals such as speech-language pathologists to verbally articulate the percept of the voice of a person with a communication disorder. The many adjectives used by speech-language pathologists to describe the sound of the voice go beyond descriptions of loudness and pitch, commonly acknowledged attributes of a person's speech (Behrman, 2007). Examples of pathological voice quality descriptors include *rough*, *hoarse*, *breathy*, and *strained*, to name a few. An accurate description of a client's voice quality, along with their case history, are believed to aid in the differential diagnosis process (see for example Darley, Aronson & Brown, 1975).

Unfortunately, it can be argued that perceptual descriptors are subjective in nature, and can be influenced by personal experience, preference and culture (Colton, Casper, & Leonard, 2011). Acoustic instrumental measurements have been proposed as potential objective correlates for perceptual judgments of voice quality (Colton et al., 2011). However, these acoustic analysis methods require further exploration and study in order to determine their usefulness, and how they might be applied in diagnosing and treating voice disorders.

Anatomy and Physiology of the Vocal Mechanism

In order to understand the origins of voice quality, it is important to have knowledge of the laryngeal mechanism. The act of phonation involves the musculo-cartilaginous structure that is the larynx, and results in the physiological process of vocal fold vibration that produces sound. The larynx is positioned in the anterior portion of the neck at cervical vertebrae four, five, and six. It extends from its superior boundary at the root of the tongue inferiorly to the first tracheal ring. When all the components of the larynx are healthy and functioning properly, normal voicing is generated (Boone, McFarlane, Von Berg, & Zraick, 2010).

The intrinsic muscles of the larynx have their insertions and origins within the laryngeal structures. Together, these muscles aid in the process of phonation by adducting (bringing together) the vocal folds to create sufficient subglottic pressure necessary for voicing, abducting (drawing apart) the vocal folds to cease the phonation process, and regulating the length and tension of the vocal folds during voice production. The complexities of these actions are exemplified by an antagonist and agonist relationship (the opposing muscles must relax to permit the acting muscle to complete its contraction)

among the sets of muscles. The intrinsic muscles are innervated by the recurrent laryngeal branch of the vagus (X) and the spinal accessory (XI) nerve, except the cricothyroid, which is innervated by the superior laryngeal branch of the vagus and the spinal accessory nerve (Duffy, 2005). Muscles that are directly associated with adduction and abduction of the vocal folds include the interarytenoids, the lateral cricoarytenoids, and posterior cricoarytenoids.

The interarytenoid muscles are partially responsible for medial compression of the vocal folds (Colton et al., 2011). An unpaired muscle, the interarytenoid can be found between the two pyramid-shaped arytenoid cartilages. The interarytenoid or arytenoideus muscle is comprised of fibers that course in two different directions, transverse and oblique. The transverse portion of the interarytenoid originates at the lateral margin of the posterior arytenoid cartilage. With a lateral course, the muscle inserts at the lateral margin of the posterior surface of the opposite arytenoid cartilage (Seikel, King, & Drumright, 2010). The origin of the oblique portion is the posterior base of the muscular process of the arytenoid cartilage, superficial to the transverse portion. Coursing upward at an angle, the oblique arytenoid muscle inserts at the opposite arytenoid cartilage's apex (Seikel et al., 2010). By contracting the transverse and oblique fibers of the interarytenoid, the muscle moves the arytenoid cartilages, and by association the vocal folds, in a medial direction, thus closing posterior aspect of the glottis (the cartilaginous glottis).

With the interarytenoids, the lateral cricoarytenoid muscles complete the task of bringing the vocal folds to midline and compressing them. The lateral cricoarytenoid muscle is a paired muscle, with its origin at the superior-lateral surface of the cricoid

cartilage. Coursing superiorly and posteriorly, the lateral cricoarytenoid muscle inserts at the arytenoid cartilage's muscular process. These origin and insertion sites result in a medial tipping or tilting of the vocal processes of the arytenoid cartilages when the lateral cricoarytenoid muscle is contracted, thus causing the true vocal folds to adduct and the membranous glottis to close. (Colton et al., 2011)

In order to bring the act of phonation to an end, the posterior cricoarytenoid muscle must abduct the vocal folds. The origin of the posterior cricoarytenoid muscle is the posterior lamina of the cricoid cartilage. From the posterior cricoid lamina, the muscles course upward at an outward angle for an attachment on the posterior surface of muscular process of the arytenoid cartilage. Contraction of the posterior cricoarytenoids causes the gliding and tilting of the vocal process of the arytenoids in a lateral direction, therefore resulting in the opening of the glottis, or abduction of the vocal folds (Colton et al., 2011; Behrman, 2007). Of course, this assumes that the interarytenoid and lateral cricoarytenoid muscles just discussed have relaxed.

The length and tension of the vocal folds are regulated by other intrinsic muscles of the larynx, specifically the cricothyroid and thyroarytenoid muscles. The cricothyroid muscle regulates gross tension of the vocal folds while the thyroarytenoid muscle controls the 'fine-tuning' of vocal fold tension (Zemlin, 1998; Hixon, Weismer & Hoit, 2008). The cricothyroid muscle is comprised of two different heads: the pars recta and the pars oblique. The pars recta's origin is on the external anterior surface of the cricoid cartilage superior to the cricoid cartilage arch. Pars recta, the medial-most portion of the cricothyroid, courses upward and outward to insert at the inferior surface of the thyroid lamina. The origin of the pars oblique can be found at the cricoid cartilage, lateral to the

pars recta. The direction of the pars oblique's fibers is obliquely upward with an insertion between the laminae and inferior horns of the thyroid cartilage (Seikel et al., 2010). When the cricothyroid muscle contracts, it decreases the distance between the inferior border of the thyroid and the superior margin of the cricoid cartilage (Hixon et al, 2008). As a result, the distance between the thyroid cartilage anteriorly and the arytenoid cartilages posteriorly increases. This elongation of the vocal folds decreases their mass and increases vocal fold tension. (Colton et al., 2011)

The thyroarytenoid muscle is a controversial structure. Some believe it includes two separate portions: the thyrovocalis and the thyromuscularis muscles (Hixon et al., 2008). However, according to Hixon et al. (2008), this division of the thyroarytenoid is not universally accepted. Some argue that dissections of the thyroarytenoid muscle have not shown separating fascial sheaths that would distinguish the thyrovocalis from the thyromuscularis (Hixon et al. 2008; Zemlin, 1998). But more recent research, as cited by Hixon et al. (2008), has suggested that there are histological differences between the thyrovocalis and thyromuscularis portions of the thyroarytenoid muscle, and that the differences in cellular structure and function support the concept of their differential actions during pitch change. Therefore, this paper will adhere to the theory that the thyrovocalis and thyromuscularis are two functionally distinct parts of the thyroarytenoid muscle, and are able to contract independently to effect pitch.

The lateral portion of the thyroarytenoid is considered to be the thyromuscularis muscle. The origin of the thyromuscularis is the inner surface of the thyroid cartilage, in close proximity to the thyroid cartilage notch. With a posterior course, the thyromuscularis' insertion attachments are the muscular process and also the base of the

arytenoid cartilages. The contraction of the thyromuscularis results in the relaxation of the vocal folds (Seikel et al., 2010). This is due to the muscle's drawing the arytenoids anteriorly, which results in the decreasing tension on the vocal ligament and the thyrovocalis muscles (Behrman, 2007). This assumes that the cricothyroid muscle is in a steady state of resistance, and has not elongated or shortened.

The medial thyroarytenoid, thyrovocalis, is also considered to originate on the inner surface of the thyroid cartilage. More specifically, the thyrovocalis' origin is near the notch of the thyroid cartilage, near the origin of thyromuscularis. With a posterior course, the thyrovocalis inserts on the lateral surface of the vocal process of the arytenoids. When the thyrovocalis muscle contracts in concert with the cricothyroid muscle, tension of the vocal folds increases both longitudinally and medially. Contraction of the thyrovocalis in isolation (with the cricothyroid muscle not active) probably aids the thyromuscularis with shortening and loosening the innermost part of the vocal folds or the vocal ligament (Seikel et al., 2010).

The above description of laryngeal anatomy provides a foundation to understand the complex relationship of the laryngeal structures and how they function together to produce sound. The most basic theory of voice production is the myoelastic-aerodynamic theory (Zemlin, 1998; van den Berg, 1958). This theory states that the act of sustained vocal fold vibration is dependent upon the elastic characteristics of the muscles and soft tissues of the vocal folds, and the airflow and pressure that pass between them as they protrude into the airway (Seikel et al., 2010). When healthy elastic vocal folds are approximated in the presence of continual airflow and pressure, voicing will occur.

Bringing together or adducting the vocal folds is the first step of accomplishing phonation. This action is made possible through the contraction of both the interarytenoids and lateral cricoarytenoid muscles (Behrman, 2007), while the posterior cricoarytenoid, a laryngeal abductor, is relaxed. During the entire vibratory cycle, the vocal folds are held in close proximity by the interarytenoid and lateral cricoarytenoid muscles, until the posterior cricoarytenoid muscle contracts and abducts the vocal folds at the conclusion of the phonation. At the initiation of phonation, the interarytenoid muscles is the first to contract at 0.5 to 0.3 seconds before sound is produced, followed by the activation of the lateral cricoarytenoid muscles 0.1 seconds after the interarytenoid (Colton et al., 2011). In order to produce gentle vocal onset, which is the least damaging type of phonation initiation, the vocal folds must rest in the adducted position as airflow between them is initiated. Subglottic pressure begins to build up as the speaker begins to exhale. Once pressure below the glottis is powerful enough to surmount the resistance from the vocal folds (a minimum of 3-5 cm of H_20), the vocal folds are forced open (Boone et al., 2010; Seikel et al., 2010). The vocal folds open in an inferior to superior direction. As the vocal folds are forced open at the inferior margins via subglottal pressure, the superior margins are passively dragged apart due to tissue elasticity as well as subglottic pressure. (Behrman, 2007; Boone et al., 2010)

When the vocal folds are blown apart in an outward direction, the tissue is stretched. The vocal folds start to recoil and return to midline, their resting state, due to the tissue-restoring forces (tissue elasticity; Seikel et al., 2010). The final closing of the vocal folds to aid their return to the resting state occurs because of the Bernoulli effect. The Bernoulli effect states that given a constant volume flow of air or a fluid in a tube,

there will be an increase in the velocity of flow at a point of constriction in the tube, along with a decrease in pressure perpendicular to the flow (Seikel et al., 2010). Once a critical distance between the vocal folds is reached, they create a sufficient constriction within the vocal tract to cause the air from the lungs to increase in its velocity as the air particles deviate around the constriction (Zemlin, 1998). Due to the increase of velocity, a drop of pressure perpendicular to the vocal folds occurs across the medial surface of the vocal folds. The perpendicular pressure reduction against the medial surface of the vocal folds creates suction between the vocal folds, causing closure of the glottis that travels in the direction of the airflow: inferior to superior (Behrman, 2007).

The process of subglottic pressure buildup beneath the adducted vocal folds begins again, repeating the myoelastic aerodynamic process until the speaker ceases phonation. Vocal fold vibration will come to an end when the medial compressor muscles (interarytenoids and lateral cricoarytenoids) relax and the posterior cricoarytenoid muscles contract to open the vocal folds (Seikel et al., 2010), or if airflow ceases. As stated previously, the action of the posterior cricoarytenoid muscles, when contracted, is rotation of the vocal process of the arytenoid cartilages in a lateral direction. This movement of the arytenoids causes the true vocal folds to abduct and open the glottis (Colton et al., 2011; Behrman, 2007).

During sustained vocal fold phonation, the pitch and loudness of the voice can be increased and decreased via differential contractional forces of the intrinsic laryngeal muscles. A two-part adjustment of cricothyroid and thyroarytenoid muscle stiffness is generally considered necessary to increase and decrease pitch. When the cricothyroid muscle contracts, it applies external stretching forces to the vocal folds, increasing their

length and raising pitch (Colton et al., 2011). Further adjustment of pitch at that particular vocal fold length is accomplished by the contraction of the thyroarytenoid muscle, to properly adjust the stiffness of the vocal folds to reach the desired frequency. This internal contractile force slightly increases or decreases during speech, depending on whether the thyrovocalis portion, thyromuscularis portion, or both, contract (Hixon et al., 2008). When the stiffness of the vocal folds is maximized at a given length, the thyrovocalis relaxes and the cricothyroid muscle contracts again, stretching the vocal folds further to increase pitch. Again, at the new length, thyrovocalis contracts or relaxes to adjust the desired higher pitch (Colton et al., 2011). This relationship of internal (thyrovocalis) and external (cricothyroid) forces is referred to as a stair-step adjustment of pitch (Hixon et al., 2008).

A louder voice, or higher intensity, is achieved by increasing subglottic pressure in conjunction with increased tension in the muscles of medial compression (Colton et al., 2011). Subglottic pressure must increase in order to overcome the increased laryngeal resistance. Increased laryngeal resistance will result in a longer closed phase and shorter opening phase during vocal fold vibration, creating an increase in amplitude of vibration (Zemlin, 1998). Healthier intensity increase comes from increased subglottic airflow rather than increased vocal fold tension, as the former prevents muscle fatigue and abrasions on the vocal fold margin. If a speaker wants to decrease the volume of his/her voice, the vocal folds must decrease their medial compression, which will increase the relative airflow due to the absence of laryngeal resistance (Colton et al, 2011). The vibratory cycle will have a shorter or absent closed phase and somewhat more balanced

opening and closing phases, resulting in decreased subglottic pressure, amplitude of vibration, and vocal loudness (Zemlin, 1998).

Characteristics of Normal Vocal Fold Vibration

To maintain continuous vocal fold vibration, three forces are necessary: laryngeal resistance, airflow, and subglottic pressure. The laryngeal resistance, created by contraction of the interarytenoid and lateral cricoarytenoid muscles as well as the tension in the thyroarytenoids (R_L), must equal the subglottic pressure (P_{SG}) divided by the airflow (U). This balance of forces can be written as the following equation: $R_L=P_{SG}/U$ (Zemlin, 1998). The minimum amount of subglottic pressure (P_{SG}) needed to initiate vocal fold vibration is *phonation threshold pressure* (Colton et al., 2011). As stated earlier, a minimum of 3-5 cm of H_20 is needed to reach the phonation threshold pressure and therefore force the vocal folds apart. If balance of the vibratory forces is maintained, phonation occurs with little effort, vocal fold oscillation is self-sustaining, and phonation threshold pressure is minimal (Colton et al., 2011). Inefficient airflow, excessive laryngeal resistance, and/or failure to achieve phonation threshold pressure cause vibratory pattern disintegration (Zemlin, 1998).

The quality of sustained phonation is contingent on the integrity of the vocal folds' body and cover. The vocal folds are comprised of five layers: an epidermal cell layer, a superficial layer of the lamina propria, an intermediate layer of the lamina propria, a deep layer of the lamina propria, and the thyroarytenoid muscle (Hirano, 1974). The first three layers make up the cover of the vocal folds while the last two form the vocal folds' body. Ciliated, columnar epithelium covers the lamina propria and protects the vocal folds from collision forces and friction during vocal fold vibration. Next comes

the lamina propria, a connective tissue. The superficial layer of the lamina propria is comprised of loose elastin fibers that can be easily stretched, while the intermediate layer of the vocal folds' cover is composed of densely arranged elastin fibers. The deep layer of lamina propria, also the first layer of the body of the vocal folds, is thick and "cotton"-like, consisting of collagen fibers. Lastly, the thyroarytenoid muscle makes up the bulk of the body of the vocal folds (Behrman, 2007). The thyroarytenoid muscle is the only part of the vocal folds that can contract, which results in the connective tissue layers being either "bunched" or stretched, depending on the vocal fold length.

The degree of coupling, or connection between the body and cover of the vocal folds, changes as pitch increases and decreases. When the vocal folds are short in length and vibrate at a lower frequency, the body and cover of the vocal folds are loosely coupled. That is, the connective tissue cover is "bunched" on the top of the contracted thyroarytenoid muscle, and tends to have its own vibratory pattern superimposed on the vibrating thyroarytenoid. This loose coupling of body and cover creates a visible mucosal wave during vibration. Ideally, a mucosal wave ripples smoothly across the medial to lateral dimension of the vocal fold. Pathology, scarring, or uneven stiffness can cause disturbances of the mucosal wave. A tightly coupled body and cover occur at higher frequencies, when both structures are stretched by the external force of the cricothyroid (Colton et al., 2011). At high frequencies, the cover and body vibrate in synchrony and the mucosal wave is less present (Ferrand, 2012).

Healthy vocal folds will also have a layer of mucus (a mucosereous blanket) across their surfaces. This substance is secreted by mucus glands below the vocal folds and within the vestibule (Behrman, 2007). This cover of mucus ensures moist and

lubricated vocal folds necessary for vocal fold vibration. Adequate hydration is needed to maintain sufficient mucus production (Gary, 2000). If the body, cover, and mucus secretions of the vocal folds are affected due to dehydration, changes to the tissue, or lesions, then vocal fold vibratory patterns will be altered (Colton et al, 2011).

Deviant Laryngeal Qualities

If proper vocal fold vibration is disturbed, deviant laryngeal quality can result. Multiple factors can disrupt the appropriate vibration of the vocal folds, including uneven weighting of the vocal folds, too much vocal fold compression, inadequate or incomplete vocal fold closure, or insufficient airflow (Ferrand, 2012; Zemlin, 1998). Such physiological disturbances have perceptually deviant laryngeal quality correlates: *rough, breathy, hoarse, aphonic breaks,* and *strained.* Each laryngeal quality descriptor, at least theoretically, is based on disruptions of normal vocal fold vibration.

A rough deviant laryngeal quality has noise elements that are perceived as crackling or popping present in the voice (Ferrand, 2012). The physiological bases for roughness in a voice are believed to be uneven weighting of the vocal folds and/or excessive tension. Aperiodic or irregular vibrations of the vocal folds occur when they are unevenly weighed. Uneven weighting of the vocal folds can be attributed to several factors, for example swelling of the vocal folds due to phonotrauma and/or organic factors, weight-increasing lesion(s) on one or both vocal folds, or muscle atrophy of one vocal fold (Ferrand, 2007). Any of these conditions may cause aperiodic vocal fold vibration and interference with the mucosal wave, which in turn causes the perception of a rough voice (Ferrand, 2007).

Rough laryngeal quality can also be caused by increased tension in the muscles of medial compression. When there is too much medial compression or force to close, the vocal folds can vibrate aperiodically, and the interarytenoid and lateral cricoarytenoid muscles eventually fatigue. Additionally, the muscles of respiration have to increase their effort in order to create enough subglottic pressure to overcome the increased laryngeal tension. If laryngeal resistance is too great, the balance of vibratory forces (R_L, P_{SG} & U) is disturbed, and voice quality can sound rough, forced and strained. (Colton et al., 2011)

Significant air leakage is the salient feature of a breathy voice. If medial compression occurs with normal force, deviant vocal quality can still occur if the vocal folds are unable to fully close (Ferrand, 2007). If an individual develops a lesion on the margin of one vocal fold (e.g., a polyp), the vocal folds are unable to fully approximate for phonation, resulting in excess air loss (Zemlin, 1998). Another cause of incomplete vocal fold closure is a posterior gap between the vocal folds, or glottal chink. This condition occurs when the lateral cricoarytenoid muscles contract but the interarytenoid muscles do not contract enough or at all, thus allowing the cartilaginous glottis to remain open. The lateral cricoarytenoid muscles may exert extra force to compensate for the loss of air created by the interarytenoid muscles' failing to contract, resulting in a forced, tense breathy quality. Finally, the presence of lesions on both vocal folds that articulate with each other can also create a posterior and/or anterior chink, as the vocal folds close at the point where the lesions meet, with open space anterior and posterior to the meeting point of lesions. Both anterior and posterior glottal chinks lead to excess air escapage during the process of phonation, and the perception of a breathy voice (Zemlin, 1998).

Hoarseness of the voice arises from a combination of rough and breathy deviant laryngeal qualities (Ferrand, 2012). Either quality can be the predominant percept of a hoarse voice. Hoarseness may occur with lesions of the margins of the vocal folds (Zemlin, 1998). This pathology prevents complete closure of the vocal folds, resulting in air leakage (the "breathy" component). Furthermore, the vocal folds are likely to be unevenly weighted because of the lesion(s), which consequently causes aperiodic vibrations (the "rough" component). Occasionally, hoarseness is accompanied by a wet, gurgly sound that results from excess mucus on the vocal folds or in the pyriform sinuses (Ferrand, 2012). Furthermore, compensating for a breathiness voice quality by increasing tension can lead to a hoarse sounding voice.

Complete cessation of voice for a short duration of time, when only a whisper is produced, is described as an *aphonic break*, or voice arrest (Colton et al., 2011). Vocal arrests occur when a severe imbalance of the vibratory forces (R_L, P_{SG} & U) prohibits the vocal folds from sustaining sufficient vibration. This severe imbalance may be due to a weight-increasing lesion on one or both vocal folds, inadequate vocal fold tension (Colton et al., 2011), or inadequate airflow.

Extreme tension in the voice with occasional stoppages are the characteristics of a *strained* quality (Ferrand, 2012). If the muscles of medial compression exert too much contractional force (increased R_L), the subglottic pressure needs to increase in relation to airflow in order force the vocal folds apart, resulting in a forced, strained vocal quality (Zemlin, 1998). This excessive tone in the muscles of medical compression is most often due to a neurological disorder.

Voice Quality Measurement

In order to measure vocal qualities, we first need to identify aspects of an aberrant voice that are amenable to quantitative assessment. For example, if roughness is accepted as a result of aperiodic vocal fold vibration, then roughness might be measured acoustically as cycle-to-cycle differences of vocal fold vibration. When great enough, these cycle-to-cycle variations would presumably generate random, aperiodic noise energy in the voice and alter perceived vocal quality (Colton et al., 2011).

Another origin of noise generation is at the level of or near the vocal folds, which can arise from air rushing through the glottis and against the vocal fold margin (Colton et al., 2011). This additional noise creates inharmonic partials, which could be measured on a frequency-by-amplitude spectrum (Baken & Orlikoff, 2000).

Based on the concepts of cycle-to-cycle differences in vocal fold vibrations and inharmonic partials, two general analysis approaches have been developed to measure the noise components in voice: perturbation measures and noise measures. Perturbation measures include cycle-to-cycle differences in frequency (*jitter*) and cycle-to-cycle differences in amplitude (*shimmer*; Behrman, 2007). Different algorithms for calculating both jitter and shimmer have been established in order to make these measures more robust to frequency and speaker artifacts. For example, *Jitter Ratio* is the mean perturbation in milliseconds divided by mean period and multiplied by 100, which attempts to normalize for the speaker's fundamental frequency of the production (F_0; Baken & Orlikoff, 2000). The frequency-based counterpart of jitter ratio, *Jitter Factor*, is the mean difference between the frequencies of adjacent cycles divided by mean fundamental frequency, multiplied by 100 (Baken & Orlikoff, 2000). Speakers have a

general tendency to increase frequency over the duration of a phonation; therefore *Relative Average Perturbations* (RAP) measures how much period-to-period difference exists if period durations are smoothed over three adjacent cycles (Baken & Orlikoff, 2000). An additional frequency perturbation measurement, *Pitch Period Perturbation Quotient* (PPQ), also relatively evaluates the period-to-period variability of pitch but with a higher smoothing factor than RAP (smoothing over five adjacent cycles; KayPENTAX, 2008). The higher smoothing factor leaves PPQ less sensitive to period-to-period variations; however it is believed to be more effective documenting pitch instability over the duration of a prolonged vowel than unsmoothed measures (KayPENTAX, 2008).

Cycle-to-cycle differences in amplitude are typically measured in either decibels (dB) or percent. *Shimmer in dB* is the period-to-period variability of peak-to-peak amplitude based on the dB ratio scale of amplitudes, making Shimmer in dB independent of absolute amplitude (Baken & Orlikoff, 2000). Measured as a percent of the amplitude of the total wave rather than in dB, *Shimmer Percent* is highly sensitive to amplitude variations, but pitch extraction errors may affect Shimmer Percent greatly (KayPENTAX, 2008). Similar to PPQ, *Amplitude Perturbation Quotient* (APQ) measures cycle-to-cycle differences in the context of a smoothing function of five cycles (KayPENTAX, 2008). It is normal to have some irregularity in the cycles of vocal fold vibration, but it is assumed that a noisy voice will have greater variations (Zemlin, 1998).

Elements of noise are naturally found in between the harmonics of the laryngeal spectrum; however, when the amplitude of the noise elements approaches the amplitude of the harmonic elements of a voice sample, they can obscure the harmonics, resulting in a perceived distortion of pitch and periodicity. Such noise can also be evaluated via time-

based measures but in the spectral domain. This type of spectral measurement includes *Voice Turbulence Index* (VTI) and *Noise to Harmonic Ratio* (NHR). VTI measures the relative energy level of high-frequency noise, presumably generated during incomplete closure of the vocal folds, compared to the energy level in the low-frequency harmonic components of a voice (KayPENTAX, 2008). NHR measures the amount of energy in the noise elements in a lower frequency range compared to VTI, divided by the amount of energy in the low-frequency harmonic components of the voice, with increased values reflecting increased spectral noise (Baken & Orlikoff, 2000). This analysis occurs for each pitch period, with results averaged over pitch periods, to obtain the NHR (or VTI) of the entire signal. Occasionally, NHR is expressed as *harmonic to noise ratio* (HNR), with increased values interpreted as decreased spectral noise. These forms of quantitative assessments are purported to measure aperiodic and inharmonic partials in a voice signal (Behrman, 2007).

One weakness of both perturbation measures and the noise measures described thus far is that all of them are dependent on an initial time-based analysis to separate a voiced signal into discrete pitch periods. Even in a prolonged vowel, analysis errors in accurately determining the length of each pitch period can occur occasionally or even frequently if the voiced signal has significant noise elements in it, disrupting periodicity. The need for an unchanging pitch makes perturbation measures inappropriate for use with connected speech samples, where frequency is constantly changing. Noise measures require a consistent vocal tract posture (so that harmonics can be determined), as well as an unchanging SFF. Thus, in order to get more accurate measures of noise in a voice, an analysis approach that is *not* time-dependent would be preferable.

Long-term spectral analysis has the potential to measure the sound spectrum of an entire moderately long speech sample. In this type of analysis, a *discrete Fourier transform* (DFT) is used to calculate a long-term averaged spectrum integrated over a whole phonated sample, and converts the speech signal from the time domain to the frequency domain (see Fig. 1-1; Baken & Orlikoff, 2000). DFT is a log power spectrum that presents energy at harmonically related frequencies by separating out the frequency and amplitude components of a complex time-by-amplitude wave (Hillenbrand & Houde, 1996). Using DFT as a foundation, *cepstral analysis* is a DFT of a DFT. Cepstral analysis is a magnitude-by-"quefrency" (time) spectrum, measured in decibels and milliseconds respectively. By creating a log power spectrum of a log power spectrum, cepstral analysis can show a well-defined harmonic structure with a strong fundamental frequency component and reduced noise in both sustained vowels and continuous speech samples produced by a normal speaker (see Figure 1-2; Hillenbrand & Houde, 1996). A linear regression line is then computed of the relationship between quefrency and cepstral magnitude in order to normalize the overall amplitude of the signal (Hillenbrand & Houde, 1996). Periodic signals are associated with more prominent or high-amplitude cepstral peaks compared to the regression line, while an aperiodic signal results in a decrease in amplitude of the cepstral peak in relation to the regression line (Awan, 2011). Thus, the amount of noise or energy in a connected speech signal can be quantified by measuring the distance between the most prominent cepstral peak and the regression line. This measure is called *cepstral peak prominence*, or CPP (see Figure 1-2).

Figure 1-1: Long-term averaged spectrum converted from time domain to frequency domain via discrete Fourier transform (DFT).

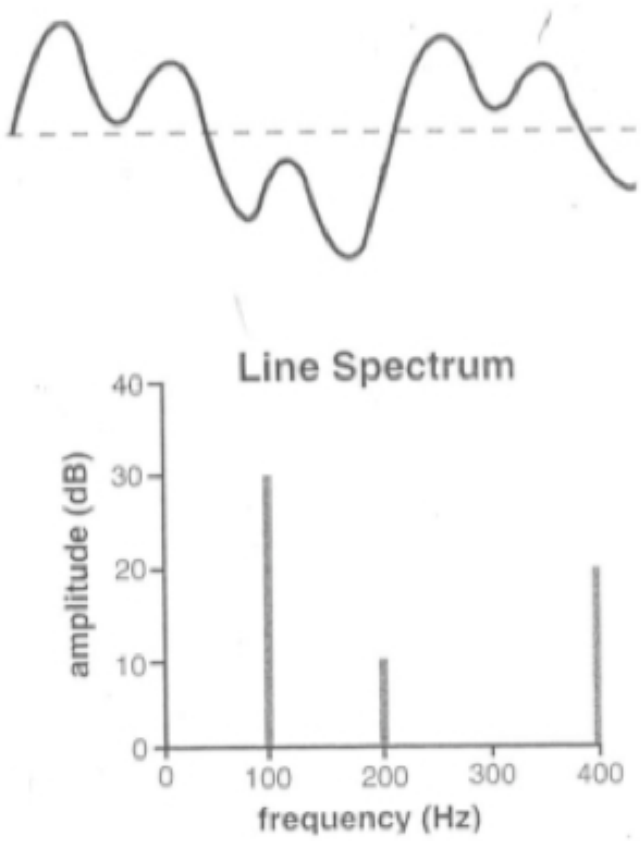
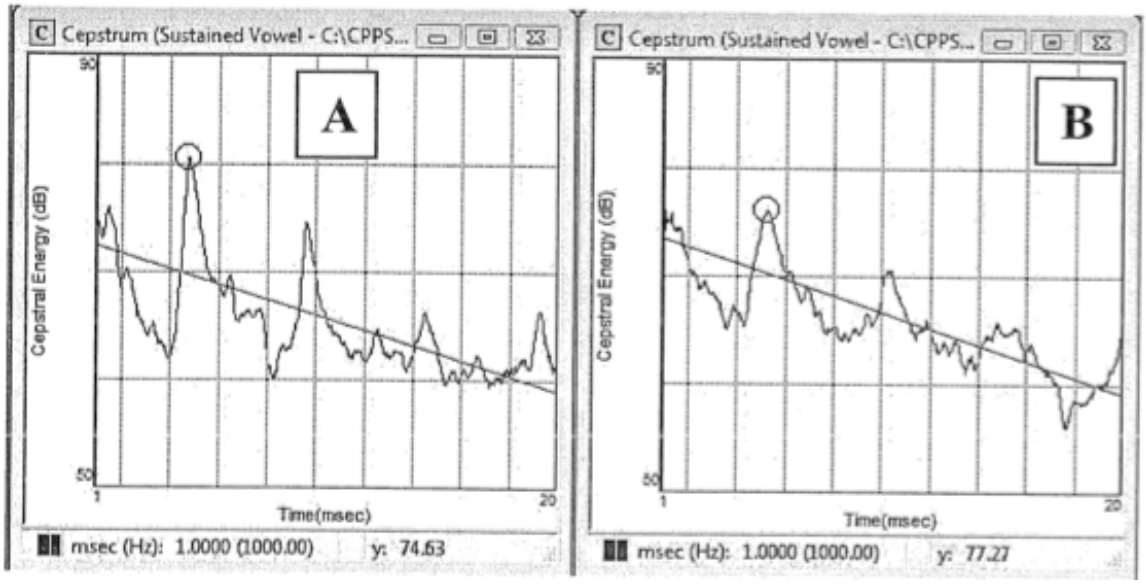


Figure 1-2: Normal female cepstrum (A) vs. a moderately breathy female cepstrum (B).



In addition to CPP, *low-versus high-spectral ratio* (L/H Ratio) has been shown to be useful in predicting perceived dysphonia severity in connected speech (Hillenbrand & Houde, 1996; Awan, Roy & Dromey, 2009; Watts & Awan, 2011). A ratio of low to high (L/H) spectral energy is calculated by comparing the average energy in the entire speech signal below 4 kilohertz (kHz) to the average energy above 4 kHz in a long-term spectral analysis (Hillenbrand & Houde, 1996). The L/H spectral ratio for normal voices tends to be increased, with more energy in the low frequencies, where the SFF and its harmonics are strongest, whereas deviant voice qualities tend to have a decreased L/H ratio, with more energy in the high-frequency noise range (Awan, 2011). Overall, cepstral and long-term averaged spectral measurements have less possibility for error due to the fact that a sample is integrated over its entire length, compared to time-based cycle-to-cycle difference measurements that are based on a large collection of individual pitch periods (Awan, 2011). Therefore, spectral/cepstral analysis algorithms theoretically appear to have a lot of promise for application to connected speech, unlike measures based on cycle-to-cycle differences (jitter, shimmer, NHR and VTI). More important is whether any of these acoustic measures can offer objective clinical measures that are useful in diagnosing and treating voice disorders.

Review of the Literature

The clinical usefulness of time-based jitter, shimmer and NHR measurements has been researched many times in years past by correlating the results of these measurements with perception of voice quality. However, even carefully done studies with naïve listeners who had at least 30 minutes of training have found only moderate correlations between time-based acoustic measures and perception of severity of

dysphonia and dysphonia type. For example, Wolfe, Fitch, and Cornell (1995) examined the predictability of perceived dysphonia severity from acoustic measures. They collected voice samples of the vowel /a/ from 20 normal speakers (10 female students and 10 male students), ranging in age between 18 and 30 years, and compared them to voice samples of 60 patients that were referred by otolaryngolgists for voice treatment. The pathological voice set consisted of 9 men and 51 women, ranging in ages 23-65 years, with a mean age of 45 years. Speakers with deviant voice qualities were placed in three diagnostic groups (20 subjects per group): vocal nodules, vocal fold paralysis, and functional dysphonia. Both the normal and abnormal speakers were asked to phonate the vowel /a/ for several seconds. The phonatory samples were acoustically analyzed using four measures: average fundamental frequency, relative average pitch perturbation (RAP), shimmer (in dB), and HNR.

In preparation for the perceptual evaluation, the 22 students attended a 30-minute training which required them to evaluate the severity of sustained /a/ vowels a week prior to the evaluations. After the students completed the training session, they were asked to rate the experimental phonatory samples using a 7-point equal-appearing interval scale with 1 denoting normal phonation and 7 denoting severely abnormal phonation. The study revealed a moderate correlation between shimmer and severity judgment ($r = .54$, $r^2 = 29\%$, $p < .01$); a moderately low correlation between HNR and severity judgment ($r = -.32$, $r^2 = 10\%$, $p < .01$); and no significant correlation between jitter and severity judgment ($r = .2$, $r^2 = 4\%$, $p > .01$). When the four acoustic measures were combined through a stepwise regression analysis, only shimmer and fundamental frequency contributed to the prediction of perceived severity ($R^2 = 31\%$, $p = .018$). In general,

Wolfe et al. (1995) found low correlation between severity of dysphonia and acoustic measures, ranging from $r^2 = 4\%$ to $r^2 = 29\%$. Furthermore, even when acoustic measures were combined, there was limited predictability of perceptual judgments based on a regression analysis ($R^2 = 31\%$).

Martin, Fitch and Wolfe (1995) looked at acoustic correlates of perceived severity for various types of dysphonia. They included 60 subjects with voice disorders, and 20 normal speakers (10 males and 10 females). The sustained vowel /a/ was judged by 29 naïve listeners who were trained for 30 minutes with synthesized voice signals. Samples were perceptually classified as either predominantly *breathy*, *hoarse*, *rough*, or *normal*, and then were rated by listeners for overall severity. Results showed that the severity of *rough* voices was best indicated by HNR ($r = .85$, $r^2 = 73\%$, $P = 0.0016$); HNR, Jitter and shimmer best predicted the severity of *breathy* of voices ($r = .86$, $r^2 = 74\%$, $P = 0.007$); but no correlates were found for the severity of *hoarse* and *normal* voices. Results from this study show better correlations between acoustic measures and some perceptual dysphonia categories (*rough* and *breathy*), but no acoustic correlates were found for others (*hoarse* and *normal*).

Wolfe, Fitch, and Martin (1997) studied acoustic correlates of various voice types and perceived severity of pathologic voices in the sustained vowels /a/ and /i/ produced at conversational loudness and pitch. Samples of fifty-one speakers with voice disorders (20 males, 31 females) were judged by two listener groups. The first listener group consisted of 21 listeners who categorized voice type after a 30-minute training session with synthetic prototypes. Listeners categorized voice type as *rough*, *breathy* or both. A second listener group judged dysphonic severity after two 30-minute training sessions.

Both /a/ and /i/ sustained vowel samples were rated on a 7-point scale (1 = normal, 7 = severely dysphonic).

Sustained vowel samples were acoustically analyzed using all Multi Dimensional Voice Program (MDVP; KayPENTAX, 2008) measures. For overall severity, regression showed best predictability with NHR combined with shimmer and an amplitude perturbation measure ($R = .63$, $R^2 = 40\%$). In regards to voice types, *rough* was best predicted by fundamental frequency variation, peak amplitude variation in percent, and fundamental frequency tremor frequency in Hertz ($R = .76$, $R^2 = 58\%$). Shimmer in dB, jitter and fundamental frequency tremor frequency in Hertz best predicted *hoarse* voice type ($R = .68$, $R^2 = 46\%$), and *breathy* was best predicted by shimmer ($R = .67$, $R^2 = 45\%$). Overall, this study only modestly predicted dysphonic severity, with 40-50% of the variance accounted for. An additional issue was that the acoustic correlates of dysphonic voice types identified in this study were not consistent with these identified by Martin et al. (1995).

Acoustic correlates of hoarseness and breathiness in addition to differences between gender and age groups were examined by Gorham-Rohan and Laures-Gore (2006). Productions of the sustained vowel /a/, elicited from 112 normal speakers including both young (mean age of 25 years) and elderly (mean age of 70 years) males and females, were judged by 10 naïve listeners (8 female, 2 male). Perceptions of *breathiness* correlated with fundamental frequency standard deviation for elderly men ($r = .48$, $r^2 = 23\%$), while perceptions of *hoarseness* correlated with NHR for elderly women ($r = .41$, $r^2 = 16\%$) and elderly men ($r = .45$, $r^2 = 20\%$). Perceptions of *hoarseness* also correlated with APQ for young men ($r = .5$, $r^2 = 25\%$) and elderly women ($r = .41$, r^2

= 16%). In general, low correlations between acoustic and perceptual measures for various gender and age groups resulted from this study. Further, it can be seen that acoustic correlates identified for the various perceptual dysphonia qualities are, again, not consistent with those identified in previous studies.

In summary, multiple problems were apparent in all the above studies of time-based measures and their ability to predict dysphonia type and severity. Low correlations and levels of predictability were shown for all studies except Martin et al. (1995), who found moderate correlations. Most predictability results (r^2 values) ranged from 4% to 58%. In addition to low correlations and levels of predictability, there appeared to be no good theoretical rationale for the acoustic correlates identified for particular dysphonic voice qualities. For example, the Wolfe et al. (1997) study found “fundamental frequency tremor frequency in Hertz” as an acoustic correlate for severity of hoarseness. However, hoarseness results from turbulent breathiness at the vocal fold level and aperiodic vocal fold vibration. Fundamental frequency tremor frequency in Hertz measures the regular tremor oscillations imposed on the vibratory pattern of the vocal folds. It is difficult to see how this measure would relate to perceived hoarseness. Lastly, acoustic correlates for particular voice qualities are not consistent throughout the time-based studies. For example, roughness was best predicted by HNR in the study conducted by Martin et al., (1995), while Wolfe et al. (1997) found fundamental frequency variation, peak amplitude variation in percent, and fundamental frequency tremor frequency in Hertz to be the best predictors of roughness. In another example, breathiness was correlated with measures of shimmer (Wolfe et al., 1997), with HNR, jitter and shimmer (Martin et al., 1995) and

with fundamental frequency standard deviation (Gorham-Roman and Laures-Gore, 2006).

Other researchers have noted that time-based acoustic methods for voice analysis are not generally effective with more severely dysphonic vowel samples (Awan, 2011; Carding, Steen, Webb, MacKenzie, Deary & Wilson, 2004). Jitter, shimmer, and HNR measures require a signal characterized by definite pitch periods, however very dysphonic voices do not have regular pitch periods that are easily identified by computer algorithms. Furthermore, time-based acoustic methods lack validity in the analysis of continuous speech because of the rapidly changing frequency, intensity and spectral characteristics of a connected speech sample (Awan, 2011). All these factors, combined with low correlations with perceptual judgments and predictability, suggest that there are considerable limitations to time-based measures. Research into acoustic measures that transcend these limitations is likely to be more clinically useful.

In recent years, research has begun exploring cepstral measures for possible clinical usefulness regarding discrimination of normal versus dysphonic speakers, and severity of dysphonia. Lowell, Colton, Kelley, and Hahn (2011) investigated spectral- and cepstral-based acoustic measures (the spectral mean, standard deviation, skewness, and kurtosis of a *Long-Term Average Spectrum* analysis [LTAS]) in terms of each measurement's ability to distinguish normal versus dysphonic speakers and overall severity. They examined 27 dysphonic voice samples (produced by speakers ages 19-86 years with a mean age of 41 years; 14 women and 13 men) and 27 normal voice samples (produced by speakers ages 26-55 years with a mean age of 39 years; 11 women and 16 men) that were selected from a published database recorded by Massachusetts Ear and

Eye Institute. The dysphonic speakers had the following primary disorders: mass lesions of the vocal folds (8), paresis/paralysis (13), keratosis/leukoplakia (3), vocal fold edema (1), presbyphonia (1), and laryngeal web (1).

Both dysphonic and normal speakers were asked to read the Rainbow Passage out loud. Lowell and colleagues then edited the samples to produce three comparison stimuli: first sentence of the Rainbow passage (17 words), second sentence of the Rainbow passage (12 words), and a constituent phrase within the second sentence (first six words). The voice samples were analyzed using LTAS for spectral and cepstral measures (Cepstral Peak Prominence, or CPP, and smoothed CPP or CPPS) using the Computerized Speech Lab (CSL) 4500 system (KayPENTAX, 2008). Prior to the analysis, unvoiced segments and pauses were edited out of each sample since the measures resulting from these algorithms are likely to be affected by unvoiced portions of the signal. Three judges with extensive experience in voice disorders rated the speech samples for overall dysphonia severity using the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). This screening test asks listeners to rate features of voice quality by marking a 100mm line (Visual Analog Scale). Prior to completing their ratings, judges attended a 1.5-hour training session. Perceptual judgments of the study samples were conducted in a separate session where judges were asked rate each sample for the presence of the three voice quality features as defined on the CAPE-V: *roughness*, *breathiness*, and *strain*. Anchor examples were presented at the beginning of the rating session and every 10 subsequent samples. Judges were able to re-listen to the anchors at any time during the session and were also able to repeat the sample being rated as often as they liked.

Four repeated measures analyses of variance were applied to test for speaker group differences (normal versus dysphonic) for spectral mean, spectral standard deviation (SD), CPP, and CPPS. Skewness and kurtosis acoustic measures were analyzed using the Mann-Whitney *U* test and the Wilcoxon Matched Pairs test, as these measurements were not normally distributed across all tasks. Within-speaker consistency was determined by correlation coefficients, which assessed relationships of sentence 1 to sentence 2 and sentence 2 to the phrase that consisted of half of sentence 2. Last, to determine the relationships between spectral mean, spectral SD, CPP, CPPS and perceptual judgments, Pearson *r* correlation analyses were performed. Because of the previously-mentioned distribution issues, a Spearman's rho analysis was utilized for correlations between skewness and kurtosis and perceptual measures.

Results indicated that three of the four spectral measures and both cepstral measures showed significant differences between dysphonic speakers and normal speakers. There was a significantly lower spectral mean, and significantly greater skewness and kurtosis in the dysphonic group compared to the normal group. The fourth spectral measure, spectral SD, also had lower values in the dysphonic group, but it did not meet the corrected alpha level for significance. CPP and CPPS values were significantly lower for the dysphonic group due to the cepstral peak and the average energy level of the cepstrum in the dysphonic group being smaller than in normal speaker group (see Figure 1-2).

Lowell et al. (2011) also found that in addition to cepstral and spectral measures being able to differentiate between normal and dysphonic voices, individual sentences for both speaker types were highly correlated with themselves. High consistency between

sentences 1 and 2 for both speaker types was found with correlation coefficients ranging from r = 0.889 to r = 0.973. Further, consistency was high between sentence 2 and the constituent phrase for both dysphonic and normal speakers, with correlation coefficients ranging from r = 0.898 to r = 0.962.

Of more interest to speech-language pathologists, Lowell et al. (2011) found that spectral and cepstral measures showed moderate to strong correlations with overall perceived voice severity. Moderate or greater correlations were indicated for spectral mean (rho = -0.64, rho^2 = 41%, $P < 0.001$); spectral skewness (rho = 0.71, rho^2 = 50%, $P < 0.001$); and spectral kurtosis (rho = 0.67, rho^2 = 45%, $P < 0.001$). Moderate to high correlations with perceived voice severity were indicated for CPP (rho = -0.78, rho^2 = 61%, $P < 0.001$) and CPPS (rho = -0.72, rho^2 = 52%, $P < 0.001$). Only spectral SD was minimally correlated to voice severity (rho = -0.26, rho^2 = 7%, $P < 0.056$). Overall, Lowell et al.'s study found that cepstral measures were generally better predictors of judgments of dysphonic severity than both time-based measures and the LTAS spectral measures they used. These findings suggest that cepstral-based measures would be helpful during the diagnostic process of a client with a possible speech disorder by helping the speech-language pathologist quantify a normal versus a dysphonic voice in addition to overall severity.

Watts and Awan (2011) also showed good potential clinical usefulness for cepstral measures in terms of differentiating between normal versus dysphonic speakers. Unlike Lowell et al. (2011) who studied only continuous speech samples, Watts and Awan performed cepstral measurements on both continuous speech and vowel prolongations. Sixteen hypofunctional speakers (mean age of 52 years; 11 females, 5

males) and 16 normal speakers (mean age of 53 years; 11 females, 5 males) were asked to sustain /a/ and read the Rainbow Passage. Two speech-language pathology graduate students served as perceptual judges. The students identified the speakers' voice quality type (*normal*, *breathy*, *rough*, or *hoarse*) and rated the severity on a 100-point visual analog scale that had labels for *mild*, *moderate*, and *severe*. The middle 1-s steady-state portion of the sustained vowel was isolated for spectral/cepstral analyses. Acoustic measures for continuous speech were centered on the second sentence of the passage. Cepstral analysis provided the acoustic measures CPP, and CPP standard deviation (CPP sd). Spectral analyses utilized an algorithm that had not been previously used by Lowell et al. (2011): L/H spectral ratio, and L/H spectral ratio standard deviation (L/H spectral ratio sd). Low to High (L/H) spectral energy ratio compares the average energy in the entire speech signal below 4 kHz to the average energy above 4 kHz in a long-term spectral analysis. Among the measures used in this study, CPP and L/H spectral ratio showed significant differences between groups in both speaking conditions. By demonstrating CPP and L/H spectral ratio as effective discriminatory measures of normal versus abnormal voice qualities, this study provides further evidence of the clinical value of cepstral/spectral-based measures.

In addition to being able to differentiate normal from dysphonic voices, acoustic measurements need to be sensitive to varying degrees of dysphonia severity. Awan et al. (2009) were able to identify spectral/cepstral measures that most effectively predicted dysphonia severity in pre- and post-treatment continuous voice recordings of female speakers. Pre- and post-treatment speech samples were selected from an archival database of patients with muscle tension dysphonia, with 104 female speakers chosen for analysis

(mean of 46.4 years of age). Voice therapy for the patients consisted of a single extended session of manual laryngeal reposturing maneuvers and/or circumlaryngeal massage, which stimulated an improved voice. The female speakers were asked to read the Rainbow Passage at a comfortable pitch and loudness. Afterwards the speech samples were edited to include only the 2nd and 3rd sentences.

All samples were analyzed for CPP, low/high (L/H) spectral ratio that the authors referred to as the DFT ratio (DFTR), and DFTR standard deviation (DFTR SD). Five master's degree students in communication disorders served as auditory-perceptual judges of the 104 speakers, with 208 total samples judged. Judges were asked to rate the continuous speech samples on a 100-point visual analogue scale. One end of the scale was labeled *normal*, and the opposite side was labeled *profoundly abnormal*, with higher numbers suggesting increased severity of dysphonia.

Step-wise linear regression analysis revealed a three-factor model consisting of CPP, DFTR SD, and DFTR, strongly correlating with perceived dysphonia severity (mean of $R = .85$; $R^2 = 73\%$). CPP was the strongest contributor to the three-factor predictive model, in addition to being the strongest individual correlate of listener perceived dysphonia severity ($r = -.81$; $r^2 = 66\%$). Paired t-tests were conducted to establish whether significant pre- versus post-treatment changes occurred in any of the three spectral/cepstral-based components (CPP, DFTR, and DFTR SD) of the predictive dysphonia severity model. Results indicated significant differences in all pre- versus post-treatment comparisons, with significant increases in all variables following treatment.

An additional series of paired t-tests was performed to determine whether significant differences existed between pre- versus post-treatment mean perceived

severity ratings and pre- versus post-treatment predicted severity ratings. In both instances, post-treatment mean perceived severity and post-treatment predicted values were significantly lower than pre-treatment observations. Last, treatment change scores were computed by subtracting post-treatment from pre-treatment ratings, showing a reduction in dysphonia severity. This study shows that strong predictions of listener-perceived dysphonia severity can be made from L/H spectral ratio and cepstral measures.

Awan, Roy, Jette, Meltzner, and Hillman (2010) also found promising results with spectral/cepstral-based measures for predicting dysphonia severity. Awan and colleagues studied dysphonia severity using spectral/cepstral-based acoustic measures in sustained vowel and continuous speech contexts. The study found strong relationships between perceptual and acoustic estimates of dysphonia severity. They collected speech samples from 24 dysphonic individuals (12 males and 12 females, between 21-78 years of age), which were divided equally into *mild*, *moderate*, and *severe* dysphonia severity categories. Eight normal speakers were also chosen: 4 males and 4 females between 25-32 years of age. Both speaker groups (dysphonic and normal) were asked to participate in the following select CAPE-V voice and speech tasks: sustained /a/, “an easy onset of phonation” sentence (‘How hard did he hit him?’), a sentence containing all voiced sounds (‘We were away a year ago’), a sentence targeted at eliciting hard glottal attack (‘We eat eggs every Easter’), and a sentence weighted with voiceless plosives (‘Peter will keep at the peak’). The perceptual judging group was composed of 25 speech pathology graduate students who were asked to rate the speech samples using a computerized graphical user interface version of four CAPE-V scales (perceptual attributes of *overall severity*, *roughness*, *breathiness*, and *strain*).

All four acoustic variables utilized in this study (CPP, CPP sd, L/H spectral ratio, and L/H spectral ratio sd) significantly combined in a four-factor regression model which correlated moderately with listener perceived severity for continuous speech with R = 0.81 (R^2 = 65%; Adjusted R^2 = 64%). CPP sd was first to enter the stepwise regression procedure, however CPP was found to have the strongest beta coefficient, signifying CPP as the strongest contributor to the overall R^2. For the analysis of continuous speech, both sentence type and gender proved to be non-significant contributors to the final multiple regression model. In terms of the sustained vowel context, all four acoustic variables along with gender significantly combined in a four-factor model which correlated with listener perceived severity with R = 0.96 (R^2 = 90%; Adjusted R^2 = 90%). Once again, CPP was observed to have the strongest beta coefficient and the strongest contribution to the overall R^2. Overall, CPP was the strongest predictor dysphonic severity for both continuous speech and sustained vowel contexts with CPP sd, L/H spectral ratio, and L/H ratio sd further strengthening the predictions. This study supported the usefulness of spectral/cepstral-based acoustic measures for objective voice assessment, especially for severity sensitivity.

The research cited above shows the strengths and weaknesses of using acoustic measurements to quantify perceptual impressions of voice quality for clinical assessment and treatment. Studies of time-based measures (e.g., jitter, shimmer, NHR) revealed low correlations between such measures and judgments of dysphonia severity and type (Wolfe et al., 1995; Wolfe et al., 1997; Gorham-Rowan & Laures-Gore, 2006), along with inconsistent acoustic correlates (Martin et al., 1995; Wolfe et al., 1997; Gorham-Rowan & Laures-Gore, 2006). In addition, there were no theoretical rationales presented

for the time-based acoustic correlate results (e.g., Wolfe et al., 1997). Overall, this type of analysis is limited to sustained vowels and is not effective with more severely dysphonic vowel samples.

In contrast, the cepstral- and spectral-based measure research reviewed above showed strong correlations with perceptually rated voice quality overall. This improvement over time-based methods may be due to the fact that spectral/cepstral measures are based on a long-term averaged spectrum of the integrated phonated sample rather than individual pitch periods. Cepstral- and spectral-based measures also reduce the noise of the sample while in turn strengthening the prominence of the fundamental frequency. Furthermore, cepstral- and spectral-based acoustic measurements can be used to evaluate voice quality in everyday speaking patterns. This is due to their suitability for application to connected speech contexts in addition to sustained vowels. Lowell et al. (2011) and Watts and Awan (2011) both found that cepstral- and spectral-based measures provided excellent discrimination of dysphonic and normal voices. In relation to predictability of dysphonia severity, Awan et al. (2009) and Awan et al. (2010) found strong correlations between perceptual severity ratings and cepstral and spectral measures. Awan et al. (2009) found CPP, L/H spectral ratio sd and L/H spectral ratio to account for 73% of the variability for connected speech, while CPP, CPP sd, L/H spectral ratio, L/H spectral ratio sd accounted for 90% of the variability for sustained vowels in the study by Awan et al. (2010). Not only were cepstral- and spectral-based measures highly correlated with perceptual severity ratings in Awan et al. (2009) and Awan et al. (2010), CPP had the strongest beta coefficient for all regression analyses computed in both studies. Based on the review of the literature, cepstral- and spectral-based acoustic

measures show significant promise as an objective measure of dysphonia for clinic utilization. Unfortunately, the small numbers of normal subjects in previous research and wide age ranges prevent a good estimation of the performance of normal speakers of various ages on these measures.

The next step in providing research-based support for clinical application of spectral/cepstral measures is to collect normative data on these measures. In collecting normative samples, non-overlapping age groups of approximately one decade should be utilized, in order to see if age effects are present in spectral and cepstral measures. Analysis methods similar to those used by Awan et al. (2010) should also be incorporated into further research, since the program used in that research was based on commercially-available software that incorporated stimuli from the CAPE-V, a standardized voice quality screening instrument used by voice clinicians.

Purpose

The purpose of this study was to begin to establish necessary baseline data for Long-Term Average spectral- and cepstral-based measures for both men and women. Specific research questions included the following: 1) What are the expected CPP, L/H spectral ratio, and CPP Fundamental Frequency (F_0) measures for men with normal voices, ages 20-30 years and 40-50 years?; 2) What are the expected CPP, L/H spectral ratio, and CPP F_0 measures for women with normal voices, ages 20-30 years and 40-50 years?; 3) Are there significant differences in CPP, L/H spectral ratio, and/or CPP F_0 as a function of gender, age, or an age x gender interaction for the vowels /a/ and /i/?; 4) Are there significant differences in CPP, L/H spectral ratio, and/or CPP F_0 as a function of gender, age, or an age x gender interaction for the four connected speech segments

elicited using the CAPE-V stimuli, plus sentences 2 and 3 of the Rainbow Passage? The eventual goal of this research is to provide normative data that will be helpful to clinicians assessing and treating voice disorders.

Method

Participants

This study included sixty participants consisting of fifteen males and fifteen females, ages 20-30 years, and fifteen males and fifteen females, ages 40-50 years. A number of exclusionary criteria were applied to all participants. No smokers were included. Participants were also excluded if they had a history of voice disorders, neurological disorders, or speech-language therapy. To be included in the study, all participants had to be healthy on the day of recording, and be native speakers of American English. A hearing screening was administered to ensure normal hearing in all subjects (ASHA, 1997). Along with normal hearing, all participants had to demonstrate normal speech and voice as determined by a screening with a modified version of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kramer, & Hillman, 2009).

Potential participants were recruited through a variety of methods, including announcements in undergraduate and graduate level classes, personal contacts, phone and email correspondence with local community churches, businesses, agencies and schools, and fliers posted at various campus locations. Interested individuals were told to contact the principal investigator for more information. During the initial contact, potential

participants were asked a series of questions regarding eligibility criteria (see Appendix A).

Participant Selection Procedures

If participants passed the initial eligibility screening, they were invited to come to the University of Wisconsin-Milwaukee Speech and Language Clinic to participate in the study. Once seated in a quiet laboratory, participants read a consent form educating them on the risks and benefits of the study (see Appendix B). All subjects included in the study agreed to and signed a consent form prior to participation.

Subjects' hearing and voice were screened to verify eligibility. A hearing screening was first administered (see Appendix C) to ensure that each participant's hearing was within normal limits according to ASHA (1997) criteria. Subjects first answered questions related to hearing and hearing loss. Next, the investigator viewed each participant's external auditory meatus and tympanic membrane with an otoscope, to ensure that the tympanic membrane was not occluded by wax or otherwise abnormal. Finally, over-the-ear headphones were placed on the participants' ears. Pure tones at 25 dB Hearing Level (HL) at the frequencies of 1000, 2000, and 4000 Hertz (Hz) were presented. Participants passed the hearing screening if there were no abnormalities or blockages of the tympanic membrane, and if reliable responses to the pure tones were obtained in both ears. The voice-screening tool utilized for this study was a modified version of the CAPE-V (Kempster et al., 2009; see Appendix D). The CAPE-V voice screening tool has been constructed by an international group of voice scientists with the goal of creating a standardized approach to evaluating and documenting auditory perceptions of voice quality. The authors of this screening tool believe that the CAPE-V

has high concurrent validity with older, less comprehensive voice rating scales. In the present study, individuals' voices were perceptually rated in terms of overall severity, roughness, breathiness, and strain. Pitch, loudness, and resonance of the voice were also evaluated according to the CAPE-V. In order to assess the adequacy of potential subjects' articulation, several items were added to the CAPE-V to draw the investigators' attention to the most frequently-misarticulated phonemes (see Appendix D). A graduate student who completed a graduate level voice course and voice clinicals, under the supervision of a speech-language pathologist with more than 25 years of voice and speech experience, both rated the voices. Ratings from the CAPE-V were based on two sustained vowels, six sentences, and a spontaneous speech sample. Speakers who had normal hearing and whose voices and speech sound production were perceptually judged to be within normal limits were included in the study.

Instrumentation and Materials

A Shure Model SM58 unidirectional dynamic microphone attached to an Audio Buddy Dual Mic Preamp was used to collect all samples from the speakers. Recording, storage, and analysis were executed on a Dell Optiplex 980 desktop computer. The computer was installed with the Kay-PENTAX Multi-Speech (Model 3700) software running the subprogram Analysis of Dysphonia in Speech and Voice (ADSV; Model 5109, version 3.4.1). The intensity of the participants' productions was monitored by a RadioShack Sound Level Meter (Catalogue Number 33-2055). The Kay-PENTAX Real-Time Pitch subprogram (RTP; Model 5121, version 3.4.1) was also used to independently assess the fundamental frequency of each speech sample.

Voice Recording Procedures

Voice recordings of the participants were conducted in a sound-treated booth with a noise level of less than 50 dB sound pressure level (SPL). The participants stood in front of a sound level meter and microphone. For all recordings, the sound level meter and microphone were each positioned at a 45-degree angle from the speakers' mouth, one on each side of the participant, with 6-inch mouth-to-microphone distances. Mouth-to-microphone distance was maintained during all recordings using a measuring device that each participant held against his or her chin. One experimenter started and stopped the data collection, while the other monitored production of the speakers in order to ensure appropriate intensity levels. Speakers were asked to sustain the vowels /a/ and /i/ for about 3 sec at a 75 dB (± 2 dB) intensity level. Participants were also asked to read out loud four CAPE-V stimulus (see Appendix E) and the 2nd and 3rd sentence of the Rainbow Passage (Fairbanks, 1960; see Appendix F), which is consistent with the stimuli the ADSV program was developed to analyze. Once again, the speakers were asked to produce the connected speech samples at a peak level of 75 dB (± 2 dB). Stimuli for sustained vowels and connected speech productions were visually displayed for participants to read aloud. Both the sustained vowel and connected speech samples were saved to a removable disk upon completion, and were stored securely.

Data Analysis Procedures

Prior to analysis, the speech sample was displayed onscreen, and the most stable one-second portion of each sustained vowel sample was isolated for spectral/cepstral analyses. For connected speech samples, the onset and offset of the sample was marked, as specified by Awan (2011). To ensure adequately loud samples, one-third to one-half of

the intensity range of the ADSV program had to be utilized (Awan, 2011). In addition, silence and low-level noise that occurred before and after any recording (both sustained vowel and connected speech) were removed prior to analysis. Finally, for the connected speech samples, the "vocalic detection" routine was utilized. This command removed unvoiced portions from the connected speech samples. This routine was done based on the observation of Lowell et al. (2011) that removing unvoiced portions from connected speech had a substantial effect on spectral measures. Awan (2011) suggested that removing unvoiced portions of the sample may give ADSV analysis results more "face validity" (p. 37), since those portions do not contribute to the spectral and cepstral analyses but can introduce artifacts into the data. However, use of the vocalic detection routine is not currently part of the ADSV default settings, and some earlier studies were completed without it.

After selecting the analysis portion of the samples, as defined above, data were obtained using the ADSV statistical analysis program. For each speaker, three dependent variables were recorded (Cepstral Peak Prominence, or CPP; Low-to-High Spectral Ratio, or L/H spectral ratio; and the fundamental frequency of the Cepstral Peak Prominence, or CPP F_0), in seven different contexts (two sustained vowels, and five connected speech segments). CPP F_0 is defined by Awan (2011) as the mean frequency of the cepstral peaks that were identified by the ADSV analysis between 60 Hz to 300 Hz (the pitch extraction range).

In order to cross-check the fundamental frequency determined by the ADSV subprogram (CPP F_0), with an accepted analysis system for frequency, a separate analysis of selected speech samples was completed using the Real-Time Pitch (RTP) subprogram

of Multi-Speech. The fundamental frequency provided by this program was recorded for use in post-hoc analyses. Analysis range for RTP was adjusted as needed to remove frequency outliers from each sample, with the pitch smoothing level set to medium.

Statistical Analysis

Independent variables of this study included gender (2 levels: male and female), and age (2 levels: 20-30 years and 40-50 years). Summary statistics were calculated for each dependent variable (CPP, L/H spectral ratio, and CPP F_0) for each level of the independent variables and for each vowel (2) and each connected speech segment (5). Two Multivariate Analyses of Variance (MANOVAs) were planned to assess the significance of differences observed based on the between-subjects variables of age, gender, and the age x gender interaction. One MANOVA was applied to the vowel data for each dependent variable, (CPP for /a/, CPP for /i/, L/H spectral ratio for /a/, L/H spectral ratio for /i/, CPP F_0 for /a/, CPP F_0 for /i/) as a function of age and gender. The second MANOVA was used to assess the connected speech segment data (5 connected speech segments, each associated with a CPP, L/H spectral ratio and CPP F_0) as a function age and gender. A probability level of $\alpha = .05$ was chosen as the criterion necessary to establish statistical significance between or among the variables.

Results

Descriptive Statistics

To establish baseline data for Long-Term Average spectral- and cepstral-based measures for both men and women in two different age groups, descriptive statistics for Cepstral Peak Prominence (CPP), Low-to-High Spectral Ratio (L/H spectral ratio), and

Cepstral Peak Prominence Fundamental Frequency (CPP F_0) were calculated. Tables 3-1 through 3-21 show summary statistics for vowels and connected speech segments as a function of age and gender. Averaged results across age, gender, and age and gender combined are also presented, along with standard deviations.

Visual inspection of Tables 3-1 through 3-6 showed multiple differences between the dependent variables as a function of subjects' *gender* for the vowel speech samples. Noticeably higher CPP and L/H spectral ratio values for both vowels were observed for males compared to females. Male and female CPP F_0 values for vowels were in the expected frequency ranges, with male fundamental frequencies about an octave below females. For females, however, there seemed to be a drop in CPP F_0 from the young age group to the older age group, while for males, a slight rise in CPP F_0 was seen for the vowel /a/ for older subjects compared to younger. However, in general *age* did not appear to have a notable effect on the dependent variables for vowel voice quality results.

In addition to vowels /a/ and /i/, descriptive data were also collected from four different sentence types elicited by the CAPE-V, and the 2nd and 3rd sentence of the Rainbow Passage. Summary statistics for connected speech segments are presented in Tables 3-7 through 3-21. It will be noted that the CPP values in connected speech segments are considerable lower than in isolated vowels. This may be because of the variability of frequency and intensity within connected speech, which reduces the prominence of the cepstral peaks. In general, the female subjects, regardless of age, appeared to have slightly higher CPP values than males in connected speech segments (in contrast to the vowel data). Younger speakers also had slightly higher CPP values for connected speech than older speakers, although not markedly so. For L/H spectral ratio,

male speakers generally were a little higher than females. Consistent age effects for L/H spectral ratio were not observed. As expected, both genders had appropriate gender-specific CPP F_0 values. Interestingly for males, CPP F_0 was higher in the older group than the younger for all five connected speech segments. This is consistent with the trend for CPP F_0 observed for males on /a/. Lastly, the older females consistently had decreased CPP F_0 values compared to the younger females.

Table 3-1. Results for CPP for /a/ as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	12.942 (1.462) [N=15]	10.965 (1.153) [N=15]	11.953 (1.639) [N=30]
40-50 yrs	12.145 (2.044) [N=15]	10.894 (1.751) [N=15]	11.520 (1.976) [N=30]
Both Age Groups	All Males	All Females	All Participants
	12.544 (1.792) [N=30]	10.929 (1.457) [N=30]	11.736 (1.813) [N=60]

Table 3-2. Results for CPP for /i/ as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	11.470 (2.021) [N=15]	7.339 (1.401) [N=15]	9.404 (2.708) [N=30]
40-50 yrs	9.489 (2.184) [N=15]	7.398 (1.879) [N=15]	8.443 (2.267) [N=30]
Both Age Groups	All Males	All Females	All Participants
	10.479 (2.300) [N=30]	7.369 (1.629) [N=30]	8.924 (2.523) [N=60]

Table 3-3. Results for L/H spectral ratio for /a/ as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	34.968 (3.297) [N=15]	31.309 (3.613) [N=15]	33.138 (3.875) [N=30]
40-50 yrs	34.871 (4.241) [N=15]	31.881 (6.073) [N=15]	33.376 (5.366) [N=30]
Both Age Groups	All Males	All Females	All Participants
	34.919 (3.733) [N=30]	31.595 (4.918) [N=30]	33.257 (4.642) [N=60]

Table 3-4. Results for L/H spectral ratio for /i/ as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	33.730 (4.183) [N=15]	29.157 (4.004) [N=15]	31.443 (4.647) [N=30]
40-50 yrs	33.116 (4.988) [N=15]	29.948 (3.904) [N=15]	31.532 (4.686) [N=30]
Both Age Groups	All Males	All Females	All Participants
	33.423 (4.534) [N=30]	29.552 (3.906) [N=30]	31.488 (4.627) [N=60]

Table 3-5. Results for CPP F_0 for /a/ as a function of gender and age, in Hz. Averaged results across gender combined are also presented. Standard deviations are in parentheses. Genders are not averaged, since males and females have distinctive fundamental frequencies.

Age	Males	Females	Both Genders
20-30 yrs	113.613 (18.743) [N=15]	243.205 (23.300) [N=15]	--
40-50 yrs	118.904 (17.784) [N=15]	218.920 (33.109) [N=15]	--
Both Age Groups	All Males	All Females	All Participants
	116.258 (18.153) [N=30]	231.063 (30.722) [N=30]	--

Table 3-6. Results for CPP F_0 for /i/ as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses. Genders are not averaged, since males and females have distinctive fundamental frequencies.

Age	Males	Females	Both Genders
20-30 yrs	122.870 (17.333) [N=15]	248.286 (20.591) [N=15]	--
40-50 yrs	122.918 (20.911) [N=15]	226.499 (32.520) [N=15]	--
Both Age Groups	All Males	All Females	All Participants
	122.894 (18.871) [N=30]	237.392 (28.948) [N=30]	--

Table 3-7. Results for CPP for Connected Speech Segment 1: “How hard did he hit him?” as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	6.355 (1.219) [N=15]	6.468 (0.931) [N=15]	6.412 (1.067) [N=30]
40-50 yrs	4.851 (0.841) [N=15]	5.962 (0.618) [N=15]	5.407 (0.919) [N=30]
Both Age Groups	All Males	All Females	All Participants
	5.603 (1.282) [N=30]	6.215 (0.818) [N=30]	5.909 (1.110) [N=60]

Table 3-8. Results for CPP for Connected Speech Segment 2: “We were away a year ago.” as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	7.982 (1.375) [N=15]	8.250 (0.985) [N=15]	8.116 (1.183) [N=30]
40-50 yrs	7.120 (1.229) [N=15]	7.898 (0.721) [N=15]	7.509 (1.066) [N=30]
Both Age Groups	All Males	All Females	All Participants
	7.551 (1.354) [N=30]	8.074 (0.867) [N=30]	7.812 (1.158) [N=60]

Table 3-9. Results for CPP for Connected Speech Segment 3: “We eat eggs every Easter.” as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	5.437 (0.955) [N=15]	6.342 (0.845) [N=15]	5.890 (0.998) [N=30]
40-50 yrs	4.521 (0.838) [N=15]	6.176 (0.844) [N=15]	5.349 (1.179) [N=30]
Both Age Groups	All Males	All Females	All Participants
	4.979 (0.998) [N=30]	6.259 (0.834) [N=30]	5.619 (1.117) [N=60]

Table 3-10. Results for CPP for Connected Speech Segment 4: “Peter will keep at the peak.” as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	6.402 (1.551) [N=15]	7.017 (1.193) [N=15]	6.709 (1.395) [N=30]
40-50 yrs	4.816 (0.753) [N=15]	6.834 (0.695) [N=15]	5.825 (1.249) [N=30]
Both Age Groups	All Males	All Females	All Participants
	5.609 (1.444) [N=30]	6.925 (0.964) N=30]	6.267 (1.386) [N=60]

Table 3-11. Results for CPP for Connected Speech Segment 5: the 2nd and 3rd sentences of the *Rainbow Passage* as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	6.510 (1.060) [N=15]	7.532 (0.972) [N=15]	7.021 (1.126) [N=30]
40-50 yrs	5.326 (0.587) [N=15]	7.038 (0.640) [N=15]	6.182 (1.06) [N=30]
Both Age Groups	All Males	All Females	All Participants
	5.918 (1.035) [N=30]	7.285 (0.847) [N=30]	6.602 (1.164) [N=60]

Table 3-12. Results for L/H spectral ratio for Connected Speech Segment 1: "How hard did he hit him?" as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	32.112 (2.820) [N=15]	29.686 (3.138) [N=15]	30.899 (3.180) [N=30]
40-50 yrs	30.896 (2.270) [N=15]	29.874 (2.719) [N=15]	30.385 (2.515) [N=30]
Both Age Groups	All Males	All Females	All Participants
	31.504 (2.590) [N=30]	29.780 (2.886) [N=30]	30.642 (2.854) [N=60]

Table 3-13. Results for L/H spectral ratio for Connected Speech Segment 2: “We were away a year ago.” as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	37.838 (2.206) [N=15]	34.323 (2.341) [N=15]	36.080 (2.862) [N=30]
40-50 yrs	37.484 (3.265) [N=15]	34.949 (2.694) [N=15]	36.216 (3.211) [N=30]
Both Age Groups	All Males	All Females	All Participants
	37.660 (2.744) [N=30]	34.636 (2.501) [N=30]	36.148 (3.017) [N=60]

Table 3-14. Results for L/H spectral ratio for Connected Speech Segment 3: “We eat eggs every Easter.” as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	31.983 (2.077) [N=15]	29.241 (2.738) [N=15]	30.612 (2.766) [N=30]
40-50 yrs	31.703 (2.774) [N=15]	30.057 (2.593) [N=15]	30.880 (2.768) [N=30]
Both Age Groups	All Males	All Females	All Participants
	31.843 (2.412) [N=30]	29.649 (2.653) [N=30]	30.746 (2.746) [N=60]

Table 3-15. Results for L/H spectral ratio for Connected Speech Segment 4: "Peter will keep at the peak." as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	33.802 (2.190) [N=15]	29.331 (3.244) [N=15]	31.566 (3.545) [N=30]
40-50 yrs	34.440 (2.916) [N=15]	31.000 (3.257) [N=15]	32.720 (3.505) [N=30]
Both Age Groups	All Males	All Females	All Participants
	34.121 (2.555) [N=30]	30.166 (3.305) [N=30]	32.143 (3.543) [N=60]

Table 3-16. Results for L/H spectral ratio for Connected Speech Segment 5: the 2nd and 3rd sentences of the *Rainbow Passage* as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses.

Age	Males	Females	Both Genders
20-30 yrs	34.168 (1.715) [N=15]	32.042 (2.327) [N=15]	33.105 2.281 [N=30]
40-50 yrs	33.354 (2.175) [N=15]	32.099 (2.5) [N=15]	32.727 (2.389) [N=30]
Both Age Groups	All Males	All Females	All Participants
	33.761 (1.968) [N=30]	32.071 (2.373) [N=30]	32.916 (2.323) N=60]

Table 3-17. Results for CPP F_0 for Connected Speech Segment 1: "How hard did he hit him?" as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses. Genders are not averaged, since males and females have distinctive fundamental frequencies.

Age	Males	Females	Both Genders
20-30 yrs	113.617 (13.018) [N=15]	221.166 (16.778) [N=15]	--
40-50 yrs	117.417 (15.923) [N=15]	202.430 (16.105) [N=15]	--
Both Age Groups	All Males	All Females	All Participants
	115.517 (14.420) [N=30]	211.798 (18.759) [N=30]	--

Table 3-18. Results for CPP F_0 for Connected Speech Segment 2: "We were away a year ago." as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses. Genders are not averaged, since males and females have distinctive fundamental frequencies.

Age	Males	Females	Both Genders
20-30 yrs	110.808 (15.576) [N=15]	212.874 (11.758) [N=15]	--
40-50 yrs	114.202 (12.477) [N=15]	194.998 (18.293) [N=15]	--
Both Age Groups	All Males	All Females	All Participants
	112.505 (13.973) [N=30]	203.936 (17.633) [N=30]	--

Table 3-19. Results for CPP F_0 for Connected Speech Segment 3: "We eat eggs every Easter." as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses. Genders are not averaged, since males and females have distinctive fundamental frequencies.

Age	Males	Females	Both Genders
20-30 yrs	117.437 (15.565) [N=15]	218.726 (13.054) [N=15]	--
40-50 yrs	121.955 (13.763) [N=15]	199.834 (16.001) [N=15]	--
Both Age Groups	All Males	All Females	All Participants
	119.696 (14.618) [N=30]	209.280) (17.268) [N=30]	--

Table 3-20. Results for CPP F_0 for Connected Speech Segment 4: "Peter will keep at the peak." as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses. Genders are not averaged, since males and females have distinctive fundamental frequencies.

Age	Males	Females	Both Genders
20-30 yrs	128.695 (18.229) [N=15]	243.824 (12.487) [N=15]	--
40-50 yrs	131.040 (15.674) [N=15]	230.810 (18.949) [N=15]	--
Both Age Groups	All Males	All Females	All Participants
	129.867 (16.747) [N=30]	237.317 (17.100) [N=30]	--

Table 3-21. Results for CPP F_0 for Connected Speech Segment 5: the 2nd and 3rd sentences of the *Rainbow Passage* as a function of gender and age. Averaged results across age, gender and age and gender combined are also presented. Standard deviations are in parentheses. Genders are not averaged, since males and females have distinctive fundamental frequencies.

Age	Males	Females	Both Genders
20-30 yrs	108.605 (15.551) [N=15]	206.647 (12.727) [N=15]	--
40-50 yrs	109.680 (16.3620) [N=15]	185.415 (18.446) [N=15]	--
Both Age Groups	All Males	All Females	All Participants
	109.143 (15.694) [N=30]	196.031 (18.948) [N=30]	--

Inferential Statistics

To determine the significance of the differences observed in CPP, L/H spectral ratio and/or CPP F_0 as a function of gender, age, or a gender x age interaction for the vowels /a/ and /i/, a single multivariate analysis of variance (MANOVA) was performed. Results are presented in Table 3-22. The MANOVA included all the dependent variables (CPP, L/H spectral ratio, and CPP F_0) coded into each vowel, with gender and age as the between-subjects independent variables. Results of the MANOVA for the vowels /a/ and /i/ can be seen in Tables 3-22. The analysis revealed *gender* as a significant main effect for all dependent variable/vowel combinations ($p \geq 0.05$). For the voice quality measures of CPP and L/H spectral ratio, men's acoustic measures were significantly higher than women's for both vowels. As expected, there were significant differences between men and women on CPP F_0, with women having significantly higher fundamental frequencies than men. There were no statistically significant differences as a function of age for any of the dependent variable/vowel combinations ($p \geq 0.05$). Two significant interactions

occurred between age and gender: one for CPP F_0 in the production of /a/, and one for CPP in the production of /i/. The descriptive statistics in Figure 3-1 show that the older female group had a lower CPP F_0 for the vowel /a/ than the younger female group, while older males had a slightly increased in CPP F_0 than younger males. In regard to CPP for the vowel /i/, males' CPP measures decreased with age, while females' CPP measures increased slightly with age (see Figure 3-2).

Table 3-22. MANOVA Results for the vowels /a/ and /i/.
Test of Between-Subjects Effects.

Source	Dependent Variable	F	Significance
Gender	CPP /a/	14.596	.000*
	LH /a/	8.417	.005*
	CPP F_0 /a/	342.828	.000*
	CPP /i/	40.474	.000*
	LH /i/	12.203	.001*
	CPP F_0 /i/	354.447	.000*
Age	CPP /a/	1.053	.309
	LH /a/	.043	.836
	CPP F_0 /a/	2.346	.131
	CPP /i/	3.859	.054
	LH /i/	.006	.937
	CPP F_0 /i/	3.194	.079
Gender x Age	CPP /a/	.737	.394
	LH /a/	.085	.772
	CPP F_0 /a/	5.689	.020*
	CPP /i/	4.350	.042*
	LH /i/	.403	.528
	CPP F_0 /i/	3.222	.078

* Significant Difference or alpha ≥ .05

Figure 3-1. Estimated Marginal Means of CPP F_0 for the vowel /a/. Gender 1 = males speakers, Gender 2 = female speakers. For age groups, 1 = 20-30 years of age, 2 = 40-50 years of age.

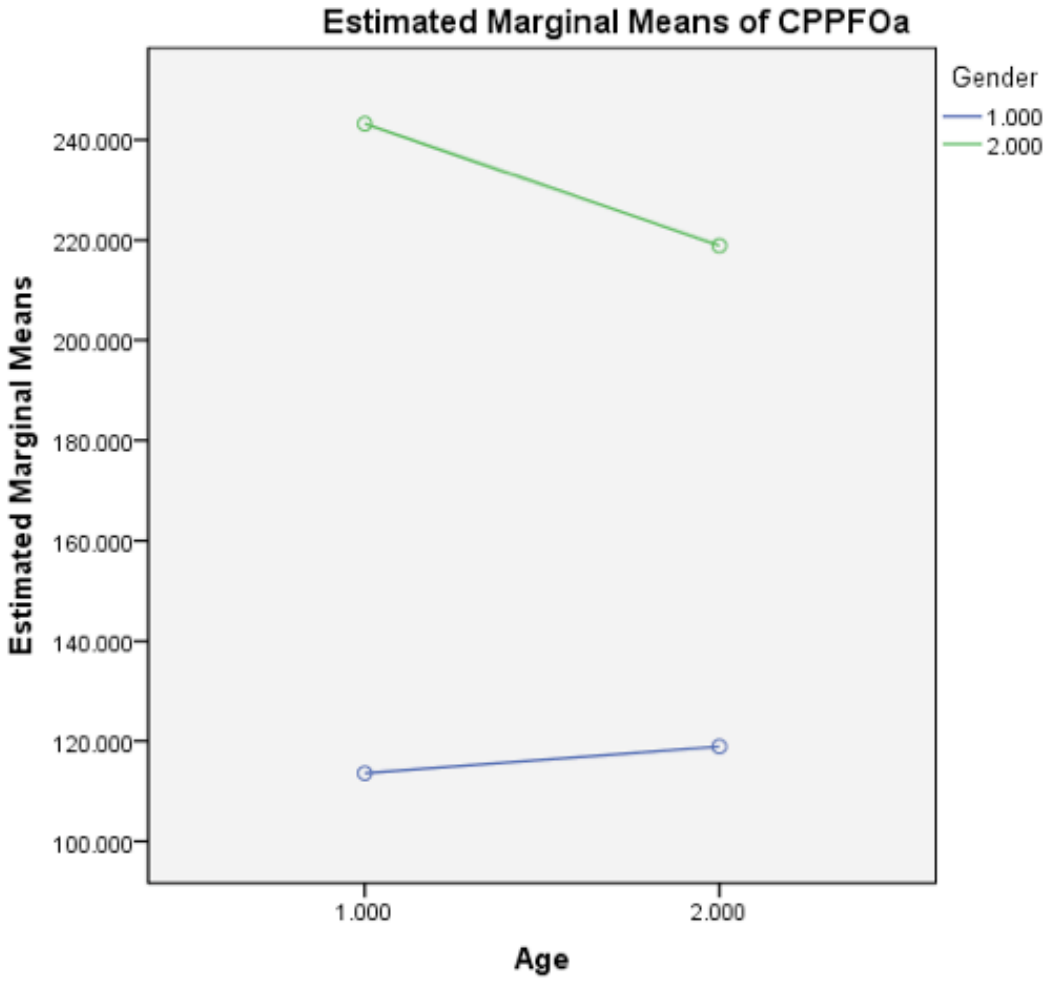
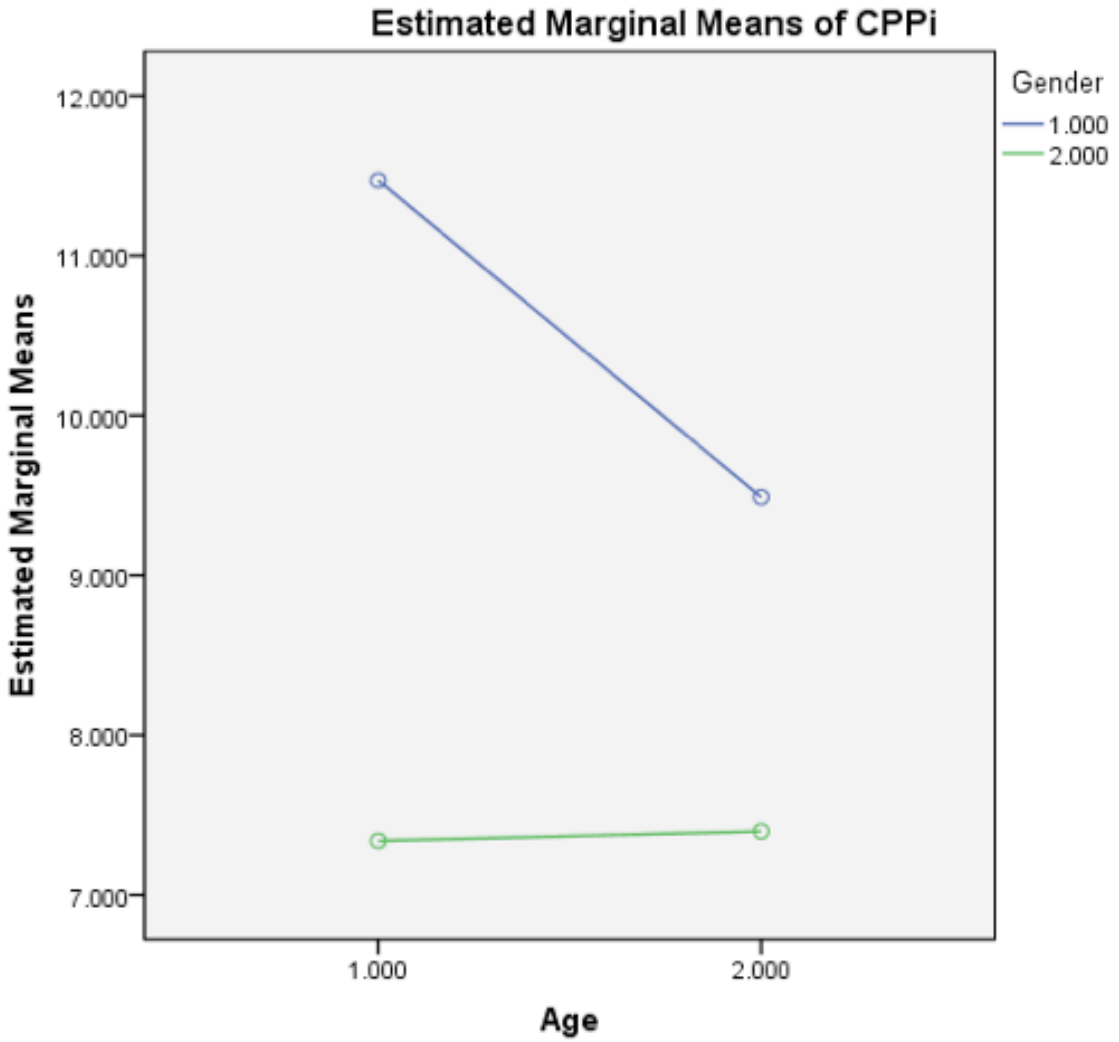


Figure 3-2. Estimated Marginal Means of CPP for the Vowel /i/. Gender 1 = males speakers, Gender 2 = female speakers. For age groups, 1 = 20-30 years of age, 2 = 40-50 years of age.



To determine the significance of the differences in CPP, L/H spectral ratio, and/or CPP F_0 as a function of gender, age, and the gender x age interaction for the five connected speech segments, another MANOVA was performed. Tables 3-23a-c show the results of this procedure for all the dependent variables – CPP, L/H spectral ratio, and CPP F_0, coded in each connected speech segment – with gender and age as the between-subjects independent variables. Once again, *gender* was found to be statistically significant for all dependent variables except for the CPP in connected speech segment 2

("We were away a year ago"). For all the significant CPP and L/H spectral ratio relationships for gender, women had significantly better CPP values than men, while men had significantly better L/H spectral ratio values than women.

Results of the MANOVA for the five different connected speech segments (Table 3-23b) revealed a significant main effect for *age* in almost half of the dependent variables: CPP for connected speech segments 1-5, and CPP F_0 for connect speech segment 5. Descriptive data on Tables 3-7 to 3-11 show that for all 5 connected speech segments elicited by the CAPE-V, CPP become slightly worse with age. The significant age effect for CPP F_0 for connected speech segment #5 appeared to be related to women's CPP F_0 for that speech segment decreasing significantly with age.

Six significant *age x gender* interactions were revealed for the five different connected speech segments (CPP for segment 1 and 4, and CPP F_0 for segment 1-3 and 5; see Table 3-23c). For connected speech segment 1 and 4, females had better CPP values than males, with both worsening with age. However, males decreased more with age compared to females (see Figures 3-3 and 3-4). According to Figures 3-5, 3-6, 3-7, and 3-8, there is an expected significant CPP F_0 difference between genders for segments 1-3 and 5. A significant gender x age interaction occurred because male CPP F_0 increased slightly with age while female values decreased more drastically.

Table 3-23 MANOVA Results for the 5 Connected Speech Segments.
Table 3-23a. Test of Between-Subjects Effects for the Five Different Connected Speech Segments with Gender as the Independent Variable.

Source	Dependent Variable	F	Significance
Gender	CPP segment 1	6.533	.013*
	LH segment 1	5.876	.019*
	CPP F_0 segment 1	577.047	.000*
	CPP segment 2	3.359	.072
	LH segment 2	19.417	.000*
	CPP F_0 segment 2	575.744	.000*
	CPP segment 3	32.315	.000*
	LH segment 3	11.010	.002*
	CPP F_0 segment 3	561.132	.000*
	CPP segment 4	21.307	.000*
	LH segment 4	27.264	.000*
	CPP F_0 segment 4	633.790	.000*
	CPP segment 5	39.743	.000*
	LH segment 5	8.868	.004*
	CPP F_0 segment 5	447.706	.000*

* Significant Difference or alpha ≥ .05

Table 3-23b. Test of Between-Subjects Effects for the Five Different Connected Speech Segments with Age as the Independent Variable.

Source	Dependent Variable	F	Significance
Age	CPP segment 1	17.604	.000*
	LH segment 1	.522	.473
	CPP F_0 segment 1	3.472	.068
	CPP segment 2	4.520	.038*
	LH segment 2	.039	.844
	CPP F_0 segment 2	3.611	.063
	CPP segment 3	5.771	.020*
	LH segment 3	.165	.687
	CPP F_0 segment 3	3.612	.063
	CPP segment 4	9.605	.003*
	LH segment 4	2.322	.133
	CPP F_0 segment 4	1.562	.217
	CPP segment 5	14.972	.000*
	LH segment 5	.445	.508
	CPP F_0 segment 5	6.024	.017*

* Significant Difference or alpha ≥ .05

Table 3-23c. Test of Between-Subjects Effects for the Five Different Connected Speech Segments with Gender x Age as the Independent Variables.

Source	Dependent Variable	F	Significance
Gender x Age	CPP segment 1	4.341	.042*
	LH segment 1	.975	.328
	CPP F_0 segment 1	7.904	.007*
	CPP segment 2	.797	.376
	LH segment 2	.510	.478
	CPP F_0 segment 2	7.789	.007*
	CPP segment 3	2.770	.102
	LH segment 3	.688	.410
	CPP F_0 segment 3	9.579	.003*
	CPP segment 4	6.048	.017*
	LH segment 4	.463	.499
	CPP F_0 segment 4	3.238	.077
	CPP segment 5	2.536	.117
	LH segment 5	.590	.446
	CPP F_0 segment 5	7.377	.009*

* Significant Difference or alpha ≥ .05

Figure 3-3. Estimated Marginal Means of CPP for Connected Speech Segment 1. Gender 1 = males speakers, Gender 2 = female speakers. For age groups, 1 = 20-30 years of age, 2 = 40-50 years of age.

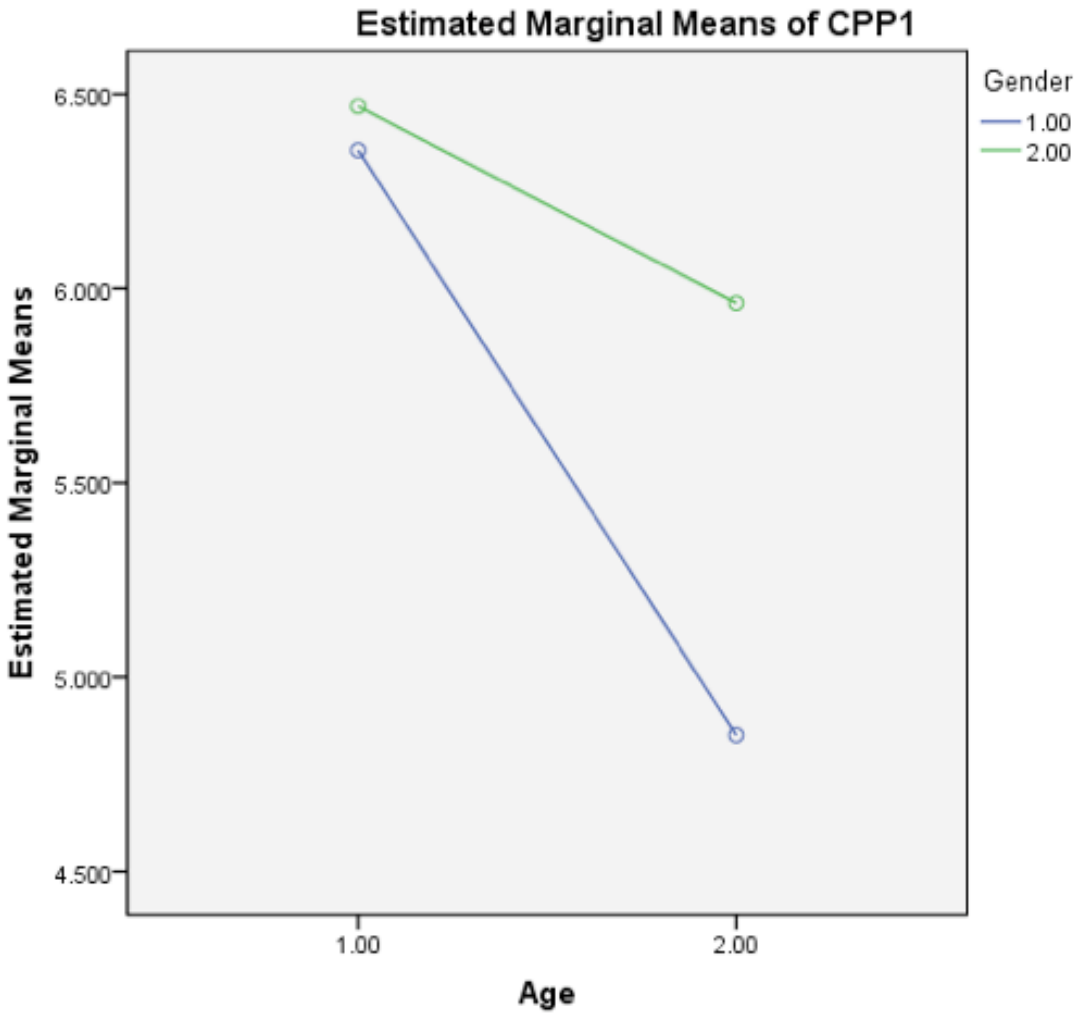


Figure 3-4. Estimated Marginal Means of CPP for Connected Speech Segment 4. Gender 1 = males speakers, Gender 2 = female speakers. For age groups, 1 = 20-30 years of age, 2 = 40-50 years of age.

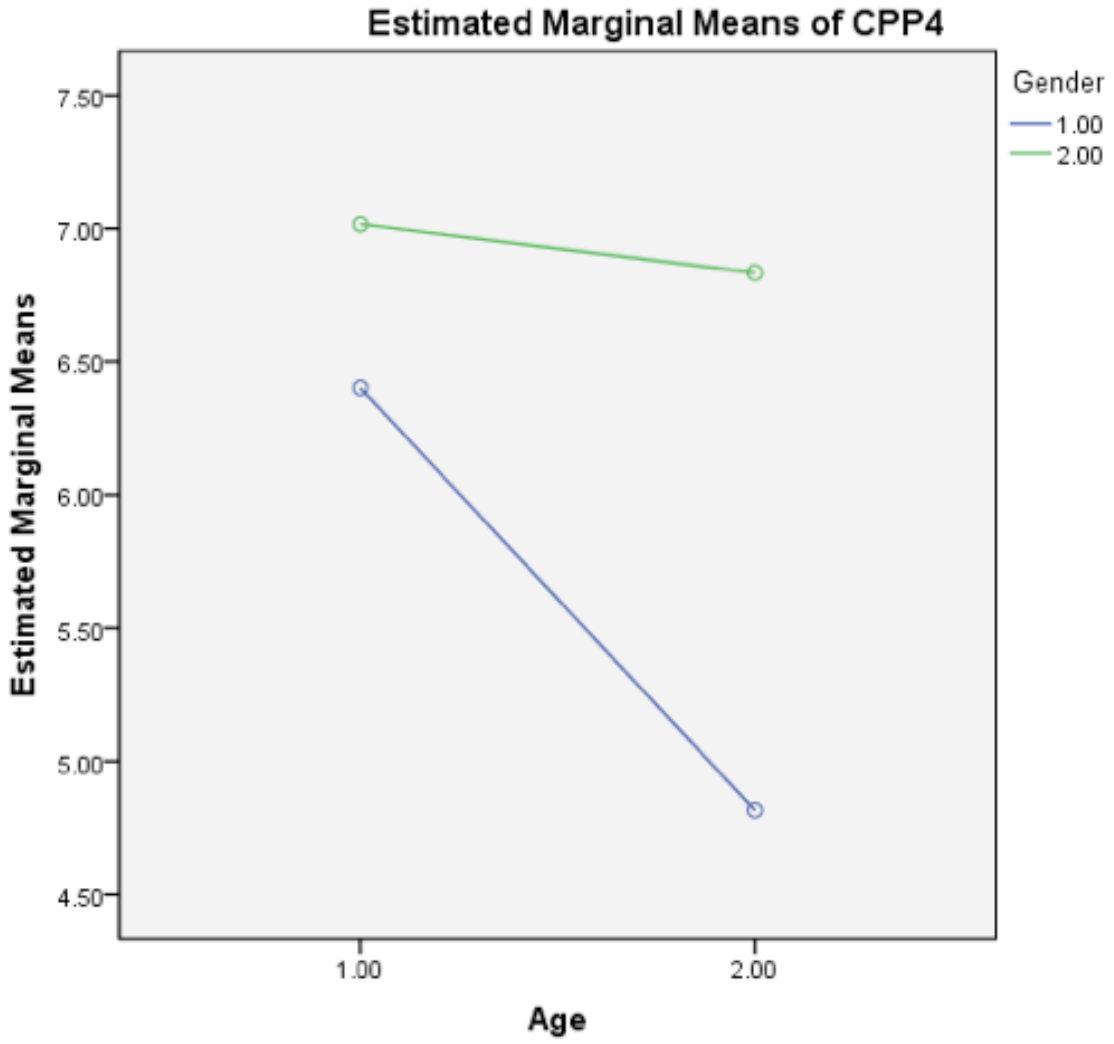


Figure 3-5. Estimated Marginal Means of CPP F_0 for Connected Speech Segment 1. Gender 1 = males speakers, Gender 2 = female speakers. For age groups, 1 = 20-30 years of age, 2 = 40-50 years of age.

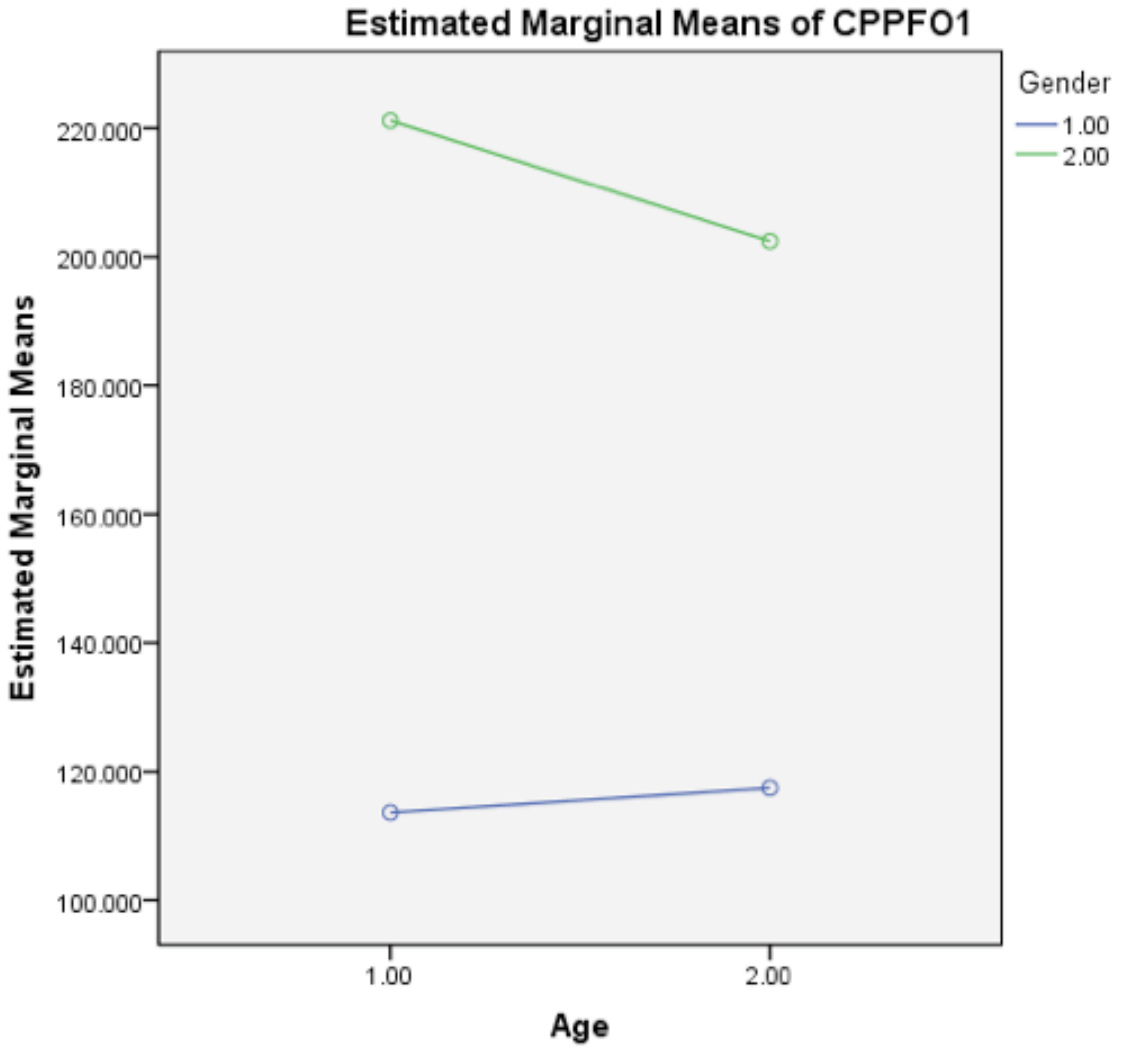


Figure 3-6. Estimated Marginal Means of CPP F_0 for Connected Speech Segment 2. Gender 1 = males speakers, Gender 2 = female speakers. For age groups, 1 = 20-30 years of age, 2 = 40-50 years of age.

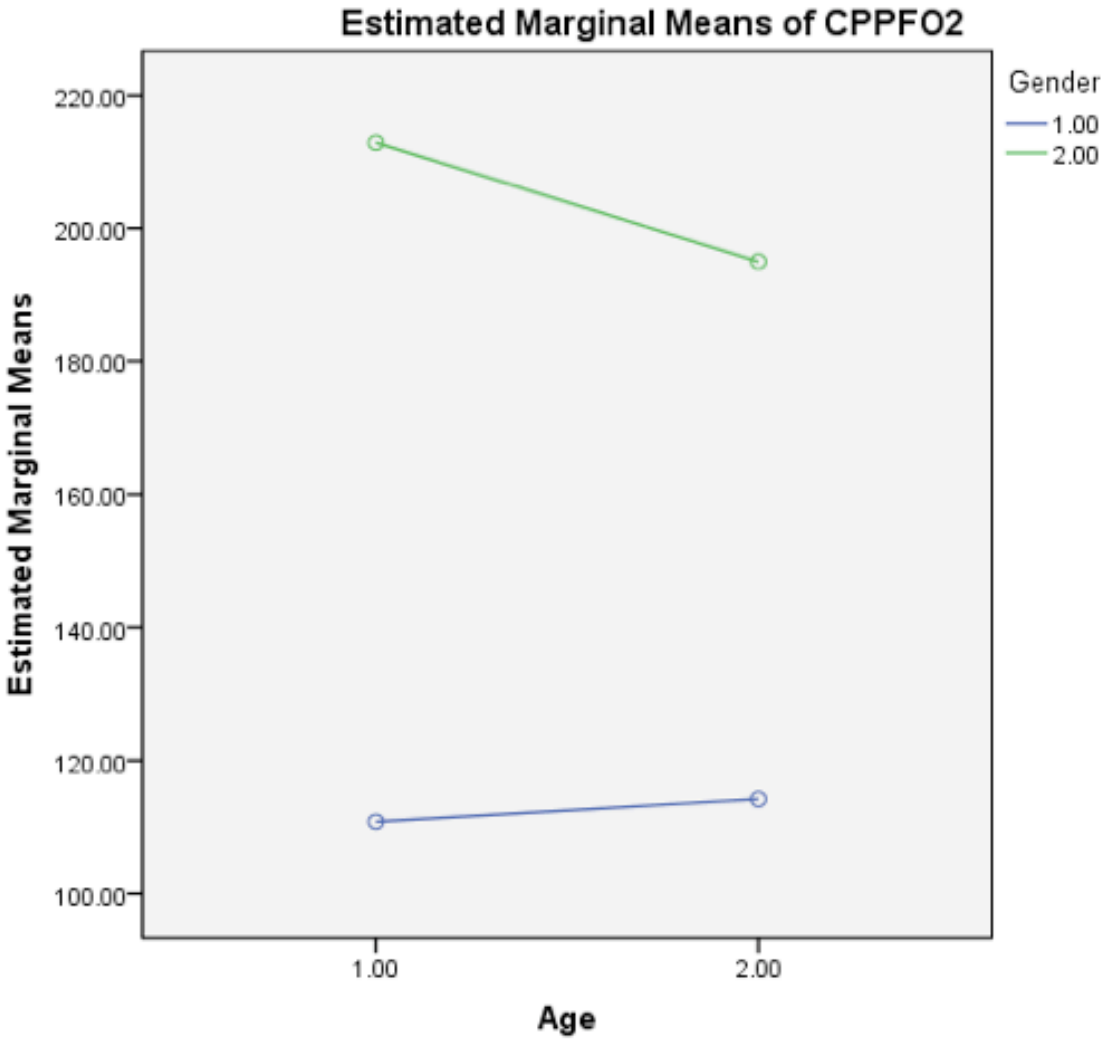


Figure 3-7. Estimated Marginal Means of CPP F_0 for Connected Speech Segment 3. Gender 1 = males speakers, Gender 2 = female speakers. For age groups, 1 = 20-30 years of age, 2 = 40-50 years of age.

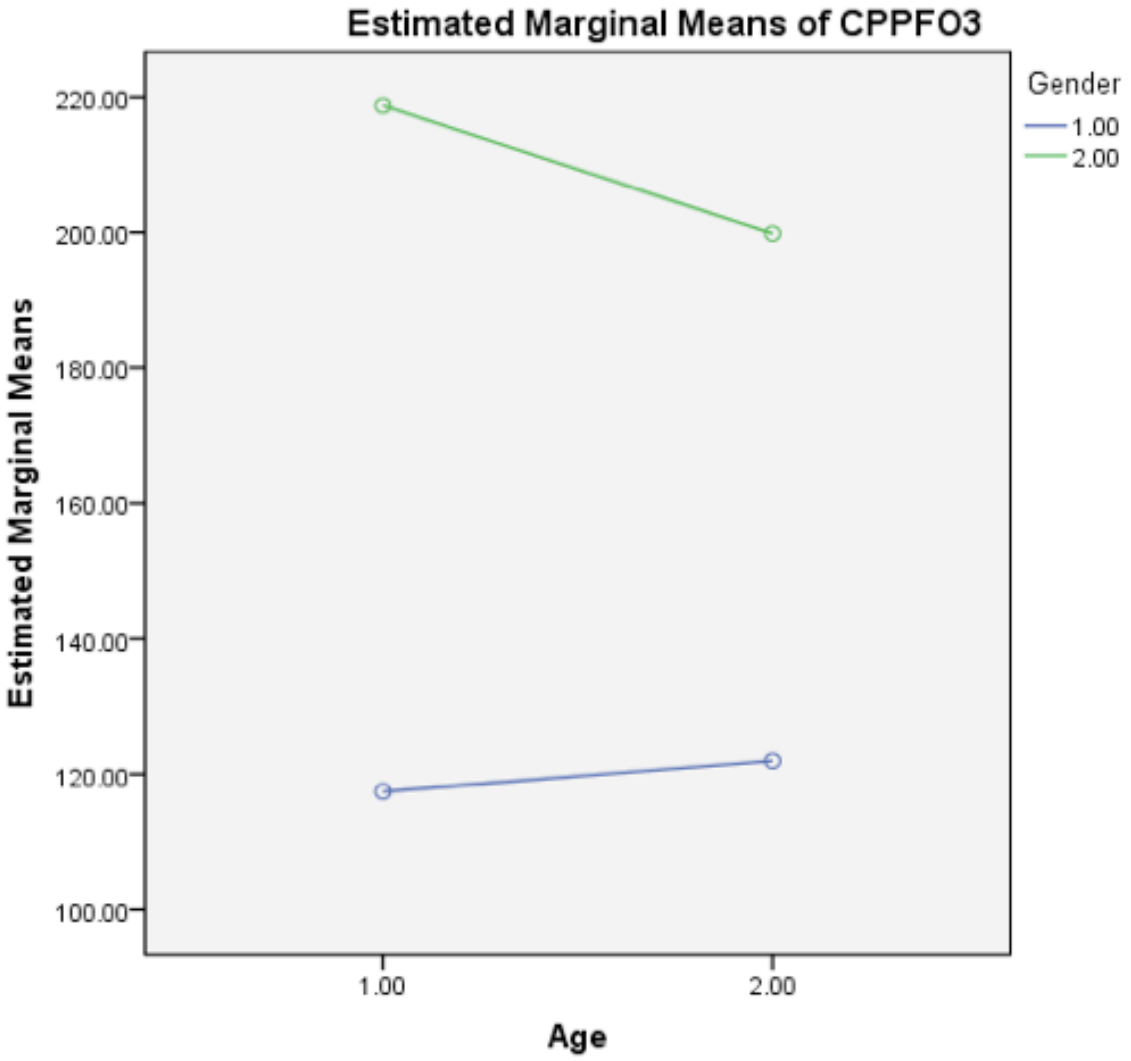
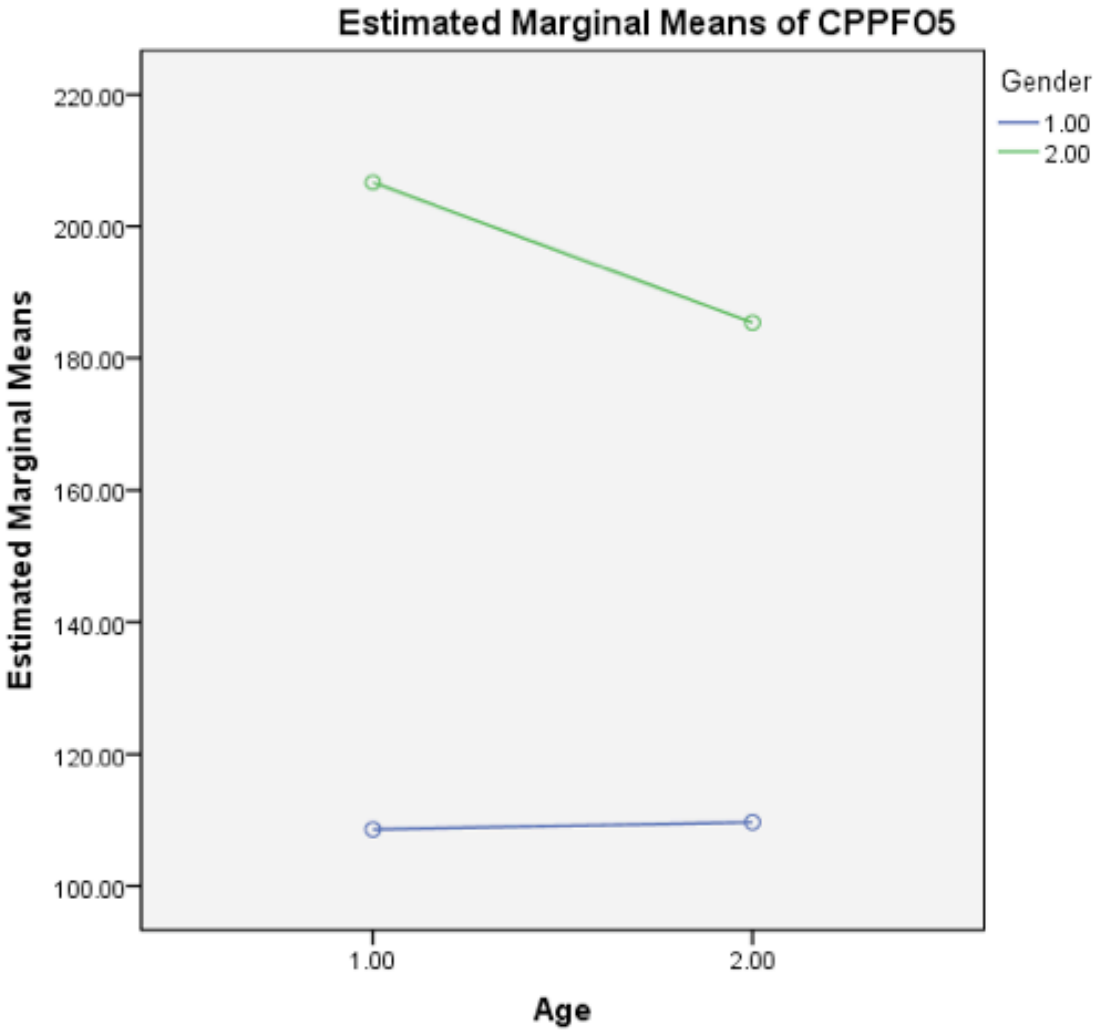


Figure 3-8. Estimated Marginal Means of CPP F_0 for Connected Speech Segment 5. Gender 1 = males speakers, Gender 2 = female speakers. For age groups, 1 = 20-30 years of age, 2 = 40-50 years of age.



Discussion

A review of recent literature suggested that cepstral- and spectral-based acoustic measures showed good potential as objective measures of dysphonia for clinical application. Therefore, the purpose of this study was to provide normative data for Long-Term Average spectral- and cepstral-based measures for both men and women in two different age groups to aid clinicians with assessing and treating voice disorders.

The first two research questions asked what the expected values of Cepstral Peak Prominence (CPP), Low-to-High Spectral Ratio (L/H spectral ratio), and Cepstral Peak Prominence Fundamental Frequency (CPP F_0) were for men and women with normal voices, ages 20-30 and 40-50 years. The third and fourth research questions addressed whether or not significant differences were present in CPP, L/H spectral ratio and CPP F_0 as a function of age, gender, or an age x gender interaction.

Results of this study showed that *gender* affected all the above-mentioned dependent variables, for both vowels and connected speech segments. Vowel results were more clear: male participants had significantly better voice quality as measured by CPP and L/H spectral ratio for both the vowels /a/ and /i/. Connected speech results were somewhat more difficult to interpret. In general, women had higher CPP values, denoting better voice quality in females; while men had higher L/H spectral ratio values, denoting better voice quality in males. It is not certain why these discrepant results were obtained, although they may relate to methodological factors (see Limitations below).

Age did not appear to have a significant effect on the dependent measures for the vowels /a/ and /i/; however, for connected speech, age appeared to have a significant effect on CPP for all 5 connected speech segments. Specifically, CPP was significantly better for younger speakers compared to older speakers, indicating better voice quality in the younger age group. This finding is generally consistent with previous research (see Relationship to Previous Research below), although it was not consistent with the investigator's perceptual impressions, especially of generally poor quality in young females (see discussions below).

Although not a voice quality measure, CPP F_0 was also significantly affected by age in connected speech segment 5 and there were several significant *age x gender* interactions for CPP F_0. In general, for both vowels and connected speech segments, younger women had markedly higher CPP F_0 values than older women, while older men had slightly higher CPP F_0 values compared to younger men. It is not surprising that younger women had higher fundamental frequencies in vowels and connected speech compared to older women, but the finding that younger men had slightly lower fundamental frequencies than older men of 40-50 years of age in vowels and connected speech was somewhat surprising. This finding is not consistent with previous research (see Relationship to Previous Research below).

Relationships Between Informal Perceptual Assessment and Acoustic Measures

During the subject selection phase of the study, both the investigator and the advisor independently rated the voice quality of all speakers on a scale of 1 (smooth, resonant) to 7 (rough, breathy). For each speaker, each speech sample was rated (2 vowels, 5 connected speech segments), and then all 7 ratings were averaged for the advisor and investigator separately, and compared. The purpose of this procedure initially was to ensure that all speakers had normal voice quality as rated by both the investigator and advisor. However, given the results of this study for acoustic measures, the investigator's and advisor's perceptual ratings were averaged together, to see if they could provide some insight into the acoustic results of the study.

Averaged perceptual ratings for the four groups were as follows: young females – 1.35; older females – 1.23; young males – 1.13; older males – 1.14. These preliminary perceptual ratings support the results of acoustic analyses showing that men had better

vocal quality than women, at least in vowels. However, these preliminary perceptual ratings do not support the finding that younger speakers generally had better voice quality than older speakers in connected speech. Clearly, more research in the relationship between measures of CPP and L/H spectral ratio compared to perceptual judgments in needed.

Limitations

As with every research study, there were some limitations in the current research. One concern was the voice quality of the young female group, which was perceptually poorer as judged by the investigator and advisor than any of the other three groups. Five young female participants had to be replaced due to deviant laryngeal quality. Of the other age/gender groups, replacements for dysphonic voices occurred only twice for the young males, once for the older males, and not at all for the older females. Even with the replacement of the worst five young female speakers with other subjects with better voice quality, the average voice quality rating for the young female group was worse (1.35) than any of the other groups (1.23, 1.13 and 1.14 respectively for older females, young males and older males). Coaching during the recording process, which included reminders about proper breath support, was necessary for majority of young female subjects.

The poor voice quality seen in young females appeared to be related to habitual use of vocal fry phonation. This observation is consistent with reports by other researches within the voice community (Wolk, Abdelli-Beruh, & Slavin, 2011; Gottliebson, Lee, Weinrich, & Sanders, 2007). Alternatively, it may be that although all speakers were

recruited randomly, perhaps our young females subjects were not representative of normal population due to the prevalence of overall deviant voices.

Another issue that arose during recording of the subjects concerned maintaining adequate intensity to meet the requirements of the ADSV program, while attempting to control speakers' intensity level. All recordings utilized for the study were produced by speakers at a 75 dB (± 2 dB) peak intensity level as measured by a sound level meter set on Weighting Network C, in order to ensure that one-third to one-half of the intensity range of the ADSV program was utilized, per manual instructions (Awan, 2011). In fact, many speakers were often too loud for the ADSV dynamic range, with intensities beyond the maximum limits of the system. Male subjects had greater difficulty staying within the dynamic range than females, with older males having the most difficulty. In addition, some connected speech segments were more difficult than others for all subjects to keep from exceeding the maximum ADSV intensity range (on e.g., "How hard did he hit him" and the Rainbow Passage), possibly due to the phonetic characteristics of the stimuli. This resulted in multiple recording re-takes to reach the optimal intensity range. While controlling speaking intensity was seen as important in obtaining valid data that permits comparison between subjects, this procedure may have introduced the risk of changing the speakers' normal productions.

Another limitation was in relation to the data analysis parameter specification for ADSV. The recommendation of Awan (2011) was to use an extraction range of 0-300 Hz (the default) for both men and women. However, after reviewing the normative data that was included in the manual of the program (Awan, 2011) along with our own results, it was apparent that the connected speech CPP F_0 values of male participants cited in the

manual were both inconsistent and higher than what was expected based on vowel data. For example, male CPP F_0 for the vowel /a/ was 110.89 Hz (well within normal limits), while for connected speech segments 3, 4, and 5, CPP F_0 values were 143.24 Hz, 160.51 Hz, and 133.12 Hz respectively (Awan, 2011). The latter three values are considerably above the typical "average" values for a group of males. No mention of these discrepant "normative" values was made in the manual; however it was noted that males with low-frequency voices might require adjustment of the maximum value of the default extraction range from 300 Hz to 200 Hz (Awan, 2011, p. 40). Our preliminary data analysis showed that the 0-200 Hz setting seemed to produce more consistent CPP F_0 results across speech samples than the 0-300 Hz setting did. Therefore, to ensure accurate CPP F_0 data for all of our male speakers, the maximum value of the default extraction range was set to 200 Hz, although the ultimate effects of this change on subsequent CPP data were unknown.

In order to assess validity of the CPP F_0 measure for males in connected speech, the present investigator and advisor analyzed 36.7% of the male connected speech data with the Real Time Pitch (RTP) sub-program of Multi-Speech to obtain a second measure of fundamental frequency (F_0) on a well-known and well-accepted pitch extraction program. This RTP F_0 value was correlated with the CPP F_0 value. The resulting correlation coefficient between the two fundamental frequencies was $r = .918$ ($p < .01$). This absolute value of the differences between RTP F_0 and CPP F_0 was 5.54 Hz. Thus, changing the upper limit of the extraction range from 300 Hz to 200 Hz appeared to result in more relatively accurate F_0 data (compared to the data presented in the manual). However, as mentioned above, the full effect of changing the maximum extraction range

is not known. The change in extraction range parameters may have led to unintended alterations in CPP and L/H spectral ratio values. More research is needed in this area.

Relationship to Previous Research

Normative data results. Normative data for cepstral- and spectral-based measures have not been previously established. The only other available form of normative data for these measurements were the non-peer reviewed data sets included in the ADSV manual (Awan, 2011). To begin with, Awan (2011) did not include data for the sustained vowel /i/, and grouped all females (N=50) together regardless of age. The same procedure was followed for male subjects (N=50). The only information provided about the data collection was that the subjects originated from North America and ranged from ages 21 to 45 years (both males and females), and that the default settings were used during analysis.

Despite the differences in methods between Awan (2011) and the current study, for the vowel /a/, the results of Awan (2011) and present research were similar. Awan (2011) showed females averaging a CPP of 10.74 dB, and males averaging a CPP of 13.03 dB. The data from this study had females and males averaged across age as 10.929 dB and 12.544 dB respectively. L/H spectral ratio values were consistent between both studies as well, with males averaging higher than females. Awan (2011) found a L/H ratio value of 32.99 dB for females in /a/, and 38.12 dB for males; while in the present study, females averaged over age had an L/H ratio of 31.595 for /a/, with males measuring 34.919 dB. The major difference between the two studies that was consistent for all 5 connected speech segments was that Awan's (2011) males obtained better CPP *and* L/H spectral ratio values than his female subjects. In the current study, females

consistently performed higher for CPP while males performed higher for L/H spectral ratio for the connected speech stimuli. However, when comparing averages, no markedly different values were observed between the two studies. The values obtained in the two studies were generally within one standard deviation of one another.

In addition to Awan (2011), descriptive statistics of control groups were provided in a few previous studies concerning cepstral and spectral measures for dysphonic speakers. As shown in Table 4-1, Watts and Awan (2011) had comparable values to those found in this research: the control group (including both males and females) had a CPP average of 11.08 dB for the vowel /a/ and L/H spectral ratio average of 32.60 dB, while this study had a CPP average of 11.736 dB and an L/H spectral ratio average of 33.257 dB, when male and female results were averaged. Although their data from the Rainbow Passage included only the 2nd sentence, Watts and Awan (2011) obtained results similar to those of the present study. For all participants of the present research, the CPP average was 6.602 dB and L/H spectral ratio was 32.906 dB for the Rainbow Passage, with Watts and Awan (2011) reporting a CPP value of 5.42 dB and a L/H spectral ratio value of 30.74 dB. For other normative data on CPP, see Table 4-1.

Table 4-1. Cepstral Peak Prominence (CPP) values for speakers with normal voices. Standard deviation in parentheses.

Study and Subjects	/a/	2nd Sentence of Rainbow Passage
Watts & Awan (2011) 5 males, 11 females mean age = 53 yrs	11.08 dB (1.91)	5.42 dB (1.38) (without vocalic detection)
Lowell et al. (2011) 16 males, 11 females mean age = 39 yrs	--	7.81 dB (0.77) (with vocalic detection)
		6.35 (0.69) (without vocalic detection)
Garrett (2013) 30 males, 30 females mean age = 34 years	11.74 dB (1.81)	2nd & 3rd Sentences of the Rainbow Passage
		6.60 dB (1.16) (with vocalic detection)

Results for age effects. According to the present study, CPP was significantly better for younger speakers compared to older speakers, which indicates better voice quality in the younger age group. To corroborate this finding, research from time-based studies was examined. In one example, Gorham-Rowan and Laures-Gore (2006) compared voice qualities using Noise-to-Harmonic ratio (NHR), Amplitude Perturbation Quotient (APQ), and H1-A1 measures of 28 young women (age = 24.7), 28 young men (age = 25.4), 28 elder women (age = 70.7), and 28 elderly men (age = 69.6). Although results were not completely consistent across age and gender groups, younger subjects in general had better voice quality than the older subjects as measured by cycle-to-cycle measurements by Gorham-Rowan and Laures-Gore (2006).

Results for changes in fundamental frequency. After reviewing the CPP F_0 results of the present study, it was surprising to find that younger men had slightly lower fundamental frequencies than older men in both vowels and connected speech. Studies summarized by Baken & Orlikoff (2000) show that males' Speaking Fundamental

Frequency (SFF) decreases from young adulthood through the ages of 50-60 years, and subsequently increases. For example, Hollien and Shipp (1972) found that men ages 20-29 years had a mean SFF of 119.5 Hz, while men ages 40-49 years had a mean SFF of 107.1 Hz. Based on this research, we might have expected CPP F_0 to show a decrease between the younger male group in this study (20-30 yrs) and the older male group (40-50 yrs), but in fact CPP F_0 was seen to increase between the younger and older male groups in this study. The reason for this unexpected finding may be related to some idiosyncratic characteristic of the current study's male subjects (higher education level?), but remains unknown.

Clinical Implications

First it is the recommendation of this study that clinicians consider changing the maximum limit of the ADSV extraction range for male participants from 300 Hz to 200 Hz for connected speech readings. We make this recommendation because according to the present study, using the 200 Hz limit results in accurate F_0 data (i.e., CPP F_0 for males correlated strongly with RTP F_0 data for males), whereas using the 300 Hz limits results in apparently questionable CPP F_0 data as presented by Awan (2011). Since it would seem important to get an accurate CPP F_0 measurement in order to accurately identify CPP, we recommend using the 200 Hz limit. The clinician, however, would need to be careful to apply the most relevant norms. If in a clinical setting there is any question the accuracy of CPP F_0, RTP should also be utilized to measure fundamental frequency, and if there are discrepancies, the results of the ADSV analysis should be considered conditional.

Second, care needs to be taken to not exceed amplitude limits of ADSV. Clinicians may need to change mouth-to-microphone distance and/or alter the input volume on the preamplifier or sound card, in order to stay in what Awan (2011) assumes is the most optimal intensity range for ADSV analysis. Third, due to the limited research on the effects of using the vocalic detection procedure, data should be analyzed both with and without vocalic detection until it becomes clear which one is more valid. Although using vocalic detection as part of the analysis procedure may increase face validity, it may cause the analysis to be less accurate in assessing noise in the voice.

The results of the present study suggested significant differences in CPP and L/H spectral ratio based on both gender (primarily) and age (to a lesser degree). Therefore separate normative data for all four age/gender groups should be used in clinical applications. As shown in Tables 3-1 through 3-21, normative data should be organized by vowels and connected speech segments as a function of age and gender. Furthermore, within each dependent variable/vowel combinations, it may be helpful to average across age, gender, and age and gender combined, along with standard deviations.

Implications for Future Research

This study has provided important data for cepstral- and spectral-based normative measures for both men and women. However, further research is needed to investigate whether or not higher CPP and L/H spectral ratio values for male voices compared to female voices during the vowel productions is a robust effect. Perceptual judgments of the researchers support this finding, but further perceptual ratings by a larger group would be beneficial. Moreover, additional studies should further examine the discrepancy between CPP values and L/H spectral ratio values in connected speech for males and

females, to determine whether consistent differences in CPP, L/H spectral ratio and perceived voice quality are present in connected speech, and which gender has the better ratings.

In addition, more research is needed to examine the potential usefulness of the vocalic detection routine. It is possible that use of the program setting of vocalic detection might have removed the deviant voice qualities of some speakers, which may have caused the resulting measures to be unrepresentative of the speaker's true voice quality. This could be the reason why the young females obtained similar CPP values as their older counterparts, even though the perceptual ratings of the researchers suggested poorer voice quality. Finally, there is a need to investigate the effect of changing the extraction range for male speakers. Changing the extraction range may have led to unknown alterations in CPP and L/H spectral ratio values. The use of spectral and cepstral measures in clinical voice analysis appears to be promising, but many procedural questions remain before these measures can be confidently used by clinicians.

References

American Speech-Language-Hearing Association (ASHA; 1997). Guidelines *for audiological screening.* Rockville, MD: Author.

Awan, S. N. (2011). *Analysis of dysphonia in speech and voice: an application guide.* Montvale, NJ: KayPENTAX.

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

Awan, S. N., Roy, N., Jette, M. E., Meltzner, G. S., & Hillman, R. E. (2010). Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: comparisons with auditory perceptual judgments from the CAPE-V. *Clinical Linguistics & Phonetics 24*(9): 742-758.

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

Behrman, A. (2007). *Speech and voice science.* San Diego, CA: Plural Publishing, Inc.

Boone, D. R. (1991). *Is your voice telling on you?* San Diego, CA: Singular Publishing Group, Inc.

Boone, D. R. McFarlane, S. C., Von Berg, S. L., Zraick, R. L. (2010). *The voice and voice therapy* (8th ed.). Needham Heights, MA: Allyn and Bacon.

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

Colton, R., Casper, J. & Leonard, R. (2011). *Understanding voice problems: A physiological perspective for diagnosis and treatment* (4th ed.). Philadelphia: Lippincott Williams and Wilkins.

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

Duffy, J. R. (2005). *Motor speech disorders: substrates, differential diagnosis, and management* (2nd ed.). St Louis: Mosby.

Fairbanks, G. (1960). Voice and articulation drill book. New York: Harper and Brothers.

Ferrand, C. T. (2007). *Speech science: an integrated approach to theory and clinical practice* (2nd ed.). Boston: Allyn and Bacon.

Ferrand, C. T. (2012). *Voice disorders: scope of theory and practice.* Boston, Allyn and Bacon.

Gary, S. D. (2000). Cellular physiology of the vocal folds. *Otolaryngologic Clinics of North America 33*(4): 679-98.

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

Gottliebson, R. O., Lee, L., Winrich B., & Sanders, J. (2007). Voice problems of future speech-langauge pathologists. *Journal of Voice 21*: 699-704.

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

Hirano, M. (1974). Morphological structure of the vocal cord as a vibrator and its variations. *Falia Phoniatrica et Logopedica 26,* 89-94.

Hixon, T. J., Weismer, G., & Hoit, J. D. (2008). *Preclinical speech science: anatomy, physiology, acoustics, perception.* San Diego, CA: Plural Publishing.

Hollien, H. & Shipp, T. (1972). Speaking fundamental frequency and chronological age in males. *Journal of Speech and Hearing Research 15,* 155-159.

KayPENTAX. (2008). *The software instruction manual of the multi-dimensional voice program (MDVP) model 5105*. Lincoln Park, NJ: KayPENTAX.

Kempster, G. B., Gerratt, B. R., Verodolini Abbott, K., Barkmeier-Kramer, 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. (2000). *Voice quality measurement*. San Diego, CA: Singular Publishing Group.

Lowell, R. Y., Colton, R. H., Kelley, R. T., Hahn, Y. C. (2011). Spectral- and cepstral-based measures during continuous speech: capacity to distinguish dysphonia and consistency within a speaker. *The Journal of Voice 25* (5), 223-232.

Martin, D., Fitch, J., & Wolfe, V. (1995). Pathologic voice type and the acoustic prediction of severity. *Journal of Speech and Hearing Research 38*(4): 765-772.

Seikel, J. A., King, D. W., & Drumright, D. G. (2010). *Anatomy & physiology for speech, language and hearing* (4th ed.). Clifton Park, NY: Thomson Delmar Learning.

Van den Berg, J. (1958). Myoelastic-aerodynamic theory of voice production. *Journal of Speech and Hearing Research 3*(1): 227-44.

Watts, C. R., & Awan, S. N. (2011). Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts. *Journal of Speech, Language, and Hearing Research 54*: 1525-1537.

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

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

Wolk, L., Abdelli-Beruh, N. B., & Slavin, D. (2011). Habitual use of vocal fry in young adult female speakers. *Journal of Voice 26*, 3: 111-116.

Zemlin, W. R. (1998). *Speech and hearing science anatomy and physiology* (4th ed.). Boston: Allyn and Bacon.

APPENDIX A

Participant Eligibility Criteria

Participant Name: _____ Date: _____

Gender: M / F

Yes___ No ___ Are you between the ages of 20-30 years?

Yes___ No ___ Are you between the ages of 40-50 years?

Yes___ No ___ Are you a native speaker of English?

Yes___ No ___ Do you have any history of voice problems, such as hoarseness or loss of voice?
If yes, please provide an explanation, the type, and the time frame:

Yes___ No ___ Do you have any history of speech and/or language therapy?
If yes, please provide an explanation, the type, and the time frame:

Yes___ No ___ Do you have any history of neurological problems, such as a head trauma, stroke or an aneurysm?
If yes, please provide an explanation, the type, and the time frame:

Yes___ No ___ Do you have any history of hearing loss?
If yes, please describe:

Yes___ No ___ Do you currently smoke?

Yes___ No ___ Have you ever smoked?
If yes, when and how long:

PASS FAIL

Date of Appointment:
Extra Credit:
Notes:

APPENDIX B

Consent Form

UNIVERSITY OF WISCONSIN – MILWAUKEE
CONSENT TO PARTICIPATE IN RESEARCH
SPEAKER PARTICIPANT CONSENT

THIS CONSENT FORM HAS BEEN APPROVED BY THE IRB FOR A ONE YEAR PERIOD

General Information

Study title: Cepstral- and Spectral-Based Acoustic Measures of Normal Voices

Person in Charge of Study (Principal Investigator):
My name is Rachel Garrett, and I am a graduate student in the Department of Communication Sciences and Disorders. I am completing this study for my thesis research. My advisor is Dr. Marylou Pausewang Gelfer, a faculty member in the department. We will be the people interacting with you during your participation in this study.

Study Description

You are being asked to participate in a research study. Your participation is completely voluntary. You do not have to participate if you do not want to.

Study description:
The purpose of this study is to learn more about the digital measures we can make that describe a person's voice. For example, we might want to use digital measures to determine how hoarse a person's voice is. Normative data must be collected so speech-language pathologists can compare clients with potential voice disorders to the data of normal speakers. In this study, we want to determine what measures are typical of normal voices.

This research is being done to help speech-language pathologists use digital measures of voice to better diagnose and treat individuals with voice disorders. This study, along with other similar studies, will serve to provide an evidence base for the practice of speech-language pathology.

This study will be conducted at the UWM Speech and Language Clinic on the 8th floor of Enderis Hall. Approximately 60 adults ages 20-30 years and 40-50 years will participate in the study. Your participation in the study will take about 30-40 minutes in total, over the course of one day.

Study Procedures

What will I be asked to do if I participate in the study?
If you agree to participate, you and other adult speakers will be asked to do the following:

1. Participate in a hearing screening: This will consist of answering some questions about your hearing, having the Student Principle Investigator look in your ears with a small flashlight, and raising your hand in response to a series of quiet sounds.
2. Participate in a voice screening: This will involve saying the vowels "ee" and "ah," reading six sentences, and providing a short speech sample.
3. Participate in the experimental procedure: This will involve again saying the vowel "ee" and "ah" for about 3 seconds each at a specific loudness level; reading four sentences out loud, also at a specific loudness level; and reading a 2-sentence passage at a specific loudness level.

With your permission, we will digitally record your voice during the activities on a computer. The recording will be done to make sure we can accurately measure your voice.

Risks and Minimizing Risks

What risks will I face by participating in this study?
The potential risks for participating in this study are minimal to none – no greater than what you would experience during everyday speech.

1. **Psychological:** There is a small possibility that you may feel embarrassed by providing the voice samples. However, you can be sure that we will keep your data confidential, and that only myself and my Faculty Advisor will have access to it.
2. **Psychological:** You may feel concerned if you fail either the hearing screening test or the voice screening test. To address your concerns, we can refer you for further evaluation and possible services to Community Audiology Services (for hearing concerns) or the UWM Speech and Language Clinic for voice concerns.

Benefits

Will I receive any benefit from my participation in this study?
There are no direct benefits to you other than to further research.

Study Costs and Compensation

Will I be charged anything for participating in this study?
You will not be responsible for any cost of taking part in this research study.

Are subjects paid or given anything for being in the study?
Your instructor may choose to give you extra credit for participating in this study, but many instructors do not offer this option.

Confidentiality

What happens to the information collected?
All information collected about you during the course of this study will be kept confidential to the extent permitted by law. We may decide to present what we find to others, or publish our results in scientific journals or at scientific conferences, but your data will never be linked to your name or any other information about you. Only the PI and her faculty advisor will have access to your personal information. When the study is over, all your personal information will be destroyed, and the files of your voice recordings will be deleted.

Alternatives

Are there alternatives to participating in the study?
There are no known alternatives available to you other than not taking part in this study.

Voluntary Participation and Withdrawal

What happens if I decide not to be in this study?
Your participation in this study is entirely voluntary. You may choose not to take part in this study, or if you decide to take part, you can change your mind later and withdraw from the study. You are free to not answer any questions or withdraw at any time. Your decision will not change any present or future relationships with the University of Wisconsin Milwaukee. The investigator may stop your participation in this study if we feel it is necessary to do so.

If you decide to withdraw or if you are withdrawn from the study before it ends, we will use the information we collected up to that point.

Questions

Who do I contact for questions about this study?
For more information about the study or the study procedures or treatments, or to withdraw from the study, contact:

Marylou Pausewang Gelfer, Ph.D.
Department of Communication Sciences and Disorders
University of Wisconsin – Milwaukee
P.O. Box 413
Milwaukee, WI 53201
(414) 229-6465

Who do I contact for questions about my rights or complaints towards my treatment as a research subject?
The Institutional Review Board may ask your name, but all complaints are kept in confidence.

Institutional Review Board
Human Research Protection Program
Department of University Safety and Assurances
University of Wisconsin – Milwaukee
P.O. Box 413
Milwaukee, WI 53201
(414) 229-3173

Signatures

Research Subject's Consent to Participate in Research:
To voluntarily agree to take part in this study, you must sign on the line below. If you choose to take part in this study, you may withdraw at any time. You are not giving up any of your legal rights by signing this form. Your signature below indicates that you have read or had read to you this entire consent form, including the risks and benefits, and have had all of your questions answered, and that you are 18 years of age or older.

_____
Printed Name of Subject/ Legally Authorized Representative

_____ _____
Signature of Subject/Legally Authorized Representative Date

Research Subject's Consent to Audio/Video/Photo Recording:

It is okay to audiotape me and use my audiotaped data in the research.

Please initial: ____Yes ____No

Principal Investigator (or Designee)
I have given this research subject information on the study that is accurate and sufficient for the subject to fully understand the nature, risks and benefits of the study.

_____ _____
Printed Name of Person Obtaining Consent Study Role

_____ _____
Signature of Person Obtaining Consent Date

APPENDIX C

Hearing Screening

Participant Name: _____ Date: _____

Case History-circle appropriate answers

Do you think you have a hearing loss?			Yes	No
Have hearing aid(s) ever been recommended for you?			Yes	No
Is your hearing better in one ear?			Yes	No
If yes, which is the better ear?	Left	Right		
Have you ever had a sudden or rapid progression of hearing loss?			Yes	No
If yes, which ear?	Left	Right		
Do you have ringing or noises in your ears?			Yes	No
If yes, which ear?	Left	Right		
Do you consider dizziness to be a problem for you?			Yes	No
Have you had recent drainage from your ear(s)?			Yes	No
If yes, which ear?	Left	Right		
Do you have pain or discomfort in your ear(s)?			Yes	No
If yes, which ear?	Left	Right		

Have you received medical consultation for any of the above conditions?

PASS **REFER**

Visual/Otoscopic Inspection

Referral for cerumen management_____ Referral for medical evaluation_____

PASS **REFER**

Pure-Tone Screen (25 db HL) (R=Response, NR = No Response)

Frequency	1000 Hz	2000 Hz	4000 Hz
Right Ear			
Left Ear			

PASS **REFER**

APPENDIX D

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

Participant Name: _____ Date: _____

Please complete the following tasks:

1. Hold out the vowels /a/ and /i/ for 3-5 seconds each.
2. Say the following sentence:
   a. The blue spot is on the key again.
   b. How hard did he hit him?
   c. We were away a year ago.
   d. We eat eggs every Easter.
   e. My mama makes lemon muffins.
   f. Peter will keep at the peak.
3. Provide a ~15 seconds long response to "Tell me about your major."

Check all that apply:

___Appropriate pitch

___Appropriate loudness

___Appropriate resonance

___Appropriate laryngeal quality

Problems noted:

___Rough

___Breathy

___Strained

___Appropriate articulation

Problems noted:

___ s, z

___ r

___ l

___ sh, ch

PASS **FAIL**

APPENDIX E

Sentence Stimuli

Say the following sentences:

1. How hard did he hit him?
2. We were away a year ago.
3. We eat eggs every Easter.
4. Peter will keep at the peak.

APPENDIX F

Rainbow Passage Stimuli

Say the following sentences:

The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon.