12-9-2019 2:00 PM

VOice analysis with Iphones: a low Cost Experimental Solution

Benjamin van der Woerd
The University of Western Ontario

Supervisor
Fung, Kevin
The University of Western Ontario Co-Supervisor
Doyle, Philip
The University of Western Ontario Co-Supervisor
Parsa, Vijay
The University of Western Ontario

Graduate Program in Surgery
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science

Recommended Citation
van der Woerd, Benjamin, "VOice analysis with Iphones: a low Cost Experimental Solution" (2019). *Electronic Thesis and Dissertation Repository*. 6719.
https://ir.lib.uwo.ca/etd/6719

Abstract

Background:

Acoustic voice analysis requires a resource intensive setup, including a soundproof booth. This project evaluates smartphone microphone and recording environment impacts on voice sample collection for acoustic voice analysis. A proprietary analysis algorithm is presented for validation.

Methods:

Microphone and recording environment were evaluated using previously collected voice samples presented in four conditions to test two microphones and two recording environments. Prospective samples were used to test the proprietary algorithm, whereby samples were analyzed using this and Praat.

Results:

Microphone and recording environment had small, clinically unimportant impacts on most measurements. The proprietary algorithm reliably analyzed sustained vowels, with strong correlation to the Praat results. Continuous speech analysis was less reliable.

Conclusion:

Smartphone microphones are adequate for voice sample collection. Quiet, non-soundproof settings can be used for voice collection. The proprietary algorithm represents a reliable method to analyze sustained vowel samples. Some improvements are necessary before continuous speech analysis can be considered valid.

Lay Summary

Acoustic voice analysis is a part of a thorough voice examination, which can be used in conjunction with aerodynamic measurements, auditory-perceptual analysis, and patient reported outcomes. It is a method of objectively quantifying normal and pathologic voices. It relies on physical properties of voice and the resultant vocal signal.

The existing methods of voice sample collection and analysis are often resource intensive, creating barriers to access. In the current work, we sought to empirically examine a multi-phase study evaluating smartphones as an avenue to perform voice sample acquisition and on-device acoustic voice analysis.

In phase one, microphones for voice sample collection and the influence of recording environment were evaluated. This showed small, clinically unimportant differences in most acoustic measures for both the microphone and the recording environment. These results indicate that current barriers to voice signal acquisition can be relaxed upon, without degrading the ability to measure microacoustic changes. In conclusion, the results of this phase indicate that robust acoustic voice analysis is feasible on current day smartphones with samples collected under non-soundproof conditions.

In phase two, a proprietary application, developed for the purposes of this study, was compared against the current standard used for acoustic voice analysis (Praat). The results of this study provide data, based on a systematic acquisition methods, which supports future use of the current proprietary algorithm. In doing so, the current findings indicate that the present smartphone application could be used reliably by patients, healthcare providers, and researchers for acoustic voice analysis.

Keywords

Voice; Acoustic Voice Analysis; Voice Analysis; Mobile Health; mHealth

Acknowledgments

First and foremost, I want to thank my wife, Carlye, for her support of my goals in the pursuit of my master's. You have been incredibly gracious with your time and flexible as my schedule has shifted to accommodate this additional workload. I would not have been able to carry out this work without your support and love through it all.

To Dr. Kevin Fung, I want to thank you for your encouragement and excitement in this project. I am so grateful to you for taking on this additional role to supervise and mentor me through these past few years.

To Dr. Philip Doyle, your guidance and patience has been exceptional. I thank you for all the time and resources you provided over the past year. I am thankful for the opportunity to have worked together on this project.

To Dr. Vijay Parsa, without your help and guidance throughout the development of this application, this project would not have been possible. Thank you for all of the time spent together completing this experiment and the countless hours you've spent finetuning this application and algorithm.

To Min Wu, thank you for your work on the development of this application. Without your assistance, this project could not have been completed. Thank you for being gracious with your time and teaching me basics of SWIFT development.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1 Introduction

Voice disorders are a common and important clinical concern, affecting more than 10% of the population (Martins et al., 2016; Roy et al., 2004). The prevalence of voice disorders reportedly approaches 50% in professional voice users (Martins et al., 2016) although the frequency and severity of such disorders is undefined. The professional voice user group includes occupations where the voice is considered to be of primary importance (e.g., teachers, salespeople, etc.), a population that includes 5-10% of the general population (Titze, Lemke, & Montequin, 1997). Furthermore, clinical voice disorders evolve secondary to both benign and malignant conditions with benign lesions forming the substantial majority. Benign conditions include structural changes to the tissue of the vocal folds (e.g., vocal fold nodules, cysts, and polyps) which occur as a result of hyperfunctional disorders (Hillman, Holmberg, Perkell, Walsh, & Vaughan, 1989), as well as those for which the etiology is neurological (e.g., spasmodic dysphonia). Many of these pathologies may involve treatment with behavioral voice therapy, for which acoustic analysis can provide an objective measure of potential improvement (Hillenbrand, 2011; Holmberg, Hillman, Hammarberg, Södersten, & Doyle, 2001; Kono et al., 2016; Stepp, Merchant, Heaton, & Hillman, 2011). As for malignant conditions, laryngeal cancers represent the most common malignant lesions of the head and neck (Kasper et al., 2011). Importantly, voice changes are often one of the earliest signs of primary and recurrent laryngeal tumors (Chu & Kim, 2008). In Canada, the estimated incidence of new laryngeal cancers is 1,150 per year (Canadian Cancer Society, 2017). Perhaps of equal importance, the treatment of laryngeal cancers pays great attention to functional preservation, vocal function being of critical importance.

Acoustic voice analysis and auditory-perceptual measures are important parts of the clinical assessment for diagnostic and/or treatment progress tracking purposes (Behrman, 2005; Hillenbrand, 2011). Acoustic measures of voice identify vibratory abnormalities. This method of analyzing the vocal signal is a non-invasive, objective method to measure normal and pathologic voices and document potential abnormality in the vocal signal. Because normal voice is characterized by a quasi-periodic signal, several micro-acoustic measures have been described, including time-based (e.g., jitter, shimmer, and harmonic-to-noise ratio) and spectral-based (e.g., cepstral peak prominence, low frequency spectral energy : high frequency spectral energy ratio) (Awan & Roy, 2006, 2009).

The process of gathering signal samples for rigorous acoustic analysis must consider a variety of issues specific to signal collection; this includes consideration of the ambient noise within the recording environment, microphone calibration, and mouth-to-microphone distances, as well as other factors (ASHA, 2002; Zraick et al., 2011). Further, ideally the acquisition of voice samples may carry the requirements for a trained individual for voice sample collection, the need for a soundproof (or sound-treated) booth prepared for sample capture, and expensive instrumentation and software to perform the analyses. These concerns may, at times, limit the regular use of acoustic voice analysis and may often only be used to gather baseline measures. Clearly, a single baseline measure presents substantial limitations in the context of the ability to compare the vocal signal clinically in an effort to more comprehensively assess treatment outcomes.

Numerous studies identify acoustic measures as robust, objective ways to screen, diagnose, and track treatment progress of voice disorders (Awan & Roy, 2006, 2009; Behrman, 2005; Hillenbrand, 2011).

Mobile health (mHealth) is a rapidly growing field with broad applications to patient care (Ali, Chew, & Yap, 2016; Steinhubl, Muse, & Topol, 2015). Significant technological advances in mobile computing has made the requisite hardware readily available to clinicians and researchers, which may then serve to alleviate some of the current barriers to voice analysis.

Voice evaluation involves the comprehensive assessment of vocal capacity that includes acoustic analysis, aerodynamic assessments, auditory-perceptual analysis, behavioral, and self-reported outcome measures. However, each of these measures may play different roles in developing a well-rounded assessment of voice and isolated measures may be exploited in some circumstances to provide documentation of voice change. Samples collected for acoustic analysis may also be used in the collection of further voice information through these other forms of assessment. That is, data obtained from acoustic measures of voice may provide the impetus to pursue additional, specific measures of vocal function. Acoustic analysis represents an important piece of a comprehensive voice assessment. When used in combination with other voice assessment tools, a well-rounded evaluation can be accomplished. Objective measures that correlate strongly with perceptual scores can be used to quantify findings in dysphonic voices and clinical voice disorders.

1.1 Clinical Voice Disorders

Benign voice disorders include a broad range of etiologies; including structural pathologies, autoimmune disorders, and neurological conditions, each of which impact the voice through different mechanisms (Verdolini, Rosen, & Branski, 2014). Voice production depends on ability to generate a steady subglottic pressure, maintain adequate glottic closure, and to finely coordinate the structural vibratory properties of the vocal fold (e.g., pitch changes). Each of these properties of voice can be impacted by the presence of mass changes in the form of benign or malignant lesions. Increased mass of the vocal fold(s) by neoplastic lesions can impact the quasi-periodic vibration required for voice, as can irregularity or stiffness of the vocal mucosa (Orlikoff & Kraus, 1996; Starmer, Tippett, & Webster, 2008). These changes can impact the differential characteristics of the vocal folds themselves which can result in asynchrony of vibration of each vocal fold, as well as potentially leading to glottic incompetence due to a mass lesion limiting vocal fold contact, or other interruptions to fold adduction. Neoplastic processes leading to unilateral changes result in irregularities between the two vocal folds, impacting the contact, vibration, and closure of the glottis with resultant changes in voice quality (Orlikoff & Kraus, 1996). Patient’s with these changes often experience hoarseness, decreased volume, and breathiness, which may be easily detected by the listener, including those who are naïve. With changes to the vocal mechanism, increased effort may be required to achieve vocal fold adduction in an effort to compensate, a volitional change that can lead to muscle strain and odynophonia (i.e., muscle tension dysphonia) (Van Houtte, Van Lierde, & Claeys, 2011). Similar to hyperfunctional changes that can be identified, vocal fold hypofunction is a frequent problem in the postoperative larynx (Doyle, 1997). “Hypofunctional” voice disorders describe conditions where the voice is “underpowered”, which leads to a weak or insufficient voice (Doyle 1997). “Hypofunctional” voices can include compensatory behaviours following laryngeal surgery, respiratory conditions, amongst others (Doyle, 1997; Titze, 1988). The pathophysiology of these conditions is typically from weak glottal approximation and low airflow (Doyle, 1997; Titze, 1988). Appropriate diagnosis of hyper- and hypofunctional disorders is important in guiding the approach to voice rehabilitation (Doyle, 1997)

A “hyperfunctional” voice disorder refers to the misuse of the laryngeal musculature in the production of voice sounds. These muscle use patterns can lead to structural pathologies over time. The repeated trauma to the vocal folds can contribute to the development of nodules, polyps, contact ulcers, edema, and vocal fold thickening (Hillman et al., 1989). Vocal hyperfunction results in increased muscle tension and increases the collision forces of the vocal folds. This creates “strain” in the voice initially, but the repetitive trauma created by these vocal habits cause changes to the mucosa and vocal fold tissues. The changes can result in changes to vibration, such as an inability to synchronously vibrate, or cycle to cycle variability, seen in measures of perturbation. In those instances where hyperfunction results in tissue change that does not resolve with a reduction of abnormal voice production behaviors, the developmental sequence over time will involve increasing levels of compensation. Consequently, increasing levels of hyperfunction in an attempt to continue voicing will further exacerbate the problem with further changes to vocal fold tissue.

Early in the development of lesions that result from hyperfunction, many are reversible by changing the way voice is produced to decrease the repetitive trauma. Voice therapy has been shown to improve and resolve many reactive lesions (Holmberg et al., 2001; Stepp et al., 2011; Ylitalo & Hammarberg, 2000). If they persist, they can result in irreversible changes that require surgical intervention to reduce mass changes on the vocal fold. (Hillman et al., 1989).

Structure-based voice disorders include the development of benign laryngeal masses (e.g., subepithelial cysts), and reactive lesions (e.g., sulcus vocalis), amongst others. Changes to the mass or stiffness of the vocal folds can be caused by nodules, polyps, scars, or simply the presence of edema (Hirano, 1981a). An important physical property of the vocal folds is the entrainment phenomenon, where oscillating or vibrating systems will assume the same vibratory period when they interact (Pippard, 2007). This property is disrupted by asymmetric mass or stiffness changes to one vocal fold. Asymmetric change in one vocal fold will introduce increasing levels of aperiodicity into the vocal signal by preventing the two folds to adopt the same vibratory period. The two folds no longer have nearly identical properties and thus respond to the subglottic pressure differently without assuming the same vibratory period. Increasing levels of aperiodicity are a result of increased chaotic vibration at the level of the glottis, which represents a noise component in the vocal signal. Vocal signals consist of two components, the harmonic component and the noise component (Baken & Orlikoff, 2000). These two components are determined by the degree of periodicity in the signal. Once the fundamental frequency of the signal is known, changes in vocal fold vibration can be detected via microacoustic measures that seek to determine cycle-by-cycle changes in the frequency and amplitude domains; thus, these fine changes in vibration can be quantified using measures of perturbation, referred to as jitter and shimmer respectively (Baken & Orlikoff, 2000). Additionally, unilateral lesions can prevent complete closure of the glottis, which contributes to the turbulent airflow that makes up the aperiodic component of the signal. The aperiodic component “contaminates” the vocal signal, decreasing the quality of the signal. When compared to the harmonic component, we can calculate a measure of vocal quality called harmonic-to-noise ratio.

Voicing requires complex control of the larynx and neurologic deficits can impair this control mechanism (M. Smith & Ramig, 1994). Common neurological conditions of the larynx are peripheral nerve injuries, spasmodic dysphonia, amyotrophic lateral sclerosis (ALS), and Parkinson disease, amongst others (M. Smith & Ramig, 1994).These conditions can lead to issues of hyperfunction, hypofunction, discoordination, and instability of the voice (M. Smith & Ramig, 1994).

There are a broad range of conditions that affect the voice, from primary laryngeal disorders to systemic diseases. These disorders impact the function of the vocal mechanism in a number of different ways, including hyperfunctioning and hypofunctioning, which impacts the treatment paradigms for each disorder. Many of these changes to the larynx result in changes to the vocal signal which can be objectively evaluated with acoustic assessment. Assessment and diagnosis of these conditions relies on a multi-dimensional approach including aerodynamic studies, auditory-perceptual analysis, self-reporting, and acoustic analysis.

1.2 Current Approaches to Assessment and Diagnosis of Voice Disorders

Currently, patients who present with voice disorders are typically assessed by an otolaryngologist and/or speech-language pathologist. This includes a complete history and physical examination of the head and neck. Indirect assessments of the larynx are possible through flexible nasopharyngoscopy and stroboscopy. These physical exam techniques allow visualization of the vocal folds, vocal tract, and assessment of the mucosal waveform. Furthermore, patient symptom scores (Voice Handicap Index, Reflux Symptom Index, etc.) can provide information regarding the severity of voice-related symptoms (Belafsky, Postma, & Koufman, 2002; Jacobson et al., 1997) and the impact such changes have on the individual.

Indirect assessments of vocal function include auditory-perceptual evaluation of voice and acoustic analysis. Auditory-perceptual assessment of voice involves a rigorous and multi-dimensional evaluation of the voice on a number of perceptual elements (Eadie & Doyle, 2005; Zraick et al., 2011). This is usually done through application of psychophysical scaling methods (Carterette, 1974; Stevens & Marks, 2017; Toner & Emanuel, 1989). Using scaling methods, voices are subjectively assessed by either naïve or trained listeners. Important to the validity of auditory-perceptual, and other forms of voice analysis, is the selection of the appropriate scaling method (Eadie & Doyle, 2002; Schiavetti, Metz, & Sitler, 1981). Work done in the assessment of perceptual phenomena has suggested that there are two forms of scaling that can be applied to the data: prothetic and metathetic (Stevens & Marks, 2017). These forms of scaling are dictated by the property that is being scaled (Stevens & Marks, 2017). A prothetic scale can be applied to data or a perceptual attribute that is additive in nature (e.g., loudness) and cannot be subdivided into equal parts. These are best represented with direct magnitude estimates. Whereas metathetic scaling represents a qualitative continuum, where the groupings can be subdivided into equal subgroupings. These are best represented with equal appearing intervals (e.g., visual analog scale). Stevens et al suggest a protocol to determine the appropriate scaling method for any particular perceptual attribute (Eadie & Doyle, 2002; Stevens & Marks, 2017). Equal appearing interval scales are regularly found in the auditory-perceptual literature and the results of these studies need to be evaluated with caution (Eadie & Doyle, 2002).

Acoustic voice analysis can be used to evaluate the vocal signal using established normative values and ranges for objective comparison. However, this represents one part of a comprehensive assessment of voice, which includes auditory-perceptual assessments, self-reporting measures, history, and clinical exam amongst others. Each part contributes important data to the diagnosis, monitoring, and documentation of clinical voice disorders.

1.3 Acoustic Voice Analysis

Acoustic voice analysis is a method to quantify both the normal voice and abnormalities of dysphonic voice. These measurements are achieved in a non-invasive manner and can be compared against established normative values and ranges. Many measures of acoustic analysis rely on assumptions of signal periodicity, where the periodic time intervals are extracted from the vocal signal. This process relies on the features of the signal, which is quasi-periodic in nature; meaning that it follows some repeating pattern. The vocal signal is a complex waveform that results from the sum of multiple simple sine waves. In a perfect system, a sustained sound would produce a perfectly repeated period, but given the imperfect mechanics of the larynx, this only approaches periodicity in a normal voice (Parsa & Jamieson, 1999). After determining the normal period of the signal, further "microacoustic" measures can be calculated. These calculations estimate cycle-to-cycle variability (e.g., jitter, shimmer) and aperiodic noise components (e.g., harmonic-to-noise ratio) of the signal (Hillenbrand, 2011; Parsa & Jamieson, 2014; Yumoto, Gould, & Baer, 1982). This provides a potential opportunity to classify voices as normal or abnormal for one or more measures (Boersma, 2002; Eadie & Doyle, 2005; Parsa & Jamieson, 2001; Umapathy, Krishnan, Parsa, & Jamieson, 2005).

Although there has been much research within this field, acoustic measures have not always been found to account for the variance that is seen in perceptual judgments of voice, particularly in grossly aperiodic voices (Eadie & Doyle, 2005; Eskenazi, Childers, & Hicks, 1990; Gelfer, 1993; Kempster, Kistler, & Hillenbrand, 1991; Murry, Singh, & Sargent, 1977). By relying on correct determination of the fundamental period of the vocal signal for the calculation of perturbation measures, some acoustic measures become increasingly unreliable with increasing aperiodicity (Eadie & Doyle, 2005). If the fundamental period is incorrectly identified, there is a domino effect on further acoustic measures calculated from this (Eadie & Doyle, 2005). Furthermore, acoustic measurements do not always correlate with auditory-perceptual assessments, which becomes increasingly true with severely dysphonic voices (Rabinov, Kreiman, Gerratt, & Bielamowicz, 1995). Measures of perturbation are considered measures of aperiodicity, but unfortunately, they lack reliability in aperiodic signals and some researchers have suggested they cannot be reliably used for even slightly aperiodic signals (Bielamowicz, Kreiman, Gerratt, Dauer, & Berke, 1996). Ideally, an acoustic measure is able to withstand the aperiodicity associated with dysphonic voices (Heman-Ackah et al., 2003). Work has been done to account for these issues and a battery of acoustic measures that includes time- and spectral-based analyses has been suggested to overcome this (Awan & Roy, 2006, 2009).

1.4 Auditory-perceptual Voice Analysis

Auditory-perceptual evaluation is the most common method of voice assessment and it is simple to perform, robust, and inexpensive (Kent, 1996; Oates, 2009). Judgements of voice are inherently perceptual in nature, in that judgments are made as to what is pleasant to listen to (Eadie & Baylor, 2006; Shrivastav, Sapienza, & Nandur, 2005). With the current limitations of acoustic and other forms of objective assessments, such as disagreement on the sensitivities of acoustic measurements, auditory-perceptual voice analysis is often seen as the gold standard in evaluating voices (Kent, 1996; Oates, 2009).

Inter-judge variability may be accounted for by what is determined to be normal to one listener as opposed to the next (Kreiman, Gerratt, Precoda, & Berke, 1992). That being said, perceptual assessments of voice are more likely to be shared by a wide variety of people, including trained and novice listeners: perceived features of voice will be similar amongst listeners of different levels of training (Oates, 2009; Wuyts, De Bodt, & Van de Heyning, 1999). Furthermore, because voice quality is perceptual in nature, communicating our auditory judgments of voice is often intuitive across groups of listeners (Oates, 2009). Work by Kreiman et al showed that people develop their own internal reference standards with inherent biases that impact the judgement of future voices (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). These can vary over time within an individual and may also be different between two listeners. This points to the major critique of auditory-perceptual assessments, namely, the reliability of such measures. This has been extensively researched and have been found, with control for confounding variables, to be a robust form of voice assessment (Dejonckere, Obbens, de Moor, & Wieneke, 1993; Karnell et al., 2007; Oates, 2009; Webb et al., 2004).

Without accounting for statistical validity or reliability, providing a common language to describe voices is a contributing factor to the overall popularity of auditory-perceptual voice analysis (Hirano, 1981b; Kent, 1996; Kreiman et al., 1993; Oates, 2009; Oates & Russell, 1998; Shrivastav et al., 2005; Webb et al., 2004; Wuyts et al., 1999). Auditory-perceptual assessments are also an inexpensive manner of evaluating voice. Beyond the recording equipment, there are no further cost requirements. Auditory-perceptual analyses can identify clinically relevant changes in the voice in an inexpensive, reliable, and simple to carry out fashion. For these reasons, this method of voice assessment represents the gold standard of voice analysis. A clinically evident voice change that is documented through auditory-perceptual assessment scales provides a reliable, repeatable, and non-obtrusive measure to compare the voice to in future encounters (ASHA, 2002; Kent, 1996; Kreiman et al., 1993). A change to the voice that can be heard and evaluated likely represents a clinical entity and auditory-perceptual scales provide a means to document these.

1.5 Technological Advances and mHealth

Smartphones, after becoming commonly available over the past decade, now represent a focus of many health-related interventions (Ali et al., 2016; Munnings, 2019; Steinhubl et al., 2015; Wong & Fung, 2015; Yang, Wang, Yang, Hu, & Xiong, 2018). Applications of this type of technology include text message related interventions, smoking cessation and other habit-forming applications, along with more sophisticated interventions such as using the device camera to assess artery patency (Di Santo et al., 2018; Steinhubl et al., 2015; Yang et al., 2018). Smartphone microphones have been shown to be useful in the recording of voice samples to be analyzed for microacoustic measures of voice (Lee et al., 2016; Lin, Hornibrook, & Ormond, 2012; Uloza et al., 2015). Previous studies have evaluated either sustained vowels or sentences, but not both, using their own proprietary assessments showing smartphone based analyses to be feasible (Fujimura et al., 2019; Mat Baki et al., 2013, 2015). A gap in the literature is the lack of comparison against a cost-effective alternative (Praat) (Boersma, 2002; Mat Baki et al., 2013, 2015). Mat Baki et al performed on-device analysis, but the high initial cost and ongoing subscription fees, as well as inability to access this application currently has prohibited others from repeating their study (Mat Baki et al., 2013, 2015). Given the feasibility of on-device analysis, ubiquitous availability of smartphones, amongst other reasons, smartphone collected voice recordings may represent an opportunity to increase access to acoustic and auditory-perceptual voice analysis.

The advantages to mobile analysis are primarily found in the ease of access. Acoustic analysis relies on signal acquisition methods, microphone effects, and microphone setup, which each play a role in the barriers to access (Doherty & Shipp, 1988; Karnell, 1991;

Parsa & Jamieson, 2014; Titze & Winholtz, 1993). Much of the current standardization in acoustic analysis is achieved using specific recording setups and sound-controlled environments, such as soundproofed booths. With current hardware in smartphones, a mobile device and a standardized recording protocol, including microphone-to-mouth distances, can lessen many of the obstacles to performing acoustic analysis. Providing clear instructions for recording tasks will limit the loss associated with having end-users provide the sample collection themselves. Mouth-to-microphone standardization may be replaced by biofeedback with sampling loudness and target ranges to achieve. Each of the aforementioned barriers are in-place for standardization, some of which can likely be relaxed upon without losing validity in the acoustic measures.

With the wide availability of smartphones, the access can be further maximized by enabling patients, clinicians, and researchers to perform analysis more frequently between formal scheduled follow up visits, which can be a resource limited factor. In this way, voice over time can be assessed by the clinician or researcher. Many of the foreseen advantages are associated with the possibility of frequency of testing, which could lead to greater within-subject comparisons. Within-subjects comparisons improve our ability to detect small, consistent changes by being able to set patient specific normative values. By repeating the test regularly, the patient specific known values are determined and changes to these can be evaluated for clinical significance through follow up assessment or alternative voice assessments (i.e., auditory-perceptual, self-reporting, etc.).

Acoustic analysis evaluates the vocal signal, including changes on a “microacoustic” level, as well as the harmonic and noise components. This requires high quality recordings with limited to no ambient noise. This is typically achieved by recording in a soundproof setting, using professional equipment, in a standardized setup. The quality, directionality, and positioning of the microphone, which is easily controlled for in a standardized setup, may be decreased by using a smartphone microphone and an untrained user. This may result in increased background noise identified in the signals, compromising the ability to detect small changes in the vocal signal. However, previous research has shown smartphone microphones to be of high quality and reliable for these purposes (Guidi et al., 2015; Lin et al., 2012; Uloza et al., 2015). These advantages are likely to outweigh the effects of decreasing the control of the recording setup.

Additionally, capturing signals at large in an uncontrolled environment that is not controlled for ambient noise will further compromise the background qualities of the signal. The variability in ambient noise from one quiet environment to the next is a significant consideration, but the absolute difference is limited and if the environment is selected with care, the recordings gathered will still be of high enough quality to perform reliable acoustic analysis. The control of the standardized nature in which samples are collected is lessened by the absence of a trained observer and this may introduce variables such as differing mouth-to-microphone distances amongst multiple samples collected by prospective users. With clear instruction, this concern can be mitigated, and the recordings can be performed without observation with high repeatability and reliability.

Through limiting the restrictions of voice sample capture, the opportunities to collect higher numbers of samples for analysis increases. If an increasing number of samples can be obtained, this may then open the opportunity for more extensive within-subjects analysis, an issue which is constrained under the current barriers to acoustic analyses.

Recordings over time offer a chance to evaluate voice change on a granular level as patients receive voice therapy, medical, or surgical treatment. This can also be achieved through short- and long-term data collection. But with the desire to obtain an increased number of samples over time, the environment in which such samples are obtained becomes of importance. Of primary concern is the issue of variations in environmental noise during sample acquisition. Based on our evaluation of the current state-of-the-art methods, the impacts of ambient noise and necessity of an observer to perform robust acoustic analysis are barriers that can be diminished through validation of a mobile voice analysis application. With repeated measures through increased access, these analyses could be used as a form of biofeedback as they progress through therapy or recover following surgical intervention. Furthermore, following treatment for malignant diseases of the larynx, analysis may prompt early reassessment, if subclinical change is identified. The ability to enhance post-treatment monitoring regardless of underlying etiology holds substantial clinical promise.

1.6 Objectives

This project sought to gather information focused on multiple objectives. More specifically, this project was designed to:

1. Confirm the iPhone microphone sample capture ability for microacoustic measurements in optimized settings;

2. Identify the capacity of the iPhone microphone to capture samples in non-optimized settings for acoustic analysis; and finally,

3. Determine the computational abilities of the iPhone for the purpose of immediate, on-device voice analysis.

Chapter 2

2 Review of the Literature

There are a number of well-researched and documented acoustic measures reported in the literature. For the purpose of this project, we selected five common acoustic measures to use, including time- and spectral-based measures. Those measures are fundamental frequency, jitter, shimmer, harmonic-to-noise ratio, and cepstral peak prominence. Each of these measures will be briefly described in the sections to follow.

2.1 Fundamental Frequency (F_0)

Vocal pitch has long been agreed to be an important factor when evaluating voice (Baken & Orlikoff, 2000; Cooper & Yanagihara, 1971; Murry, 1978). Traditionally, it was evaluated according to gross practitioner perception with categorical assessment based on whether the listener perceived the speaker's pitch to be normal for age and gender, too high, too low, or erratic in its consistency (Baken & Orlikoff, 2000; Laguaite & Waldrop, 1964). While this method could identify gross pitch attributes of the voice, it proved to be too crude and unstandardized to be a reliable way to evaluate voice, particularly when an abnormality existed. Furthermore, it may be influenced by the listener's personal preferences or assessments, as to what is considered 'too high' or 'too low', or abnormal in its consistency. Objective measures of voice were developed to introduce inter-rater reliability and determine unbiased normative values (Baken & Orlikoff, 2000). In doing so, data reported in the literature has provided well-established points of reference for pitch with consideration of age and gender. Because pitch is a psychological attribute of voice as well as a physical one, the ability to characterize vocal pitch can be done through both subjective perceptual evaluation and objective acoustic assessment.

Fundamental frequency (F_0) is a measure of the vibratory rate at which a waveform is repeated per time frame, traditionally measured per second in Hertz (Hz) (Baken & Orlikoff, 2000). F_0 is the measurable parameter that corresponds to the perceptible attribute pitch (Baken & Orlikoff, 2000). As F_0 increases, pitch increases, though this is not a linear relationship and listeners tend to be more sensitive to changes in the lower frequencies than higher frequencies (Baken & Orlikoff, 2000; Stevens & Volkmann, 1940; Stevens, Volkmann, & Newman, 1937). It is also important to note that despite differences in speaker F_0 ranges in absolute values, the ability to normalize values obtained in order to determine a speaker's functional range needs to occur (Britto & Doyle, 1990).

F_0 can be calculated in a number of ways, many of which have been facilitated and optimized in ease by the introduction of computers (Anderson, 1925; Baken & Orlikoff, 2000; Metfessel, 1928; Parsa & Jamieson, 2014). There are two ways to classify F_0 extraction algorithms: event detection (peak picking, zero crossing) and averaging methods that rely on waveform matching approaches (Baken & Orlikoff, 2000; Parsa & Jamieson, 1999). Event detections were the earliest methods devised to extract F_0 from a vocal signal (Anderson, 1925; Metfessel, 1928). These include methods such as peak picking and zero crossing by using complicated photographic and light flashing devices that relied on the waveform generated from vibrations produced from vocal signals (Anderson, 1925; Metfessel, 1928). For example, in one of these devices, called the phonodeik, a mirror was used to reflect the vibratory pattern of a wire onto a photographic film strip (Anderson, 1925). Peak picking relies upon the pattern of waves and the resultant peaks per cycle, positive and negative (Baken & Orlikoff, 2000). To calculate F_0, and the number of peaks in one direction (positive or negative) per second are counted (Baken & Orlikoff, 2000). The harmonic peaks of the vocal signal need to be counted, and this is easily accomplished using a computer by setting a threshold for detection. Once the threshold is breached, the peak is recorded and the wave is converted to series of dots indicating a threshold crossing, and the number of times per second is then recorded (Baken & Orlikoff, 2000). Another event detection method of calculating F_0 is zero-crossing. Similarly, this relies on the properties of waves, which cross the zero-voltage line twice (once up-going, once down-going) per cycle (Baken & Orlikoff, 2000). The signal is converted from a waveform to a series of pulses, which is subsequently run through a high-pass filter, which only captures peaks above the arbitrary threshold, converting the pulses to a series of positive or negative spikes (Baken & Orlikoff, 2000). Each of the spikes in one direction are counted in a one second time interval (i.e., time-based) to calculate the F_0 value (Baken & Orlikoff, 2000).

More recently, averaging methods have been developed to stand up to the presence of signal noise and amplitude variations that are often found in pathologic samples (Parsa & Jamieson, 2014). That is, as vocal signals become more aperiodic, the likelihood of misidentifying a “period” of vibration is increased with a corresponding influence on the accuracy of the measure obtained. In an experiment by Parsa and Jamieson (2014), a number of these methods were compared in normal and pathologic conditions, and waveform matching was found to be the most robust at providing reliable results in the setting of jitter, shimmer, and background noise. Waveform matching uses a least-squared difference between two adjacent waveforms (Milenkovic, 1987; Titze & Liang, 1993). This least-squared method is accomplished by marking the absolute low-point of one waveform, a second point is placed in an adjacent trough to accommodate the lowest mean squared error possible, and this process is repeated until all waveforms have been evaluated (Milenkovic, 1987; Parsa & Jamieson, 2014; Titze & Liang, 1993).

Fundamental frequency, which is perceived as pitch, is an important acoustic measure of voice. There are well established normative ranges for gender and age. Determining the fundamental period accurately is important for calculation of measures of perturbation, which evaluate cycle-to-cycle changes to the fundamental period. Correct determination of the fundamental period sets the stage for frequency and amplitude measures, by which if the fundamental period is incorrectly determined both jitter and shimmer will not produce reliable results. The downstream effects of correct period identification are critical for further calculations and waveform matching methods are the most effective at doing this in dysphonic voice samples.

2.2 Frequency and Amplitude Perturbation Measures (Jitter and Shimmer)

Measures of signal perturbation form another well-researched area of voice science with the objective of estimating the cycle-to-cycle differences in the fundamental period (Hillenbrand 1987). Small differences, or perturbations, between adjacent cycles are always present, even when trying to produce a steady sound-state such as that which occurs with a sustained vowel (Maryn, Corthals, De Bodt, Van Cauwenberge, & Deliyski, 2009). These variations are reflective of the imperfect, quasi-periodic mechanism of voice production (Baken & Orlikoff, 2000). Two common measures termed jitter and shimmer, are used to calculate pitch and amplitude variations, respectively (Kitajima & Gould, 1976; Koike, 1969; Lieberman, 1963). In a perfectly stable production system, jitter and shimmer would equal zero (Baken & Orlikoff, 2000). That is, no differences in either the frequency or amplitude domain would exist with a pure tone due to the fact that each adjacent vibratory cycle is identical. Consequently, an increase in jitter and/or shimmer is a measurable representation of inconsistent vibratory patterns (Baken & Orlikoff, 2000). Typically, increases in measures of perturbation are associated with worsening overall voice quality, an acoustic phenomenon that is perceived as dysphonia (Maryn et al., 2009).

As noted, frequency perturbation is more commonly known as jitter (Baken & Orlikoff, 2000) and this measure represents frequency variability of the fundamental period. It a measurement of the frequency changes in the periods that are immediately next to each other, and in this way is measured irrespective of cycles in non-adjacent vocal sounds, which may represent voluntary changes to the frequency (Baken & Orlikoff, 2000; Boersma, 2009; Parsa & Jamieson, 2014). Jitter is typically presented as a percentage of the fundamental frequency. Higher measured levels of jitter will correspond to increased levels of vocal abnormality that is represented in the frequency domain.

Amplitude perturbation is termed vocal shimmer and this measure quantifies short term variability in signal amplitude (Baken & Orlikoff, 2000; Wendahl, 1966). Calculation of shimmer is reliant upon the peak amplitude of each adjacent vocal period (Baken & Orlikoff, 2000; Horii, 1980). Shimmer can be calculated for two or more adjacent periods and the result is typically presented in decibels (dB) (Baken & Orlikoff, 2000). Similar to jitter, increasing variation in the amplitude of a vocal signal will negatively impact perceptions of vocal quality and it has been found to be an important contributor to the perception of hoarseness (Eskenazi et al., 1990; Wendahl, 1966, 2009).

Jitter and shimmer are used as one important part of an acoustic assessment of voice. Increased values of either have been shown to represent increased disorder in the vibratory pattern of voice production at the level of the glottis or vocal folds. In a perfect mechanism, there would be no cycle-to-cycle variability. In the same way, as the vocal mechanism behaves in a more orderly fashion, which we attribute to normal function, measures of perturbation decrease. However, as mentioned above, even when producing relatively steady state sounds (e.g., a sustained vowel), there are always variabilities between adjacent cycles due to the imperfect mechanics of the larynx. Therefore, these measurements have been shown to increase in pathologic voices and a combination of these measures can be used as an overall correlate of vocal quality (Werth, Voigt, Döllinger, Eysholdt, & Lohscheller, 2010). This capacity allows for these measures to serve as a valuable means of monitoring and indexing vocal change over time.

2.3 Harmonic-to-Noise Ratio (HNR)

“Hoarseness”, while generically non-descript, is a common symptom of altered voice quality, including a perceptual judgment that includes changes to pitch, loudness, and quality (Schwartz, Rosenfeld, Dailey, & Cohen, 2009). It is a term to describe a rough or noisy voice that is frequently associated with poor voice quality; hoarseness has often been used to broadly categorize the presence of dysphonia (Schwartz et al., 2009). Harmonics-to-noise ratio (HNR) was developed by Yumoto et al (1982) as a measure to identify pathology based on the detection of hoarseness. It is well established that the sound of hoarseness results from the replacement of harmonic energy with noise energy (Isshiki, Yanagihara, & Morimoto, 1966; Yanagihara, 1967). Hoarseness signifies an increase in the aperiodic nature of the vibrations, as well as serving to quantify associated turbulence replacing the quasi-periodic nature of harmonic voice (Baken & Orlikoff, 2000; Eadie & Doyle, 2005; Isshiki et al., 1966; Yanagihara, 1967). The development of this ratio as an acoustic measure of voice seeks to quantify the spectrographic features of the voice (i.e., all energy components that comprise the vocal signal) that accompany hoarseness (Baken & Orlikoff, 2000; H. Kojima, Gould, & Lambiase, 1979; Hisayoshi Kojima, Gould, And, & Isshiki, 1980; Yumoto et al., 1982).

The theory behind the HNR measure relies on the idea that there are two major components to voice: near periodic and aperiodic (Baken & Orlikoff, 2000). The near periodic or quasi-periodic component makes up the harmonic aspect of voice, whereas aperiodicity contributes random noise into the signal through sound pressure variation (Baken & Orlikoff, 2000; Eadie & Doyle, 2005). With increasing hoarseness, the periodic signal is subsequently contaminated by this random noise (Baken & Orlikoff, 2000; Kim, Kakita, & Hirano, 1982; Hisayoshi Kojima et al., 1980; Yumoto et al., 1982). The degree of contamination is expressed as a ratio of the harmonic energy to the noise components (Baken & Orlikoff, 2000). This can be calculated by taking the averaged waveform, which is the harmonic component, and subtracting it from the non-averaged or absolute waveform (Baken & Orlikoff, 2000). This results in the noise component being left over. A ratio of the averaged waveform and the noise waveform values results in the harmonic-to-noise ratio (Baken & Orlikoff, 2000).

Increased turbulence at the level of the glottis leads to a diminished harmonics-to-noise ratio as the turbulence elevates the noise component in the signal (Ferrand, 2002). From an auditory-perceptual perspective, HNR represents a broad measure of voice quality (Ferrand, 2002). Studies evaluating vocal quality have found HNR to be an important correlate as an acoustic measure (Eskenazi et al., 1990). The voice is a complex waveform, made up of the quasi-periodic and aperiodic components. The quasi-periodic nature of the voice allows the subtraction of the averaged waveform, which represents the harmonic sound, from the complex signal leaving only the noise component. This component is the aperiodic vibration that contributes noise to the signal. Increasing noise in the signal decreases the ratio, representing a decrease in overall vocal quality. In this way, HNR can be used as a correlate of overall voice quality.

2.4 Cepstral Peak Prominence (CPP)

An ideal measure of voice is one that is reliable in the context of aperiodicity (Heman-Ackah et al., 2003). In that context, a measure can evaluate even the most chaotic vibratory patterns that are associated with severely disordered voices. Measures of perturbation, jitter and shimmer, rely on the accurate detection of the fundamental period, which becomes increasingly difficult in severely aperiodic voices (Heman-Ackah et al., 2003). A powerful additional method of calculating the fundamental frequency was first described in 1964 by Noll et al; a measure relying on Fourier analysis of the vocal signal (Baken & Orlikoff, 2000; Hillenbrand & Houde, 1996; Noll, 1964). Using these concepts, Hillenbrand et al (1994) developed an acoustic measure to predict dysphonia severity, termed cepstral peak prominence (Heman-Ackah et al., 2003; Hillenbrand, Cleveland, & Erickson, 1994; Hillenbrand & Houde, 1996).

Cepstral peak prominence (CPP) is a somewhat newer acoustic measure, one which evaluates the vocal signal without relying on identifying the fundamental period (Heman-Ackah et al., 2003, 2014). A spectrographic representation of the vocal signal is a result of a Fourier transform of the vocal signal's frequency wave (Heman-Ackah et al., 2003). This takes the signal from the amplitude (dB) and time (s) domains and produces a waveform of the intensity (dB) and the frequency (Hz) (Heman-Ackah et al., 2003). In doing so, the converted signal is a spectral representation, which results in a logarithmic representation of the amplitudes, increasing the visibility of small differences in amplitude (Heman-Ackah et al., 2003). With respect to voice, the fundamental frequency typically represents the largest frequency amplitude, followed by multiples of harmonic frequencies achieved in the resonant tract of the upper airway (Heman-Ackah et al., 2003). A cepstral representation, called a cepstrum, is achieved by performing a second Fourier transform; the signal is now presented in a waveform of magnitude (dB) and quefrency (ms) (Heman-Ackah et al., 2003; Noll, 1964). "Quefrency" is a term coined by Noll to avoid confusion of transforming the signal back into the time domain, where the unit of measurements is seconds (Heman-Ackah et al., 2003; Noll, 1964). The Fourier transformation of the spectral wave transforms the spectrum from the frequency domain to the time domain (Heman-Ackah et al., 2003). It serves to further clarify the harmonic components of the voice, whereby the most prominent peak corresponds to the fundamental frequency (Heman-Ackah, Michael, & Goding, 2002).

Voice production generates a complex waveform where the most prominent peak of a spectral graph represents the fundamental frequency. The subsequent, less prominent, peaks represent harmonic frequencies which are typically multiples of the fundamental frequency (Heman-Ackah et al., 2003). When converted to the cepstrum, these harmonic peaks are referred to as “rhamonic” peaks, which occur at multiples of the fundamental period (Heman-Ackah et al., 2003; Noll, 1964, 1967). A linear regression line is then calculated to represent the average sound energy through the graphic representation, which is used to calculate the cepstral peak prominence (Heman-Ackah et al., 2003). This normalizes the variability between amplitude in voice, accounting for variations in the volume the sample was captured at by determining the magnitude of the peak against the loudness at which it was produced (Heman-Ackah et al., 2003). The linear regression line is calculated from the quefrency to cepstral magnitude together, and the cepstral peak prominence is calculated as the amplitude from the peak to the corresponding value on the regression line (Heman-Ackah et al., 2002). Evaluating this measure is uncomplicated, in that the higher the prominence, the more periodic the voice sample that is being analyzed. For this reason, a highly periodic voice will result in a high CPP, which indicates highly organized harmonic structure (Heman-Ackah et al., 2003, 2002). Similarly, voices with a significant noise component will have poorly organized harmonic structures resulting in a flat appearing cepstrum, the peak from the linear regression line will be lower accordingly (Heman-Ackah et al., 2002).

Cepstral peak prominence has been shown to reliably correlate with perceptual measures of dysphonia (Awan, Roy, & Dromey, 2009; Heman-Ackah et al., 2003, 2002, 2014). CPP holds some of the benefits required of an optimal measure of voice in that it is calculated independent of the fundamental period, which in turn can become increasingly unreliable in severely dysphonic voices (Heman-Ackah et al., 2003, 2014; Noll, 1967). Furthermore, it is unimpacted by the recording technique and volume (Heman-Ackah et al., 2014). That is, it is performed without requiring detection of periodicity, and by using a linear regression approach the value is independent to the loudness the sample is captured at (Heman-Ackah et al., 2003, 2002). Similar to HNR, CPP evaluates the harmonic sound in the vocal signal and compares this to the noise component (Heman-Ackah et al., 2003, 2002). This work by Hillenbrand has produced a reproducible measure for detection of dysphonic voices and a normative range of values has since been established (Awan et al., 2009; Heman-Ackah et al., 2003; Hillenbrand et al., 1994; Hillenbrand & Houde, 1996). For normal voices, a CPP will be greater than 4.0 (Heman-Ackah et al., 2014). CPP has been established as an acoustic measure that reliably predicts dysphonia severity (Heman-Ackah et al., 2003, 2002, 2014). It does this by quantifying the harmonic organization of a voice sample in a way that is not reliant on identification of the fundamental period, making it reliable in the setting of increasingly aperiodic voices (Heman-Ackah et al., 2003, 2014; Hillenbrand et al., 1994; Hillenbrand & Houde, 1996).

2.5 Smartphone Voice Analysis

Acoustic voice analysis has long been seen as an important part of a thorough voice assessment. Objective measures can be collected as part of a well-rounded approach that is complemented by rigorous auditory-perceptual evaluation (Eadie & Doyle, 2005). However, barriers exist to collecting adequate voice samples for analysis, including sound-controlled environments, costly set ups, analysis software, and requirement for a trained observer to collect the sample (Kisenwether & Sataloff, 2015; Parsa & Jamieson, 2001; Parsa, Jamieson, & Pretty, 2001; Zraick et al., 2011). With a substantial increase in the widespread availability of smartphones, new avenues for data collection and mobile analysis are available to potentially alleviate some of these concerns (Burdette, Herchline, & Oehler, 2008; Manfredi et al., 2017; Park & Chen, 2007; A. Smith, 2012; Steinhubl et al., 2015). Over the past several years, research has been completed to indicate the validity of these recordings for assessment and the analyses that can be completed on a smartphone (Uloza et al., 2015).

Much of the research has been directed toward evaluating the quality of smartphone voice recordings for the purposes of micro-acoustic analysis (Guidi et al., 2015; Lin et al., 2012; Uloza et al., 2015). Lin et al evaluated the iPhone for use in recording voice samples for acoustic analysis, finding good reliability when compared to a traditional set up for some acoustic measures (Lin et al., 2012). Uloza et al specifically evaluated the correlation of the results when recorded with a high-quality microphone (AKG Perception 220) compared to a smartphone microphone (Samsung Galaxy Note 3); their data showed high correlation between the analysis results from both groups, indicating reliability in the smartphone captured samples (Uloza et al., 2015). The findings of these two studies show that these types of recordings are now of high enough quality that valid acoustic analysis can be performed and using these analyses could improve early diagnosis of laryngeal diseases (Uloza et al., 2015). This alone may reduce some of the barriers to acoustic analysis by providing a low-cost option for recording the voice sample and analyzing the data through existing software programs.

Beyond the capabilities of voice capture and offline analysis, there have been several studies that assessed the ability to perform on-device analysis (Fujimura et al., 2019; Mat Baki et al., 2013, 2015; Siau, Goswamy, Jones, & Khwaja, 2017). With significant increases in computing power, smartphone devices are now able to carry out analysis without first exporting the voice samples (Manfredi et al., 2017). That is, they now have the computational abilities to perform the complex mathematics without the use of external software programs. The capacity for smartphone voice sample capture and analysis could improve access to ongoing, acoustic voice analysis by minimizing the steps required to offer acoustic analysis (Manfredi et al., 2017). This advantage is of particularly importance for patients who may be distant from their otolaryngology health care provider or come from resource constrained areas. By decreasing barriers to access, repeated measures (i.e., multiple ongoing recordings and analysis) are a notable benefit, leading to increased data points for evaluation in clinical and research settings (Manfredi et al., 2017). Furthermore, Manfredi et al went on to test the limits of currently available smartphones and their respective microphones (low cost vs. high cost) and found limited difference between the two levels of devices (Manfredi et al., 2017).

Smartphone technology has increased over the past decade and is now frequently an area of clinical interest (Burdette et al., 2008; Steinhubl et al., 2015). Within voice science, there has been similar interest and much of the groundwork for validation has been completed (Fujimura et al., 2019; Lin et al., 2012; Manfredi et al., 2017; Mat Baki et al., 2015; Uloza et al., 2015). The introduction of new validated tools may lead to wider adoption and decrease the aforementioned barriers to acoustic voice analysis.

Chapter 3

3 Materials and Methods

This study was carried out in two phases. First, the microphones and environments were evaluated using pre-recorded samples. Second, the analysis capabilities of the smartphone were evaluated using prospectively collected samples. The study procedures will be discussed in detail in the following sections.

3.1 Power Calculation

A power calculation for analysis was carried out a priori (Appendix B). For comparison of means, in a two-level between groups independent variable analysis, 17 persons per group was required. A total of 51 samples were collected to achieve this requirement.

3.2 Volunteer Recruitment

The study design and rationale were described in detail to volunteers at the Otolaryngology – Head and Neck Surgery Clinic at London Health Sciences Centre. A 5-page study-information package was provided to interested persons to review independently prior to agreeing to participate (Appendix A). Informed consent was collected from all participants who volunteered to provide a voice sample. The demographic data of included participants can be found in Table 1.

Table 1 – Participants' Demographic Data.

Group	Mean Age	Age Range	Male	Female	Pathologic
Yeti AB	40.2	23-81	7	10	1
iPhone AB	34.9	21-72	12	5	3
iPhone NAB	46.5	24-74	8	9	5
Overall	40.5	21-81	27	24	9*

* Pathologies included: unilateral vocal fold paralysis, hyperfunctional voice disorders, lung-related voice pathology, vocal fold neoplasm.

3.3 Eligibility

Inclusion criteria included English-speaking persons over the age of 18.

Exclusion criteria included pediatric patients and non-English speaking persons.

3.4 Ethics Approval

This study was approved by the Health Sciences Research Ethics Board at Western University on February 28, 2019. The letter of information and consent can be reviewed in Appendix A.

3.5 Sources of Funding

This study was supported by the Peter Cheski Innovative Resident Research Fund as well as through internal funding from the Department of Otolaryngology – Head and Neck Surgery.

3.6 Application Development

The VOICES application was developed for the proprietary purposes of this study by Dr. Benjamin van der Woerd, Min Wu, and Dr. Vijay Parsa. It was developed using XCode 10.0+ running on MacOS 10.13.6 High Sierra or higher. The programming language was Swift 5.0. Analysis algorithm development was completed using MATLAB. Validation of the algorithm relied on Praat to generate acoustic analysis values for comparison.

3.7 Recording Hardware Specifications and Settings

Two microphones were used for the collection of the audio samples: 1) an iPhone 7 Plus Internal Microphone and 2) a Blue Yeti Ultimate USB Microphone. The iPhone 7 plus internal microphone is an omnidirectional transducer, quantized at 16 bits per sample, and sampling rate of 44,100 Hz. Samples were recorded in mono. The Blue Yeti Ultimate USB Microphone records at a sampling rate of 48,000 Hz and quantized at 16 bits per sample. The polar patterns for audio capture on the Yeti microphone was set to cardioid for each sample, which consisted of a sustained vowel and continuous speech segment.

Each sample was transferred to the secure S-Drive at London Health Sciences Centre for storage.

3.8 Pre-recorded Voice Samples

To evaluate the impact of the microphone and the recording environment on the calculation of microacoustic measures, pre-recorded, high-quality samples were used. For the vowel analysis, twenty (n=20) sustained vowel (/a/) samples were included. This set of samples included normal (n=10) and dysphonic (n=10) voices; the quality of dysphonic samples ranged from mild to severe. The sustained vowel sample set included male and female voices. The continuous speech samples were based on informal perceptual evaluation of the experimenter. The sentence samples were comprised of the second sentence of The Rainbow Passage: “The rainbow is a division of white light into many beautiful colors.” (Fairbanks, 1960). This set of voices included a series of both male (n=12) and female (n=12) voices for analysis. Similar to the vowel samples, sentence samples were also judged to range in perceptual quality from those that were normal to those that were severely pathologic.

3.9 Mannequin Recording Process

A standardized process was used to capture samples of the pre-recorded voice recordings under varying conditions: Blue Yeti Microphone Soundproof Room (Yeti AB), Blue Yeti Microphone Non-Soundproof Room (Yeti NAB), iPhone Microphone Soundproof Room (iPhone AB), and iPhone Microphone Non-Soundproof Room (iPhone NAB). A calibrated mannequin speaker, which is standard practice in telecommunications and audiology research, was used to present the pre-recorded vocal signals, as seen in Figure 1. This was completed at a fixed distance of 15 inches (38.1 cm). The signal volume was calibrated to 69 dBA using a sound level meter. This resulted in four identical sets of recordings: Yeti AB, Yeti NAB, iPhone AB, and iPhone NAB.

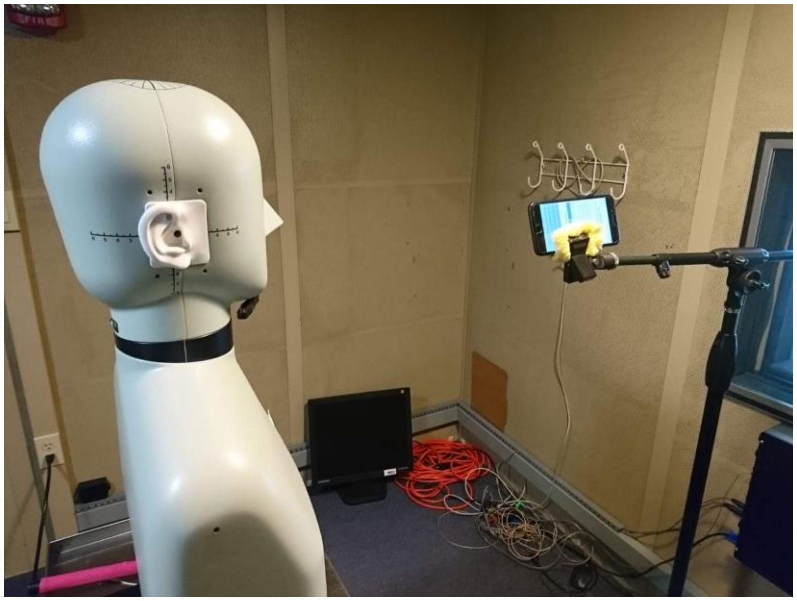


Figure 1. Mannequin speaker and iPhone microphone setup.

3.10 Part 1: Mannequin Voice Sample Analysis

The four sets of identical recordings were time-windowed to ensure the same time signal was analyzed. Windowing of signals was carried out using a MATLAB script, which truncated the recordings to include the same length of each vowel, starting and ending at the same time points of each recording for each of the four conditions. These sets of recordings were then analyzed individually using a Praat script developed by Sebastiano Failla, MSc, from the Voice Production and Perception Laboratory, Western University.

This script automated the calculation of fundamental frequency (Hz), jitter (%), shimmer (dB), harmonic-to-noise ratio (dB), and smoothed cepstral peak prominence (dB). Harmonic-to-noise ratio and cepstral peak prominence were calculated for both the sustained vowel and sentence samples, the remainder were only calculated on the sustained vowel samples. Results were output to an Excel file for later comparison.

3.11 Part 1: Mannequin Statistical Analysis

A repeated measures ANOVA was used to analyze the four sets of identical recordings from the pre-recorded signals. We report descriptive statistics, including the mean, median, standard deviation and standard error of the mean values. All statistics were completed using SPSS version 23.0.0.0.

3.12 Prospective Voice Sample Collection

After signing an informed consent, volunteers were enrolled in a convenience pattern to one of three recording configurations: Blue Yeti Microphone Soundproof Room (Yeti AB), iPhone Microphone Soundproof Room (iPhone AB), or iPhone Microphone Non-Soundproof Room (iPhone NAB). The recording protocol was explained to them in detail and examples of the volume and pitch were shown prior to starting. They were given the opportunity to ask questions prior to initiating the recording process. The selected microphone was placed 6 inches from the volunteer’s mouth at a 45-degree angle. Distance was measured using a small ruler and volunteers were instructed to maintain this position throughout the recording process.

The recording protocol was a three-task process: sustained vowels, CAPE-V sentences, and a free running speech sample. In each task, the participants were to speak at a comfortable, conversational pitch and volume.

For the sustained vowel task, volunteers were asked to say /a/ as in "dot", /i/ as in "deep", and /u/ as in "boot" for six seconds each. This task was repeated three times.

The sentences were taken from the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) for the purposes of testing specific aspects of laryngeal behaviour. Those sentences are:

1. "The blue spot is on the key again"
    a. Includes all vowel sounds in the English language
2. "How hard did he hit him"
    a. Easy onset with /h/ sound
3. "We were away a year ago"
    a. All voiced sounds
4. "We eat eggs every Easter"
    a. Hard glottal attack
5. "My mama makes lemon jam"
    a. Contains nasal sounds

6. “Peter will keep at the peak”

    a. Voiceless plosive sounds

The third task asked the participant to speak about a hobby or special interest for 20 seconds of natural speech.

From each of the sustained vowel samples, a steady portion of sustained /a/ was extracted for analysis. The six CAPE-V sentences were extracted together. With this, a sustained vowel and standardized continuous speech sample was available for analysis from each prospective sample.

3.13 Part 2: Praat vs. On-device Voice Sample Analysis

Following the collection of the voice samples, each sample was exported for long-term storage. The sustained vowel /a/ segments and CAPE-V segments were extracted from each sample. The resultant sustained vowel (n=51) and CAPE-V (n=51) samples were analyzed for the collection of seven variables using Praat and the proprietary on-device analysis. The previously mentioned script was used to automate the calculation of fundamental frequency (Hz), jitter (%), shimmer (dB), harmonic-to-noise ratio (dB), and cepstral peak prominence (dB). Fundamental frequency, jitter, and shimmer were calculated only for sustained vowel samples. Results were output to an excel file for comparison. For the on-device analysis, each of the samples were loaded onto an iPhone 7 Plus and analysis was completed using our in-house developed algorithm. Each of the same variables were calculated and transferred to an excel file for comparison.

3.14 Part 2: Praat vs. On-device Statistical Analysis

All statistical analysis was completed using SPSS. Means were compared between groups using paired sample t-tests. Mean, median, standard deviation, and standard error of the mean are also reported.

Chapter 4

4 Results

This study was carried out in two phases. First, we evaluated the impact of the microphone of choice on our ability to calculate our chosen microacoustic measures: F_0, jitter, shimmer, HNR, and CPP. In the assessment of sustained vowel samples, all five acoustic measures were used. For the evaluation of the CAPE-V sentences, only HNR and CPP were used. To differentiate the HNR and CPP measures for each of the samples, they have been labeled as HNR-V, CPP-V, HNR-S, and CPP-S to reflect vowel and sentence analyses respectively. At the same time, we evaluated the impact of the recording environment on the same measures. Next, we prospectively collected voice samples to analyze in the traditional, gold standard method of analysis (Praat), and the proprietary, on-device analysis algorithm developed for this project (Boersma, 2002). In the following sections, the results of the microphone impact, recording environment impact, and comparison of the proprietary algorithm and Praat on sustained vowel and continuous speech samples will be discussed.

4.1 Microphone Impact

During the first phase, microphone impact was evaluated for the calculation of fundamental frequency (F_0), jitter, shimmer, HNR-V, CPP-V, HNR-S, and CPP-S. A repeated measures ANOVA was used to evaluate the different microphone conditions. A statistically significant difference was identified between the two microphones for the calculation of shimmer, HNR-V, CPP-V, HNR-S, and CPP-S (see Table 2). Calculation of F_0 and jitter were not significantly different when comparing the choice of microphone in this study (Table 2). An interaction between the variables, microphone choice and recording environment, was present for the measurement of shimmer, HNR-V, CPP-V, and CPP-S.

Table 2 - Repeated Measures ANOVA Results.

F_0	Microphone: p = .564 Booth: p = .013
Jitter	Microphone: p = .519 Booth: p = .516
Shimmer*	Microphone: p < .001 Booth: p < .001
HNR-V*	Microphone: p < .001 Booth: p < .001
CPP-V	Microphone: p < .001 Booth: p < .001
HNR-S	Microphone: p < .001 Booth: p = .167
CPP-S*	Microphone: p < .001 Booth: p = .002

Note. *indicates interaction between variables.

Mean, median, standard deviation, and standard error of the mean of each measurement across the four differing conditions are presented in Tables 3-9. These show a small overall difference for F_0, jitter, CPP-V, HNR-S, and CPP-S. A more substantial difference was found in the calculation of shimmer and HNR-V. Scatter plots of each measurement can be found in Figures 2-8.

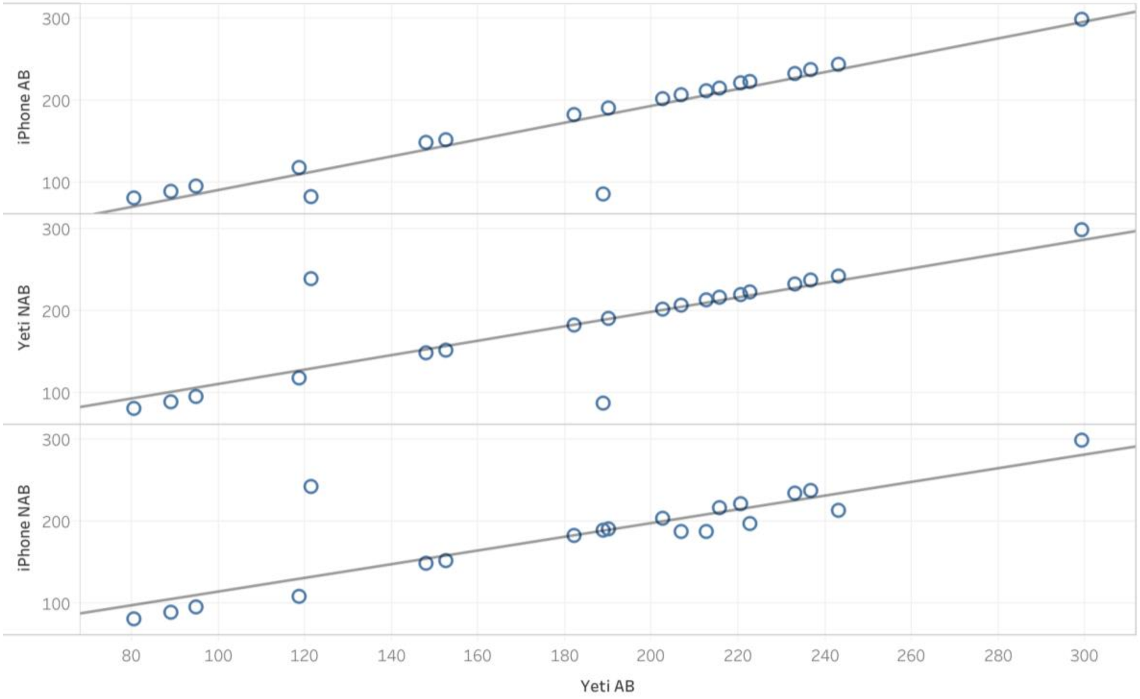


Figure 2. F_0 scatter plot comparison: Yeti AB: iPhone AB/Yeti NAB/iPhone NAB.

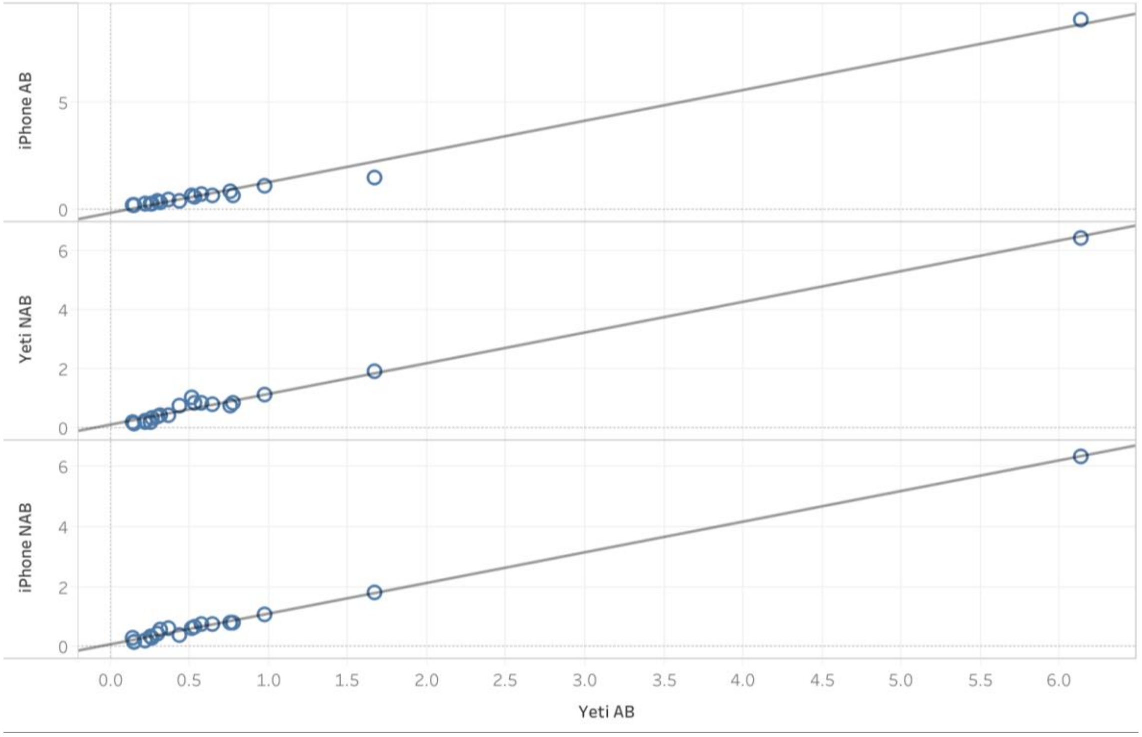


Figure 3. Jitter scatter plot comparison: Yeti AB: iPhone AB/Yeti NAB/iPhone NAB.

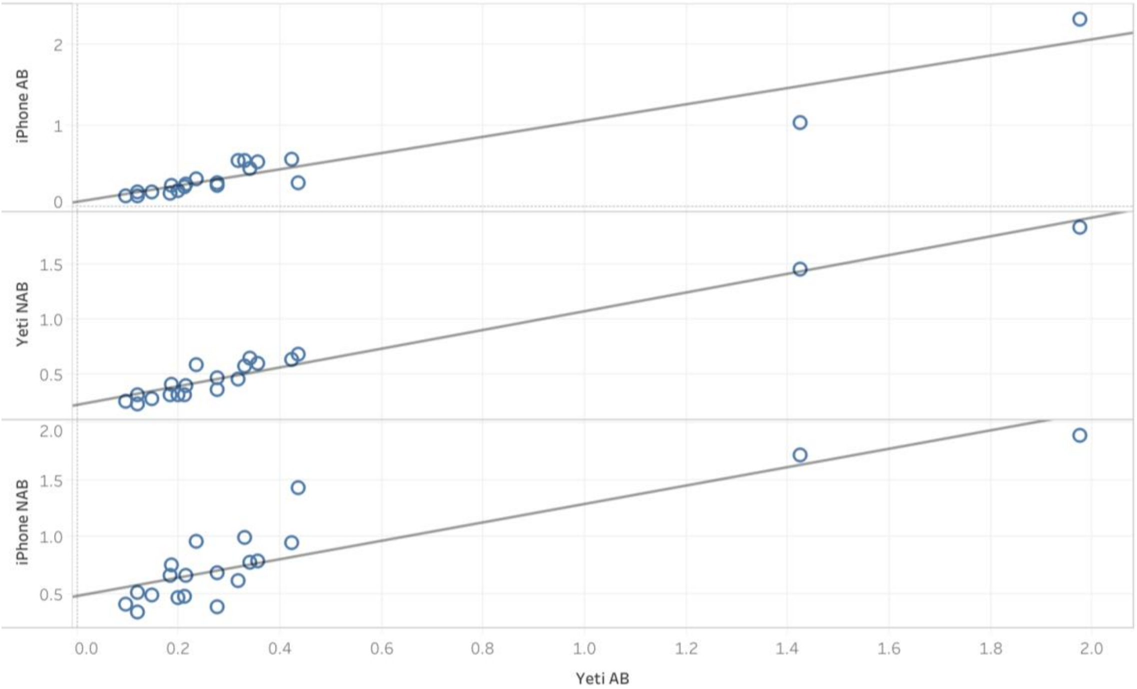


Figure 4. Shimmer scatter plot comparison: Yeti AB: iPhone AB/Yeti NAB/iPhone NAB.

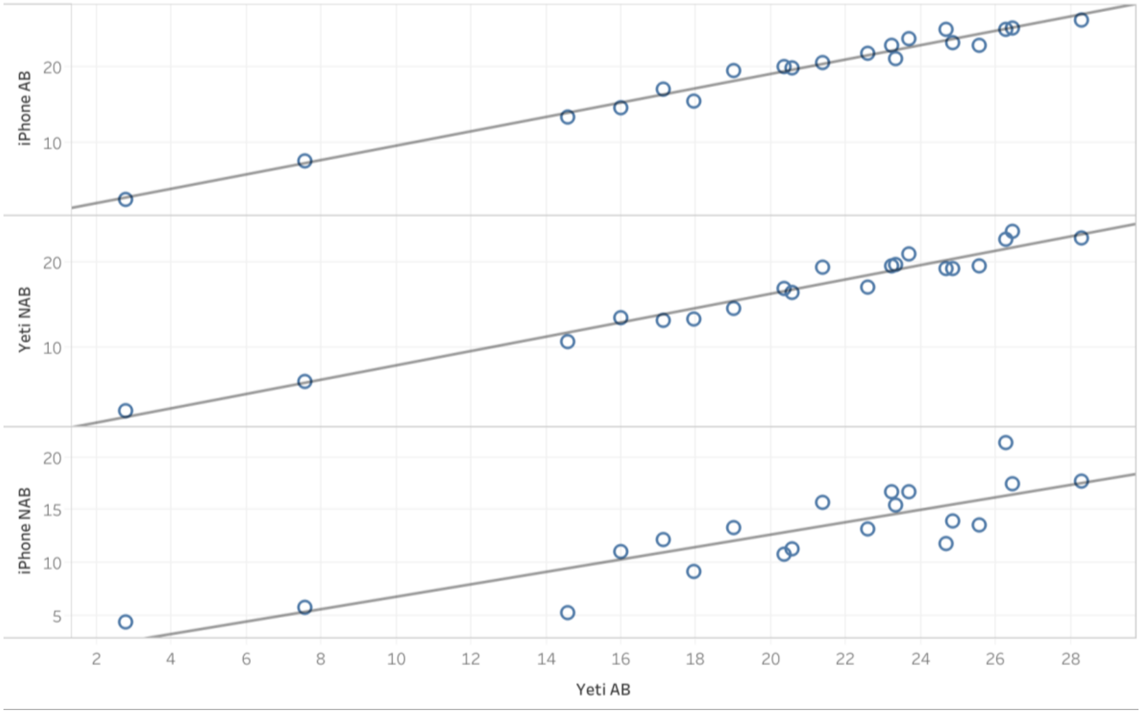


Figure 5. HNR Vowels scatter plot comparison: Yeti AB: iPhone AB/Yeti NAB/iPhone NAB.

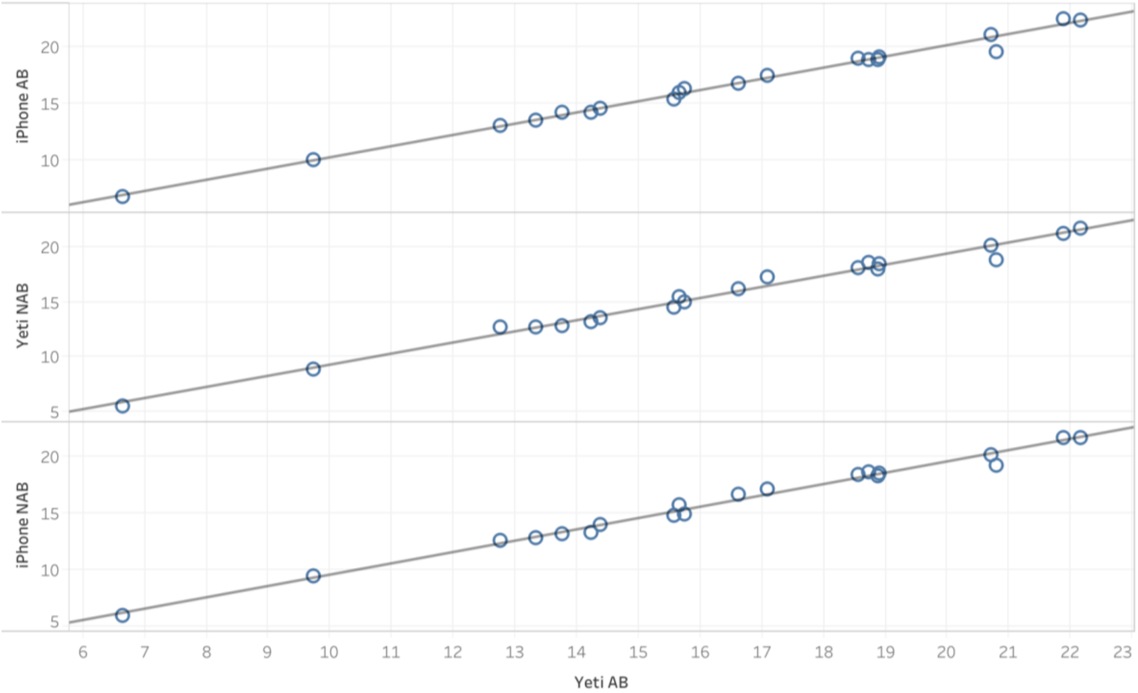


Figure 6. CPP Vowels scatter plot comparison: Yeti AB: iPhone AB/Yeti NAB/iPhone NAB.

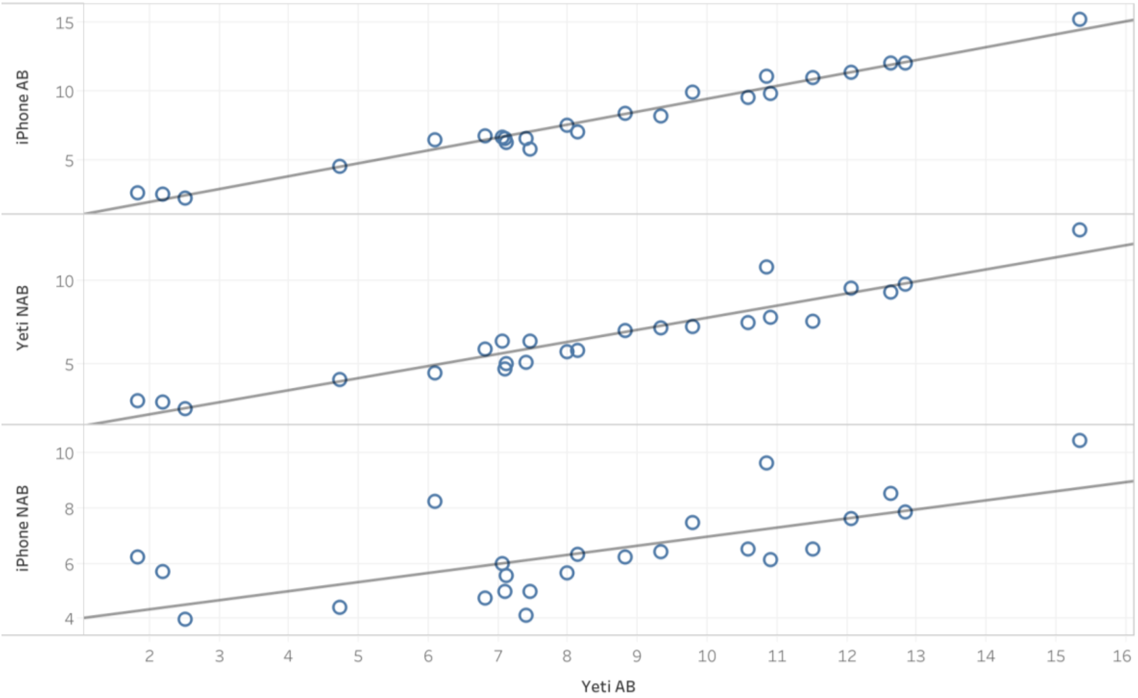


Figure 7. HNR Sentences scatter plot comparison: Yeti AB: iPhone AB/Yeti NAB/iPhone NAB.

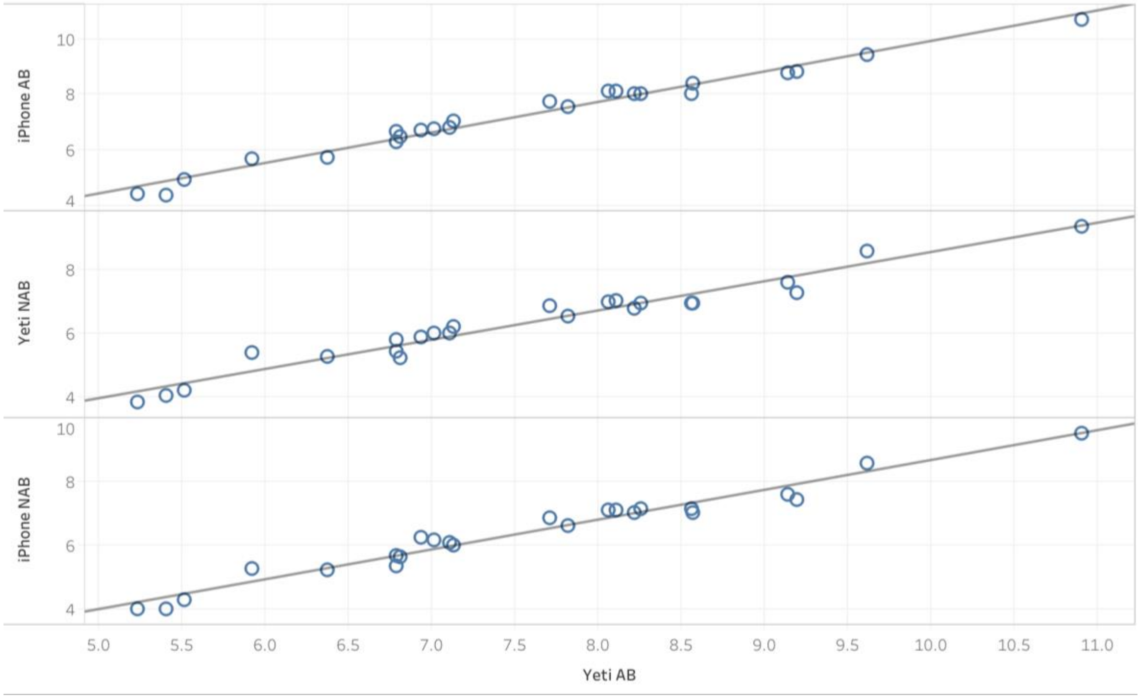


Figure 8. CPP Sentences scatter plot comparison: Yeti AB: iPhone AB/Yeti NAB/iPhone NAB.

4.2 Recording Environment Impact

In the same phase, we evaluated the recording environment's impact on calculation of the same seven variables. A repeated measures ANOVA was used. A statistically significant difference was identified for the calculation of F_0, shimmer, HNR-V, CPP-V, and CPP-S (see Table 2). Using this data set, the calculation of jitter and HNR-S were not statistically different when comparing the recording environment of a soundproof booth to a quiet office (see Table 2).

Mean, median, standard deviation, and standard error of the mean for each recording condition are reported in Tables 3-9. These indicate limited differences with respect to the calculation of F_0, jitter, CPP-V, HNR-S, and CPP-S. A larger impact was identified for calculation of shimmer and HNR-V.

Table 3. Mean, median, standard deviation, standard error of the mean for F_0 as measured across the following four conditions: Yeti AB, iPhone AB, Yeti NAB, and iPhone NAB.

F_0	Yeti AB	iPhone AB	Yeti NAB	iPhone NAB
Mean	183.04	175.81	183.84	183.48
Median	196.43	196.43	204.89	189.27
Standard Deviation	58.81	65.15	62.52	56.87
Standard Error of the Mean	13.15	14.57	13.98	12.72

Table 4. Mean, median, standard deviation, standard error of the mean for jitter as measured across the following four conditions: Yeti AB, iPhone AB, Yeti NAB, and iPhone NAB.

Jitter	Yeti AB	iPhone AB	Yeti NAB	iPhone NAB
Mean	0.77	0.91	0.90	0.86
Median	0.40	0.39	0.59	0.60
Standard Deviation	1.32	1.90	1.37	1.34
Standard Error of the Mean	0.29	0.43	0.31	0.30

Table 5. Mean, median, standard deviation, standard error of the mean for shimmer as measured across the following four conditions: Yeti AB, iPhone AB, Yeti NAB, and iPhone NAB.

Shimmer	Yeti AB	iPhone AB	Yeti NAB	iPhone NAB
Mean	0.39	0.45	0.55	0.80
Median	0.26	0.28	0.43	0.67
Standard Deviation	0.47	0.49	0.40	0.43
Standard Error of the Mean	0.10	0.11	0.09	0.10

Table 6. Mean, median, standard deviation, standard error of the mean for HNR Vowel as measured across the following four conditions: Yeti AB, iPhone AB, Yeti NAB, and iPhone NAB.

HNR Vowel	Yeti AB	iPhone AB	Yeti NAB	iPhone NAB
Mean	20.31	19.41	16.50	12.82
Median	22.00	20.87	18.07	13.21
Standard Deviation	6.40	6.18	5.47	4.40
Standard Error of the Mean	1.43	1.38	1.22	0.98

Table 7. Mean, median, standard deviation, standard error of the mean for CPP Vowel as measured across the following four conditions: Yeti AB, iPhone AB, Yeti NAB, and iPhone NAB.

CPP Vowel	Yeti AB	iPhone AB	Yeti NAB	iPhone NAB
Mean	16.31	16.47	15.62	15.83
Median	16.18	16.52	15.77	16.15
Standard Deviation	4.00	3.99	4.05	4.01
Standard Error of the Mean	0.89	0.89	0.91	0.90

Table 8. Mean, median, standard deviation, standard error of the mean for HNR Sentences as measured across the following four conditions: Yeti AB, iPhone AB, Yeti NAB, and iPhone NAB.

HNR Sentences	Yeti AB	iPhone AB	Yeti NAB	iPhone NAB
Mean	8.38	7.90	6.55	6.42
Median	8.06	7.25	6.34	6.23
Standard Deviation	3.45	3.27	2.65	1.66
Standard Error of the Mean	0.70	0.67	0.54	0.34

Table 9. Mean, median, standard deviation, standard error of the mean for CPP Sentences as measured across the following four conditions: Yeti AB, iPhone AB, Yeti NAB, and iPhone NAB.

CPP Sentences	Yeti AB	iPhone AB	Yeti NAB	iPhone NAB
Mean	7.55	7.23	6.28	6.38
Median	7.42	7.30	6.36	6.43
Standard Deviation	1.40	1.55	1.31	1.34
Standard Error of the Mean	0.29	0.32	0.27	0.27

4.3 Prospective Recordings - Sustained Vowels

The second phase of the present study evaluated prospectively collected voice samples that were analyzed using two methods: 1) via the proprietary algorithm and 2) using Praat. Prospectively recorded voice samples yielded a sustained vowel and continuous speech sample for analysis. This resulted in two paired data sets (n=51), one set analyzed by each algorithm. This data was compared using paired sample t-tests in SPSS.

Analysis revealed a statistically significant difference in the calculation of F_0, Shimmer, HNR-V, and CPP-V. The correlations were strong between each of these measures (see Tables 10, 12-14). The differences between the calculated values for jitter were not found to be statistically significant. The correlation for jitter was moderate (see Table 11). In order to visually evaluate the present data, scatter plots comparing the two groups with respect to these five measures are graphically represented in Figures 9-13.

Table 10. Mean, standard deviation, standard error of the mean, and correlation coefficient for F_0 as calculated by Praat and Proprietary algorithms.

F_0	Mean	N	St. Dev	Std. Error Mean	Correlation (p-value)
Praat	142.43	51	50.26	7.04	0.879 (p<.001)
Proprietary	153.25	51	48.97	6.86	

Table 11. Mean, standard deviation, standard error of the mean, and correlation coefficient for jitter as calculated by Praat and Proprietary algorithms.

Jitter	Mean	N	St. Dev	Std. Error Mean	Correlation (p-value)
Praat	0.54	51	0.32	0.05	0.355 (p=.110)
Proprietary	0.43	51	0.85	0.12	

Table 12. Mean, standard deviation, standard error of the mean, and correlation coefficient for shimmer as calculated by Praat and Proprietary algorithms.

Shimmer	Mean	N	St. Dev	Std. Error Mean	Correlation (p-value)
Praat	0.40	51	0.23	0.03	0.829 (p<.001)
Proprietary	0.52	51	0.32	0.04	

Table 13. Mean, standard deviation, standard error of the mean, and correlation coefficient for HNR vowel as calculated by Praat and Proprietary algorithms.

HNR Vowel	Mean	N	St. Dev	Std. Error Mean	Correlation (p-value)
Praat	19.11	51	4.27	0.60	0.964 (p<.001)
Proprietary	16.71	51	3.90	0.55	

Table 14. Mean, standard deviation, standard error of the mean, and correlation coefficient for CPP vowel as calculated by Praat and Proprietary algorithms.

CPP Vowel	Mean	N	St. Dev	Std. Error Mean	Correlation (p-value)
Praat	19.99	51	3.64	0.51	0.881 (p<.001)
Proprietary	23.39	51	3.92	0.55	

The mean F_0 calculated using Praat was 142.43 Hz, compared to 153.25 Hz when calculated with the proprietary application and correlation between the calculations was 0.879 (p < .001) (see Table 10). This indicates a reliable identification of the fundamental period, which is critical for the calculation for both jitter and shimmer, as previously discussed. Interestingly, mean jitter calculated in Praat was 0.54 %, compared to 0.43 % when calculated with the proprietary application. Here, the correlation between the calculations was only moderate at 0.355 (p = .011) (see Table 11). Next, the correlation of shimmer was strong, at 0.829 (p < .001) (see Table 12). The mean shimmer calculated in Praat was 0.40 dB, compared to 0.52 dB when calculated with the proprietary application. Once again there was a strong correlation between the two methods of measurement for HNR-V (0.964, p < .001) and CPP-V (0.881, p < .001). The mean HNR-V calculated in Praat was 19.11 dB, compared to 16.71 dB when calculated with the proprietary application. Finally, the mean CPP-V calculated in Praat was 18.99 dB, compared to 23.39 dB when calculated with the proprietary application.

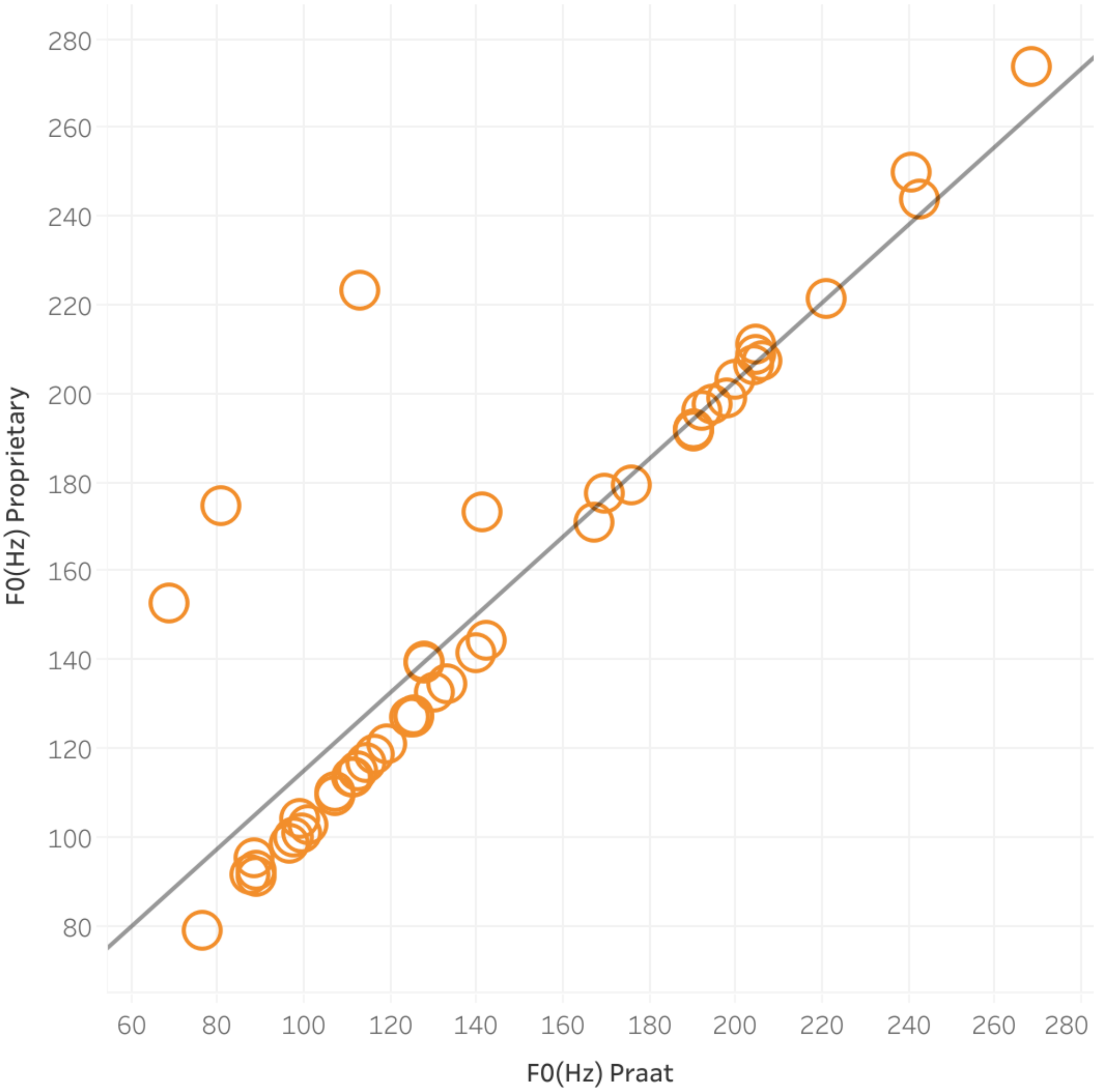


Figure 9. F_0 scatter plot comparison: Proprietary algorithm vs. Praat.

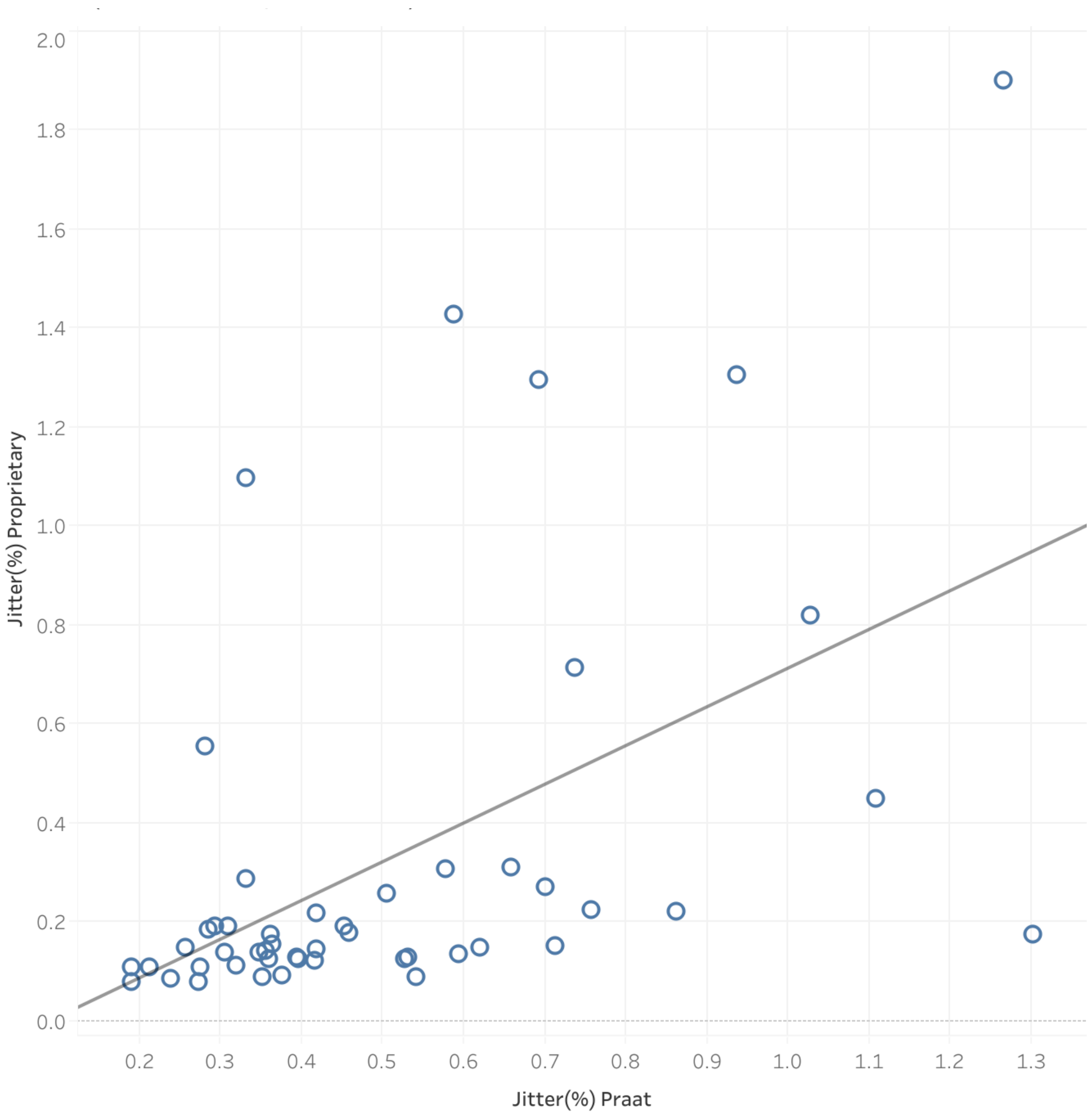


Figure 10. Jitter scatter plot comparison: Proprietary algorithm vs. Praat.

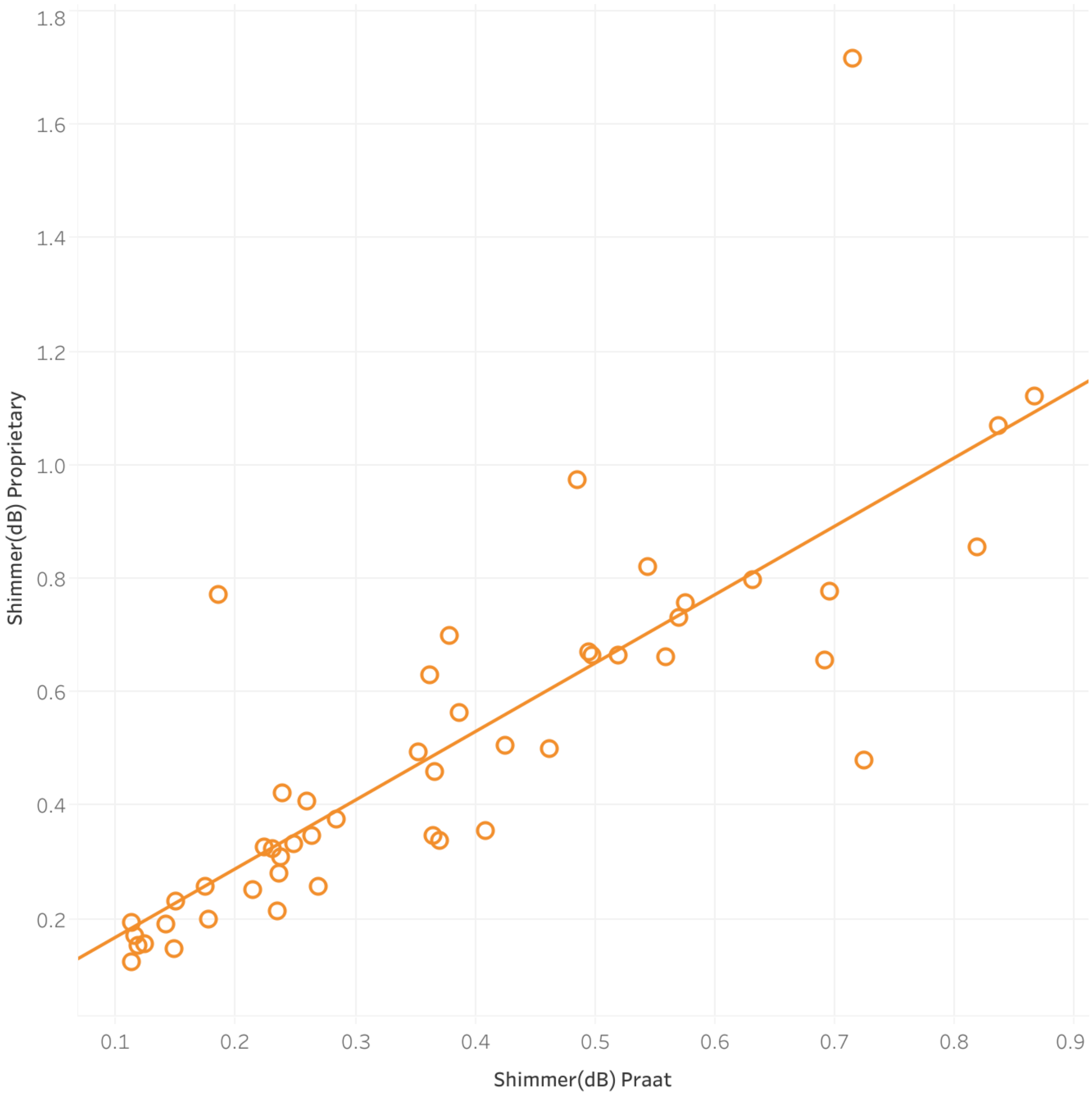


Figure 11. Shimmer scatter plot comparison: Proprietary algorithm vs. Praat.

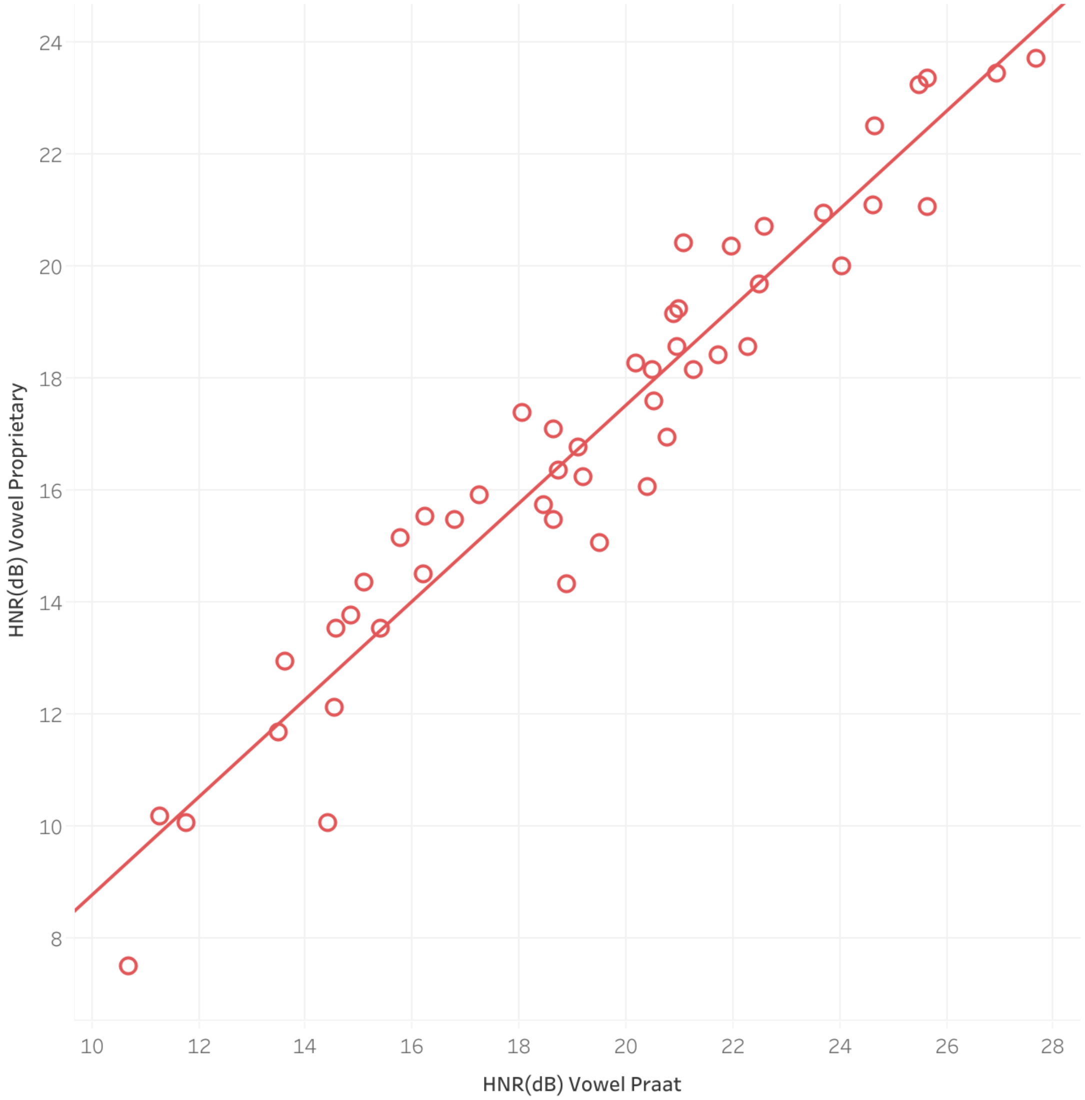


Figure 12. HNR Vowel scatter plot comparison: Proprietary algorithm vs. Praat.

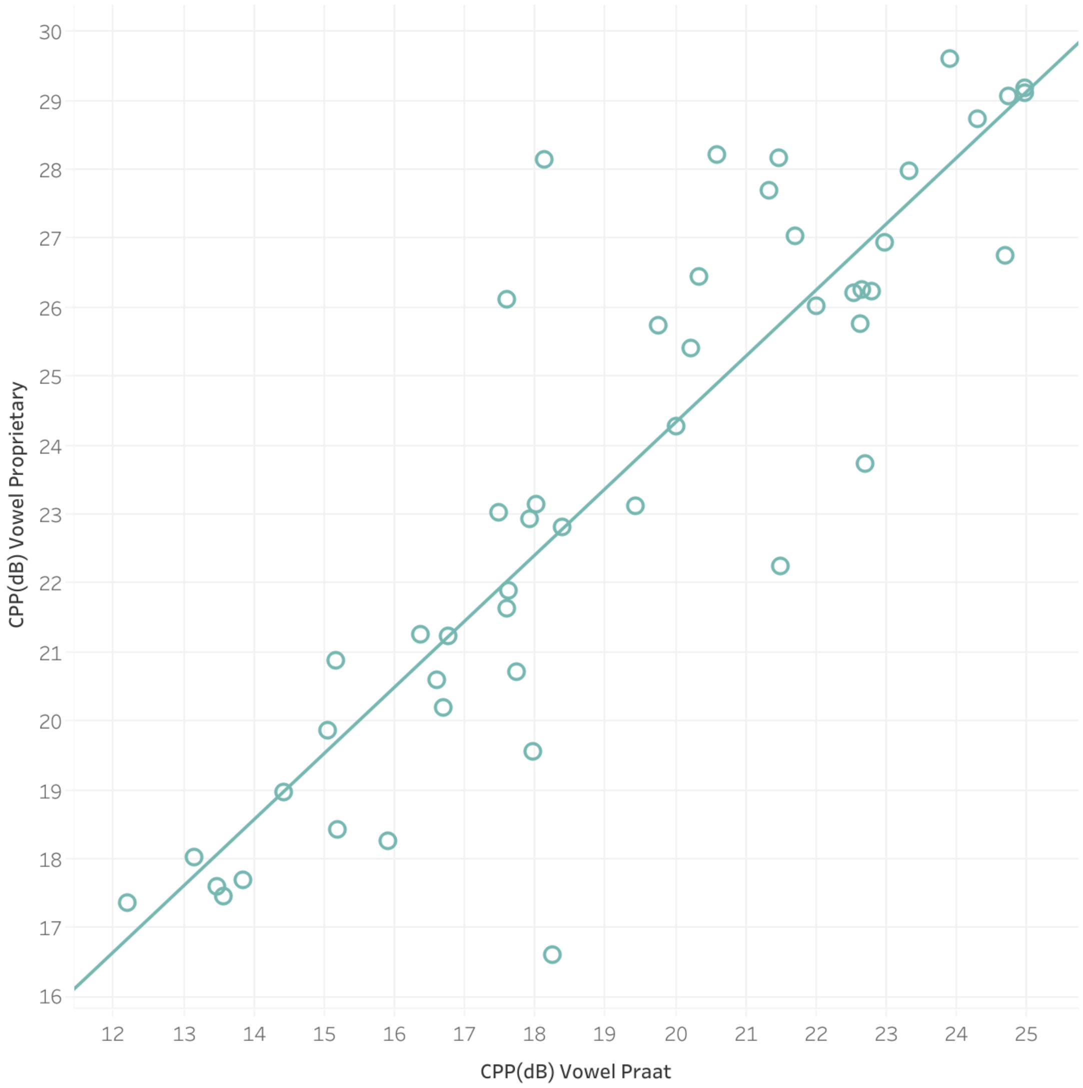


Figure 13. CPP Vowel scatter plot comparison: Proprietary algorithm vs. Praat.

4.4 Prospective Recordings - Continuous Speech

In addition to the above sustained vowel analysis, the same two paired data sets of 51 continuous speech samples were analyzed using Praat and the proprietary developed application. These data were evaluated using paired sample t-tests in SPSS. Because of the nature of continuous speech and the increased complexity of the signal, only HNR-S and CPP-S were calculated for these samples.

In the comparison of means, the differences in CPP-S and HNR-S were found to be statistically significant (p < .001) as seen in Tables 15 and 16. In this analysis, the differences between the values were substantially larger. The mean CPP-S calculated in Praat was 9.197 dB, compared to 16.985 dB when calculated with the proprietary application (see Table 16). Similarly, for HNR-V, the mean calculated in Praat was 12.440 dB, compared to 4.132 dB when calculated with the proprietary application (see Table 15). These differences were reflected in the correlations between the values, which was strong for CPP-S (0.632, p < 0.001), and moderate for HNR-S (0.299, p = .033) (see Tables 15-16). Scatter plots comparing the two groups with respect to these two measures are also visually presented in Figures 14-15.

Table 15. Mean, standard deviation, standard error of the mean, and correlation coefficient for HNR sentences as calculated by Praat and Proprietary algorithms.

HNR Sentences	Mean	N	St. Dev	Std. Error Mean	Correlation (p-value)
Praat	12.44	51	1.86	0.26	0.299 (p=.033)
Proprietary	4.13	51	1.92	0.27	

Table 16. Mean, standard deviation, standard error of the mean, and correlation coefficient for CPP sentences as calculated by Praat and Proprietary algorithms.

CPP Sentences	Mean	N	St. Dev	Std. Error Mean	Correlation (p-value)
Praat	9.20	51	1.20	0.17	0.632 (p<.001)
Proprietary	16.98	51	0.92	0.13	

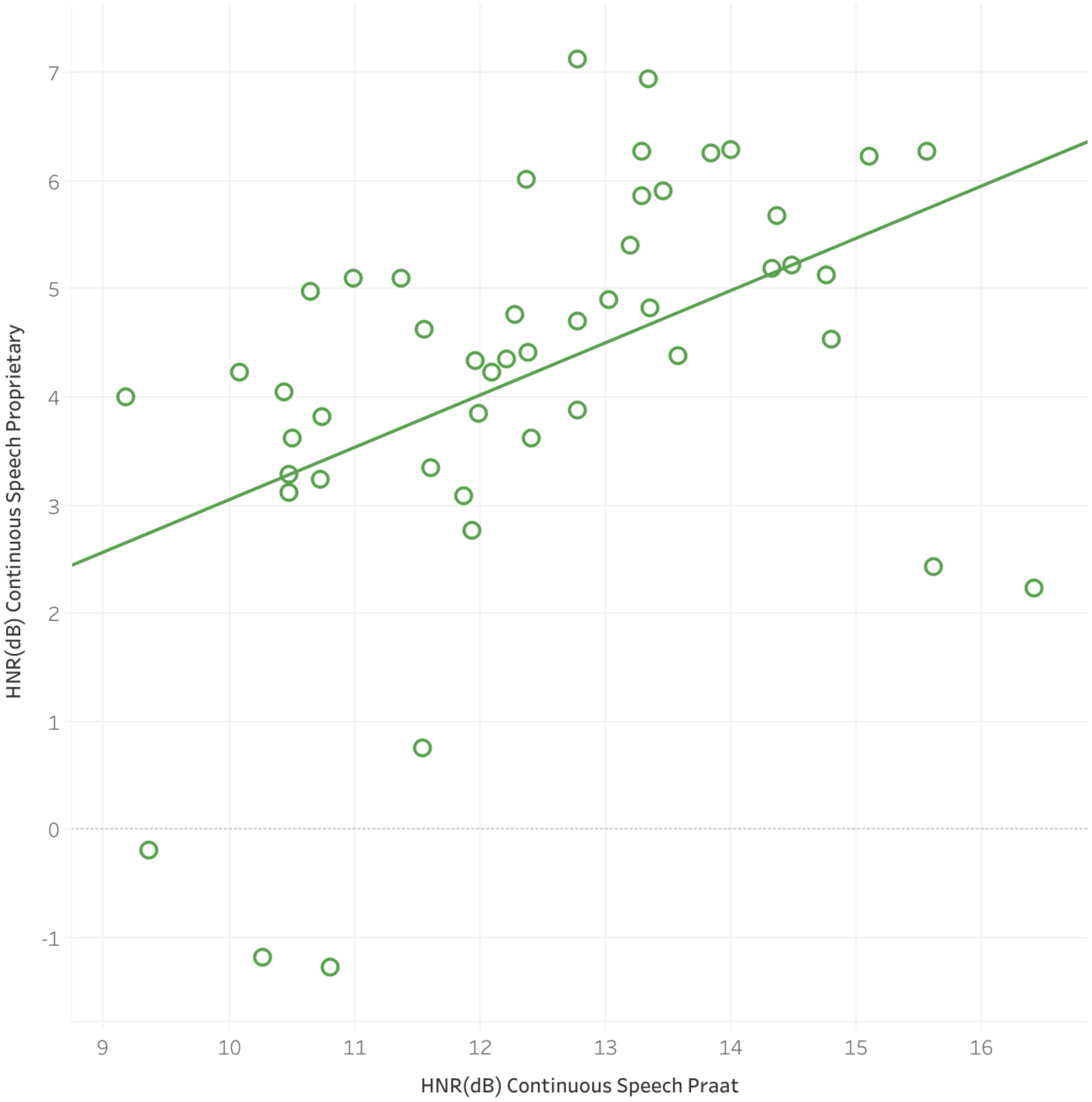


Figure 14. HNR Sentences scatter plot comparison: Proprietary algorithm vs. Praat.

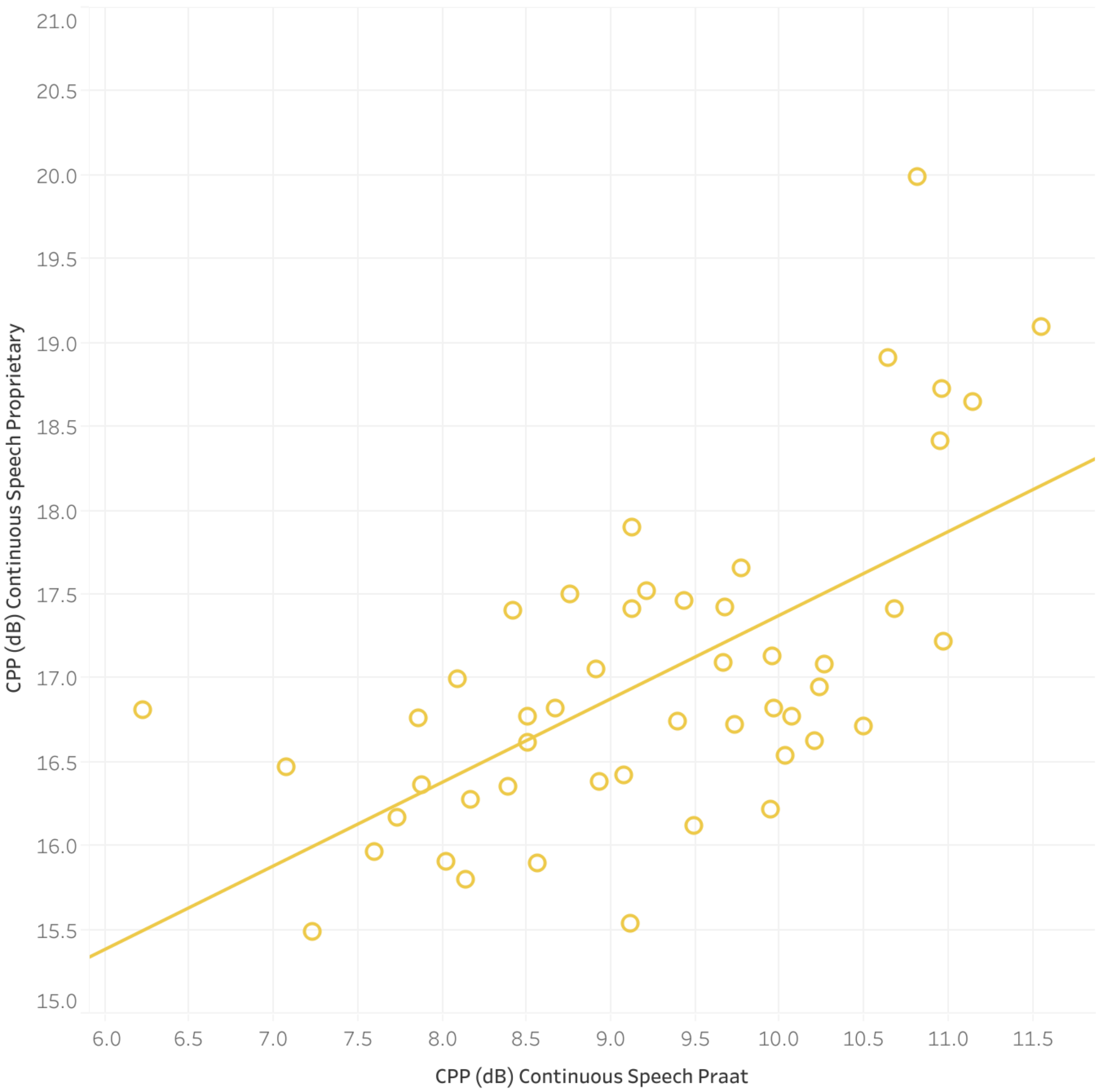


Figure 15. CPP Sentences scatter plot comparison: Proprietary algorithm vs. Praat.

Chapter 5

5 Discussion and Conclusion

Voice analysis is an important part of a thorough voice examination. The comprehensive evaluation can include auditory-perceptual analysis, aerodynamic studies, and acoustic voice analysis, amongst others. Acoustic analysis can provide objective measurements for diagnostic and ongoing monitoring documentation and such data can complement auditory-perceptual assessments or prompt further investigations. The existing methods of voice sample collection and analysis are often resource intensive, creating barriers to access. In the current work, we sought to empirically examine a multi-phase study evaluating smartphones as an avenue to perform voice sample acquisition and on-device acoustic voice analysis. In the sections to follow, a discussion of the following areas of interest will be addressed. First, the assessment of the microphones and recording environments, as completed in phase one will be detailed. Second, we will discuss the results of the prospectively collected samples and the validity of the proprietary algorithm. Next, limitations of the study will be addressed. Clinical implications of this study and future study directions will each be discussed separately.

5.1 Mannequin Study

In phase one of this study, microphones for voice sample collection and the influence of recording environment were evaluated. Pre-recorded samples were captured under four different conditions (i.e., Yeti microphone in a soundproof booth, iPhone microphone in a soundproof booth, Yeti microphone in a quiet office, and iPhone microphone in a quiet office) to reflect the two different microphones and recording environments that were evaluated. The resultant data revealed that statistically significant impacts from both the microphone and the environment do exist on the acoustic measurements, albeit limited. The differences in values that were produced were typically of small magnitude, that would not represent a clinically important difference in most cases. For example, the difference between the mean F_0 in the gold standard scenario (Yeti microphone in a soundproof room) and the least controlled setting (iPhone microphone in a quiet office) revealed a 0.9 Hz difference. Normative ranges for F_0 are approximately 120 Hz for men and 210 Hz for women, where a half hertz difference would not make an impact on clinical decision making. Similarly, small differences were found in the calculation of jitter and CPP-V, whereby clinical interpretation would not be affected. By introducing a lower quality microphone, many, but not all of the acoustic measures evaluated were found to be significantly different. Similarly, by recording in a less controlled environment, there were impacts identified specific to the calculation of microacoustic measures. That is, the calculation of F_0, shimmer, CPP-V, and CPP-S all showed statistically significant differences when comparing the recording environments. The differences that were identified are consistent as expected, but small in overall effect. The magnitude of difference between the soundproof room and the quiet office was modest in the calculation of F_0, jitter, CPP-V, HNR-S, and CPP-S. Despite the statistically significant differences observed, these impacts do not appear to represent a clinically important difference in the ability to measure these selected acoustic measures.

A repeated measures ANOVA indicates statistically significant differences that are present consistently between the variables being assessed. However, it does not indicate what the difference between the variables is. That is, whether one variable being assessed was higher or lower than the other cannot be verified using an ANOVA. To understand the identified results, the mean values were graphed in a scatter plot to evaluate which microphone had an effect, what the impact was, and the degree of difference between the microphones. Based on evaluation of the means of each group, it was clear that the effect of the microphone was limited and largely viewed to be clinically acceptable (Tables 2-8). However, calculation of shimmer and HNR-V had more significant effects from both the microphone and recording environment; therefore, the interpretation of these results should be undertaken with caution in a clinical setting. Though formal minimal clinically important differences are not known for either measurement in Praat, the normative values for each are informative in this context. The normative mean value for HNR is 11.9 dB (CI 7.0-17.0) (Yumoto et al., 1982). The normative range for shimmer is typically greater than 0.350 dB as a pathologic threshold (Praat Documentation, 2003). Given the greater mean differences observed in each of these values, the likelihood of a false result is elevated. For these reasons these results should be contextualized within the conditions the voice sample was collected in.

5.2 Proprietary Application Analysis

Phase two of this study evaluated the proprietary application against the current standard used for acoustic voice analysis (Praat) (Boersma, 2002). Prospective voice samples were collected from volunteers. From these, a sustained vowel and continuous speech sample were collected and relevant components of these samples were extracted for analysis.

For the vowel analysis, the proprietary algorithm was observed to perform well when compared to the data generated using Praat. The correlations of the measures across the two platforms ranged from moderate-to-strong (0.355 to 0.964). Jitter results were the least correlated measure based on the current results. Looking at the jitter scatter plot (Figure 10), one can observe that there are 6 major outliers. Evaluation of the raw data reveals the selected fundamental period to be substantially different between the two analyses. As an example, one such non-dysphonic female sample had her F_0 as measured by Praat to be 112.92 Hz, measured as 223.43 Hz by the proprietary algorithm. Closer examination of the pitch track reveals this to have been incorrectly selected by Praat (see Figure 16). This explains the variability in the jitter, which relies on fundamental period calculation to accurately determine. Jitter is a measure of cycle-to-cycle period variability. It compares the cycle-to-cycle differences to the observed fundamental period. If the fundamental period is incorrectly identified, then its utility as the comparator group is negated. The subsequent cycles will inevitably be highly variable, and the percentage differences will rise inappropriately leading to high measured jitter.

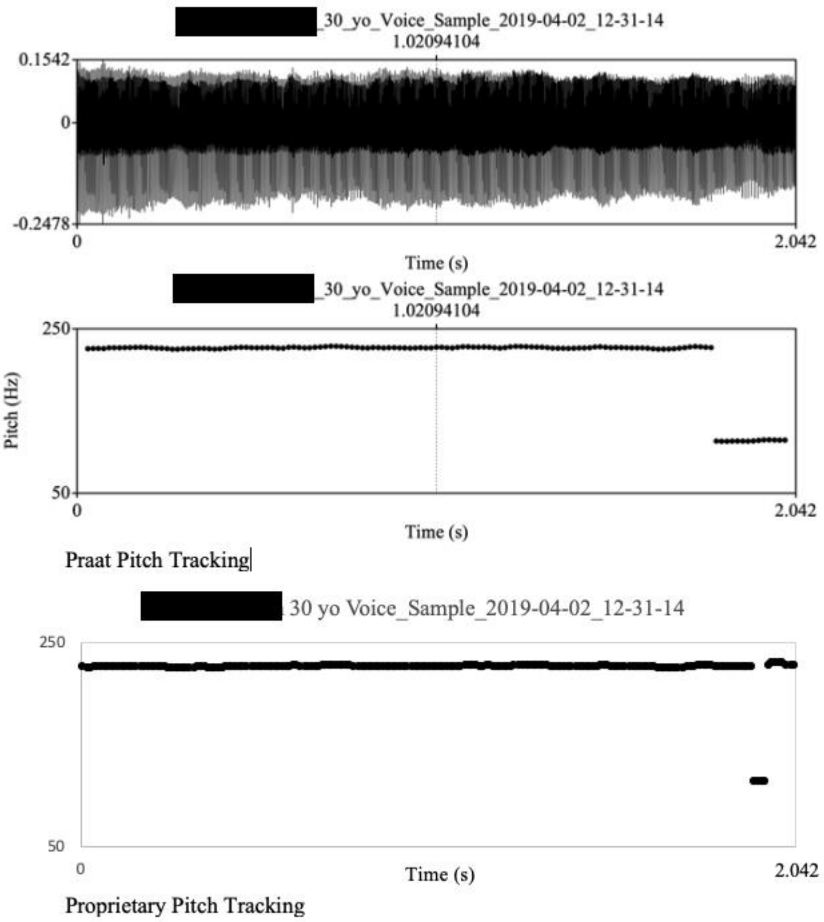


Figure 16. Pitch tracking example; Praat vs. proprietary algorithm.

When reviewing the sustained vowel analysis, these data provided validation that the proprietary algorithm developed for this study can reliably measure our chosen acoustic variables. Furthermore, the fundamental frequency was correctly identified in some cases that appeared to have been miscalculated by Praat, leading to some of the discrepancies in the comparison of means. Importantly, this impacts not only the calculation of the mean F_0 values that were compared but has downstream impacts on the calculation of both jitter and shimmer. Furthermore, this is an indication of the robustness of the algorithm, whereby the accuracy was maintained in noisier samples, even in the context of miscalculations by the gold standard (Praat).

The sustained vowel analysis resulted in highly valid results based on the data in this study. However, analysis of continuous speech is a significantly more challenging task. More specifically, running speech signals are vastly more complex in the range of frequency and amplitude, as well as in the temporal domain (e.g., changes in speech rate, pauses, voice breaks between segments, etc.). The comparison of HNR-S and CPP-S measures for the continuous speech samples reflects these challenges. HNR-S, when calculated by the proprietary algorithm, had a moderate correlation (0.299) when compared to the same measure calculated by Praat. As discussed above, this acoustic measure is calculated by taking the smoothed waveform to represent the harmonic signal, subtracted from the complex overall signal, and the remaining waveform reflects the background noise. By not concatenating the continuous speech sample, there are breaks that include ambient noise. Ambient noise is the background noise in the samples. Specifically, in the context of samples collected outside of a soundproof setting, this background noise affects the calculation of microacoustic measures. It contributes noise to the signal, which affects the residual waveform when the harmonic signal is removed in the calculation of the HNR. Furthermore, the linear regression line for the calculation of CPP, which is based on the amplitude of the signal, will be affected by the presence of background noise. Ambient noise between voice segments was not accounted for in the proprietary analysis, which likely explains the reasons for the significantly different values; in this instance, this may reflect an altered estimation of the presence of dysphonia.

Despite this, in comparison to the HNR evaluation data, CPP performed better and demonstrated a strong correlation of 0.632. Once again, however, there was a significant discrepancy between the means of the comparator groups. This suggests that although this measure was shown to be reliable, it was different from the same measures provided using Praat. Thus, method of analysis remains an important area of consideration relative to acoustic measures of voice.

These collective results bring forth an important consideration when evaluating measures of acoustic voice analysis. More specifically, these data verify the importance of setting normative values for the analysis conditions (microphone, algorithm, and recording environment) (Watts, Awan, & Maryn, 2017). This is always an important consideration when comparing acoustic measures collected through different methods, which can have a variety of different algorithmic variances that can lead to different results (Watts et al., 2017). While differences across the analysis methods of the same measure are not to be unexpected, these findings stress the continued importance of establishing a normative database for both normal and disordered voices with any given system. As further support to this suggestion Watts et al (2017) compared two validated methods of calculating CPP and showed both resulted in reliable, though different, measurements. Yet most importantly, though the two methods are valid, they cannot be directly compared (Watts et al., 2017). With respect to this study, the measurement of CPP in continuous speech sample held strong correlation to Praat.

In summary, the present project has validated a new proprietary algorithm for acoustic voice analysis of sustained vowels; in some cases, outperforming the gold standard method. Furthermore, it has resulted in reliable measurements of CPP-S. This study has highlighted the importance of establishing normative values for each condition of acoustic voice analysis, to be used as context for interpreting clinical results. Finally, the results of the sentence analysis confirmed the findings that reliable measures of the same variable cannot necessarily be compared across different measurement methods.

5.3 Study Limitations

It is important to contextualize the results of this study, as well as to recognize the limitations of acoustic analysis. As previously stated, acoustic measurements do not always correlate with auditory-perceptual assessments. Understanding this phenomenon when introducing new methods of measurement requires the clinical context to recognize that it is important to correlate the findings not just with previous acoustic methods, but also with auditory-perceptual assessments directly.

Furthermore, the proprietary algorithm used was developed for the unique purposes of this study. Although it performed well in the sustained vowel analysis component of the work, the correlations were less convincing in analysis of continuous speech. With these discrepancies in measurement, there always will be changes to the algorithmic approaches moving forward. These changes will serve to improve the sensitivity and specificity of the measurements. Therefore, any changes to the algorithm may result in different results if this experiment was to be repeated. However, if measures were to be reanalyzed within-system (i.e., the algorithm) we would expect uniform results.

Additionally, the majority of prospectively collected samples represented non-dysphonic voices (n=42). In theory, these are easier to analyze than moderate or severely disordered voices. Of major concern here is the issue of how a period is extracted, which in turn overlays to other microacoustic measures. As such, this may to some extent represent a gap in our validation. This concern may then raise support for efforts moving forward which seek to improve the algorithm in future studies.

With respect to the study design, there were three recording subgroups for the assessment of the proprietary algorithm (Yeti AB, iPhone AB, iPhone NAB). The design was powered to allow for subgroup analyses of the individual groups, but these analyses were not completed. All participant samples (n=51) were combined and analyzed under both algorithms, before the results were compared with a paired-sample t-test. Though the correlation results were excellent when comparing the groups at this level, subgroup analyses may have revealed discrepancies that were otherwise not identified for the larger group analysis.

Finally, the comparison of the proprietary analysis in this study was completed with correlational statistics. To assess whether a new method of measurement is adequate for the purposes of replacing an old one, a limits of agreement analysis (Bland-Altman plot) should be performed (Bland & Altman, 1999). When measuring the same feature or factor in two different ways, there is bound to be a high level of correlation. Limits of agreement analysis is likely to reveal that two different methods do not perfectly agree, even in the setting of high correlation coefficients. For this reason, such analysis would appear to be important for determining whether the different measures can be used equally without causing errors in clinical judgement.

5.4 Clinical Implications

This study was designed to potentially reduce barriers to access for acoustic voice analysis. This was carried out in two phases, each phase structured to address specific barriers. First, we evaluated smartphone microphone utility for the purposes of acoustic voice analysis. This was done to systematically confirm the results of previous studies (Lin et al., 2012; Uloza et al., 2015). In the same phase, we also evaluated the effects of a non-soundproof environment on acoustic analysis measurements. With limited effect from each of these variables, the current results indicate that voice samples collected in quiet, non-soundproof, environments can acceptably be captured using smartphones for the purposes of performing acoustic analysis.

Statistically significant differences were identified on repeated measure ANOVAs, which looks for trends of differences between data sets and whether differences lead to a consistent effect. This statistical test does not indicate degree or amount of impact the variable being assessed had on the outcome, but rather, just that the effect was consistent and attributable to the variable rather than a result of random chance. With respect to clinical decision making, degree of impact is considerably more important than the presence of an impact alone. In the case of the voice measures included in the present study, there are no defined minimal clinically important differences (MCID). As a threshold for clinical significance, the normative ranges and pathologic limits were used as comparators. Accordingly, we looked at whether or not a mean difference identified represented a large or small amount of change compared to the normative range.

This brings up an important point, namely, that the normative values for any set of recording conditions needs to be set individually. This would include consideration of a different algorithm, microphone, or recording specifications, such as mouth-to-microphone distance, amongst others. Thus, these normative values are specific to a particular set of conditions and, therefore, any new measures would need to be established for any new procedure prior to further comparisons. This is particularly important when comparing results of one analysis environment to another as evidenced by work by Watts et al. (2017) where they showed that previously validated algorithms measuring cepstral peak prominence resulted in different absolute values. With this in mind, the normative values or ranges for each may not necessarily be equivalent when evaluating new conditions and algorithms, as was done in the present validation of the proprietary analysis algorithm.

Based on our results, the data suggests that voice samples can be collected from patients and research participants in a wider variety of settings. More specifically, this lowers the resource intensity requirements and can enable sample collection in remote locations where a soundproof setting is not available. And, of particular importance, these measures may provide reliable and valid indices of voice characteristics in both normal and dysphonic speakers. One setting where this can be directly utilized is in otolaryngology practices with a widely distributed patient population. In these settings, patients often are required to travel significant distances for (re)assessment, limiting the availability of resource intense testing practices. Remote collection of voice samples in a reliable manner under standardized collection protocols can limit the necessity of travel required for acoustic voice analysis. Additionally, patients may be able to collect a greater number of samples for assessment. These also can be collected in quiet spaces, which can include a home environment. This clearly opens the possibility of greater within subjects' analyses moving forward using the present smartphone application.

In the second phase of this project, a proprietary algorithm was compared to a well-established and widely used analysis procedure (Praat). This revealed strongly correlated assessments for sustained vowel analysis and moderate-to-strongly correlated results for continuous speech analysis. With this mobile acoustic voice analysis algorithm validation, this study indicates that both remote collection and analysis is feasible.

Introduction of a mobile acoustic voice analysis application will provide patients, practitioners, and researchers with a highly reliable, low-cost method of performing acoustic voice analysis. This increases the previously mentioned benefits, by further increasing the accessibility to a useful format of objectively assessing voice. Furthermore, within-subjects' assessments could be used as a comparator group to which other samples can be evaluated against while receiving treatment. This may in some manner provide the opportunity for direct biofeedback measures as a patient receives voice therapy. Furthermore, it could potentially indicate changes to the voice following surgical management of benign or malignant laryngeal tumors, whereby a change could prompt a direct reassessment by an otolaryngologist. During the development of a patient tool whereby voice samples would be collected and stored for analysis, significant consideration must be given to confidentiality of personal health information. Standards exist in the governing of the storage of personal health information (PHIPA, 2004) and these would need to be strictly adhered to for the safety of the end users.

5.5 Directions for Future Research

The current study was the first introduction of a new proprietary algorithm to perform acoustic voice analysis on a mobile platform. It performed well, but there is room for improvement and refinement. Most specifically, improvements in the ability of the application to analyze the complex signals of continuous speech is required. Following these improvements, validation of the algorithm for use in continuous speech will be necessary and ideally, experimental replication within a clinical environment will validate the process in the future.

Additionally, it is important to mention that this study evaluated voice samples gathered in non-soundproof environments for voice analysis. A future study should seek to systematically evaluate environmental factors to define the limits of ambient noise on reliable acoustic assessments. Efforts of this type hold considerable value in confirming that voice measure can be accurately obtained and analyzed in a reliable manner.

Next, one of the benefits of this format of acoustic voice analysis pertains to within subject analyses; a future study should be designed to investigate this further as a means of tracking voice over time. As well, this algorithm could be evaluated for specificity with respect to specific voice disordered populations. Ultimately, the ability of the smartphone system to gather data over time provides the capacity to generate a time series of voice measures that can be compared over both short- and longer-term periods. This would be beneficial for both those who may await treatment, as well as those who are receiving treatment or post-treatment.

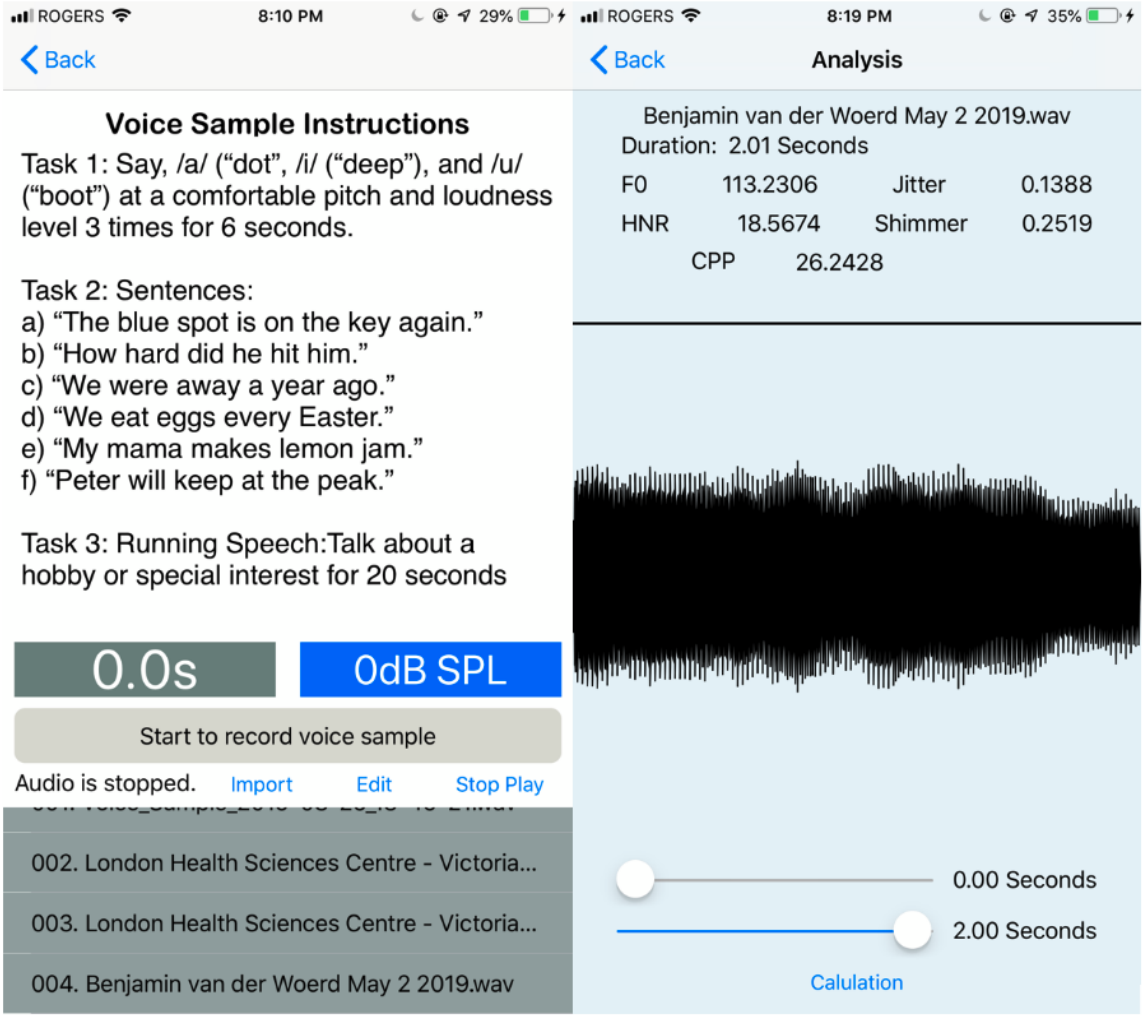
5.6 Conclusion

Smartphone based voice analysis has recently become a real possibility with improvements in mobile technology (Fujimura et al., 2019; Lin et al., 2012; Mat Baki et al., 2013, 2015; Uloza et al., 2015). The utility of these systems has been hypothesized by Manfredi et al (2017) but has not yet been fully achieved at present. No widespread adoption of a mobile platform has occurred, but with increased availability and validation of these tools, there is likely to be new benefits realized. Continued research and refinements based on new data will serve to optimize both data acquisition and the comprehensive analyses of samples gathered.

The results of this study confirm the validity of previous findings reported in the literature that smartphone voice recordings are adequate for acoustic voice analysis (Lin et al., 2012; Uloza et al., 2015). The present study found the impact of a smartphone microphone on acoustic voice measures to be limited in magnitude and unlikely to impact clinical interpretation. Similarly, it is well-recognized that a soundproof setting is ideal for controlling the ambient noise effects relative to the acoustic analysis of voice, however, this is not always feasible. Therefore, the results of this study indicate that a quiet, non-soundproof setting is adequate for voice sample collection. While care must always be taken in gathering samples in a consistent fashion, the ability to provide such data on a within-subject basis serves to mitigate this influence.

Furthermore, the prospectively captured samples were used as part of the validation for the proprietary analysis algorithm. The proprietary algorithm in its current state can reliably analyze sustained vowel samples on par with Praat. The analysis of continuous speech is more complex with variability in the range of frequency, amplitude, and temporal (e.g., rate of speech, pauses, voice breaks, etc.) domains. Reliable mobile analysis is possible, based on the present study, but the existing algorithm requires modifications before its reliability meets that of existing available options such as Praat.

In conclusion, the results of this study indicate that robust acoustic voice analysis is feasible on current day smartphones with samples collected under non-soundproof conditions. These results provide an opportunity to decrease existing obstacles to acoustic voice analysis as part of a routine clinical voice assessment. The objectives of using mobile technology to perform acoustic analysis aim to increase accessibility and remove barriers to use. The results of this study provide data, based on a systematic acquisition methods which supports future use of the current proprietary algorithm. In doing so, the current findings indicate that the present smartphone application, as seen in Figure 17, could be used reliably by patients, healthcare providers, and researchers for acoustic voice analysis.



Voice Sample Instructions

Task 1: Say, /a/ ("dot", /i/ ("deep"), and /u/ ("boot") at a comfortable pitch and loudness level 3 times for 6 seconds.

Task 2: Sentences:
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 jam."
f) "Peter will keep at the peak."

Task 3: Running Speech:Talk about a hobby or special interest for 20 seconds

Audio is stopped.

002. London Health Sciences Centre - Victoria...
003. London Health Sciences Centre - Victoria...
004. Benjamin van der Woerd May 2 2019.wav

Benjamin van der Woerd May 2 2019.wav
Duration: 2.01 Seconds

F0	113.2306	Jitter	0.1388
HNR	18.5674	Shimmer	0.2519
CPP	26.2428		

Figure 17. Screen shots showing recording and analysis screens of proprietary mobile application.

References

Ali, E. E., Chew, L., & Yap, K. Y.-L. (2016). Evolution and current status of mhealth research: a systematic review. *BMJ Innovations*, *2*(1), 33–40. https://doi.org/10.1136/bmjinnov-2015-000096

Anderson, S. H. (1925). Design and Calibration of a Phonodeik. *Journal of the Optical Society of America*, *11*(1), 31. https://doi.org/10.1364/JOSA.11.000031

ASHA. (2002). Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *American Speech-Language-Hearing Association Special Interest Division 3, Voice and Voice Disorders*.

Awan, S., & Roy, N. (2006). Toward the development of an objective index of dysphonia severity: A four-factor acoustic model. *Clinical Linguistics & Phonetics*, *20*(1), 35–49. https://doi.org/10.1080/02699200400008353

Awan, S., & Roy, N. (2009). Outcomes Measurement in Voice Disorders: Application of an Acoustic Index of Dysphonia Severity. *Journal of Speech, Language, and Hearing Research*, *52*(2), 482–499. https://doi.org/10.1044/1092-4388(2009/08-0034)

Awan, S., Roy, N., & Dromey, C. (2009). Estimating dysphonia severity in continuous speech: Application of a multi-parameter spectralcepstral model estimating dysphonia severity in continuous speech. *Clinical Linguistics and Phonetics*, *23*(11), 825–841. https://doi.org/10.3109/02699200903242988

Baken, R. J. (Ronald J. ., & Orlikoff, R. F. (2000). *Clinical measurement of speech and voice*. Retrieved from https://books.google.ca/books/about/Clinical_Measurement_of_Speech_and_Voice.html?id=ElPyvaJbDiwC&redir_esc=y

Behrman, A. (2005). Common Practices of Voice Therapists in the Evaluation of Patients. *Journal of Voice*, *19*(3), 454–469. https://doi.org/10.1016/J.JVOICE.2004.08.004

Belafsky, P. C., Postma, G. N., & Koufman, J. A. (2002). Validity and Reliability of the Reflux Symptom Index (RSI). *Journal of Voice*, *16*(2), 274–277. https://doi.org/10.1016/S0892-1997(02)00097-8

Bielamowicz, S., Kreiman, J., Gerratt, B. R., Dauer, M. S., & Berke, G. S. (1996). Comparison of voice analysis systems for perturbation measurement. *Journal of Speech and Hearing Research*, *39*(1), 126–134. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8820704

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*. Retrieved from http://dare.uva.nl/search?arno.record.id=109185

Boersma, P. (2009). Should jitter be measured by peak picking or by waveform matching? *Folia Phoniatrica et Logopaedica*, Vol. 61, pp. 305–308. https://doi.org/10.1159/000245159

Britto, A. I., & Doyle, P. C. (1990). A Comparison of Habitual and Derived Optimal Voice Fundamental Frequency Values in Normal Young Adult Speakers. *Journal of*

Speech and Hearing Disorders, *55*(3), 476–484. https://doi.org/10.1044/jshd.5503.476

Burdette, S. D., Herchline, T. E., & Oehler, R. (2008). Surfing The Web: Practicing Medicine in a Technological Age: Using Smartphones in Clinical Practice. *Clinical Infectious Diseases*, *47*(1), 117–122. https://doi.org/10.1086/588788

Canadian Cancer Society. (2017). Laryngeal cancer statistics. Retrieved August 23, 2019, from Canadian Cancer Statistics 2017 website: https://www.cancer.ca/en/cancer-information/cancer-type/laryngeal/statistics/?region=ab

Carterette, E. (1974). *Psychophysical Judgment and Measurement.* Retrieved from https://books.google.ca/books?hl=en&lr=&id=ANMzgCmA4agC&oi=fnd&pg=PA361&dq=+Handbook+of+Perception+E.+C.+Caterette+and+M.+P.+Friedman~Academic,+New+York!,+Vol.+2,+pp.+22–+40.&ots=Ip-srmdLxZ&sig=-oU2RaqLLZr7Fa59gnMA-OQ5dJo#v=onepage&q&f=false

Chu, E. A., & Kim, Y. J. (2008). Laryngeal Cancer: Diagnosis and Preoperative Work-up. *Otolaryngologic Clinics of North America*, *41*(4), 673–695. https://doi.org/10.1016/j.otc.2008.01.016

Cooper, M., & Yanagihara, N. (1971). A study of the basal pitch level variations found in the normal speaking voices of males and females. *Journal of Communication Disorders*, *3*(4), 261–266. https://doi.org/10.1016/0021-9924(71)90032-3

Dejonckere, P. H., Obbens, C., de Moor, G. M., & Wieneke, G. H. (1993). Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatrica*, *45*(2), 76–83. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8325573

Di Santo, P., Harnett, D. T., Simard, T., Ramirez, F. D., Pourdjabbar, A., Yousef, A., … Hibbert, B. (2018). Photoplethysmography using a smartphone application for assessment of ulnar artery patency: a randomized clinical trial. *Canadian Medical Association Journal*, *190*(13), E380–E388. https://doi.org/10.1503/cmaj.170432

Doherty, E. T., & Shipp, T. (1988). Tape Recorder Effects on Jitter and Shimmer Extraction. *Journal of Speech, Language, and Hearing Research*, *31*(3), 485–490. https://doi.org/10.1044/jshr.3103.485

Doyle, P. C. (1997). Voice Refinement Following Conservation Surgery for Cancer of the Larynx. *American Journal of Speech-Language Pathology*, *6*(3), 27–35. https://doi.org/10.1044/1058-0360.0603.27

Eadie, T. L., & Baylor, C. R. (2006). The Effect of Perceptual Training on Inexperienced Listeners' Judgments of Dysphonic Voice. *Journal of Voice*, *20*(4), 527–544. https://doi.org/10.1016/j.jvoice.2005.08.007

Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *The Journal of the Acoustical Society of America*, *112*(6), 3014–3021. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12509023

Eadie, T. L., & Doyle, P. C. (2005). Classification of Dysphonic Voice: Acoustic and Auditory-Perceptual Measures. *Journal of Voice*, *19*(1), 1–14. https://doi.org/10.1016/j.jvoice.2004.02.002

Eskenazi, L., Childers, D. G., & Hicks, D. M. (1990). Acoustic Correlates of Vocal Quality. *Journal of Speech, Language, and Hearing Research*, *33*(2), 298–306.

https://doi.org/10.1044/jshr.3302.298

Fairbanks, G. (1960). The rainbow passage. In *Voice and Articulation Drillbook*. Retrieved from https://scholar.google.ca/scholar?hl=en&as_sdt=0%2C5&q=fairbanks+1960+rainbow+passage&btnG=#d=gs_cit&u=%2Fscholar%3Fq%3Dinfo%3ATsXnIGHR1XIJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den

Ferrand, C. T. (2002). Harmonics-to-Noise Ratio: An Index of Vocal Aging. *Journal of Voice*, *16*(4), 480–487. https://doi.org/10.1016/S0892-1997(02)00123-6

Fujimura, S., Kojima, T., Okanoue, Y., Kagoshima, H., Taguchi, A., Shoji, K., … Hori, R. (2019). Real-Time Acoustic Voice Analysis Using a Handheld Device Running Android Operating System. *Journal of Voice*. https://doi.org/10.1016/J.JVOICE.2019.05.013

Gelfer, M. P. (1993). A Multidimensional Scaling Study of Voice Quality in Females. *Phonetica*, *50*(1), 15–27. https://doi.org/10.1159/000261923

Guidi, A., Salvi, S., Ottaviano, M., Gentili, C., Bertschy, G., de Rossi, D., … Vanello, N. (2015). Smartphone Application for the Analysis of Prosodic Features in Running Speech with a Focus on Bipolar Disorders: System Performance Evaluation and Case Study. *Sensors*, *15*(11), 28070–28087. https://doi.org/10.3390/s151128070

Heman-Ackah, Y. D., Michael, D. D., Baroody, M. M., Ostrowski, R., Hillenbrand, J., Heuer, R. J., … Sataloff, R. T. (2003). Cepstral Peak Prominence: A More Reliable Measure of Dysphonia. *Annals of Otology, Rhinology & Laryngology*, *112*(4), 324–333. https://doi.org/10.1177/000348940311200406

Heman-Ackah, Y. D., Michael, D. D., & Goding, G. S. (2002). The Relationship Between Cepstral Peak Prominence and Selected Parameters of Dysphonia. *Journal of Voice*, *16*(1), 20–27. https://doi.org/10.1016/S0892-1997(02)00067-X

Heman-Ackah, Y. D., Sataloff, R. T., Laureyns, G., Lurie, D., Michael, D. D., Heuer, R., … Hillenbrand, J. (2014). Quantifying the cepstral peak prominence, a measure of dysphonia. *Journal of Voice*, *28*(6), 783–788. https://doi.org/10.1016/j.jvoice.2014.05.005

Hillenbrand, J. (2011). Acoustic Analysis of Voice: A Tutorial. *Perspectives on Speech Science and Orofacial Disorders*, *21*(2), 31. https://doi.org/10.1044/ssod21.2.31

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research*, *37*(4), 769–778. https://doi.org/10.1044/jshr.3704.769

Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech, Language, and Hearing Research*, *39*(2), 311–321. https://doi.org/10.1044/jshr.3902.311

Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M., & Vaughan, C. (1989). Objective assessment of vocal hyperfunction: an experimental framework and initial results. *Journal of Speech and Hearing Research*, *32*(2), 373–392. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2739390

Hirano, M. (1981a). *Clinical examination of voice*. Retrieved from https://trove.nla.gov.au/work/25613022?q&versionId=30851490

Hirano, M. (1981b). *Clinical examination of voice*. Retrieved from https://books.google.ca/books?id=TAYKAQAAMAAJ&dq=Clinical+Examination+of+Voice+by+Minoru+Hirano&hl=en&sa=X&ved=0ahUKEwiNr86hnoPkAhVGK80KHUnND-QQ6AEIKjAA

Holmberg, E. B., Hillman, R. E., Hammarberg, B., Södersten, M., & Doyle, P. (2001). Efficacy of a Behaviorally Based Voice Therapy Protocol for Vocal Nodules. *Journal of Voice*, *15*(3), 395–412. https://doi.org/10.1016/S0892-1997(01)00041-8

Horii, Y. (1980). Vocal Shimmer in Sustained Phonation. *Journal of Speech, Language, and Hearing Research*, *23*(1), 202–209. https://doi.org/10.1044/jshr.2301.202

Isshiki, N., Yanagihara, N., & Morimoto, M. (1966). Approach to the Objective Diagnosis of Hoarseness. *Folia Phoniatrica et Logopaedica*, *18*(6), 393–400. https://doi.org/10.1159/000263069

Jacobson, B. H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., Benninger, M. S., & Newman, C. W. (1997). The Voice Handicap Index (VHI). *American Journal of Speech-Language Pathology*, *6*(3), 66–70. https://doi.org/10.1044/1058-0360.0603.66

Karnell, M. P. (1991). Laryngeal Perturbation Analysis. *Journal of Speech, Language, and Hearing Research, 34*(3), 544–548. https://doi.org/10.1044/jshr.3403.544

Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (2007). Reliability of Clinician-Based (GRBAS and CAPE-V) and Patient-Based (V-RQOL and IPVI) Documentation of Voice Disorders. *Journal of Voice*, *21*(5), 576–590. https://doi.org/10.1016/j.jvoice.2006.05.001

Kasper, C., Schuster, M., Psychogios, G., Zenk, J., Ströbele, A., Rosanowski, F., …

Haderlein, T. (2011). Voice handicap index and voice-related quality of life in small laryngeal carcinoma. *European Archives of Oto-Rhino-Laryngology*, *268*(3), 401–404. https://doi.org/10.1007/s00405-010-1374-0

Kempster, G. B., Kistler, D. J., & Hillenbrand, J. (1991). Multidimensional scaling analysis of dysphonia in two speaker groups. *Journal of Speech and Hearing Research*, *34*(3), 534–543. https://doi.org/10.1044/jshr.3403.534

Kent, R. D. (1996). Hearing and Believing. *American Journal of Speech-Language Pathology*, *5*(3), 7–23. https://doi.org/10.1044/1058-0360.0503.07

Kim, K. M., Kakita, Y., & Hirano, M. (1982). Sound spectrographic analysis of the voice of patients with recurrent laryngeal nerve paralysis. *Folia Phoniatrica*, *34*(3), 124–133. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7129283

Kisenwether, J. S., & Sataloff, R. T. (2015). The Effect of Microphone Type on Acoustical Measures of Synthesized Vowels. *Journal of Voice*, *29*(5), 548–551. https://doi.org/10.1016/J.JVOICE.2014.11.006

Kitajima, K., & Gould, W. J. (1976). Vocal Shimmer in Sustained Phonation of Normal and Pathologic Voice. *Annals of Otology, Rhinology & Laryngology*, *85*(3), 377–381. https://doi.org/10.1177/000348947608500308

Koike, Y. (1969). Vowel Amplitude Modulations in Patients with Laryngeal Diseases. *The Journal of the Acoustical Society of America*, *45*(4), 839–844. https://doi.org/10.1121/1.1911554

Kojima, H., Gould, W. J., & Lambiase, A. (1979). Computer analysis of hoarseness. *The Journal of the Acoustical Society of America*, *65*(S1), S67–S67.

https://doi.org/10.1121/1.2017385

Kojima, Hisayoshi, Gould, W. J., And, A. L., & Isshiki, N. (1980). Computer Analysis of Hoarseness. *Acta Oto-Laryngologica*, *89*(3–6), 547–554. https://doi.org/10.3109/00016488009127173

Kono, T., Saito, K., Yabe, H., Uno, K., Ogawa, K., T., K., … K., U. (2016). Comparative multidimensional assessment of laryngeal function and quality of life after radiotherapy and laser surgery for early glottic cancer. *Head and Neck*, *38*(7), 1085–1090. https://doi.org/http://dx.doi.org/10.1002/hed.24412

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual Evaluation of Voice Quality. *Journal of Speech, Language, and Hearing Research*, *36*(1), 21–40. https://doi.org/10.1044/jshr.3601.21

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual Differences in Voice Quality Perception. *Journal of Speech, Language, and Hearing Research*, *35*(3), 512–520. https://doi.org/10.1044/jshr.3503.512

Laguaite, J. K., & Waldrop, W. F. (1964). ACOUSTIC ANALYSIS OF FUNDAMENTAL FREQUENCY OF VOICES BEFORE AND AFTER THERAPY. *Folia Phoniatrica*, *16*, 183–192. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14153457

Lee, M., Ryu, C. H., Chang, H. W., Kim, G. C., Kim, S. Y. S. W. S. Y. S. W., Kim, S. Y. S. W. S. Y. S. W., … Kim, S. Y. S. W. S. Y. S. W. (2016). Radiotherapy-associated furin expression and tumor invasiveness in recurrent laryngeal cancer. *Anticancer Research*, *36*(10), 5117–5125.

https://doi.org/http://dx.doi.org/10.21873/anticanres.11081

Lieberman, P. (1963). Some Acoustic Measures of the Fundamental Periodicity of Normal and Pathologic Larynges. *The Journal of the Acoustical Society of America*, *35*(3), 344–353. https://doi.org/10.1121/1.1918465

Lin, E., Hornibrook, J., & Ormond, T. (2012). Evaluating iPhone Recordings for Acoustic Voice Assessment. *Folia Phoniatrica et Logopaedica*, *64*(3), 122–130. https://doi.org/10.1159/000335874

Manfredi, C., Lebacq, J., Cantarella, G., Schoentgen, J., Orlandi, S., Bandini, A., & DeJonckere, P. H. (2017). Smartphones Offer New Opportunities in Clinical Voice Research. *Journal of Voice*, *31*(1), 111.e1-111.e7. https://doi.org/10.1016/J.JVOICE.2015.12.020

Martins, R. H. G., do Amaral, H. A., Tavares, E. L. M., Martins, M. G., Gonçalves, T. M., & Dias, N. H. (2016). Voice Disorders: Etiology and Diagnosis. *Journal of Voice*, *30*(6), 761.e1-761.e9. https://doi.org/10.1016/J.JVOICE.2015.09.017

Maryn, Y., Corthals, P., De Bodt, M., Van Cauwenberge, P., & Deliyski, D. (2009). Perturbation Measures of Voice: A Comparative Study between Multi-Dimensional Voice Program and Praat. *Folia Phoniatrica et Logopaedica, 61*(4), 217–226. https://doi.org/10.1159/000227999

Mat Baki, M., Wood, G., Alston, M., Ratcliffe, P., Sandhu, G., Rubin, J. S., & Birchall, M. A. (2013). Comparison between OperaVOX^{TM} and MDVP: Preliminary Results. *Otolaryngology–Head and Neck Surgery*, *149*(2_suppl), P203–P204. https://doi.org/10.1177/0194599813496044a185

Mat Baki, M., Wood, G., Alston, M., Ratcliffe, P., Sandhu, G., Rubin, J. S., & Birchall, M. A. (2015). Reliability of OperaVOX against Multidimensional Voice Program (MDVP). *Clinical Otolaryngology*, *40*(1), 22–28. https://doi.org/10.1111/coa.12313

Metfessel, M. (1928). A PHOTOGRAPHIC METHOD OF MEASURING PITCH. *Science*, *68*(1766), 430–432. https://doi.org/10.1126/science.68.1766.430-a

Milenkovic, P. (1987). Least Mean Square Measures of Voice Perturbation. *Journal of Speech, Language, and Hearing Research*, *30*(4), 529–538. https://doi.org/10.1044/jshr.3004.529

Munnings, A. J. (2019). The Current State and Future Possibilities of Mobile Phone “Voice Analyser” Applications, in Relation to Otorhinolaryngology. *Journal of Voice*. https://doi.org/10.1016/J.JVOICE.2018.12.018

Murry, T. (1978). Speaking Fundamental Frequency Characteristics Associated with Voice Pathologies. *Journal of Speech and Hearing Disorders*, *43*(3), 374–379. https://doi.org/10.1044/jshd.4303.374

Murry, T., Singh, S., & Sargent, M. (1977). Multidimensional classification of abnormal voice qualities. *The Journal of the Acoustical Society of America*, *61*(6), 1630–1635. https://doi.org/10.1121/1.381439

Noll, A. M. (1964). Short-Time Spectrum and “Cepstrum” Techniques for Vocal-Pitch Detection. *The Journal of the Acoustical Society of America*, *36*(2), 296–302. https://doi.org/10.1121/1.1918949

Noll, A. M. (1967). Cepstrum Pitch Determination. *The Journal of the Acoustical Society of America*, *41*(2), 293–309. https://doi.org/10.1121/1.1910339

Oates, J. (2009). Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions. *Folia Phoniatrica et Logopaedica*, *61*(1), 49–56. https://doi.org/10.1159/000200768

Oates, J., & Russell, A. (1998). Learning voice analysis using an interactive multi-media package: Development and preliminary evaluation. *Journal of Voice*, *12*(4), 500–512. https://doi.org/10.1016/S0892-1997(98)80059-3

Orlikoff, R. F., & Kraus, D. H. (1996). Dysphonia Following Nonsurgical Management of Advanced Laryngeal Carcinoma. *American Journal of Speech-Language Pathology*, *5*(3), 47–52. https://doi.org/10.1044/1058-0360.0503.47

Park, Y., & Chen, J. V. (2007). Acceptance and adoption of the innovative use of smartphone. *Industrial Management & Data Systems*, *107*(9), 1349–1365. https://doi.org/10.1108/02635570710834009

Parsa, V., & Jamieson, D. G. (1999). A Comparison of High Precision FO Extraction Algorithms for Sustained Vowels. *Journal of Speech, Language, and Hearing Research*, *42*(1), 112–126. https://doi.org/10.1044/jslhr.4201.112

Parsa, V., & Jamieson, D. G. (2001). Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. *Journal of Speech, Language, and Hearing Research : JSLHR*, *44*(2), 327–339. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11324655

Parsa, V., & Jamieson, D. G. (2014). A Comparison of High Precision FO Extraction Algorithms for Sustained Vowels. *Journal of Speech, Language, and Hearing Research*, *42*(1), 112–126. https://doi.org/10.1044/jslhr.4201.112

Parsa, V., Jamieson, D., & Pretty, B. (2001). Effects of microphone type on acoustic measures of voice. *Journal of Voice*, *15*(3), 331–343. https://doi.org/10.1016/S0892-1997(01)00035-2

PHIPA. (2004). Personal Health Information Protection Act, 2004, S.O. 2004, c. 3, Sched. A. Retrieved December 16, 2019, from https://www.ontario.ca/laws/statute/04p03

Pippard, A. (2007). *The physics of vibration*. Retrieved from https://books.google.ca/books?hl=en&lr=&id=F8-9UNvsCBoC&oi=fnd&pg=PP1&dq=A.++B.++Pippard,The++Physics++of++Vibration&ots=KnwkljnYbu&sig=g6J3cLjs435uQV08LDQWRzsMf_4

Praat Documentation. (2003). Voice 3. Shimmer. Retrieved August 28, 2019, from Praat Documentation website: http://www.fon.hum.uva.nl/praat/manual/Voice_3__Shimmer.html

Rabinov, C. R., Kreiman, J., Gerratt, B. R., & Bielamowicz, S. (1995). Comparing reliability of perceptual ratings of roughness and acoustic measure of jitter. *Journal of Speech and Hearing Research*, *38*(1), 26–32. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7731216

Roy, N., Merrill, R. M., Thibeault, S., Parsa, R. A., Gray, S. D., & Smith, E. M. (2004). Prevalence of Voice Disorders in Teachers and the General Population. *Journal of Speech, Language, and Hearing Research*, *47*(2), 281–293. https://doi.org/10.1044/1092-4388(2004/023)

Schiavetti, N., Metz, D. E., & Sitler, R. W. (1981). Construct Validity of Direct

Magnitude Estimation and Interval Scaling of Speech Intelligibility. *Journal of Speech, Language, and Hearing Research*, *24*(3), 441–445. https://doi.org/10.1044/jshr.2403.441

Schwartz, S., Rosenfeld, R., Dailey, S., & Cohen, S. (2009). Evidence-based Management of Hoarseness (Dysphonia). *Otolaryngology–Head and Neck Surgery*, *141*(2_suppl), P16–P17. https://doi.org/10.1016/j.otohns.2009.06.035

Shrivastav, R., Sapienza, C. M., & Nandur, V. (2005). Application of Psychometric Theory to the Measurement of Voice Quality Using Rating Scales. *Journal of Speech, Language, and Hearing Research*, *48*(2), 323–335. https://doi.org/10.1044/1092-4388(2005/022)

Siau, R. T. K., Goswamy, J., Jones, S., & Khwaja, S. (2017). Is OperaVOX a clinically useful tool for the assessment of voice in a general ENT clinic? *BMC Ear, Nose and Throat Disorders*, *17*(1), 4. https://doi.org/10.1186/s12901-017-0037-9

Smith, A. (2012). 46 % of American adults are smartphone owners phones within the national adult population. *Changes*, 1–9. Retrieved from https://www.pewinternet.org/wp-content/uploads/sites/9/media/Files/Reports/2012/Smartphone-ownership-2012.pdf

Smith, M., & Ramig, L. (1994). Neurological disorders and the voice. *National Center for Voice and Speech Status and Progress Report*, *7*(December), 207–227. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.738.3828&rep=rep1&type=pdf#page=214

Starmer, H. M., Tippett, D. C., & Webster, K. T. (2008). Effects of Laryngeal Cancer on Voice and Swallowing. *Otolaryngologic Clinics of North America*, *41*(4), 793–818. https://doi.org/10.1016/J.OTC.2008.01.018

Steinhubl, S. R., Muse, E. D., & Topol, E. J. (2015). The emerging field of mobile health. *Science Translational Medicine*, *7*(283), 283rv3. https://doi.org/10.1126/scitranslmed.aaa3487

Stepp, C. E., Merchant, G. R., Heaton, J. T., & Hillman, R. E. (2011). Effects of Voice Therapy on Relative Fundamental Frequency During Voicing Offset and Onset in Patients With Vocal Hyperfunction. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1260–1266. https://doi.org/10.1044/1092-4388(2011/10-0274)

Stevens, S. S., & Marks, L. E. (2017). *Psychophysics*. https://doi.org/10.4324/9781315127675

Stevens, S. S., & Volkmann, J. (1940). The Relation of Pitch to Frequency: A Revised Scale. *The American Journal of Psychology*, *53*(3), 329. https://doi.org/10.2307/1417526

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190. https://doi.org/10.1121/1.1915893

Titze, I. (1988). Regulation of vocal power and effi- ciency by subglottal pressure and glottal width. En: Vocal physiology: voice production, mechanisms and functions. In *Vocal Fold Physiology*. Retrieved from https://ci.nii.ac.jp/naid/10007697963/#cit

Titze, I., Lemke, J., & Montequin, D. (1997). Populations in the U.S. workforce who rely on voice as a primary tool of trade: a preliminary report. *Journal of Voice*, *11*(3), 254–259. https://doi.org/10.1016/S0892-1997(97)80002-1

Titze, I., & Liang, H. (1993). Comparison of F o Extraction Methods for High-Precision Voice Perturbation Measurements. *Journal of Speech, Language, and Hearing Research*, *36*(6), 1120–1133. https://doi.org/10.1044/jshr.3606.1120

Titze, I., & Winholtz, W. (1993). Effect of Microphone Type and Placement on Voice Perturbation Measurements. *Journal of Speech, Language, and Hearing Research*, *36*(6), 1177–1190. https://doi.org/10.1044/jshr.3606.1177

Toner, M. A., & Emanuel, F. W. (1989). Direct Magnitude Estimation and Equal Appearing Interval Scaling of Vowel Roughness. *Journal of Speech, Language, and Hearing Research*, *32*(1), 78–82. https://doi.org/10.1044/jshr.3201.78

Uloza, V., Padervinskis, E., Vegiene, A., Pribuisiene, R., Saferis, V., Vaiciukynas, E., … Verikas, A. (2015). Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening. *European Archives of Oto-Rhino-Laryngology*, *272*(11), 3391–3399. https://doi.org/10.1007/s00405-015-3708-4

Umapathy, K., Krishnan, S., Parsa, V., & Jamieson, D. G. (2005). Discrimination of Pathological Voices Using a Time-Frequency Approach. *IEEE Transactions on Biomedical Engineering*, *52*(3), 421–430. https://doi.org/10.1109/TBME.2004.842962

Van Houtte, E., Van Lierde, K., & Claeys, S. (2011). Pathophysiology and Treatment of Muscle Tension Dysphonia: A Review of the Current Knowledge. *Journal of Voice*, *25*(2), 202–207. https://doi.org/10.1016/J.JVOICE.2009.10.009

Verdolini, K. (2014). Classification Manual for Voice Disorders-I. In *Classification Manual for Voice Disorders-I*. https://doi.org/10.4324/9781410617293

Watts, C. R., Awan, S. N., & Maryn, Y. (2017). A Comparison of Cepstral Peak Prominence Measures From Two Acoustic Analysis Programs. *Journal of Voice*, *31*(3), 387.e1-387.e10. https://doi.org/10.1016/j.jvoice.2016.09.012

Webb, A. L., Carding, P. N., Deary, I. J., MacKenzie, K., Steen, N., & Wilson, J. A. (2004). The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Oto-Rhino-Laryngology*, *261*(8), 429–434. https://doi.org/10.1007/s00405-003-0707-7

Wendahl, R. W. (1966). Laryngeal Analog Synthesis of Jitter and Shimmer Auditory Parameters of Harshness. *Folia Phoniatrica et Logopaedica*, *18*(2), 98–108. https://doi.org/10.1159/000263059

Wendahl, R. W. (2009). Laryngeal Analog Synthesis of Jitter and Shimmer Auditory Parameters of Harshness. *Folia Phoniatrica et Logopaedica*, *18*(2), 98–108. https://doi.org/10.1159/000263059

Werth, K., Voigt, D., Döllinger, M., Eysholdt, U., & Lohscheller, J. (2010). Clinical value of acoustic voice measures: a retrospective study. *European Archives of Oto-Rhino-Laryngology*, *267*(8), 1261–1271. https://doi.org/10.1007/s00405-010-1214-2

Wong, M. C., & Fung, K. (2015). Mobile Applications in Otolaryngology–Head and Neck Surgery. *Otolaryngology-Head and Neck Surgery*, *152*(4), 638–643. https://doi.org/10.1177/0194599815568946

Wuyts, F. L., De Bodt, M. S., & Van de Heyning, P. H. (1999). Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *Journal of Voice*, *13*(4), 508–517. https://doi.org/10.1016/S0892-1997(99)80006-X

Yanagihara, N. (1967). Significance of Harmonic Changes and Noise Components in Hoarseness. *Journal of Speech and Hearing Research*, *10*(3), 531–541. https://doi.org/10.1044/jshr.1003.531

Yang, F., Wang, Y., Yang, C., Hu, H., & Xiong, Z. (2018). Mobile health applications in self-management of patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis of their efficacy. *BMC Pulmonary Medicine*, *18*(1), 147. https://doi.org/10.1186/s12890-018-0671-z

Ylitalo, R., & Hammarberg, B. (2000). Voice characteristics, effects of voice therapy, and long-term follow-up of contact granuloma patients. *Journal of Voice*, *14*(4), 557–566. https://doi.org/10.1016/S0892-1997(00)80011-9

Yumoto, E., Gould, W. J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, *71*(6), 1544–1550. https://doi.org/10.1121/1.387808

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., … Glaze, L. E. (2011). Establishing Validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology*, *20*(1), 14–22. https://doi.org/10.1044/1058-0360(2010/09-0105)

Appendices

Power Calculation

For Comparison of Means

A total sample size (N) of 34 individuals, 17 at each level, is sufficient to detect the hypothesized effect (r^2=.25) of a two-level between-groups independent variable 90.4 percent of the time using a .05 alpha level.

Technical Summary:

N=34, f^2=0.3333, lambda=11.3333

power=.9039, critical F(1,32)=4.1491

For Correlations

A total sample size (N) of 34 individuals is sufficient to detect the hypothesized effect (r^2=.25) of a single predictor variable 90.4 percent of the time using a .05 alpha level.

Technical Summary:

N=34, f^2=0.3333, lambda=11.3333

power=.9039, critical F(1,32)=4.1491

Appendix A. Power Calculation

Information Sheet and Consent Form

Project title: VOice analysis with Iphones: a low-Cost Experimental Solution (VOICES)

Name of Principal Investigator (Study Doctor)

Dr. Kevin Fung, MD FRCSC

[redacted]

Co-Investigators

Dr. Benjamin van der Woerd

Dr. Phil Doyle

Dr. Vijay Parsa

Min Wu

This application is being developed and study by the above co-investigators. We acknowledge the potential of bias. This is an internal validation study and all data will be blinded to the analyzer (P.Doyle). This application will not be commercialized, as part of the agreement between the five co-investigators.

Statement of Voluntary Participation

1. Your participation in this study is voluntary.

   a. You may decide not to be in this study
   b. If you initially decide to be in this study, may change your might at a later date and withdraw
   c. You may leave the study at any time without affecting the care that you will receive.
   d. We will give you new information that is learned during the study that might affect your decision to stay in the study

2. You may refuse to answer any question you do not wish to answer

Conflict of Interest

There are no conflicts of interest to declare related to this study.

Introduction

You are being invited to participate in a research study that aims to validate a smartphone-based application to perform acoustic voice analysis. Specifically, this is looking at the ability of a smartphone-based application to perform acoustic voice analysis and whether this is comparable to currently used systems.

Background/Purpose

Acoustic analysis of voice is a valuable evaluation method used in the diagnosis and surveillance of both benign and malignant voice abnormalities. Typically, acoustic evaluation requires a resource intensive set-up, including a soundproof booth and a speech language pathologist, to be performed accurately. Given current advances in technology, we feel that a low-resource tool can be developed for smartphones to replicate currently available voice analysis options. Once validated, this tool may potentially make voice analysis more accessible to patients, clinicians, and researchers.

Study Design

Voice samples will be collected using smartphone application and run through a smartphone voice analysis application as well as the current computer-based voice analysis software. We aim to collect 60 voice samples. The primary domains of analysis will be focused on how voice sounds by analysing different components of the voice signal that can distinguish normal from abnormal voices. Voice samples (current gold standard) will be obtained using current assessment standards and will be run through both voice analysis application (smartphone app and computer-based voice analysis software). Participants will provide one of three types of samples for the study: in a sound proof booth with a standard microphone, in a sound proof booth with an iPhone microphone, in a non-sound proof room with an iPhone microphone. The results of each of these samples will be compared for rate of agreement to assess the new method of measurement in ability to accurately capture measures of normal and abnormal voice.

Secondly, samples from the smartphone collected data and professionally collected voice samples will be evaluated by blinded speech language pathologists for perceptual voice analysis. Auditory-perceptual methods are well-established in the experimental and clinical literature related to voice disorders. All data will by blinded to Dr. P. Doyle, who will be performing the analysis.

Procedures

The participant will be seated in a comfortable position. A standardized mouth-to-microphone distance will be maintained at 6 inches, with the microphone positioned directly at a 45-degree angle to the corner of the mouth. Digital audio samples will be recorded during three specific vocal tasks: sustained vowels, sentences, and running speech.

Task 1: Sustained Vowels: Three vowels, /a/ (as in “dot”, /i/ (as in “deep”), and /u/ (as in “boot”) will be produced at a typical and comfortable pitch and loudness level. The vowel sounds are to be produced 3 times each for approximately 6 seconds.

Task 2: Sentences: The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) developed 6 sentences to test various aspects of laryngeal behaviour.

a) “The blue spot is on the key again.” –includes all vowel sounds in the English language
b) “How hard did he hit him.” –easy onset with /h/ sound
c) “We were away a year ago.” –all voiced sounds
d) “We eat eggs every Easter.” –hard glottal attack
e) “My mama makes lemon jam.” –nasal sounds
f) “Peter will keep at the peak.” –voiceless plosive sounds

Sentences are to be read one at a time as if being spoken in conversation at one's typical pitch and loudness level. As noted, the sentences each emphasize different voice characteristics.

Task 3: Running Speech: The participant provides 20 seconds of natural speech by responding to prompting statements (e.g., "Talk about a hobby or special interest" or "Talk about what you do/did for work").

Withdrawal from Study

A participant may withdraw from the study at any time. If withdrawn from the study, a participant can request that their data be withdrawn as well. You have the legal right to inform you are leaving the study without explaining your reasons. Signing the consent form does not remove your legal rights, the consent form signed means you accept to participate in the study as a volunteer. Foreseeable reasons under which the participant may be removed from the study is meeting one or more of the exclusion criteria. This would not affect your future treatment.

Risks

There are no foreseeable risks to being involved in this study.

Benefits

No direct benefit to the participant is anticipated. Demonstrating smartphone-based acoustic voice analysis in both sustained vowels and running speech may provide future benefit to patients in time, travel, and convenience.

Furthermore, this information will be valuable for otolaryngologists when managing future voice patients. Acknowledging the current limitations in obtaining acoustic voice analysis samples may have far-reaching effects on screening, diagnostic support, evaluation of treatment effects, and assessing treatment progress in future voice patients.

Alternatives to Being in the Study

An alternative to this study is to not participate and continue on as you do currently.

Confidentiality

Research data will be stored on a password protected LHSC computer. Personal identifiers (Initials, date of birth and hospital number) will be removed prior to data analysis. The principal investigator and co-investigators will have access to the data for data analysis. In addition to principal investigators, Lawson Research Institute and Western's Health Sciences Research Ethics Board will have access to the information collected. All information collected during this study, including personal health information, will be kept confidential and will not be shared with anyone outside the study unless required by law. You will not be named in any reports, publications or presentations that may come from this study. Instead, a non-related ID number will be given to you and will be used on your file. As per Lawson Research Institute requirements, study data will be kept for 15 years following study cessation. Application data will be stored on the Lawson Research Institute REDcap Server and LHSC S-Drive. Temporary storage on the smartphone will be required prior to export to the S-Drive.

- Representatives of Western University's Health Sciences Research Ethics Board may contact you or require access to your study-related records to monitor the conduct of the research.

- Representatives of Lawson Quality Assurance (QA) Education Program may look at study data for QA purposes

Decision to leave the study

Your participation is voluntary; you are free to withdraw your participation from this study at any time. You also have the right to have your data fully removed from our database.

How the findings will be used:

The results of the study will be used for scholarly purposes only. The results from the study will be presented in educational settings and at professional conferences, and the results might be published in a professional journal in the field of surgery. Participant information will remain de-identified.

Costs

There are no costs associated with this study that you are responsible for.

Compensation

You will not be given compensation for your participation in this study.

Contact information:

If you have concerns or questions about this study Dr. Benjamin van der Woerd at [redacted] or the Primary Investigator (PI) Dr. Kevin Fung at [redacted].

Questions about the Study

If you have any questions or concerns regarding your participation in this study, please do not hesitate to contact:

If you have any questions about your rights as a research participant or the conduct of this study, you may contact The Office of Research Ethics [redacted], email: [redacted].

Principle investigator:

Dr. Kevin Fung, MD FRCSC

LHSC, Victoria Campus

Appendix B. Patient Study Information Package

Study Consent Form

Title of Research Project:

VOice analysis with Iphones: a low-Cost Experimental Solution (VOICES)

Consent:

By signing this form, I agree that:

1) You have explained this study to me. You have answered all my questions.

2) You have explained the possible harms and benefits (if any) of this study.

3) I know what I could do instead participating in this study. I also have the right to withdraw my participation from the study at any time. My decision about my participation will not affect my health care at London Health Science Centre.

4) I am free now, and in the future, to ask questions about the study.

5) I have been told that my information and medical records will be kept private except as described to me.

6) I understand that no information about my participation will be given to anyone or be published without first asking my permission.

7) I have read and understood pages 1 to 5 of this Information letter/consent form. I agree, or consent, to take part in this study.

____

____

Printed Name: Participant Participant’s signature & date

____

____

Printed Name: Person Obtaining Consent Signature & date, time

Appendix C. Consent Form

Curriculum Vitae

Name:	Benjamin van der Woerd
Post-secondary Education and Degrees:	University of Ottawa Ottawa, Ontario, Canada 2008-2012 B.Sc.
	McMaster University Hamilton, Ontario, Canada 2013-2016 M.D.
	Western University London, Ontario, Canada 2018-2019 MSc.
Honours and Awards:	Peter Cheski Innovative Research Award 2018
Related Work Experience	Otolaryngology – Head and Neck Surgery Resident Western University 2016-present

Publications:

Journal of Otolaryngology – Head & Neck Surgery, Western University

November 2018

van der Woerd, B. D., Patel, K.B., Nichols, A.C., Fung, K., Yoo, J., & MacNeil, S. D. (2018). Functional Outcomes in Early (T1/T2) Supraglottic Cancer: A Systematic Review. *Journal of Otolaryngology – Head & Neck Surgery*.

Medicine, Western University

October 2017

van der Woerd, B. D., & MacNeil, S. D. (2017). Sialocutaneous fistula to the external auditory canal repaired with superficial parotidectomy and temporoparietal flap: A case report. *Medicine, 96*(42).

International Journal of Surgery Case Reports, McMaster University

October 2015

van der Woerd, B., Robichaud, J., & Gupta, M. (2015). Parapharyngeal abscess following use of a laryngeal mask airway during open revision septorhinoplasty. *International Journal of Surgery Case Reports, 16*(198-201).