2020

Evaluating the translational potential of relative fundamental frequency

BOSTON UNIVERSITY

SARGENT COLLEGE OF HEALTH AND REHABILITATION SCIENCES

Dissertation

EVALUATING THE TRANSLATIONAL POTENTIAL OF RELATIVE FUNDAMENTAL FREQUENCY

by

YEONGGWANG PARK

B.A., Gordon College, 2015

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2020

Approved by

First Reader ____
Cara E. Stepp, Ph.D.
Associate Professor of Speech, Language, & Hearing Sciences
Associate Professor of Biomedical Engineering
Associate Professor of Otolaryngology – Head and Neck Surgery

Second Reader ____
Joseph S. Perkell, Ph.D.
Senior Research Scientist in Speech, Language, & Hearing Sciences

Third Reader ____
Daryush D. Mehta, Ph.D.
Assistant Professor of Surgery
Harvard Medical School

Fourth Reader ____
Kathleen F. Nagle, Ph.D.
Assistant Professor of Speech-Language Pathology
Seton Hall University

Acknowledgments

I would first like to thank my advisor, Dr. Cara Stepp. Her research on voice and relative fundamental frequency attracted me to start a doctoral study with her. I am very grateful for receiving training from her and working with her on the research topic that I enjoy very much. She has been a phenomenal advisor for me. Her pursuit of excellent work and professionalism has been motivated me to complete my doctoral work and become a professional researcher like her. I want to thank my other committee members, Drs. Joe Perkell, Daryush Mehta, and Kathy Nagle. They have given me huge support and their advice, comments, and suggestions have been valuable. I would like to thank our Stepp lab members, also known as Stepp-sisters and brothers, for their amazing help and support. The studies in this dissertation could not have been successfully completed without them. I also want to thank Vocal Hyperfunction Clinical Research Center team for their insights that aided this work.

I would also like to thank my family, my parents, grandparents, sister, brother-in-law, and newborn nephew in Korea for their love, prayer, and devotion for me to continue my study in the United States for 13 years now. I want to give special thanks to my girlfriend, Yeonjee, and her parents because their love and support has allowed me to continue my doctoral work. I express my sincere gratitude to my pastor Seo-Kwon Kim and his wife, Yusoon Kim, in Seoul, for their love, prayer, and message that have inspired and strengthened me. I also thank Jesus' Love Church family members in both Boston and Seoul for their continual love and prayer. Finally, I want to end this section by acknowledging my source of strength and wisdom, Jesus Christ.

EVALUATING THE TRANSLATIONAL POTENTIAL OF

RELATIVE FUNDAMENTAL FREQUENCY

YEONGGWANG PARK

Boston University Sargent College of Health and Rehabilitation Sciences, 2020

Major Professor: Cara E. Stepp, Ph.D., Associate Professor of Speech, Language, & Hearing Sciences, Sargent College of Health and Rehabilitation Sciences; Associate Professor of Biomedical Engineering, College of Engineering; Associate Professor of Otolaryngology – Head and Neck Surgery, School of Medicine

ABSTRACT

Relative fundamental frequency (RFF) is an acoustic measure that quantifies short-term changes in fundamental frequency during voicing transitions surrounding a voiceless consonant. RFF is hypothesized to be decreased by increased laryngeal tension during voice production and has been considered a potential objective measure of vocal hyperfunction. Previous studies have supported claims that decreased RFF values may indicate the severity of vocal hyperfunction and have attempted to improve the methods to obtain RFF. In order to make progress towards developing RFF into a clinical measure, this dissertation aimed to investigate further the validity and reliability of RFF. Specifically, we examined the underlying physiological mechanisms, the auditory-perceptual relationship with strained voice quality, and test-retest reliability.

The first study evaluated one of the previously hypothesized physiological mechanisms for RFF, vocal fold abduction. Vocal fold kinematics and RFF were obtained from both younger and older typical speakers producing RFF stimuli with voiceless fricatives and stops during high-speed videoendoscopy. We did not find any statistical

differences between younger and older speakers, but we found that vocal folds were less adducted and RFF was lower at voicing onset after the voiceless stop compared to the fricative. This finding is in accordance with the hypothesized positive association between vocal fold contact area during voicing transitions and RFF.

The second study examined the relationship between RFF and strain, a major auditory-perceptual feature of vocal hyperfunction. RFF values were synthetically modified by exchanging the RFF contours between voice samples that were produced with a comfortable voice and with maximum vocal effort, while other acoustic features remained constant. We observed that comfortable voice samples with the RFF values of maximum vocal effort samples had increased strain ratings, whereas maximum vocal effort samples with the RFF values of comfortable voice samples had decreased strain ratings. These findings support the contribution of RFF to perceived strain.

The third study compared the test-retest reliability of RFF with that of conventional voice measures. We recorded individuals with healthy voices during five consecutive days and obtained acoustic, aerodynamic, and auditory-perceptual measures from the recordings. RFF was comparably reliable as acoustic and aerodynamic measures and more reliable than auditory-perceptual measures.

This dissertation supports the translational potential of RFF by providing empirical evidence of the physiological mechanisms of RFF, the relationship between RFF and perceived strain, and test-retest reliability of RFF. Clinical applications of RFF are expected to improve objective diagnosis and assessment of vocal hyperfunction, and thus to lead to better voice care for individuals with vocal hyperfunction.

Preface

This dissertation contains three studies that aimed to support the clinical application of relative fundamental frequency (RFF). Chapter 1 provides the background information on vocal hyperfunction, current state of subjective and objective assessment, and previous research on relative fundamental frequency and the motivation for the three studies. Chapters 2 – 4 are three self-contained manuscripts written in preparation for publication. Due to this nature of the three chapters, there is some overlap background information between the chapters. Chapter 5 summarizes the important findings and their implications, discusses possible clinical applications of RFF, and suggests future research to further develop RFF as a clinical measure.

Following list provides the authors and the titles of the manuscripts:

Chapter 2: Park, Y., Wang, F., Díaz Cádiz, M. E., Vojtech, J. M., Groll, M. D., & Stepp, C. E. "Vocal fold kinematics and relative fundamental frequency as a function of obstruent type and speaker age," submitted to the Journal of the Acoustic Society of America.

Chapter 3: Park, Y., Díaz Cádiz, M. E., Nagle, K. F., Stepp, C. E. "Perceptual and acoustic assessment of strain using synthetically modified voice samples," submitted to Journal of Speech, Language, and Hearing Research.

Chapter 4: Park, Y., Stepp, C. E. "Test-retest reliability of relative fundamental frequency and conventional acoustic, aerodynamic, and perceptual measures in individuals with healthy voices," Journal of Speech, Language, and Hearing Research, 62(6), pp. 1707-1718, 2019.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

α .... Significance Level

AC .... Alternating Current

ANOVA .... Analysis of Variance

AVQI .... Acoustic Voice Quality Index

CAPE-V .... Consensus Auditory-Perceptual Evaluation of Voice

CI .... Confidence Interval

CPPS .... Smoothed Cepstral Peak Prominence

CSID .... Cepstral Spectral Index of Dysphonia

CT .... Cricothyroid

d' .... Sensitivity Index

dB .... Decibel

d_{off} .... Duration of Voicing Offset

d_{on} .... Duration of Voicing Onset

EMG .... Electromyography

f_o .... Fundamental Frequency

f_{ref} .... f_o of the Reference Cycle

GRBAS .... Grade, Roughness, Breathiness, Asthenia, and Strain

HNR .... Harmonics-to-Noise Ratio

HSV .... High-Speed Videoendoscopy

ICC .... Intraclass Correlation Coefficient

IPSV .... Inability to Produce Soft Voice

JND .... Just Noticeable Difference

KS .... Kinematic Estimate of Laryngeal Stiffness

LCA .... Lateral Cricoarytenoid

LPC .... Linear Predictive Coding

M .... Mean

MTD .... Muscle Tension Dysphonia

NHR .... Noise-to-Harmonics Ratio

η_p^2 .... Partial Eta Squared

Offset 10 RFF .... RFF at Offset Cycle 10

Offset RFF .... RFF at Voicing Offset

Onset 1 RFF .... RFF at Onset Cycle 1

Onset RFF .... RFF at Voicing Onset

p .... Probability Value

p_{adj} .... Adjusted Probability Value

PCA .... Posterior Cricoarytenoid

Psub .... Subglottic Pressure Estimate

PTP .... Phonation Threshold Pressure

r .... Correlation Coefficient

RFF .... Relative Fundamental Frequency

RSI .... Reflux Symptom Index

SD .... Standard Deviation

SPL .... Sound Pressure Level

ST .... Semitone

STRAIGHT .... Speech Transformation and Representation using Adaptive Interpolation of Weighted spectrum

TA .... Thyroarytenoid

t .... t-Statistic

t_{abd} .... The Start of Abduction

t_{add} .... The End of Adduction

t_{off} .... The End of Voicing Offset

t_{on} .... The Start of Voicing Onset

Θ_{max} .... The Maximum Glottal Angle

Θ_{off} .... Glottal Angle at Vocing Offset

Θ_{on} .... Glottal Angle at Vocing Onset

VCV .... Voiced Sonorant-Voiceless Consonant-Voiced Sonorant

VHI .... Voice Handicap Index

VSR .... Visual Sort-and-Rate

z .... Inverse of Normal Distribution

Chapter 1: Background

Introduction

Voice performs essential roles in human life. Voice reflects our personality and enables us to express our emotions, attitudes, and communicate our message to others (Gobl & Ní Chasaide, 2003; Scherer, 1978). Due to these essential functions of voice, harm to the voice of an individual, such as in voice disorders, can be detrimental to an individual's quality of life.

Voices disorders have been defined as "the abnormal production and/or absence of vocal quality, pitch, loudness, and/or duration which is appropriate for an individual's age and/or sex" (American Speech-Language-Hearing Association, 1993). The presentation of voice disorders can range from mild voice symptoms to complete voice loss (Ramig & Verdolini, 1998). The probability of acquiring a voice disorder was reported as 30%, and approximately 3–9% of the population of the United States has been diagnosed with a voice disorder or voice problem within the last year (Bhattacharyya, 2014; Morris, Meier, Griffin, Branda, & Phelan, 2016; Ramig & Verdolini, 1998).

Voice impairment due to voice disorders can inhibit many daily activities, which negatively affects the quality of life of individuals psychologically, economically, and socially (Bouwers & Dikkers, 2009; Chen, Chiang, Chung, Hsiao, & Hsiao, 2010; Roy, Merrill, Gray, & Smith, 2005; E. Smith, Verdolini, Gray, Nichols, Lemke, Barkmeier, Dove, & Hoffman, 1994). Individuals who are required to use a loud voice extensively for their careers (e.g., teachers) are at a higher risk of developing and suffering from the negative effects of voice disorders (Chen et al., 2010; Roy et al., 2005). Thus, the accurate assessment and effective treatment of individuals with voices disorders are critical.

Vocal Hyperfunction

One of the common features of many voice disorders is vocal hyperfunction, defined as "conditions of abuse and/or misuse of the vocal mechanism due to excessive and/or 'imbalanced' muscular forces" of the intrinsic and extrinsic laryngeal muscles (Hillman, Holmberg, Perkell, Walsh, & Vaughan, 1989). Approximately 65% of all cases of voice disorders in voice clinics in the United States have vocal hyperfunction (Brodnitz, 1966; Ramig & Verdolini, 1998). Strained voice quality, or increased vocal effort, characterizes vocal hyperfunction in individuals with voice disorders; however, their voices can vary from extremely pressed to breathy, possibly due to varying musculature or structural pathologies (Koufman & Blalock, 1991; Morrison, 1997; Roy, 2008).

Two Major Types of Vocal Hyperfunction

The existence of two major types of vocal hyperfunction have been postulated: 1) *adducted* or *phonotraumatic* vocal hyperfunction and 2) *nonadducted* or *non-phonotraumatic* vocal hyperfunction (Hillman et al., 1989; Mehta, Van Stan, Zanartu, Ghassemi, Guttag, Espinoza, Cortes, Cheyne, & Hillman, 2015). Phonotraumatic vocal hyperfunction refers to a condition in which the vocal folds contain abnormal structural features, such as vocal nodules and polyps. The relationship between the behaviors of vocal hyperfunction and these structural features is unknown, whether causative or compensatory (Belafsky, Postma, Reulbach, Holland, & Koufman, 2002b; Galindo, Peterson, Erath, Castro, Hillman, & Zanartu, 2017; Hillman et al., 1989; Johns, 2003).

Non-phonotraumatic vocal hyperfunction describes vocal hyperfunction without any abnormal structural features of the vocal folds and is diagnosed as muscle tension dysphonia (MTD; or primary MTD). MTD has been previously called hyperfunctional dysphonia, muscle misuse dysphonia, and musculoskeletal tension dysphonia (Altman, Atkinson, & Lazarus, 2005). The following section summarizes the characteristics, etiology, and pathophysiology of the two types of vocal hyperfunction.

Phonotraumatic vocal hyperfunction

Phonotraumatic vocal hyperfunction is thought to involve tightly approximated vocal folds from over-activation of the intrinsic and extrinsic laryngeal muscles during phonation. Tightly approximated vocal folds have been suspected to result in increased vocal fold collision forces and increase the risk of vocal fatigue, tissue trauma, and the development of vocal fold lesions (Hillman et al., 1989; See Figure 1.1). This speculation was partly supported by a mathematical modeling experiment that applied compensating mechanisms such as increasing the subglottic pressure and laryngeal muscle activation to the vocal fold model (Galindo et al., 2017). The authors observed that when the compensating mechanisms were applied to less adducted vocal folds to achieve the sound intensity similar to that of well-adducted vocal folds, vocal fold collision forces increased in a statistically significant manner (Galindo et al., 2017). The results of a recent study involving an in-vivo rabbit model also suggested this positive association between increased vocal fold collision forces and the development of vocal fold lesions (Rousseau, Kojima, Novaleski, Kimball, Valenzuela, Mizuta, Daniero, Garrett, & Sivasankar, 2017). If this association exists, assessing the presence of vocal hyperfunction is critical before individuals develop structural changes, which pose the danger of permanent negative impact on voice.

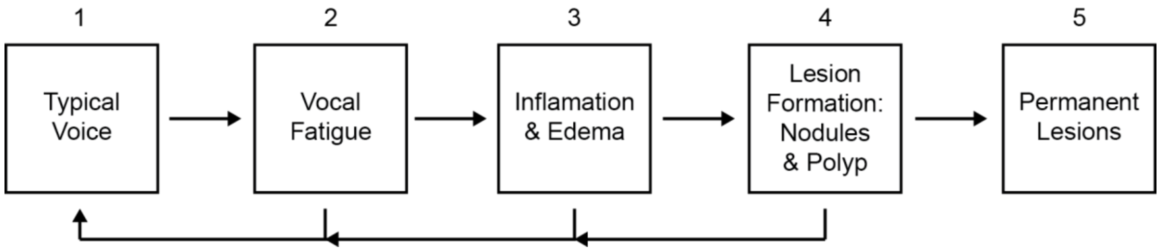

Figure 1.1: Adapted from Hillman et al. (1989). This flow chart presents the progression of phonotraumatic (adducted) vocal hyperfunction. The numbers, 1 to 5, represent the increasing order of severity, and the arrows to the right indicates progression of symptoms, whereas the arrows to the left indicate reduction of symptoms.

Non-phonotraumatic vocal hyperfunction

Non-phonotraumatic vocal hyperfunction is thought to involve voice production with stiff vocal folds but without complete glottal closure (Hillman et al., 1989). Since the vocal folds are not completely closed, no structural change is assumed to occur in non-phonotraumatic vocal hyperfunction. Individuals with non-phonotraumatic vocal hyperfunction still experience muscle fatigue and can sound both strained and breathy (Hillman et al., 1989; Koufman & Blalock, 1991; Morrison, 1997). Hillman et al. (1989) suggested that increasing severity of hyperfunctional behavior in individuals with non-phonotraumatic vocal hyperfunction can lead to aphonia, in which the vocal folds do not vibrate (Figure 1.2). The exact cause of this behavior is unknown. The symptoms of non-phonotraumatic vocal hyperfunction may be triggered by various factors including vocal misuse, compensation for changes in larynx such as laryngopharyngeal reflux or upper respiratory infection, or psychological factors such as introversion, anxiety, or stress (Van Houtte, Van Lierde, & Claeys, 2011). Non-phonotraumatic vocal hyperfunction may be a compensating mechanism for underlying glottal insufficiency (Belafsky et al., 2002b). In addition to excessive and/or imbalanced control in the laryngeal system, the respiratory, resonance, and associated voluntary muscular systems are also suspected of being poorly coordinated in this type of vocal hyperfunction (Morrison & Rammage, 1993).

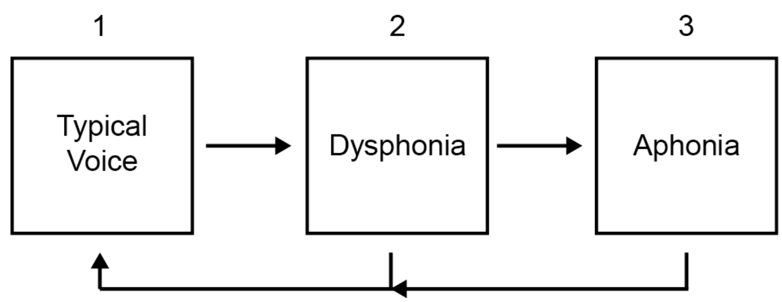

Figure 1.2: Adapted from Hillman et al. (1989). The chart presents the progression of non-phonotraumatic (nonadducted) vocal hyperfunction. The numbers, 1 to 3, represent the increasing order of severity, and the arrows to the right indicates progression of symptoms, whereas the arrows to the left indicate reduction of symptoms.

Subjective Assessment of Vocal Hyperfunction

Vocal hyperfunction is assessed clinically primarily by patient self-report and history examination, neck musculature palpation, visual inspection using videoendoscopy, and auditory-perceptual evaluation (Morrison, 1997; Roy, 2008). However, these clinical methods are subjective, and even expert clinicians may have low intra- and interrater reliability of performing these evaluations (Nawka & Konerding, 2012; Stepp, Heaton, Braden, Jette, Stadelman-Cohen, & Hillman, 2011a; Zraick, Kempster, Connor, Thibeault, Klaben, Bursac, Thrush, & Glaze, 2011). Some of these methods also lack direct associations with vocal hyperfunction. The following sections will summarize previous findings on the validity and reliability of the subjective methods used to assess vocal hyperfunction clinically.

Patient self-report and history examination

Clinical assessment of vocal hyperfunction usually begins with examining patients' self-report and their history. Individuals with vocal hyperfunction often report hoarseness, vocal fatigue and strain, and pain with speaking (Altman et al., 2005). Examining the patient's history can identify factors such as gastroesophageal reflux, high stress levels, and excessive and/or occupational voice use that are commonly associated with vocal hyperfunction (Altman et al., 2005). Previous studies that have suggested possible associations of anxiety and depression with vocal hyperfunction (Koufman & Blalock, 1991; Morrison & Rammage, 1993) lack empirical evidence (Roy, 2003). Roy, Bless, and Heisey (2000) found that extroversion was associated with phonotraumatic vocal hyperfunction, whereas introversion and neuroticism were associated with non-phonotraumatic vocal hyperfunction. However, evidence of the contributions of psychological factors to vocal hyperfunction is inconclusive.

In addition to patient self-reporting and history examination, a self-rating questionnaire is often administered with a standardized tool, such as the voice handicap index (VHI; Jacobson, Johnson, Grywalski, Silbergleit, Jacobson, Benninger, & Newman, 1997) and the voice-related quality of life (VRQOL; Hogikyan & Sethuraman, 1999). VHI has shown to improve after voice therapy in individuals with MTD (Jafari, Salehi, Izadi, Talebian Moghadam, Ebadi, Dabirmoghadam, Faham, & Shahbazi, 2017), but these self-rating questionnaires are focused on the general impacts of voice disruption on an individual's life and do not specifically address vocal hyperfunction.

Neck palpation

Neck palpation is often used to evaluate the presence and the degree of extrinsic laryngeal and other related muscle tension during voice production. Individuals with vocal hyperfunction are thought to use extrinsic laryngeal and other muscles excessively, rather than intrinsic laryngeal muscles alone, in order to produce voice. These individuals were often reported to have visible cervical neck tension, tight shoulders, and restriction of the jaw and the larynx (Altman et al., 2005; Morrison, 1997).

Although there is no standard grading tool for neck palpation, Angsuwarangsee and Morrison (2002) developed a grading tool for palpating extrinsic laryngeal and cricothyroid muscle tension in four different muscle groups: suprahyoid, thyrohyoid, cricothyroid, and pharyngolaryngeal. However, the authors found that only thyrohyoid ratings were significantly different between individuals with vocal hyperfunction and typical voices. Additionally, even though the authors reported poor-to-good reliability of this tool, Stepp et al. (2011a) observed only poor-to-moderate reliability of this tool when evaluating pre- and post-therapy changes in neck muscle tension in individuals with vocal hyperfunction. No correlation exists between the palpation ratings and surface electromyography (sEMG) measurement of the muscles that were palpated, suggesting questionable validity of this tool (Stepp et al., 2011a). Furthermore, a recent review of neck palpation technique suggests a lack of scientific evidence to support the validity and reliability of available neck palpation grading systems (Khoddami, Ansari, & Jalaie, 2015). Recently, a laryngeal palpatory scale (LPS; Jafari, Salehi, Meerschman, Izadi, Ebadi, Talebian, Khoddami, Dabirmoghadam, Drinnan, Jordens, D'Haeseleer, & Van Lierde, 2020) was developed to include more rating items than the scale from Angsuwarangsee and Morrison (2002). These additional items include the rating of infrahyoid and sternocleidomastoid muscles and the positions of the head and the shoulders. LPS showed promising interrater reliability, but no evidence yet exists to support its effectiveness in assessing the degree of vocal hyperfunction and treatment outcomes.

Visual inspection

Visual inspection of the larynx through videoendoscopy, performed using either a rigid transoral or flexible transnasal endoscope, determines the presence of any abnormal characteristics of the larynx and the vocal fold vibration. A rigid endoscope allows for the evaluation of the structures of the vocal folds, whereas a flexible endoscope allows for the evaluation of the functional aspects of voice production since it allows speech or singing tasks. Clinicians examine videoendoscopic images for the presence of supraglottic compression, asymmetry of vocal fold vibration, and a posterior glottal gap, all of which may be related to vocal hyperfunction (Hsiao, Liu, Hsu, Lee, & Lin, 2001). Grading systems that rate the videoendoscopic features (e.g., vocal fold vibration amplitude, mucosa wave, supraglottic activity, etc.) are available (Olthoff, Woywod, & Kruse, 2007; Poburka, Patel, & Bless, 2017), although are not standardized.

The reliability of rating each endoscopic feature visually varies from weak to strong (Poburka et al., 2017). Vocal fold vibratory features had especially poor reliability (Poburka et al., 2017). Poor reliability is likely due to the limitation of videostroboscopy, which estimates and constructs images of vocal fold vibration based on fundamental frequency (f_o). High-speed videoendoscopy (HSV), which can capture over 1000 frames/s, showed higher reliability (Olthoff et al., 2007; Poburka et al., 2017), but high-speed systems are not usually available in clinics and require a long time to process a large amount of data.

In addition, studies on rating tools for laryngeal videoendoscopy are scant; thus, the validity of these tools to assess vocal hyperfunction is still uncertain. Even among individuals with vocal hyperfunction, different vocal fold and vibratory characteristics have been reported, and 10% of participants with vocal hyperfunction did not show any abnormality in the assessed features (Hsiao et al., 2001). The variability of videoendoscopic features in individuals with vocal hyperfunction further suggests difficulty in assessing vocal hyperfunction with visual inspection.

Auditory-perceptual evaluation

Of all the subjective evaluation methods, auditory-perceptual evaluation has been central to clinical voice care because voice is perceived through the auditory system (Carding, Wilson, MacKenzie, & Deary, 2009; Kreiman, Gerratt, Kempster, Erman, & Berke, 1993; Oates, 2009). The development of standard grading systems such as the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009) and the Grade, Roughness, Breathiness, Asthenia, and Strain (GRBAS) scale (Hirano, 1981) has assisted clinicians in rating different dimensions of voice quality. Strain, defined as "perception of excessive vocal effort (hyperfunction)" in CAPE-V (Kempster et al., 2009), is a major perceptual feature related to vocal hyperfunction. Thus, strain assessment is critical to the effective assessment and treatment of individuals with vocal hyperfunction. However, previous studies on the reliability of the auditory-perceptual evaluation have shown that among voice qualities, the rating of strain results in poor intra- and interrater reliability (Webb, Carding, Deary, MacKenzie, Steen, & Wilson, 2004; Zraick et al., 2011). This lack of reliability is a serious problem because auditory-perceptual assessment by a single clinician is the current the gold standard to evaluate voice disorders, guide treatment, and assess treatment outcomes (Oates, 2009).

Objective Assessment of Vocal Hyperfunction

Due to the limitations of subjective evaluations, instrumental measures have aided in the clinical evaluation of voice disorders (Carding et al., 2009; Mehta & Hillman, 2008; Roy, Barkmeier-Kraemer, Eadie, Sivasankar, Mehta, Paul, & Hillman, 2013). Two instrumental methods recommended for clinical use are acoustic and aerodynamic measurements (Patel, Awan, Barkmeier-Kraemer, Courey, Deliyski, Eadie, Paul, Svec, & Hillman, 2018). Although these measurements can be performed relatively easily in clinics and have been shown to be effective in evaluating overall dysphonia (Awan & Roy, 2006; Heman-Ackah, Heuer, Michael, Ostrowski, Horman, Baroody, Hillenbrand, & Sataloff, 2003), assessing vocal hyperfunction specifically with acoustic and aerodynamic measurements has been difficult. The following sections summarize previous findings of the studies that attempted to evaluate vocal hyperfunction objectively with these two measurements.

Acoustic measurements

Fundamental frequency and sound pressure level

Fundamental frequency (f_o) and sound pressure level (SPL) are the commonly used acoustic measures of voice. These measures have direct perceptual correlates; pitch and loudness, respectively. However, in many previous investigations, neither f_o nor SPL has been able to differentiate between individuals with healthy voices and vocal hyperfunction (Belsky, Rothenberger, Gillespie, & Gartner-Schmidt, In Press; Hillman et al., 1989; Ju, Jung, Kwon, Woo, Cho, Park, Park, & Baek, 2013; Schindler, Mozzanica, Maruzzi, Atac, De Cristofaro, & Ottaviani, 2013) or between pre- and post-therapy sessions (Fex, Fex, Shiromoto, & Hirano, 1994; Holmberg, Doyle, Perkell, Hammarberg, & Hillman, 2003; Ju et al., 2013). Some of these studies have shown mixed findings as well. Hillman et al. (1989) examined individuals with different types of vocal hyperfunction and found that f_o was decreased in participants with contact ulcer, but increased in participants with MTD. The f_o values of participants with nodules and polyps were not different from those of controls (Hillman et al., 1989). Fex et al. (1994) observed statistically significant decreases in f_o after therapy sessions in female participants, and both increased and decreased f_o patterns after therapy sessions in male participants. Recent ambulatory voice monitoring studies have also observed that f_o and SPL measured throughout a day did not differ between individuals with typical voices and vocal hyperfunction (Mehta et al., 2015; Van Stan, Mehta, Ortiz, Burns, Toles, Marks, Vangel, Hron, Zeitels, & Hillman, 2020). These ambulatory findings further suggested that f_o and SPL may not be useful to assess vocal hyperfunction because ambulatory voice monitoring recorded individuals for longer periods of time throughout the day than typical laboratory recordings.

Jitter, shimmer, and harmonics-to-noise ratio

Other common acoustic measures that have been frequently used are jitter, shimmer, and harmonics-to-noise ratio (HNR), also known as perturbation measures. Jitter quantifies cycle-to-cycle variations in f_o, and shimmer quantifies cycle-to-cycle variation in the amplitude of the signals. Increases in these variations were thought to suggest irregularity in voicing, and thus, dysphonia. HNR estimates the amount of harmonic energy in relation to noise energy, which estimates the periodicity of the sound, which is also related to overall dysphonia (Awan & Roy, 2006; Heman-Ackah et al., 2003). These perturbation measures have shown to improve after therapy sessions with individuals with vocal hyperfunction (Fex et al., 1994; Lin, Sun, Yang, Zhang, Shen, Shi, Fang, & Sun, 2014; Roy & Leeper, 1993; Schindler et al., 2013), but did not respond to therapy in other studies (Gillespie, Dastolfo, Magid, & Gartner-Schmidt, 2014; Ju et al., 2013). No group differences between individuals with typical voices and MTD have been observed in these measures (Dabirmoghaddam, Aghajanzadeh, Erfanian, Aghazadeh, Sohrabpour, Firouzifar, Maroufizadeh, & Nikravesh, 2019) except when these measures were obtained from /a/. Thus, the ability of perturbation measures to assess the presence and degree of vocal hyperfunction remains unclear.

The reliability of perturbation measures is questionable. These measures are affected by sound intensity and improved when individuals produced a loud voice (Brockmann-Bauser, Bohlender, & Mehta, 2018; Brockmann, Storck, Carding, & Drinnan, 2008). Perturbation measures require accurate f_o tracking to obtain precise measures, and thus were not considered suitable for dysphonic voices in which accurate f_o tracking is not possible (Mehta & Hillman, 2008). The measures also showed poor to moderate test-retest reliability in previous studies (Bough, Heuer, Sataloff, Hills, & Cater, 1996; Carding, Steen, Webb, MacKenzie, Deary, & Wilson, 2004; Leong, Hawkshaw, Dentchev, Gupta, Lurie, & Sataloff, 2013). Perhaps due to these reasons, perturbation measures have not been recommended recently as acoustic measures for clinical use (Mehta & Hillman, 2008; Patel et al., 2018).

Cepstral peak prominence

The only acoustic measure for voice quality that was present in the recent recommendation for instrumental evaluation of voice is cepstral peak prominence (CPP; (Patel et al., 2018). CPP is obtained from cepstrum, the Fourier transformation of the power spectrum, and is the first peak amplitude at the fundamental period normalized by the cepstral amplitude of overall background signal (Hillenbrand, Cleveland, & Erickson, 1994). CPP can represent the harmonic energies presented in the signal, and thus the periodicity of the voice signal. CPP has an advantage over perturbation measures because it can be obtained from both sustained vowel and sentence context, and the accurate f_o estimation is not required (Mehta & Hillman, 2008; Patel et al., 2018). CPP has been statistically associated with breathiness and overall dysphonia since it represents the periodicity of the signal (Awan & Roy, 2006; Brinca, Batista, Tavares, Goncalves, & Moreno, 2014; Heman-Ackah et al., 2003; Lowell, Kelley, Awan, Colton, & Chan, 2012a).

However, inferences from previous conflicting evidence show that CPP may not reflect vocal hyperfunction specifically. Belasky et al. (2020) found no group differences in CPP between MTD and the control group, whereas Dabirmoghaddam et al. (In Press) and Shim, Jung, Koul, and Ko (2016) found that CPP was lower in the MTD group than the control group. Ambulatory monitoring studies observed no difference in CPP among control groups, the phonotraumatic vocal hyperfunction groups (Mehta et al., 2015; Van Stan et al., 2020), and the MTD groups (Mehta et al., 2015). CPP improved after therapy sessions for individuals with vocal lesions, but CPP did not improve in individuals with MTD (Gillespie et al., 2014). CPP has shown statistically positive (Anand, Kopf, Shrivastav, & Eddins, 2019), negative (Lowell et al., 2012a), and a lack of correlation with strain (Brinca et al., 2014; McKenna & Stepp, 2018). Additionally, the negative correlation observed in Lowell et al. (2012) may be related to overall dysphonia present in individuals with laryngeal dystonia, paralysis, papilloma, and presbyphonia, who were also included in the study.

CPP, which reflects periodicity of the signal, lacks a direct theoretical association with strained voices in vocal hyperfunction. The relationship between signal periodicity and strained voices is unknown. Van Stan et al. (2020) also suggested that CPP may not have been different between vocal hyperfunction and a control because individuals with vocal hyperfunction may use compensating mechanisms to produce overall voice quality similar to typical voices. Thus, CPP measured in individuals with vocal hyperfunction may not differ from CPP in individuals with healthy voices (Van Stan et al., 2020).

Aerodynamic measurement

Previous studies on aerodynamic measures have also presented conflicting evidence for differentiating between individuals with typical voices and vocal hyperfunction. Mean airflow rate has been observed to be lower (Carroll, Rooney, Ow, & Tan, 2018; Zheng, Zhang, Su, Gong, Yuan, Ding, & Rao, 2012 [only in males]), higher (Hillman et al., 1989), and not different for individuals with vocal hyperfunction relative to controls (Belsky et al., In Press; Espinoza, Zanartu, Van Stan, Mehta, & Hillman, 2017). Gillespie et al. (2014) observed all three patterns of airflow rate in subgroups of MTD participants. The estimate of subglottic pressure was higher in individuals with phonotraumatic vocal hyperfunction (Carroll et al., 2018; Espinoza et al., 2017), nodule and polyps (Hillman et al., 1989), and nodule only (Hillman, Holmberg, Perkell, Walsh, & Vaughan, 1990b) than controls, but was not found to be different (Hillman et al., 1989 [ulcer]; Hillman et al., 1990 [polyp and ulcer]). The estimate of subglottic pressure was also higher in non-phonotraumatic vocal hyperfunction than in controls (Espinoza et al., 2017; Gillespie, Gartner-Schmidt, Rubinstein, & Abbott, 2013; Hillman et al., 1989; Zheng et al., 2012), but not different as well (Carroll et al., 2018; Gillespie et al., 2013).

The conflicting evidence of the previous studies may be related to inherent variability in aerodynamic measures. Mean airflow rate and subglottic pressure estimations have been observed to vary even within speakers with typical voices (Holmberg, Hillman, Perkell, & Gress, 1994). Vocal fold lesions have been found to affect the aerodynamic measures depending on their locations, sizes, and the compensating mechanisms of speakers for the lesions (Hillman et al., 1990b). Thus, aerodynamic measures in individuals with vocal fold lesions may be more related to the lesions than the degree of vocal hyperfunction. Five different aerodynamic patterns were observed among individuals with

MTD without vocal fold lesions, indicating various aerodynamic mechanisms achieve voicing in the MTD population (Gillespie et al., 2013). In addition, subglottic pressure estimation also can be affected by ‘peaky’ pressure waveforms (Holmberg, Perkell, & Hillman, 1984) and are found to be only poor-to moderately reliable (Awan, Novaleski, & Yingling, 2013).

Relative Fundamental Frequency

An acoustic measure, relative fundamental frequency (RFF), has been recently proposed as a possible objective index of vocal hyperfunction. RFF has been defined as the instantaneous fundamental frequencies (f_os) sof the 10 voicing cycles prior to and after a voiceless consonant, normalized by the f_os of the steady-state portions of voiced sounds in a voiced-voiceless consonant-voiced (VCV) context (Figure 1.3; Stepp, Hillman, & Heaton, 2010). Thus, RFF quantifies the amount of short-term variation of f_o surrounding the production of the voiceless consonant. The instantaneous values of f_os are normalized to the reference values of the f_o of the cycles furthest from the consonant: the offset cycle 1 and onset cycle 10 (Figure 1.3). The f_o values of the 10 offset cycles are normalized to the f_o of the offset cycle 1 and the f_o values of the 10 onset cycles are normalized to the f_o of the onset cycle 10 from the semitone (ST) equation: $12 \times \log_2(f_o/\text{reference } f_o)$. The offset cycle 1 and onset cycle 10 are used as reference cycles because they are considered stable portions of the vowels surrounding the consonant in a VCV context. Normalizing f_o allows the comparison of RFF values between speakers with different habitual speaking f_os.

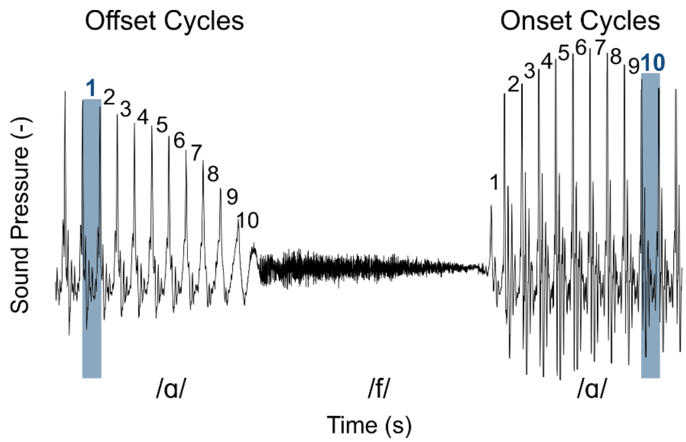

Figure 1.3: An example acoustic waveform of a voiced-voiceless consonant-voiced context. The offset cycles and onset cycles 1 – 10 are labeled. The offset cycle 1 and onset cycle 10, used as reference cycles for RFF calculation, are bolded and shaded.

This short-term variation of f_o was investigated in relation to laryngeal tension since f_o depends on the mass, length, and the tension of the vocal folds, similar to a vibrating string under tension (Zhang, 2016b). Long-term variation of f_o in the context of sentences was considered to be unsuitable for reflecting baseline laryngeal tension since f_o can be affected by voluntary modulation (e.g., prosodic control). In contrast, short-term changes in f_o during the devoicing-voicing maneuver in the production of an intervocalic obstruent were regarded as a better indicator of baseline laryngeal tension, since the changes in f_o are less likely to be affected by intentional f_o control in a prosodic manner (Stepp et al., 2010a).

Physiological Mechanisms of RFF

Although the exact way in which RFF reflects increased laryngeal tension in voice production is unknown, three physiological factors for RFF patterns were hypothesized by Watson (1998) and Stepp, Merchant, Heaton, and Hillman (2011b): cricothyroid (CT) muscle activation, abduction, and aerodynamics. The following section summarizes each hypothesized factor, its rationale, and existing evidence supporting each factor.

CT muscle activation

The hypothesis that CT muscle activation would affect RFF during intervocalic obstruent production originates from the observation of House and Fairbanks (1953) that vowels after voiceless consonants had higher f_os compared to vowels after voiced consonants. Based on this finding, Stevens (1977) postulated that this higher f_o after voiceless consonants might be due to an increase in the CT muscle activation that might aid devoicing for voiceless consonant production and carry onto the first few cycles of the following vowel. The CT muscle is known to increase the longitudinal tension of the vocal folds, and thus would increase f_o. CT muscle activation during voiceless consonant production was supported by Löfqvist, Baer, McGarr, and Story (1989) in a three-participant, hook-wire electromyography (EMG) experiment. This study demonstrated that the CT muscle showed increased activation before and after productions of a voiceless consonant in a VCV context but less activation during productions of a voiced consonant. Because of the possible increased CT muscle activation surrounding the voiceless consonant in VCV contexts, RFF is hypothesized to be increased surrounding voiceless consonants (Figure 1.4).

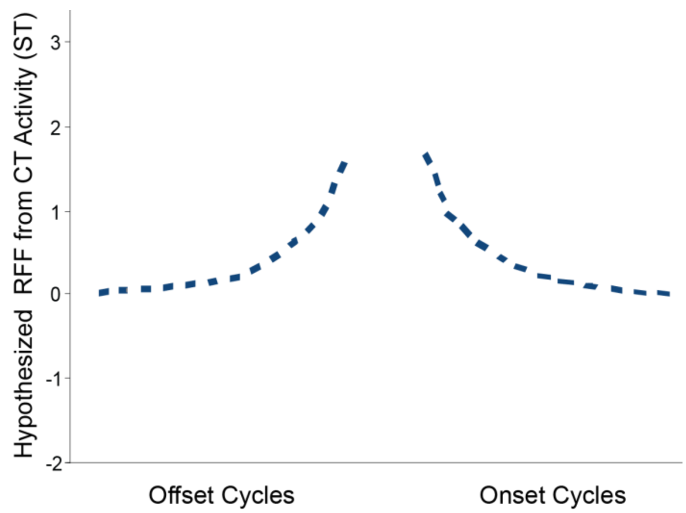

Figure 1.4: A schematic of the hypothesized RFF resulting from the increased CT muscle activity during the obstruent production in a VCV context.

A possible difference in the effect of the CT muscle activation between individuals with healthy voices and individuals with vocal hyperfunction was the proposed explanation of why RFF would be lower in individuals with vocal hyperfunction than in individuals with healthy voices (Stepp et al., 2010a). Heightened baseline laryngeal tension in individuals with vocal hyperfunction was thought to reduce the magnitude of short-term increases in vocal fold longitudinal tension by the CT muscles due to a suspected ceiling effect, and thus short-term increases in f_0 during VCV contexts (Stepp et al., 2011b). The authors did not specify the physiological mechanism of how the effect of the CT muscles would be reduced, but we suspect that the baseline CT muscle activity may increase due to vocal hyperfunction and result in a ceiling effect. We also suspect that the effect of CT muscles may be reduced due to either increased thyroarytenoid (TA) muscles that counteract the CT muscle (Chhetri, Neubauer, & Berry, 2012; Yin & Zhang, 2013) or already increased longitudinal vocal fold tension due to a raised larynx (Honda, Hirai, Masaki, & Shimada, 1999), commonly found in individuals with vocal hyperfunction (Lowell, Kelley, Colton, Smith, & Portnoy, 2012b), resulting in a ceiling effect. Stepp et al. (2011b) hypothesized that the reduced effect of the CT muscles would result in reduced increases in RFF surrounding a voiceless consonant (Figure 1.5). However, this hypothesis has not yet been directly supported by empirical evidence.

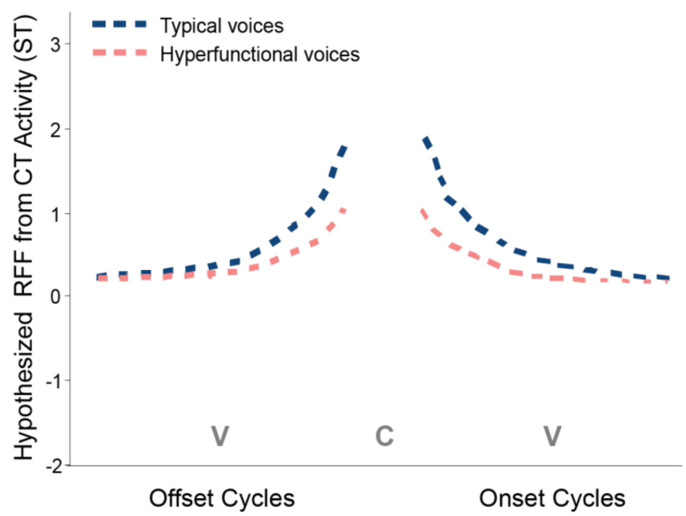

Figure 1.5: A schematic of hypothesized RFF resulting from the increased CT muscle activity during the obstruent production in a VCV context in hyperfunctional voices (pink) and typical voices (blue).

Although direct evidence for increased baseline laryngeal tension in vocal hyperfunction does not exist, indirect estimation of baseline laryngeal tension has been calculated using vocal fold kinematics from videoendoscopic images of the vocal fold movements. In the discipline of exercise physiology, the kinematic estimate of stiffness was developed to estimate the stiffness of moving muscles and was defined as the ratio of maximum velocity to the extent of the movement (J. D. Cooke, 1980; Feldman, 1980). This estimate was then applied to vocal fold adductory gestures from videoendoscopic images to estimate the vocal fold stiffness; and, this estimate was found to be higher when obtained from hard onsets than when obtained from breathy onsets (A. Cooke, Ludlow, Hallett, & Selbie, 1997), which suggested its applicability of estimating laryngeal stiffness. This kinematic estimate of laryngeal stiffness (KS) was measured from individuals with healthy voices and individuals with vocal hyperfunction when they increased the speed of producing “sniff-/i/” from comfortable to fast speed. Individuals with vocal hyperfunction demonstrated statistically lower increases in KS than individuals with typical voices when they increased the speed than individuals with healthy voices did, which suggested their already increased baseline laryngeal tension during the comfortable speed, which would result in a ceiling effect (Stepp, Hillman, & Heaton, 2010b). The authors also examined KS in a simple vocal fold trajectory model. When the stiffness values of the TA, lateral cricoarytenoid (LCA), and posterior cricoarytenoid (PCA) muscles increased in the model, the KS value also increased, supporting the association between KS and the laryngeal stiffness (Stepp et al., 2010b). McKenna, Heller Murray, Lien, and Stepp (2016) examined KS in relation to RFF values and found statistical, negative correlations between them when individuals with healthy voices modulated their vocal strain. These KS findings indirectly supported the idea that increased baseline laryngeal tension may result in decreased RFF.

Abduction

Another hypothesized physiological factor, abduction, has been thought to decrease RFF (Figure 1.6; Stepp et al., 2011b; Watson, 1998). Before RFF was explored in relation to vocal hyperfunction, it was examined in younger and older adults to explore voice change associated with aging. RFF was observed to be stable and near zero prior to a voiceless consonant in younger adults, whereas it showed a decreasing contour before the consonant in older adults (Watson, 1998). From this finding, Watson (1998) speculated that the effects of the CT muscle activation (Stevens, 1977) and vocal fold abduction, which was suspected to decrease RFF, might cancel each other out and result in RFF values at voicing offset (offset RFF) near zero for younger adults. In contrast, Watson (1998) suspected that older adults may lack the ability to use the CT muscles for devoicing due to age-related muscular and morphological changes in the larynx, and thus may rely more on abduction. Increased abduction in older adults was the suspected reason why offset RFF values might be lower in older adults compared to younger adults (Watson, 1998).

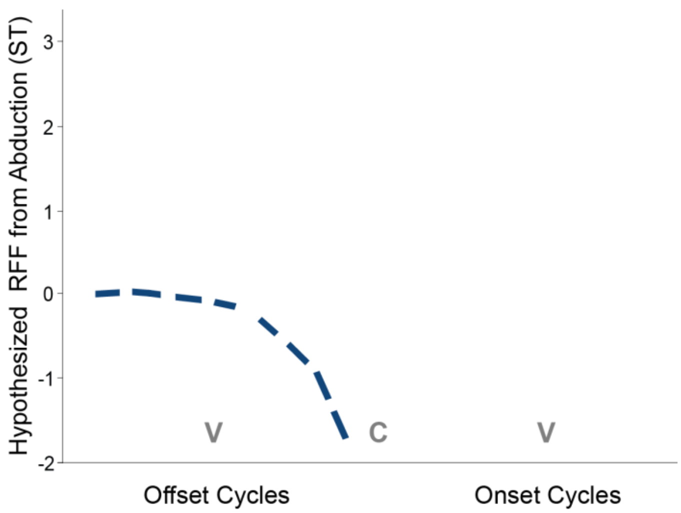

Figure 1.6: A schematic of hypothesized RFF resulting from abduction prior to the obstruent production in a VCV context.

Watson (1998) hypothesized that abduction could affect f_o at voicing offset since abduction starts during the vowel prior to a voiceless consonant (Fukui & Hirose, 1983). The contact area of the vocal fold vibratory cycle decreases during abduction as the vocal folds are moving apart. As a result, Watson (1998) suspected that both the collision force for closing the vocal folds and the recoil force for pulling the vocal folds apart in a vibratory cycle would also decrease. The decreases in the collision and recoil forces would increase both contacting and decontacting phases of a vibratory cycle (Figure 1.7). Increased contacting and decontacting phases increase the period of a vibratory cycle, which in turn decreases the f_o (Watson, 1998). Based on the same rationale, Stepp (2013) hypothesized that individuals with Parkinson's disease would have lower RFF offset values since they may start vocal fold abduction even earlier than aged-matched control groups to compensate for their difficulty of devoicing (Gallena, Smith, Zeffiro, & Ludlow, 2001). However, the exact effect of abduction on RFF is not yet supported by experiments directly examining vocal fold kinematics and RFF.

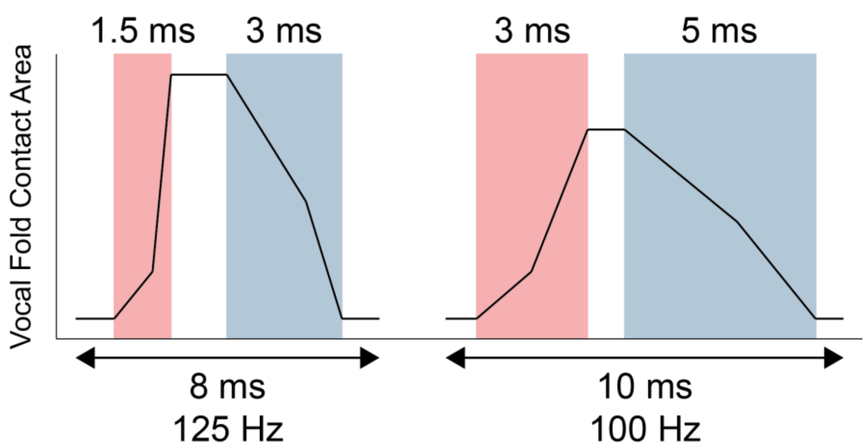

Figure 1.7: A schematized plot of vocal fold contact area as a function of time. Two vibratory cycles are presented, and the cycle on the right represents a cycle with decreased vocal fold contact area, which was suspected to result in increases in both a contacting phase (pink) and a decontacting phase (blue). Thus, the fundamental period of the cycle on the right is greater than of the one on the left, and thus the f_o is lower.

Aerodynamics

Regarding aerodynamics, Watson (1998) postulated that the increased RFF at voicing onset (onset RFF) following the voiceless consonant is due to the combination of the CT muscle activation and the increase in airflow rate from the production of the voiceless consonant. Stepp et al. (2011b) also suspected the positive relationship between airflow rate and RFF (Figure 1.8), based on Ladefoged (1967), which suggested that a higher airflow rate could result in a higher Bernoulli force (faster vocal fold closing) and a faster vibration of the vocal folds, which would result in an increase in f_o. Peak and minimum airflow rates were observed to increase following the voiceless consonant in a VCV context in a six-participant study (Löfqvist, Koenig, & McGowan, 1995). However, no empirical evidence supports the proposed direct relationship between the airflow rate and f_o.

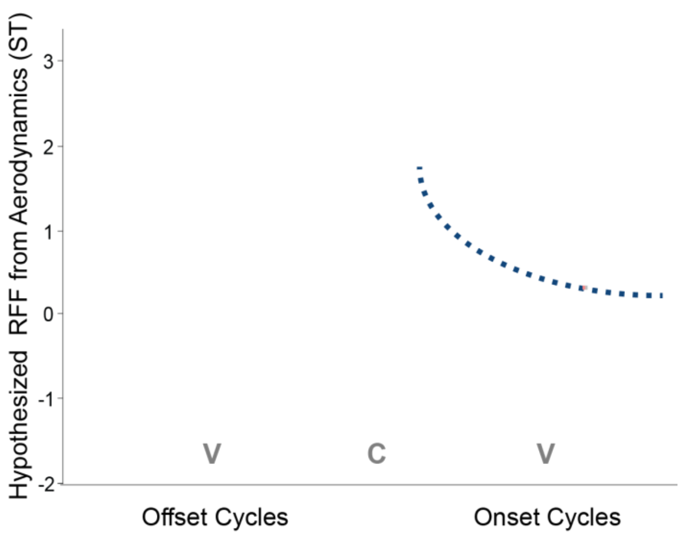

Figure 1.8: A schematic of hypothesized RFF resulting from aerodynamics after the obstruent production in a VCV context.

Figure 1.9 presents the hypothesized RFF from the three physiological factors that have explained the RFF pattern observed in typical voices in previous studies (Stepp et al., 2011b; Stevens, 1977; Watson, 1998).

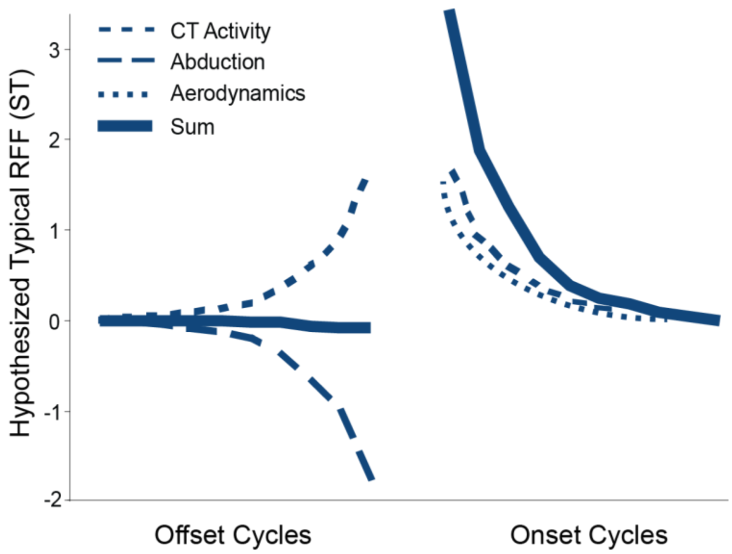

Figure 1.9: Adapted from Stepp et al. (2011b). A schematic of RFF in typical speakers resulting from the three hypothesized mechanisms behind RFF.

RFF's Validity of Assessing the Presence and the Degree of Vocal Hyperfunction

Although no studies exist that directly examine the physiological factors for RFF, previous studies have indirectly supported RFF's validity of assessing the presence and the degree of vocal hyperfunction. Based on the decreased offset and onset RFF patterns observed in individuals with Parkinson's disease (Goberman & Blomgren, 2008), Stepp et al. (2010a) postulated that the decreased RFF may be attributed to increased baseline laryngeal tension due to the overall rigidity affecting the larynx as well. The authors speculated that increased baseline laryngeal tension may reduce the f_o-increasing effect of the CT muscle activation due to a ceiling effect during production of VCV contexts (Stepp et al., 2010a). Then, the authors applied this rationale to individuals with vocal hyperfunction, who also exhibit increased baseline laryngeal tension, hypothesizing that individuals with vocal hyperfunction would also have lower RFF values than individuals with typical voices. The participants of this study included 15 young controls, 30 individuals with vocal fold polyps, 30 individuals with vocal fold nodules, and 22 individuals with MTD. This study found that individuals with voice disorders related to vocal hyperfunction had significantly lower RFF than controls (Figure 1.10; Stepp et al., 2010a). They also examined RFF from patients before and after surgery to remove vocal fold lesions to understand if the previous RFF results were due to the structural pathologies. They found that RFF did not statistically change after the surgery, which suggested that RFF would be likely to reflect habitual vocal hyperfunction that the patients still may have exhibited after the surgery.

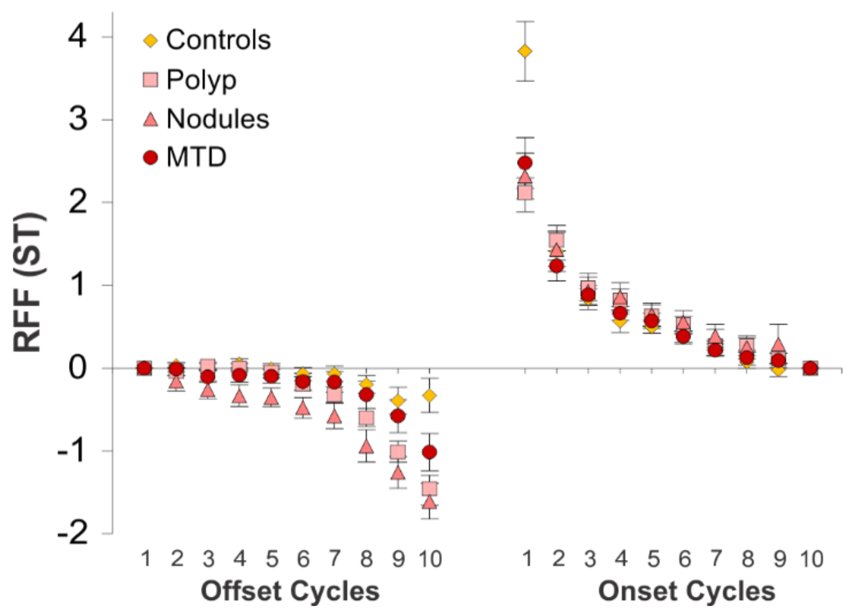

Figure 1.10: Adapted from Stepp et al. (2010a). Mean RFF from control participants and participants with polyps, nodules, and MTD. Error bars indicate standard errors.

Further research supported RFF's ability to assess increased baseline laryngeal tension. Lower RFF values in individuals with vocal hyperfunction were observed again (Stepp, Sawin, & Eadie, 2012), even in a larger group of 111 individuals with MTD (Roy, Fetrow, Merrill, & Dromey, 2016). Stepp (2013) examined RFF in individuals at various stages of Parkinson's disease and found that a group with more severe Parkinson's disease had lower RFF values compared to both less severe and age-matched control groups. Individuals with adductory laryngeal dystonia (also known as spasmodic dysphonia), who also exhibit increased laryngeal tension in their voice production, also showed lower RFF values than individuals with typical voices (Eadie & Stepp, 2013). Different RFF values were also observed between phonotraumatic and non-phonotraumatic vocal hyperfunction; the individuals with phonotraumatic vocal hyperfunction had lower RFF than individuals with non-phonotraumatic vocal hyperfunction, which was attributed to their more tightly approximated vocal folds relative to non-phonotraumatic vocal hyperfunction (Heller Murray, Lien, Van Stan, Mehta, Hillman, Pieter Noordzij, & Stepp, 2017).

Recent studies also supported RFF's ability to track changes in the degree of vocal hyperfunction. After observing that RFF values were lower in individuals with vocal hyperfunction (Stepp et al., 2010a), the same group examined acoustic recordings of 16 females with either vocal nodules or MTD who went through successful voice therapy sessions (Stepp et al., 2011b). RFF values were estimated from these recordings, and post-therapy RFF values were statistically higher than pre-therapy RFF values. RFF at offset cycle 10 (offset 10 RFF) and onset cycle 1 (onset 1 RFF) showed the most changes after the therapy among the RFF cycles (Stepp et al., 2011b) in particular. Because the goal of therapy for these individuals was to reduce excessive laryngeal tension, the results indicated that RFF may reflect the changes in baseline laryngeal tension. In a similar study, Roy et al. (2016) also found that successful voice therapy sessions increased onset 1 RFF values in a larger group of participants (n = 111) with MTD. Additionally, RFF was also observed to increase after a long-term high voice use period (Heller Murray, Hands, Calabrese, & Stepp, 2016). Heller Murray et al. (2016) recruited 12 university volleyball players and recorded their voices a week before and a week after a 10-week spring season, along with 6 controls. The volleyball season was associated with the excessive use of the players' voices. After the season, onset 1 RFF values of the volleyball players statistically decreased, whereas onset 1 RFF values of controls remained the same (Heller Murray et al., 2016). The results suggested that onset 1 RFF may reflect vocal abuse and possibly increased laryngeal tension in individuals with healthy voices.

In addition to reflecting the long-term changes in the degree of baseline laryngeal tension, RFF has also shown to reflect increased vocal effort in individuals self-modulating their vocal effort. Lien, Michener, Eadie, and Stepp (2015b) asked 12 participants with healthy voices to purposefully modulate their vocal effort from relaxed voice to maximum vocal effort while producing /pae/ trains. RFF was found to decrease as individuals purposefully increased their vocal effort (Lien et al., 2015b). They also obtained aerodynamic estimates of vocal efficiency (the ratio of sound pressure level over subglottic pressure) and observed a statistically significant but weak relationship between aerodynamic estimates of vocal efficiency and RFF. When the same correlation was examined within the individuals in this study, a stronger relationship was found, which suggested that RFF may be more useful in tracking changes in hyperfunctional behavior within individuals than in comparing the degree of vocal hyperfunction across individuals.

Perceptual studies have examined the relationship between RFF and perceived vocal effort, which is strain. Strain is a major perceptual attribute of vocal hyperfunction

(Kempster et al., 2009). Since RFF was observed to be lower in individuals with vocal hyperfunction, these perceptual studies aimed to delineate the relationship between RFF and strain. Stepp et al. (2012) performed a perceptual study in which 12 inexperienced listeners rated the strain of 10 speakers with healthy voices and 30 speakers with voice disorders related to vocal hyperfunction. The correlation analysis revealed a weak but statistically significant negative correlation between offset 10 RFF and strain, but no relationship was found between onset 1 RFF and strain (Stepp et al., 2012). In contrast, when the correlation between RFF and strain was examined in 19 individuals with adductory laryngeal dystonia, onset 1 RFF showed a statistically significant negative correlation with strain, whereas offset 10 RFF did not show a relationship. Similarly, when RFF was examined within 12 individuals with healthy voices modulating their vocal effort, a stronger correlation was found between onset 1 RFF and strain than between offset 10 RFF and strain. However, McKenna and Stepp (2018) again found from voice samples of individuals with healthy voices modulating their vocal effort that only offset 10 RFF was a significant predictor of strain when they performed mixed-effect regression analysis with other acoustic measures related to strain. In these previous experiments, the correlations observed between RFF and strain also might have been confounded by other acoustic measures, which were likely to be different in voice samples from individuals with and without vocal hyperfunction or individuals self-modulating their vocal effort. Thus, the question still remains if the listeners directly perceive RFF as changes in strain.

The results from a few perceptual studies also suggest that RFF may reflect increased baseline laryngeal tension that is not perceived by listeners as increased strain.

Stepp et al. (2012) evaluated RFF and strain in the individuals with healthy voices and individuals with vocal hyperfunction and observed that some individuals with vocal hyperfunction had similar strain ratings to those of individuals with healthy voices but had lower offset 10 RFF values. These findings may indicate that RFF may be able to detect possible underlying vocal function changes that cannot be perceived by listeners. A similar finding was observed in the study evaluating volleyball players' voices prior to and after a period of high voice use (Heller Murray et al., 2016). The experiment also included an auditory-perceptual evaluation of strain as well as RFF. Although their onset 1 RFF significantly decreased after the high voice use period, their strain ratings did not change, which also suggests that RFF may reflect small functional changes in voice that may not be perceived by listeners.

Previous Studies Optimizing the Reliability of RFF

Since RFF has shown promise as a possible clinical measure of vocal hyperfunction, researchers have sought to optimize the reliability of RFF. The first attempt was evaluating how many RFF instances should be averaged to precisely represent an individual's vocal function (Eadie & Stepp, 2013). Prior to this study, RFF estimation usually included one to six samples to be averaged to represent an individual (Goberman & Blomgren, 2008; Stepp et al., 2010a; Watson, 1998). Eadie and Stepp (2013) averaged one to nine RFF instances in 19 participants to represent an individual's RFF. For each number of RFF instances used, a coefficient of determination (R^2) between RFF and strain was calculated; and, the manner in which the increased number of samples used to obtain RFF changed the strength of correlation between RFF and strain was evaluated. The results showed that as the number of samples increased to calculate onset 1 RFF, the strength of the correlation increased, and the coefficient reached a plateau after six or more samples were averaged. This finding indicated that at least six or more samples should be averaged to obtain RFF values that can precisely reflect strain.

The phonetic context of RFF stimuli has also been evaluated because RFF has been obtained from different vowels and voiceless consonants in a VCV context. Lien, Gattuccio, and Stepp (2014) examined the effect of different voiceless consonants and stimulus types on RFF to develop an RFF stimulus that could minimize within-speaker variability. The voiceless consonants included /f, s, ʃ, p, t, k/ and the stimulus types included uniform utterances (e.g., /ifi/ and /upu/) and sentences with same (e.g., "**I te**ll you, m**y te**a is wa**y to**o warm") or different (e.g., "**A pe**nn**y ca**n only get you s**o fa**r in life") voiceless consonants in VCV contexts. Both voiceless consonants and stimulus types showed statistically significant effects on both RFF means and within-subject standard deviations with small effect sizes. Both offset 10 and onset 1 RFF means were higher in fricatives than in stops and were also higher in sentences than in uniform utterances. RFF within-subject standard deviation was lower in fricatives than in stops, and was the lowest in uniform utterances among the stimulus types. Lien et al. (2014) suspected that uniform utterances may result in less variable RFF due to a possible effect of the vowels on RFF since the uniform utterance included the same vowels surrounding a voiceless consonant, whereas the sentence stimuli had different vowels. Based on the findings of this study, Lien et al. (2014) recommended uniform utterances with voiceless fricatives to obtain reliable RFF values. However, another study had conflicting findings and did not show any significant difference between /f/ and /p/ in both offset and onset RFF values (A. B. Smith & Robb, 2013).

The effect on voiceless consonants on changes in RFF after voice therapy was also evaluated. Comparing RFF values obtained from /f, ʃ, p/ between pre- and post-therapy sessions in an MTD group, Roy et al. (2016) observed that /f/ and /ʃ/ showed greater RFF differences after the successful therapy sessions than /p/. They also observed a greater correlation between dysphonia severity and RFF obtained from /f/ and /ʃ/ than between dysphonia severity and RFF obtained from /p/ (Roy et al., 2016).

Although the uniform utterances with fricatives showed improved reliability relative to sentence stimuli, Park and Stepp (2019) examined other possible factors that could still affect the RFF mean values and within-speaker variations: stress type, vowel, baseline f_o, and loudness. The stress type showed a medium-to-large effect on both RFF means and within-subject standard deviations, and equal stress on both vowels resulted in the lowest within-subject standard deviation (Park & Stepp, 2019). In contrast, vowel type, baseline $f_{o,}$ and loudness showed minor effects on RFF means and standard deviations. Based on the results of this study, Park and Stepp (2019) added to the recommendation of Lien et al. (2014) on RFF stimuli that the stress type should be controlled with equal stresses on both vowels.

A semi-automated algorithm was also developed to estimate RFF in order to reduce the labor and subjectivity of manual RFF analysis (Lien, Heller Murray, Calabrese, Michener, Van Stan, Mehta, Hillman, Noordzij, & Stepp, 2017; Vojtech, Segina, Buckley, Kolin, Tardif, Noordzij, & Stepp, 2019). Prior to the development of the semi-automated algorithm, RFF was manually estimated by technicians. Manual estimation involves determining the last and first cycles of voicing offset and voicing onset of each RFF instance, identifying 10 cycles for both offset and onset in Praat acoustic software (Boersma & Weenink, 2019), and converting the f_o values to RFF with the RFF equation. This procedure takes an unreasonable amount of time for clinical application since six or more RFF instances were recommended to be averaged (Eadie & Stepp, 2013). Also, the determination of the last and first cycles of voicing offset and onset is subjective. The algorithm was developed to automate this process so that RFF can be estimated fast and objectively. The first step of the algorithm, identifying fricatives in the signal, involves user confirmation and correction, making this algorithm 'semi-automated.' This algorithm was validated with recordings from 154 individuals with voice disorders and 36 individuals with healthy voices. The correlation between manual and the semi-automated algorithm was high (r = 0.82 to 0.91). The findings of this study additionally suggested that RFF can be estimated from dysphonic voices with reduced periodicity. Only a small percentage of the voice samples did not have sufficient periodicity for RFF analysis in the dataset, in which one-fourth of the samples had CAPE-V overall severity ratings greater than 54.2. Attempts to improve the semi-automated algorithm have continued, and recently Vojtech et al. (2019) implemented a better f_o estimation method for RFF estimation than the f_o estimation method of the previous version, along with a feature to account for the periodicity of the sample when determining the last and first cycles of voicing offset and onset.

Purpose and Research Questions

This dissertation attempted to move forward with the previous contributions to support and optimize RFF as a clinical measure of vocal hyperfunction. Figure 1.11 summarizes key elements of clinical measures that this dissertation aimed to evaluate and outlines the gaps in the research on RFF in these elements. Although RFF's validity of assessing vocal hyperfunction has been indirectly supported by many behavioral studies, the physiological explanations of RFF still lack sufficient scientific evidence. The emerging convergent validity of RFF, which suggests a correlation between RFF and perceived strain, is also not yet clear from the previous experiments since those studies involved changes in other acoustic features of the voice samples in addition to RFF (Lien et al., 2015b; McKenna & Stepp, 2018; Stepp et al., 2012). Therefore, it is not yet known whether RFF is perceived by listeners in a direct relation to strain or if RFF is not perceived but just reflects underlying vocal hyperfunction.

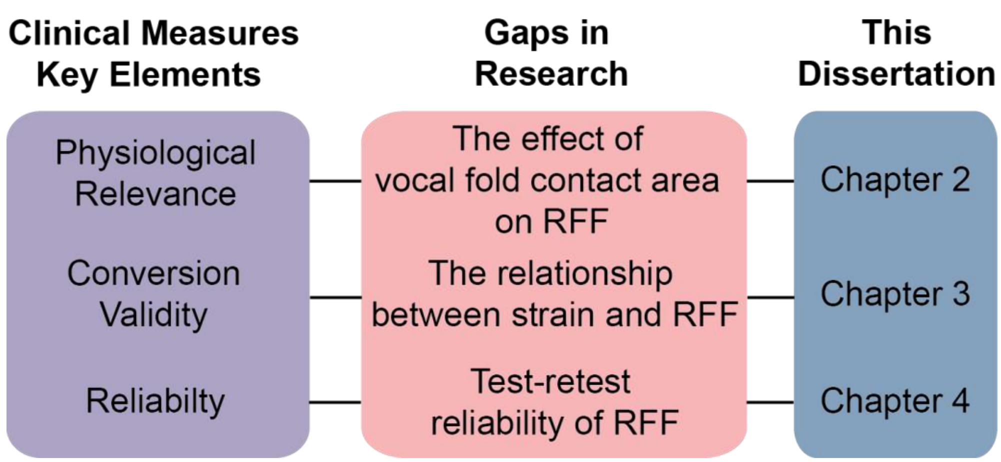

Figure 1.11: An organizational chart that summarizes the gaps in research for RFF to be applied clinically that this dissertation evaluated.

In addition to further supporting the validity of RFF, clinical measures must demonstrate appropriate reliability. The test-retest reliability of the current RFF estimation protocol must be evaluated since RFF stimuli and the estimation method have been optimized to be more reliable, faster, and objective. Thus, this dissertation includes three studies that further supported RFF's validity and reliability: 1) a high-speed videoendoscopy (HSV) study examining vocal fold kinematics and RFF as a function of obstruent type and speaker's age; 2) a perceptual study evaluating the effect of RFF synthetic modification; and 3) a test-retest reliability study of RFF.

Vocal Fold Kinematics and RFF

Chapter 2 of the dissertation aimed to provide empirical evidence for the physiological factors underlying RFF. Watson (1998) hypothesized that abduction may lead to decreases in RFF during voicing offset due to decreases in the vocal fold contact area as the vocal folds pull apart. We evaluated this hypothesis from Watson (1998) by examining the vocal fold kinematics along with RFF during the productions of voiceless fricatives and stops in younger and older speakers.

Figure 1.12 summarizes previous findings on possible differences in vocal fold kinematics between fricatives and stops and our hypotheses. For voicing offset prior to obstruents, stops require a complete oral constriction unlike fricatives, which may stop the airflow and thus make devoicing easier for stop production relative to fricative production. Thus, less devoicing strategy may be required in stop production, which was also suggested by previous studies (McGarr & Löfqvist, 1988; Yoshioka, Löfqvist, & Collier, 1982), and the vocal folds may be less abducted during voicing offset for stops relative to fricatives. For voicing onset after obstruents, stops are produced with burst, which can result in instantaneous increase in airflow rate (Löfqvist et al., 1995). This increased airflow rate may initiate vocal fold vibration when the vocal folds are farther apart, less adducted, resulting in larger glottal angle during voicing onset (Hottinger, Tao, & Jiang, 2007). We also suspected that higher airflow rate from the burst in stop production may require stiffer vocal folds in order to maintain airflow rates compatible with typical human phonation (Collier, Lisker, Hirose, & Ushijima, 1979).

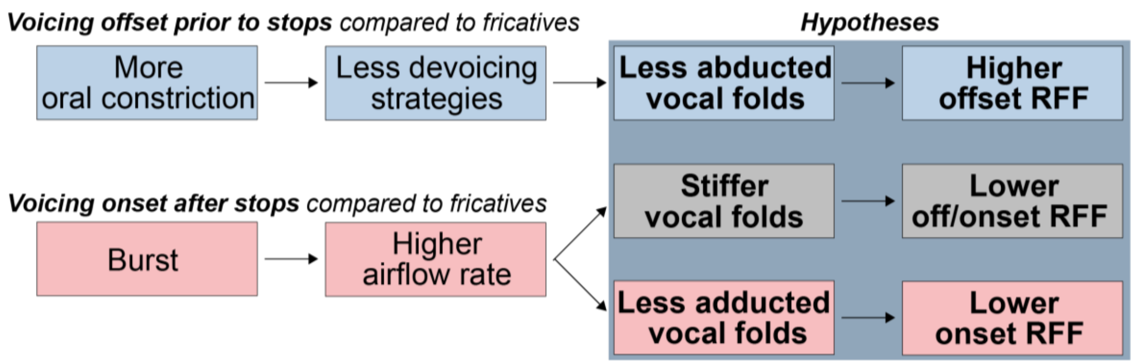

Figure 1.12: A flow chart that summarizes the differences in voicing offset and onset during intervocalic stop production relative to fricative production. Hypotheses relating to the differences in vocal fold kinematics and RFF between stops and fricatives are inside the shaded box. Blue represents voicing offset, pink represents voicing onset, and gray represents both voicing offset and onset.

We examined the differences in vocal fold kinematics and RFF between younger and older adults. Older adults are known to have increased vocal fold stiffness (Sato & Hirano, 1997; Sato, Hirano, & Nakashima, 2002), often have vocal fold atrophy or bowing (Honjo & Isshiki, 1980), and decreased type II muscle fibers, which contract faster than type I fibers (Rodeno, Sanchez-Fernandez, & Rivera-Pomar, 1993). These age-related morphological and muscular changes may result in stiffer and less adducted vocal folds during voicing transitions in older adults relative to younger adults.

Research question 1.a: How are the differences in vocal fold kinematics between voiceless fricative and stop productions reflected in RFF values?

Hypotheses 1.a: Offset RFF would not differ between the two obstruents due to the antagonistic effects of smaller glottal angles at voicing offset (increasing RFF) and stiffer vocal folds (decreasing RFF) in voiceless stops. Onset RFF would be lower in voiceless stops than in voiceless fricatives due to greater glottal angles at voicing onset and stiffer vocal folds acting together to decrease RFF during voicing onset in voiceless stops.

Research question 1.b: How are the differences in vocal fold kinematics between RFF stimuli productions of younger and older adults reflected in their RFF values?

Hypotheses 1.b: Older adults would produce lower offset and onset RFF values compared to younger adults due to age-related morphological and muscular changes, which may result in greater glottal angles and stiffer vocal folds at both voicing offset and onset.

Strain and RFF

Chapter 3 attempted to evaluate the direct contribution of RFF to the perception of strain. The previous perceptual studies of RFF cannot support a direct relationship between RFF and strain since other acoustic parameters in the recorded voice samples were also different (Figure 1.13); specifically, participants had different voice qualities (Eadie & Stepp, 2013; Stepp et al., 2012). Thus, we decided to synthetically modify only RFF in the voice samples to isolate its effect on strain.

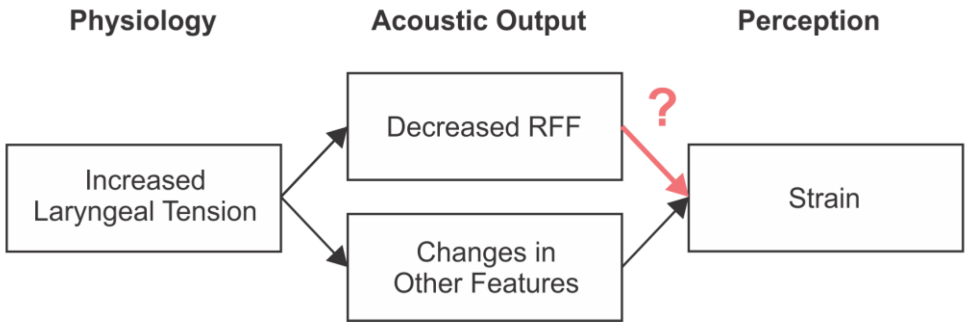

Figure 1.12: A flow chart that summarizes the research question of Chapter 3, the relationship between RFF and strain (a pink arrow with a question mark).

Research question 2: Would listeners perceive changes in strain in the voice samples in which RFF alone is only modified?

Hypothesis 2: Synthetically lowering RFF would result in an increase in the perception of strain and that synthetically raising RFF would result in a decrease in the perception of strain.

Test-Retest Reliability of RFF

Chapter 4 evaluated the test-retest reliability of RFF in comparison to conventional voice measures as appropriate test-retest reliability is required for clinical measures.

Research question 3: How does the test-retest reliability of RFF compare to the test-retest reliability of conventional voice measures?

Hypothesis 3: The test-retest reliability RFF would be comparable or slightly lower than conventional measures since RFF is thought to reflect a functional aspect of vocal production, whereas many conventional measures often reflect structural changes of the vocal folds.

Chapter 2: Vocal Fold Kinematics and Relative Fundamental Frequency as a Function of Obstruent Type and Speaker Age

Abstract

The acoustic measure, relative fundamental frequency (RFF), has been proposed as an objective metric for assessing vocal hyperfunction; however, its underlying physiological mechanisms have not yet been fully characterized. This study aimed to characterize the relationship between RFF and vocal fold kinematics. Simultaneous acoustic and high-speed videoendoscopic (HSV) recordings were collected as younger and older speakers repeated the utterances /ifi/ and /iti/. RFF values at voicing offsets and onsets surrounding the obstruents were estimated from acoustic recordings, whereas glottal angles, durations of voicing offset and onset, and a kinematic estimate of laryngeal stiffness (KS) were obtained from HSV images. No differences were found between younger and older speakers for any measure. RFF did not differ between the two obstruents at voicing offset; however, fricatives necessitated larger glottal angles and longer durations to devoice. RFF values were lower and glottal angle values were greater for stops relative to fricatives at voicing onset. KS values were greater in stops relative to fricatives. The less adducted, stiffer vocal folds and lower RFF at voicing onset for stops relative to fricatives in this study were in accordance with prior speculations that decreased vocal fold contact area and increased vocal fold stiffness may decrease RFF.

Introduction

Relative fundamental frequency (RFF) is an acoustic measure that quantifies instantaneous changes in fundamental frequency (f_o) during the transition into and out of a voiceless consonant (e.g., in a voiced sonorant–voiceless consonant–voiced sonorant, or VCV, utterance; see Figure 2.1). RFF is calculated by comparing the instantaneous f_o values of the ten voicing cycles immediately before and after to the voiceless consonant to a steady-state value. In this way, RFF examines changes in f_o as a speaker terminates (voiced sonorant into voiceless consonant; "voicing offset") and re-initiates (voiceless consonant into voiced sonorant; "voicing onset") phonation. Because RFF captures changes in instantaneous f_o, resulting RFF values reflect changes in the vibratory rate of the vocal folds during voicing offsets and onsets.

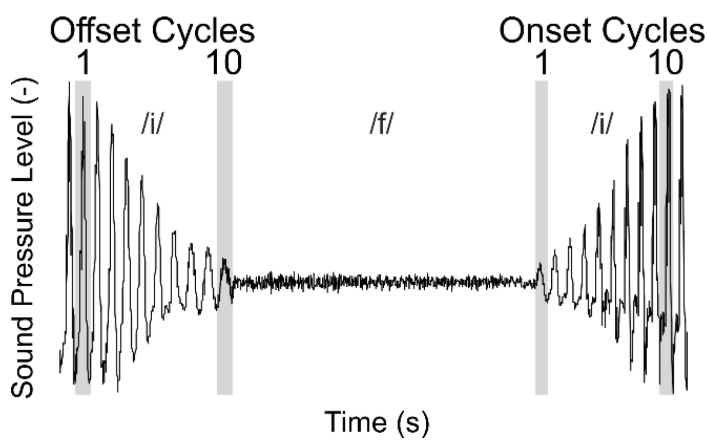

Figure 2.1: An example acoustic waveform of a voiced sonorant-voiceless consonant-voiced sonorant, from which RFF is calculated. The offset cycles and onset cycles 1 and 10 are labeled and shaded in the waveform.

RFF shows promise as a non-invasive, objective measure for assessing vocal hyperfunction (Roy et al., 2016; Stepp et al., 2010a; Stepp et al., 2011b). Vocal hyperfunction is a common feature of voice disorders, and is described as excessive or imbalanced laryngeal muscle forces due to daily vocal overuse and/or misuse (Hillman et al., 1989). Vocal hyperfunction may also present as increased vocal effort (i.e., perceived exertion of a vocalist to a perceived communication scenario; Hunter, Cantor-Cutiva, van Leer, van Mersbergen, Nanjundeswaran, Bottalico, Sandage, & Whitling, 2020) and (para)laryngeal stiffness while speaking (Morrison, 1997). Previous studies have shown that individuals with vocal hyperfunction produce lower average RFF values than those of typical speakers (Roy et al., 2016; Stepp et al., 2010a; Stepp et al., 2011b), implicating a possible relationship between laryngeal muscle tension and RFF. Indeed, prior work indicates a relationship between RFF and listener perceptions of vocal effort (Lien et al., 2015b; McKenna & Stepp, 2018; Stepp et al., 2012), and between RFF and the degree of laryngeal stiffness when estimated via vocal fold kinematics (McKenna et al., 2016). Taken together, these studies support the potential of RFF to reflect increased vocal effort and laryngeal muscle tension in individuals with vocal hyperfunction.

Physiological Mechanisms of RFF

Despite the potential for RFF as an objective tool for assessing vocal hyperfunction, the underlying physiological mechanisms are still unclear. Studies in typical speakers have shown that RFF values are generally increased in voicing cycles closest to the voiceless consonant and then decrease toward zero during voicing onset in typical speakers (House & Fairbanks, 1953; Stepp et al., 2010a; Watson, 1998). In other words, voice f_o is generally increased from steady-state values at the start of voicing onset, but normalizes to steady-state over time; this increase in f_o (Hunter et al., 2020)in voicing cycles near the voiceless obstruent may be the result of increased longitudinal tension in the vocal folds from activation of the cricothyroid (CT) muscle. Specifically, Stevens (1977) postulated that increased longitudinal vocal fold tension might aid in the devoicing needed to produce intervocalic voiceless consonants. It is possible that this tension is carried over to the first few cycles of voicing after voiceless consonants (Stevens, 1977), which would increase the f_o values—and thus RFF values—of these cycles. Indeed, contributions from the CT muscle in devoicing have been observed via laryngeal electromyographic experiments, in which CT muscle activity increased both before and during the production of voiceless consonants in VCV utterances (Löfqvist et al., 1989). In contrast, CT muscle activity did not increase during this period when the same speakers produced *voiced* consonants (Löfqvist et al., 1989). The decreased RFF values observed in individuals with vocal hyperfunction (compared to healthy speakers) are thought to be a result of increased levels of baseline laryngeal tension, which may reduce the effect of increased CT muscle activity on RFF (Heller Murray et al., 2017; Stepp et al., 2010a).

In conjunction with increased CT activity, the vocal folds abduct to assist in devoicing during voicing offset. Vocal fold abduction has been hypothesized to lead to decreases in RFF during voicing offset due to decreases in the vocal fold contact area as the vocal folds pull apart (Watson, 1998). Watson (1998) postulated that decreased vocal fold contact area might result in decreased recoil forces as well as less abrupt closure of the vocal folds. The decreased recoil forces and less abrupt vocal fold closure may lead to increases in contacting and decontacting phases of the vibratory cycle, which ultimately increase the voicing cycle period, and thus decrease f_o (Watson, 1998). Abduction, along with increased CT activity for devoicing, is thought to produce the relatively stable or slightly decreasing offset RFF values observed in young adults with healthy voices (Stepp et al., 2010a; Watson, 1998). Although the hypothesis describing the effect of decreased vocal fold contact area on RFF (Watson, 1998) is theoretically reasonable, it has not yet been supported with empirical evidence.

Relationship between RFF and Vocal Fold Kinematics

In this study, we sought to evaluate the hypothesis from Watson (1998) by examining differences in RFF and vocal fold kinematics between two intervocalic voiceless obstruents: fricatives and stops. Prior work examining voiceless obstruents within VCV productions shows that fricatives tend to result in higher RFF values at voicing onset relative to stops (Lien et al., 2014; Roy et al., 2016). The effects of intervocalic obstruents on RFF values at voicing offset are less clear: Lien et al. (2014) found fricatives to lead to higher RFF values than stops at voicing offset, whereas Roy et al. (2016) found stops to result in higher RFF values at voicing offset. We postulate that these conflicting findings may be due to differences in vocal fold kinematics during intervocalic obstruent production. Specifically, the degree of abduction and adduction are thought to differ when surrounding a voiceless fricative versus a voiceless stop (Löfqvist et al., 1995; McGarr & Löfqvist, 1988), but there has been no direct evidence for differences via vocal fold kinematic features. In the following section, we summarize previous findings on possible differences in vocal fold kinematics between fricatives and stops that motivated our research questions and hypotheses.

Glottal angles and duration of voicing transition

Two kinematic features thought to differ between voiceless fricatives and stops are glottal angle and the duration of voicing transition (i.e., duration of voicing offset or onset). Discrepancies in these features are likely the result of differences in the mechanisms necessary to devoice and reinitiate voicing when producing an intervocalic fricative rather than a stop. For instance, voiceless fricatives are produced via a partial constriction—unlike voiceless stops, which require a full oral constriction—and may require additional strategies to devoice. Continuous airflow in the absence of a complete oral constriction may keep the vocal folds vibrating longer, resulting in a greater duration of voicing offset. One strategy that may be used during intervocalic fricative production is an increased reliance on vocal fold abduction to pull the vocal folds apart and terminate vocal fold vibration: an increased reliance on vocal fold abduction may lead to a smaller vocal fold contact area, and as a result, a larger abductory angle. Indeed, voiceless fricatives have been shown to have greater maximum abductory angles than voiceless stops (McGarr & Löfqvist, 1988; Yoshioka et al., 1982). Overall, these results suggest that the transition from voiced sonorant to voiceless fricative is marked by larger glottal angles and a longer duration of voicing offset than the transition from voiced sonorant to voiceless stop. The increase in voicing offset may also decrease RFF since more vibratory cycles would be affected by the decreased vocal fold contact area; thus, longer voicing offset before fricatives than stops may be related to lower offset RFF for stops relative to fricatives.

In contrast to these features during voicing offset, it is thought that voiceless stops require a larger adductory angle and longer duration of voicing onset than voiceless fricatives. Specifically, previous work has shown that voiced sonorants are produced with larger glottal flow rates and open quotients when following a voiceless stop rather than a voiceless fricative (Löfqvist et al., 1995). These findings suggest that voicing onset occurred when the vocal folds were at a larger glottal angle, which may be the result of the instantaneous increase in airflow from the burst. This increased airflow rate may initiate vocal fold vibration when the vocal folds are farther apart (i.e., larger glottal angle), as was observed in a previous study examining *ex vivo* human larynges (Hottinger et al., 2007). Because the vocal folds are at a larger glottal angle at the start of phonation, it is possible that voicing onset necessitates a longer duration when following a voiceless stop rather than a fricative. As a result, it is thought that glottal angle and duration of voicing onset are larger when transitioning into a voiceless stop rather than a voiceless fricative; thus, RFF would be lower at voicing onset after a stop than after a fricative.

Vocal fold stiffness

In addition to glottal angle and duration of voicing transition, the kinematic estimate of laryngeal stiffness (KS) may also be used to elucidate differences in vocal fold kinematics between voiceless fricatives and stops. Previous studies used KS to indirectly estimate vocal fold stiffness during adductory gestures (Dailey, Kobler, Hillman, Tangrom, Thananart, Mauri, & Zeitels, 2005; Stepp et al., 2010b), and KS has been shown to positively correlate with stiffness values of the thyroarytenoid (TA), lateral cricoarytenoid (LCA), and posterior cricoarytenoid (PCA) muscles in a simple virtual trajectory model of vocal fold kinematics (Stepp et al., 2010b). KS has also been shown to negatively correlate with voicing offset RFF values when typical speakers were instructed to modulate their vocal effort (McKenna et al., 2016). These findings suggest that KS may be a useful tool for examining differences in vocal fold stiffness during the transition into and out of voiceless fricatives and stops.

Prior work suggests that voiceless fricatives may be produced with lower KS values than are voiceless stops. Collier et al. (1979) used hooked-wire electromyography to investigate the TA and LCA muscles during intervocalic fricative and stop productions. The authors determined that the intervocalic production of an /f/ resulted in a higher degree of reduction in the TA and LCA muscle activity before and during the consonant than an intervocalic /t/. It is possible that lower activation levels in these muscles during fricative productions relative to stop productions may not be a devoicing strategy, however, as similar TA activity was observed before intervocalic /v/ productions (Collier et al., 1979). Based on this finding, we suspect that the degree of reduction in the TA muscle activity — which would decrease laryngeal stiffness—might be one of the physiological mechanisms that differ between fricatives and stops. Intervocalic, voiceless stop production may result in greater KS values since higher TA activation is necessary to produce a stop than a fricative (Collier et al., 1979). The reduction in TA muscle activity at voicing transitions may also decrease RFF values because the TA muscle is likely to diminish the f_o-raising effect of the CT muscle, which may also assist in devoicing. TA muscles have shown to have an antagonistic effect on CT muscle in f_o control (Chhetri et al., 2012). Lower KS for stops relative to fricatives may be related to lower RFF at both voicing offset and onset for stops relative to fricatives since the reduction of the TA muscle activity was observed to occur before and during the obstruent production (Collier et al., 1979).

Effects of age on RFF and vocal fold kinematics

Age-related differences in voice production occur as a result of age-related morphological and muscular changes in the vocal folds (Honjo & Isshiki, 1980; Rodeno et al., 1993; Sato & Hirano, 1997; Sato et al., 2002; Watson, 1998). Older adults often exhibit stiffer vocal folds (Sato & Hirano, 1997; Sato et al., 2002) and may produce voice with less adducted vocal folds due to vocal fold atrophy or bowing (Honjo & Isshiki, 1980). Moreover, older adults have been reported to produce lower RFF values than younger adults, perhaps because of a greater reliance on the abductory gesture to assist in devoicing, as suggested by Watson (1998).

Purpose of the Current Study

The current study aimed to use simultaneous acoustic and high-speed videoendoscopic (HSV) recordings to determine the relationship between vocal fold kinematics and RFF of voiceless obstruents. Acoustic signals were captured using a microphone for use in manually calculating RFF. HSV images were used to examine vocal fold movement, from which glottal angles at voicing offset (Θ_{off}) and onset (Θ_{on}) as well as durations of voicing offset (d_{off}; from the start of abduction to voicing offset) and onset (d_{on}; from voicing onset to the end of adduction) were extracted. These images were also used to compute KS, which was estimated as the ratio of the maximum adductory velocity during adduction to the displacement of the glottal angle during the adductory gesture (McKenna, Diaz-Cadiz, Shembel, Enos, & Stepp, 2019).

Our hypotheses are described in detail below, as well as schematized via glottal angle waveforms in Figure 2.2. With respect to vocal fold kinematics, we hypothesized that, relative to fricatives: 1) Θ_{off} and d_{off} would be lower in voiceless stops due to a lesser reliance on abduction; 2) Θ_{on} and d_{on} would be greater in voiceless stops due to a higher airflow rate from the burst; and 3) KS would be greater in voiceless stops due to higher TA muscle activity. With respect to changes in vocal fold kinematics with age, we hypothesized that, compared to younger adults, older adults would exhibit: 1) greater Θ_{off} and d_{off} due to an increased reliance on the abductory gesture to devoice; 2) greater Θ_{on} and d_{on} due to aged-related morphological and muscular changes in the vocal folds; and 3) greater KS due to stiffer vocal folds. We further hypothesized that RFF at voicing offset (offset RFF) would not differ between the two obstruents due to the antagonistic effects of smaller glottal angle to increase RFF and stiffer vocal folds to decrease RFF in voiceless stops, whereas RFF at voicing onset (onset RFF) would be lower in voiceless stops than in voiceless fricatives due to greater glottal angles and stiffer vocal folds to decrease RFF during voicing onset. We also hypothesize that, compared to younger adults, older adults would produce lower offset and onset RFF values due to greater glottal angles and stiffer vocal folds at both voicing offset and onset.

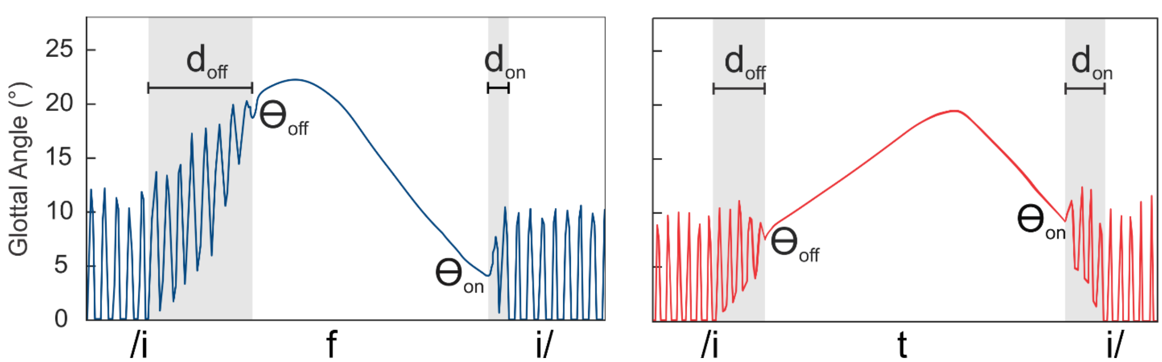

Figure 2.2: Schematics of hypothesized glottal angle waveforms extracted during /ifi/ and /iti/ VCV productions (abbreviations: Θ_{off} and Θ_{on} = glottal angles at voicing offset and onset; d_{off} and d_{on} = durations of voicing offset and onset).

Method

Participants

Twenty typical speakers were recruited to participate in the study, including 10 younger adults (5 females, 5 males; mean age = 22.7 years, range = 19 – 26 years) and 10 older adults (5 females, 5 males; mean age = 62.1 years, range = 53 – 76 years). Participants were native English speakers, were non-smokers, and had no history of speech, language, hearing, or voice disorders. All participants were screened for normal vocal function by a certified speech-language pathologist. Speakers provided written consent prior to participation in compliance with the Boston University Institutional Review Board.

Recording Procedures

Participants were trained to produce iterations of the utterances /ifi/ and /iti/. Each /ifi/ and /iti/ set consisted of eight consecutive productions with a pause in the middle (e.g., /ifi ifi ifi ifi/, pause, /ifi ifi ifi ifi/). For this study, /ifi/ was selected to examine the effects of an intervocalic, voiceless fricative (i.e., /f/) on vocal fold kinematics and resulting RFF values. Similarly, /iti/ was selected to examine the effects of an intervocalic, voiceless stop (i.e., /t/). Participants were instructed to produce these stimuli with equal stress on both vowels using a comfortable voice (Park & Stepp, 2019).

Prior to recording, a directional headset microphone (SM35 XLR, Shure) was placed 7 cm away from the corner of the mouth at a 45-degree angle from the midline. A speech-language pathologist trained to perform transnasal endoscopy examination visualized participants’ larynges using a flexible nasoendoscope. A pediatric endoscope (Pentax, Model FNL-7RP3, 2.4 mm) was used for 14 of the participants and an adult nasoendoscope (Pentax, Model FNL-10RP3, 3.5 mm) was used on the remainder based on the clearance of nasal passages and their tolerance to the endoscope. All examinations were performed in a sound-treated room at Boston University. A FASTCAM Mini AX100 camera (Model 540K-C-16GB), operated at 256 × 256 pixel resolution and at a frame rate of 1000 frames/sec, was attached to the endoscope by a 40-mm optical lens adaptor along with a steady xenon light source (300 W KayPentax Model 7162B) for high-speed visualization. The video images were obtained through Photron Fastcam Viewer software (Version 3.6.6). Although the endoscopy procedure could cause discomfort, a numbing agent was not provided, due to its effect on laryngeal sensory feedback. However, the participants were presented with the option of using a nasal decongestant. After a clear view of the larynx was obtained, the participants produced one set of eight /ifi/s followed by one set of eight /iti/s. If the participant produced stressed or glottalized vowels or if the speech-language pathologist failed to acquire clear larynx images, participants were instructed to repeat the set. The acoustic signals were recorded, amplified with Xenyx Behringer 802 Preamplifier, and digitalized at 30 kHz using a data acquisition board (DAQ; National Instruments 6312 USB) to time-synchronize acoustic recordings with the HSV images.

Data Analysis

HSV image analysis

A series of kinematic measures were extracted semi-automatically from HSV images using a graphical user interface described in Diaz-Cadiz, McKenna, Vojtech, and Stepp (2019). In brief, the algorithm uses differences in pixel intensities between the glottis and the vocal folds to estimate the glottal angle during laryngeal articulatory and/or vibratory movements as a function of time. Figure 2.3 presents a schematic of a glottal angle waveform from an /ifi/ utterance. Within the graphical user interface, a single technician examined the glottal angle waveform in order to locate a series of pertinent time points necessary to estimate vocal fold kinematics, as described below.

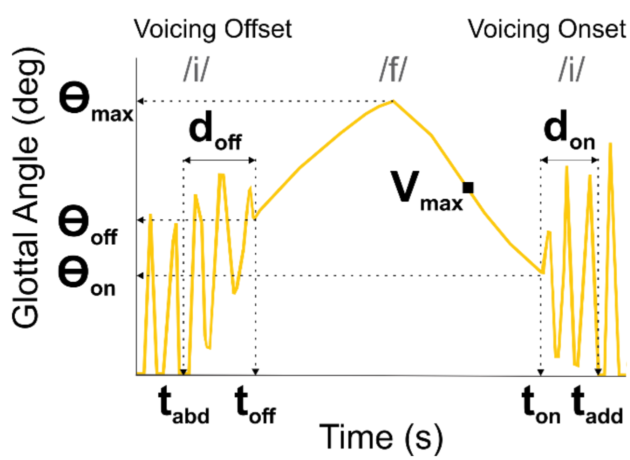

Figure 2.3: A schematized glottal angle waveform during a production of /ifi/ surrounding the obstruent (abbreviations: Θ_{max} = maximum glottal angle; Θ_{off} and Θ_{on} = glottal angles at voicing offset and onset; d_{off} and d_{on} = durations of voicing offset and onset; t_{abd} = start of abduction; t_{off} = voicing offset; t_{on} = voicing onset; t_{add} = end of adduction).

Two time points were extracted through manual examination of the glottal angle waveform and HSV images during devoicing: the start of abduction (t_{abd}) and the time of voicing offset (t_{off}). The technician marked t_{abd} as the last complete contact of the vocal folds during the voicing offset. If the vocal folds had never reached full closure during the preceding vowel, t_{abd} was marked as the time at which the last maximum closure of the vocal folds occurred before the glottal angle started to increase for abduction (18.3%). If arytenoid cartilages blocked the view of the vocal folds, t_{abd} was marked as the time at which the two arytenoid cartilages started to move apart from one another (14.1%). The location of t_{off} was considered as the termination of the last vibratory cycle prior to the obstruent. Two time points were then identified during the reinitiation of voicing: the time of voicing onset (t_{on}) and the termination of adduction (t_{add}). The location of t_{on} was marked as the time point in which the vocal folds first started to vibrate after the obstruent. The technician marked t_{add} at the time point in which the first full or maximum (22.1%) vocal fold closure during voicing onset was achieved, or, if the view of the vocal folds was blocked, when the arytenoid cartilages stopped moving toward each other (18.6%).

After locating these time points using the glottal angle waveform, the technician corroborated the indices via manual visualization of the raw HSV images; this method of verification was performed in order to minimize errors that may occur in time point identification if the glottal angle waveform failed to capture small glottal gaps during closed phases of vibratory cycles. From here, four estimates of vocal fold kinematics were obtained:

1. Θ_{off}: Glottal angle at voicing offset, extracted at t_{off}
2. d_{off}: Duration of voicing offset, calculated as $t_{off} - t_{abd}$
3. Θ_{on}: Glottal angle at voicing onset, extracted at t_{on}
4. d_{on}: Duration of voicing offset, calculated as $t_{add} - t_{on}$

Finally, a kinematic estimate of laryngeal stiffness, KS, was calculated as the ratio of the maximum adductory velocity during the adductory gesture prior to voicing onset (V_{max}) to the maximum glottal angle (or in case of incomplete vocal fold closure after voicing onset, the maximum glottal angle minus the minimum glottal angle after voicing onset; Θ_{max}), which represents the displacement of the glottal angle during the adductory gesture.

In sum, five vocal fold kinematic measures were obtained from this analysis: Θ_{off}, Θ_{on}, d_{off}, $d_{on,}$ and KS. These measures were averaged within participant for all /ifi/ and /iti/ productions to enable comparisons across participants in terms of obstruent type (i.e., /f/ or /t/) and age group (i.e., younger or older adult).

Manual RFF estimation

A single trained technician[1] carried out manual RFF estimation using Praat software (Version 6.0.21). The technician first examined each /ifi/ and /iti/ waveform to determine whether there was evidence of voicing during the associated obstruent. If there was clear evidence of voicing, such as glottal pulses present during the obstruent, the RFF production was immediately rejected (2.7%). Otherwise, the technician proceeded with RFF computation, as follows: (1) the boundary between voiced and voiceless speech segments were identified for both voicing offset and onset, (2) the voicing cycles closest to the voice offset boundary (offset cycle 10) and voice onset boundary (onset cycle 1) were located (Figure 2.1), (3) the nine voicing cycles before voice offset cycle 10 and after voice onset cycle 1 were selected via examining the general waveform shape of the vibratory cycles, (4) the instantaneous f_o of each voicing cycle was calculated as the inverse of cycle period, (5) RFF (semitones; ST) was calculated using Eq. 1:

[1] The dataset used to train individuals in manual RFF estimation is a separate dataset from that described here, and may be downloaded from: https://sites.bu.edu/stepplab/research/rff/.

$$RFF\ (ST) = 12 \cdot \log_2\left(\frac{f_o}{f_{ref}}\right) \qquad [1]$$

In Eq. 1, f_{ref} refers to the f_o of the cycles located furthest from the obstruent, assumed to be part of steady-state voicing. Specifically, the f_o of offset cycle 1 was used as f_{ref} for offset cycles and the f_o of onset cycle 10 was used as f_{ref} for onset cycles (Stepp et al., 2010a).

RFF values were then examined to determine whether offset or onset instances were valid following criteria set by Vojtech and Heller Murray (2019). If the RFF value closest to the reference cycle (offset cycle 2, onset cycle 9) was larger than 0.8 ST, this specific instance was rejected, was steady-state voicing was not achieved (1.2%). Any RFF instances with evidence of glottalization were also rejected, since accurate estimation of f_o during glottalization is not possible (2.9%). We also excluded RFF instances in which HSV images could not be analyzed due to obstruction of the view of the vocal folds or poor resolution of the images (1.9%). In total, 51 of 581 (8.8%) of offset and/or onset instances were rejected. RFF values from offset cycle 10 (offset 10 RFF) and onset cycle 1 (onset 1 RFF) values were averaged within participants for /ifi/ and /iti/. These values were considered for further examination, as they have been shown to reflect the largest differences between individuals with healthy and hyperfunctional voices (Stepp et al., 2011b).

Reliability

Prior to carrying out statistical analyses, the reliability of the extracted vocal fold kinematic measures and RFF values were assessed. To assess intrarater reliability, the primary technician reanalyzed 20% of HSV and RFF samples in a separate sitting. Interrater reliability was assessed by comparing HSV time markings and RFF measures between the primary technician and an additional technician who carried out HSV image analysis and manual RFF estimation on 20% of samples. Intrarater and interrater reliability were each calculated via two-way mixed-effects intraclass correlation coefficients (ICC) for absolute agreement (single measures).

Reliability measures are presented in Table 2.1. Overall, the reliability of HSV-based time markings were excellent (Portney & Watkins, 2000), with intrarater reliability reaching a mean ICC = 0.98 (SD = 0.02, range = 0.96 – 0.99) and interrater reliability averaging at ICC = 0.99 (SD = 0.01, range = 0.98 – 0.99). Intrarater reliability of RFF markings was good (Portney & Watkins, 2000), with a mean intrarater reliability of ICC = 0.87 and interrater reliability of ICC = 0.76.

Table 2.1 Intra- and interrater reliability values (Intraclass correlation coefficients; single measures, absolute agreement, a 2-way mixed-effects model) of all measures (95% CI in parentheses)

Measures	Intrarater Reliability	Interrater Reliability
RFF	0.87 (0.81 – 0.91)	0.76 (0.66 – 0.83)
t_{abd}	0.96 (0.89 – 0.98)	0.98 (0.96 – 0.99)
t_{off}	0.99 (0.99 – 0.99)	0.99 (0.90 – 0.99)
t_{on}	0.99 (0.99 – 0.99)	0.99 (0.99 – 0.99)
t_{add}	0.99 (0.99 – 0.99)	0.99 (0.98 – 0.99)

Statistical Analysis

Wilcoxon signed-rank tests were performed on RFF (offset 10, onset 1) and vocal fold kinematic measures (Θ_{off}, Θ_{on}, d_{off}, $d_{on,}$ and KS) to evaluate hypothesized differences in these measures between /ifi/ and /iti/ within participants. To assess within-participant relationships between RFF and kinematic variables we used repeated-measures correlations (Bakdash & Marusich, 2017), as previous studies have reported that correlations between RFF and other measures are stronger within participants than across participants (Lien et al., 2015b; McKenna et al., 2016). Mann-Whitney U tests were then performed to compare group differences in RFF and vocal fold kinematics between younger and older adults. Resulting *p* values were adjusted using the Bonferroni correction to account for 20 tests (7 Wilcoxon signed-rank tests, 6 repeated measures correlations, 7 Mann-Whitney U tests; i.e., $p = 0.05/20 = 0.0025$).

Results

Figure 2.4 presents the mean RFF and vocal fold kinematic measures for /ifi/ and /iti/ productions across all participants. Offset 10 RFF values were not statistically significantly different between /ifi/ and /iti/, whereas onset 1 RFF values were statistically significantly greater in /ifi/ productions (p = 0.002). The vocal fold kinematic measures, Θ_{off} and d_{off}, were statistically significantly greater in /ifi/ relative to /iti/ ($p < 0.001$), whereas Θ_{on} and KS were statistically greater in /iti/ than in /ifi/ (p = 0.002). Although average d_{on} values were greater in /iti/ than in /ifi/, this difference did not reach statistical significance (p = 0.007).

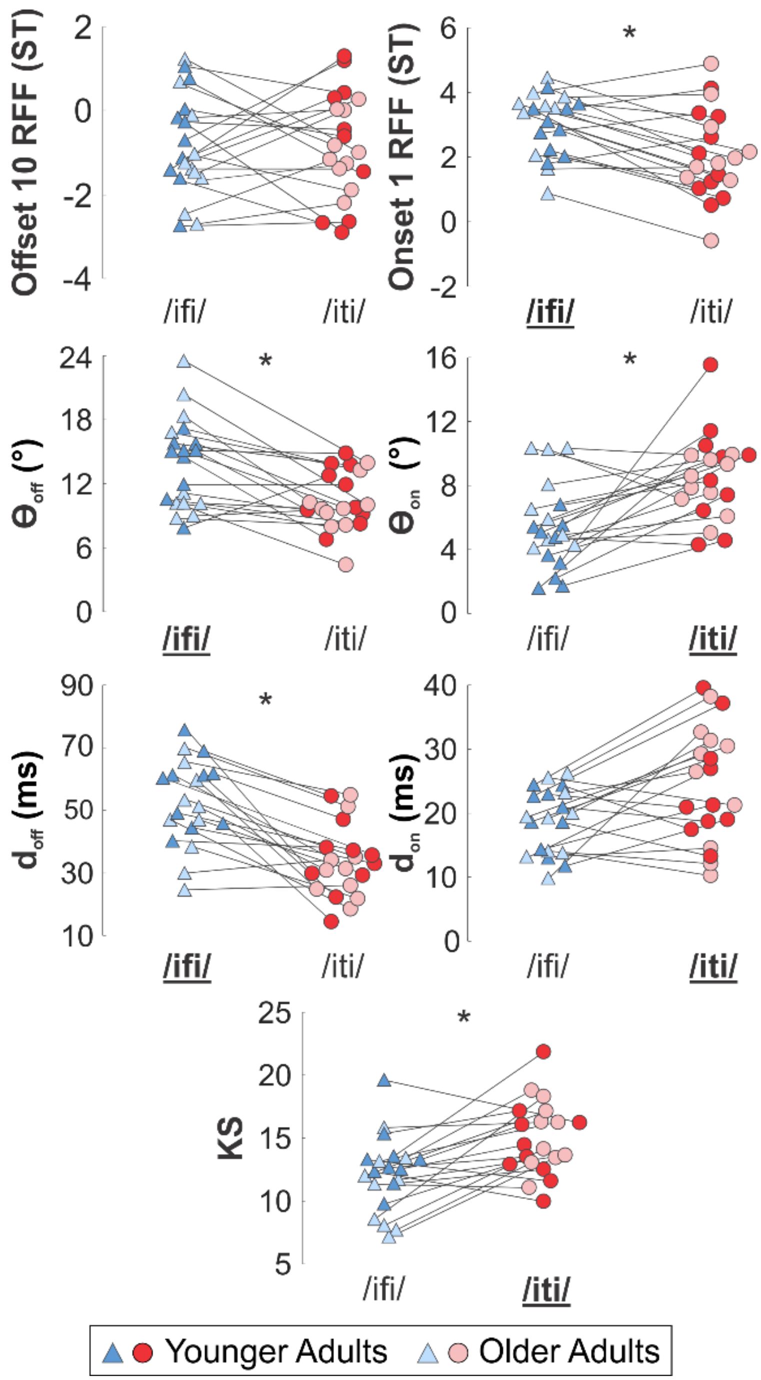

Figure 2.4: RFF and the vocal fold kinematic measures from /ifi/ and /iti/. Lines connect data from each participant. Within each measure, the bolded and underlined /ifi/ or /iti/ on the x-axis represents which utterance was hypothesized to have greater values for the given measure. Asterisks indicate statistical significance at p = 0.0025 (abbreviations: Θ_{off} and Θ_{on} = glottal angles at voicing offset and onset; d_{off} and d_{on} = durations of voicing offset and onset; KS = kinematic estimate of laryngeal stiffness).

Repeated measures correlations were performed to assess possible within-participant associations between RFF and vocal fold kinematic measures. Offset 10 RFF was not statistically significantly correlated with Θ_{off} (r = -0.20, p = 0.19), d_{off} (r = -0.24, p = 0.15), or KS (r = 0.21, p = 0.18). On the other hand, onset 1 RFF was moderately correlated with Θ_{on} (r = -0.46, p = 0.02), d_{on} (r = -0.42, p = 0.03), and KS (r = -0.36, p = 0.05); however, these relationships were not statistically significant.

Mann-Whitney U tests to compare younger and older speakers did not show statistically significant differences for any of the measures (Figure 2.4).

Discussion

The present study sought to examine the relationship between vocal fold kinematics and RFF during voicing offset and onset. To carry out this investigation, we chose two voiceless obstruents, /f/ and /t/, to characterize the effects of intervocalic fricative and stop productions, respectively, on resulting measures. In examining the impact of intervocalic fricatives versus stops on resulting RFF values, we hypothesized that onset RFF would be lower in voiceless stops than fricatives, whereas offset RFF would not statistically differ between the two obstruents. We further hypothesized that older adults would produce lower offset and onset RFF values. In assessing the impacts of intervocalic obstruents on vocal fold kinematics, we hypothesized that average glottal angle and duration at voicing offset would be lower in stops. Conversely, average glottal angle and duration at voicing onset and laryngeal stiffness were thought to be greater in stops. Finally, when compared to younger adults, we hypothesized that older adults would exhibit greater glottal angles, greater voicing onset and offset durations, greater laryngeal stiffness values, and decreased offset and onset RFF values.

RFF and Vocal Fold Kinematics during Intervocalic Fricatives and Stops

Voicing offset

As hypothesized, offset 10 RFF values were not statistically significantly different between /ifi/ and /iti/. These results may be due to the effects of less stiff vocal folds (hypothesized to increase RFF) canceling out the effects of more abducted vocal folds (hypothesized to decrease RFF) at voicing offset prior to fricative production. We suspect that our findings of lower kinematic laryngeal stiffness during /ifi/ may be due to the higher degree of relaxation of the TA muscles observed prior to and during the production of a fricative compared to the degree of relaxation of the TA muscles observed prior to and during the production of a stop consonant (Collier et al., 1979). Less TA activity may increase offset RFF prior to fricative production due to the antagonistic relationship between the TA and CT muscles in f_o control (Chhetri et al., 2012). KS showed a statistically significant, negative correlation with offset 10 RFF values in McKenna et al. (2016); however, we did not observe a statistically significant association between KS and offset 10 RFF. This may be the result of differences in methodology. McKenna et al. (2016) included recordings of speakers modulating their vocal effort level and thus had a wider range of KS (7.3 – 31.9) compared to the range of KS in our study (7.2 – 21.8). Additionally, McKenna et al. (2016) used only /ifi/ for the recording stimulus, whereas we examined two different obstruents in RFF stimuli that could vary in other dimensions (e.g., glottal angle and duration of voicing offset) that may affect RFF.

Greater glottal angles during voicing offset were observed prior to fricative production relative to stop production. This finding is consistent with previous observations of a greater degree of abduction during fricatives than during stops (McGarr & Löfqvist, 1988; Yoshioka et al., 1982). McGarr and Löfqvist (1988)—who examined transillumination images of vocal fold movements during VCV stimuli for both voiceless fricatives and stops in a single healthy speaker—demonstrated that fricatives resulted in higher maximum glottal areas and abduction velocities when compared to stops. Moreover, a hooked-wire electromyographic experiment showed that fricatives led to greater PCA muscle activity than stops (Yoshioka et al., 1982). Although this study provides additional evidence for increased reliance on abduction prior to fricative production (i.e., as the PCA is an abductor muscle), it was also only conducted in a single participant. The results of the current study provide more concrete evidence from twenty speakers that fricative production involves greater vocal fold abduction than stop production.

Greater glottal angles and durations of voicing offset resulting from greater abduction in fricative production were thought to lower offset 10 RFF based on the hypothesis from Watson (1998). However, we also hypothesized that there would be a co-occurring effect of lower KS (hypothesized to increase RFF), thereby canceling out possible RFF-decreasing effects of greater glottal angle and duration of voicing offset. Thus, we hypothesized offset 10 RFF would not be statistically different between the obstruents. Unsurprisingly, we found no statistically significant difference in offset 10 RFF values between the obstruents and no statistically significant correlations between offset 10 RFF values and glottal angle or duration of voicing offset.

Voicing onset

In the current study, onset 1 RFF values were lower and glottal angles were greater at voicing onset for intervocalic stops relative to fricatives. These results are consistent with prior work showing that onset 1 RFF is lower (Lien et al., 2014; Roy et al., 2016) and glottal angles at voicing onset are larger (Löfqvist et al., 1995) for stops relative to fricatives. Furthermore, these findings are consistent with the hypothesis from Watson (1998), which posited a negative relationship between vocal fold contact area and RFF. A larger glottal angle may be the result of higher minimum glottal flow rates, which, in turn, may indicate that vocal fold contact area is smaller when producing the intervocalic stop,

/t/, than the fricative, /f/, within a VCV production. A moderate, negative correlation was also identified between onset 1 RFF and Θ_{on} (r = -0.46, p = 0.02); however, this relationship was not statistically significant. This may have been due to the limited range of Θ_{on} (0 – 15.5°) being produced by individuals with healthy voices. Future studies should therefore examine the relationship between glottal angle and RFF at voicing onset in a wider range of vocal function, including individuals with voice disorders and/or individuals self-modulating their vocal effort.

Based on these results, we suspect that the degree of vocal fold closure at voicing offset and onset surrounding an obstruent may be an important factor that contributes to RFF. Specifically, Watson (1998) suggested that decreased vocal fold contact area in a vibratory cycle would decrease the magnitudes of both recoil forces (thus, opening the vocal folds) and abrupt closure, which would increase the period and thus decrease the frequency of the cycle. Results from prior computational modeling also support this hypothesis: when the vocal folds were positioned at a greater glottal angle prior to phonation (i.e., decreased vocal fold contact area), the experimenters observed decreases in the closed quotient of the vibratory cycle and of the resultant f_o. A decrease in the closed quotient in response to a decrease in vocal fold contact area suggests that the vocal folds may remain closed for a shorter time in the vibratory cycle from slower opening and closing of the vocal folds (Zhang, 2016a). These hypotheses would support the lower onset 1 RFF values observed in the current study.

Lower onset 1 RFF values following intervocalic stops may also be the result of a higher KS observed during stop production relative to fricative production. Lien et al.

(2014) reasoned that onset 1 RFF was higher for fricatives in their study because the TA muscles are more relaxed during fricative production than during stop productions, as shown in Collier et al. (1979). Our result showing lower stiffness during /ifi/ than /iti/ also supports the notion that the TA muscles may be less active during fricative production than during stop production. Although the rationale behind greater TA activity in stop productions is not yet known, we suspect that this may be part of the strategy to maintain typical voice quality after stop production. Specifically, the bursts of air during stop production can inhibit typical phonation by pulling the vocal folds apart; increasing TA activity (and thus, stiffness) may be a useful mechanism to maintain typical phonation. Computational modeling has demonstrated that increasing driving pressure for phonation causes stiffness parameters to increase to maintain airflow rates compatible with human phonation (Zhang, 2015). Greater TA activity during stop production could, in turn, decrease onset RFF since the CT (which aids in devoicing) and TA muscles exhibit antagonistic effects on f_o when activated together (Chhetri et al., 2012). Although we observed both higher kinematic laryngeal stiffness and lower onset 1 RFF during /iti/ productions, we did not find a statistically significant relationship between KS and onset 1 RFF. This finding is similar to the results from McKenna et al. (2016), in which only 40% of the participants showed moderate, negative relationships between KS and onset 1 RFF when the speakers modulated their vocal effort.

RFF and Vocal Fold Kinematics in Younger and Older Adults

We recruited both younger and older adults to evaluate the associations between RFF and the kinematic measures in a wide range of voices, as well as to investigate the differences between younger and older adult voices due to aging. In contrast to our hypotheses, we did not find statistically significant differences between age groups in any of our measures. Due to the observations of Watson (1998), we hypothesized that RFF would be lower in older adults. We further hypothesized that glottal angles, voicing transition durations, and KS would be higher in older adults due to increased difficulty with voicing offset and onset, as well as increased vocal fold stiffness due to age-related changes to the laryngeal system (Honjo & Isshiki, 1980; Rodeno et al., 1993; Sato & Hirano, 1997; Sato et al., 2002; Watson, 1998). The anatomical changes in aging laryngeal systems have also been mirrored by the changes in voice quality and vocal function as a result of age in previous studies (Dehqan, Scherer, Dashti, Ansari-Moghaddam, & Fanaie, 2012; Xue & Deliyski, 2001). A possible reason that we did not find differences as a function of speaker age may be because our older group was younger than the older groups recruited in previous studies. Older adults in these previous studies had an average age of roughly 75 years; however, the age of our older adult participants ranged from 53 to 76 years, with an average of only 62 years. Due to this discrepancy, the anatomical characteristics of our older adult group may be different from those of these prior studies. Futures studies should evaluate older adults aged over 75 years to assess possible differences in RFF and vocal fold kinematic measures.

Implications for Vocal Hyperfunction

Although all of the participants studied had healthy voices, our findings may be used to inform future hypotheses about potential laryngeal mechanisms underlying differences in RFF in individuals with vocal hyperfunction. RFF has been shown to be lower in individuals with vocal hyperfunction than in individuals with typical voices in several studies (Heller Murray et al., 2017; Roy et al., 2016; Stepp et al., 2010a; Stepp et al., 2011b). Yet the original physiological hypothesis about lower RFF in individuals with vocal hyperfunction was based on reduced activity of the CT muscle before and during the voiceless consonant as a result of increased baseline laryngeal tension (Stepp et al., 2010a; Stepp et al., 2011b). Heller Murray et al. (2017) hypothesized that increased baseline laryngeal tension may also result in an increased duration of voicing offset to terminate vocal fold vibration, which was thought to decrease offset RFF by Watson (1998). In line with their hypothesis, we also suspect that if the effect of the activity of the CT muscle on the vocal folds—which was hypothesized to assist in devoicing—is reduced in those with vocal hyperfunction, these individuals may require additional mechanisms to achieve devoicing. An example of this mechanism could be an increased degree of abduction, which was also observed in intervocalic voiceless fricative production (Collier et al., 1979; Yoshioka et al., 1982). Thus, individuals with vocal hyperfunction may have an increased glottal angles and durations of voicing offset, which may decrease their offset RFF. This speculation is consistent with the hypothesis proposed by Watson (1998). A possible difference in the degree of adduction between typical speakers and speakers with vocal hyperfunction may also be a factor in explaining their differences in RFF. Previous studies also suggest that increases in glottal angles and durations of voicing onset may be due to glottal insufficiency in individuals with vocal hyperfunction (Galindo et al., 2017; Woo, 2017). Future HSV studies should directly investigate these hypotheses on voicing offset and onset features of individuals with vocal hyperfunction when compared to healthy voices.

Limitations

One limitation of this study is the set order of /ifi/ and /iti/ productions during HSV. Recording order could affect vocal function if there were a difference in the effect of flexible endoscopy on the participant over time (e.g., due to discomfort). Although psychological stress has been associated with changes in vocal function (Dietrich & Verdolini Abbott, 2012; Helou, Wang, Ashmore, Rosen, & Abbott, 2013), the effect of flexible endoscopy without anesthetic on voice over time has not been studied. Considering the relatively short duration of endoscopy and possible variations in how participants tolerate the endoscope, we suspect that changing the order of recording would not substantially affect our results.

The frame rate used for videoendoscopic recordings is another possible limitation of this study. Our videos were recorded at a frame rate of 1000 frames/sec, which is well above the typical speaking f_o of adult men and women. Yet technicians manually selected t_{abd} and t_{add} on a smaller scale than a complete vibratory cycle, basing these selections on subjective evaluations of HSV images. If the actual initiation or termination of voicing were not captured due to the relatively low frame rate, adjacent cycles would have been selected, potentially leading to less precise measurements for d_{off} and d_{on}. However, the difference we observed in d_{off} between /ifi/ and /iti/ was an average of 19 ms, which is around two to four times greater than typical fundamental periods of speaking voices (5 – 10 ms). Thus, we suspect that the frame rate is unlikely to have affected the results of the study.

Conclusions

In the current study, we examined the differences in RFF and vocal fold kinematics between intervocalic voiceless fricatives and stops. Our results suggest that lower onset RFF values, smaller glottal angles and durations of voicing offset, larger glottal angles at voicing onset, and greater kinematic estimates of laryngeal stiffness are observed in the production of intervocalic stops at compared to fricatives. The lower onset RFF may be related to less adducted and stiffer vocal folds required during stop production. This work indirectly supports prior speculation of a positive relationship between the degree of vocal fold contact area and RFF. Future studies should use computational modeling to examine this hypothesis directly.

Chapter 3: Perceptual and Acoustic Assessment of Strain Using Synthetically Modified Voice Samples

Abstract

Assessment of strained voice quality is difficult due to the weak reliability of auditory-perceptual evaluation and lack of strong acoustic correlates. This study evaluated the contributions of relative fundamental frequency (RFF) and mid-to-high frequency noise to the perception of strain. Stimuli were created using recordings of speakers producing /ifi/ with a comfortable voice and with maximum vocal effort. RFF values of the comfortable voice samples were synthetically lowered and RFF values of the maximum vocal effort samples were synthetically raised. Mid-to-high frequency noise was added to the samples. Twenty listeners rated strain in a visual sort-and-rate task. The effects of RFF modification and added noise on strain were assessed using an analysis of variance; intra- and interrater reliability were compared with and without noise. Lowering RFF in the comfortable voice samples increased their perceived strain, whereas raising RFF in the maximum vocal effort samples decreased their strain. Adding noise increased strain and decreased intra- and interrater reliability relative to samples without added noise. Both RFF and mid-to-high frequency noise contribute to the perception of strain. The presence of dysphonia may decrease the reliability of auditory-perceptual evaluation of strain, which supports the need for complementary objective assessments.

Introduction

The lifetime prevalence of voice disorders in the United States is 30%, with an incidence of 7% (Roy et al., 2005). Voice disorders disrupt an individual's quality of life, affecting both economic and social activities (E. Smith, Verdolini, Gray, Nichols, Lemke, Barkmeier, Dove, & Hoffman, 1996). One of the most common features of voice disorders is vocal hyperfunction (Stemple, Roy, & Klaben, 2014), which comprises approximately 65% of all cases in voice clinics in the United States (Brodnitz, 1966; Ramig & Verdolini, 1998). Vocal hyperfunction involves excessive and/or imbalanced laryngeal and paralaryngeal muscular forces and is often associated with phonotrauma, which can result in organic changes to the vocal folds (Hillman et al., 1989). Vocal hyperfunction is also present without phonotrauma; this type of vocal hyperfunction is usually referred to as muscle tension dysphonia (MTD), which is estimated to comprise 10 – 40% of vocal disorders diagnosed clinically (Roy, 2003).

The current clinical assessment of vocal hyperfunction is primarily based on auditory-perceptual evaluation (Roy et al., 2013). Auditory-perceptual evaluation of voice quality is routinely used in clinical assessment (Carding et al., 2009; De Bodt, Van de Heyning, Wuyts, & Lambrechts, 1996) due to its convenience and efficiency (Kent, 1996). It is currently considered the gold standard for evaluating the severity of voice disorders and the outcome of voice therapy (Oates, 2009); thus, auditory-perceptual evaluation is essential to clinical management of voice disorders. A major auditory-perceptual quality associated with vocal hyperfunction is strain. It is defined as the "perception of excessive vocal effort (hyperfunction)" in a standard clinical tool, the Consensus Auditory-

Perceptual Evaluation of Voice (CAPE-V; Kempster, Gerratt, Verdolini Abbott, Barkmeier-Kraemer, & Hillman, 2009). Because strain is a major perceptual attribute of vocal hyperfunction (Kempster et al., 2009; Morrison, 1997), evaluating strain is important to guide the treatment of individuals with voice disorders related to vocal hyperfunction.

Despite the reliance on it in voice clinics, auditory-perceptual evaluation has low reliability. Because of the inherently subjective nature of perceptual evaluation, highly experienced listeners frequently disagree with one another when rating voice quality (Kreiman et al., 1993). This disagreement seems to affect the evaluation of strain more than other voice qualities such as breathiness and roughness, resulting in particularly poor intra- and interrater reliability (Webb et al., 2004; Zraick et al., 2011). Accurate evaluation of different disordered voice qualities can assist clinicians in choosing the most appropriate therapy technique targeted to an individual (Stemple, 2019), suggesting that better methods of evaluating strain have the potential to improve clinical outcomes.

Instrumental measures are often used to supplement auditory-perceptual ratings, but there is no strong acoustic correlate of strain yet available. The smoothed cepstral peak prominence (CPPS) is a cepstral peak amplitude normalized over the entire background signal amplitude calculated from the smoothed cepstrum. CPPS has been shown to correlate strongly with auditory-perceptual ratings of overall severity of dysphonia and breathiness (Awan & Roy, 2006; Heman-Ackah et al., 2003; Hillenbrand et al., 1994). Cepstral measures related to CPPS have also shown potential for assessing roughness (Awan & Awan, 2020). However, cepstral measures have shown mixed results for strain (Anand et al., 2019; Lowell et al., 2012a; McKenna & Stepp, 2018; Van Stan et al., 2020), and strain has not been strongly correlated with other conventional acoustic measures such as frequency and amplitude perturbation measures (Bhuta, Patrick, & Garnett, 2004).

One potential reason for difficulty with perceptual and acoustic evaluation of strain may be its multidimensionality. It has been shown that breathiness and roughness often accompany strain (Lowell et al., 2012a) and that ratings of strain are likely influenced by other co-occurring voice qualities (Kent, 1996). Thus, to improve the auditory-perceptual and acoustic evaluation of strain, its acoustic factors must be revealed. Three acoustic characteristics related to strain have been suggested previously: increased spectral energy at higher harmonic frequencies (Anand et al., 2019; Bergan, Titze, & Story, 2004; Klatt & Klatt, 1990; Stevens, 1977; Sundberg & Gauffin, 1978), decreased relative fundamental frequency (RFF; Stepp, Hillman, & Heaton, 2010; Stepp, Merchant, Heaton, & Hillman, 2011), and increased mid-to-high frequency noise (Hirano, 1981; Klatt & Klatt, 1990; Lowell et al., 2012a).

Increased Spectral Energy at Higher Harmonic Frequencies

Increased spectral energy at higher harmonic frequencies has been associated with strain (Anand et al., 2019). Increased energy at higher harmonic frequencies has also been associated with a pressed voice quality, which results from phonation with excessively adducted vocal folds, suggesting a similarity between strained and pressed voice qualities (Kreiman, Shue, Chen, Iseli, Gerratt, Neubauer, & Alwan, 2012). Increases in energy at higher harmonic frequencies in synthesized voice samples increased listeners' perceptions of pressed voice quality (Bergan et al., 2004). Thus, the relationship between increased spectral intensity at higher harmonic frequencies and strain is well-supported, both theoretically and empirically.

Decreased RFF

RFF has been proposed as an acoustic feature reflecting increased laryngeal tension and strain (Stepp et al., 2010a; Stepp et al., 2011b). RFF quantifies the short-term variation of fundamental frequency (f_o) in sonorant-voiceless consonant-sonorant productions. It is defined as the instantaneous f_os of the ten voicing offset and onset cycles before and after a voiceless consonant, normalized by the f_os of the cycles furthest from the consonant. Compared to the RFF of individuals with healthy voices, RFF values are lower in individuals thought to have increased laryngeal tension, including those with: vocal hyperfunction (Heller Murray et al., 2017; Roy et al., 2016; Stepp et al., 2010a; Stepp et al., 2012), Parkinson disease (Bowen, Hands, Pradhan, & Stepp, 2013; Goberman & Blomgren, 2008; Stepp, 2013), and adductory laryngeal dystonia (Eadie & Stepp, 2013). Stepp et al. (2010) hypothesized that increased baseline laryngeal tension would decrease the extent of the f_o changes before and after intervocalic voiceless consonant production.

The relationship between RFF and strain has also been evaluated and was found to be moderate in previous auditory-perceptual studies (Eadie & Stepp, 2013; Lien et al., 2015b; McKenna & Stepp, 2018; Stepp et al., 2012). However, it is not clear whether listeners responded specifically to the changes in RFF or other acoustic features that may change in concert with RFF. In addition, changes in RFF values over a period of high voice use did not result in changes in strain perceived by listeners in one study (Heller Murray, Hands, Calabrese, & Stepp, 2016). This suggests that RFF may reflect underlying laryngeal tension that may not necessarily be perceived by listeners. Thus, it is not yet clear whether

RFF is directly perceived by listeners as changes in strain or whether it co-varies with other acoustic features that are perceived by listeners.

Increased Mid-to-High Frequency Noise

Strain has also been described as containing increased mid-to-high frequency noise (Hirano, 1981) and as being often accompanied by perceived breathiness (Lowell et al., 2012a). Breathiness is known to be associated with increased aspiration noise in the mid-to-high frequency range near the third formant (Klatt & Klatt, 1990). However, if aspiration noise interferes with higher harmonic frequencies in a similar frequency range, which may also contribute to strained voice quality, it is unclear how aspiration noise actually affects strain. Kreiman and Gerratt (2012) observed that increases in noise decreased listeners' acuity to changes in the harmonic structure of the voice source. Thus, the presence of noise may also decrease listeners' acuity to the percept caused by RFF, which is also dependent on the ability of the peripheral auditory system to resolve the harmonic structures of the voice source. In summary, aspiration noise may interfere with other acoustic characteristics of strain, and it is thus unclear how aspiration noise may contribute to the perception of strain. In order to study the effect of aspiration noise, mid-to-high frequency noise can be synthetically added to voice samples.

Purpose

In this study, we aimed to understand the contributions of the two acoustic characteristics, RFF and mid-to-high frequency noise, to the auditory-perceptual measure of strain. We used synthesis techniques to precisely control these acoustic features and evaluate their direct associations with strain. The previously observed correlations with natural voice samples cannot fully elucidate the relationships between these acoustic measures and auditory-perception, as other acoustic parameters may also differ across voice samples. By modifying only the acoustic parameters of interest, we aimed to delineate more directly the roles of RFF and mid-to-high frequency noise on the auditory-perceptual evaluation of strain. We did not examine increased spectral energy at higher harmonic frequencies in this study because it has already been examined with synthesized samples and showed a statistically significant association with strain (Bergan et al., 2004).

We hypothesized that synthetically lowering RFF would result in an increase in the perception of strain and that synthetically raising RFF would result in a decrease in the perception of strain. We also hypothesized that adding mid-to-high frequency noise to speech samples would increase the perception of strain. We further hypothesized that adding noise would result in decreases in both intra- and interrater reliability of strain ratings, since noise may interfere with other acoustic characteristics of strain.

Methods

Original Voice Recordings

Voice samples of eight individuals (4 females; mean age = 32.6 years, range = 18 – 67 years) with healthy voices were selected from a database of participant recordings of RFF stimuli, /ifi/. These recordings were collected from speakers who were asked to increase their vocal effort to mild, moderate, and maximum levels. The speakers were given the instruction, “produce your voice as if you are trying to push out the air without increasing the loudness,” and the experimenter provided demonstrations of different vocal effort levels. Recordings of comfortable voice and maximum vocal effort conditions were used because they were expected to show the largest differences in strain and RFF values among all combinations of the recordings. Three /ifi/ productions were selected from each effort condition for each participant. Recordings were selected for inclusion based on three criteria to best support the study hypotheses: 1) increases in self-modulated vocal effort accompanied with decreases in RFF, 2) increases in self-modulated vocal effort accompanied with increases in listener-perceived strain, and 3) minimal listener-perceived breathiness regardless of vocal effort level. These criteria are further explained below.

Increases in self-modulated vocal effort accompanied with decreases in RFF

RFF is generally expected to decrease as vocal effort increases, although between-speaker variability has been reported (Lien et al., 2015b; McKenna & Stepp, 2018; Stepp et al., 2012). We purposefully chose voice samples that showed decreased RFF along with increased vocal effort to evaluate the contribution of RFF to strain. To achieve this aim, we planned to synthetically lower RFF values of comfortable voice samples to match the

RFF values of maximum effort samples from the same speakers and to examine whether the RFF modifications increased the strain. We also planned to synthetically raise RFF values in maximum effort samples to match the RFF values of comfortable voice samples from the same speakers and to examine whether the RFF modifications decreased the strain. RFF was manually estimated with Praat acoustic analysis software (Version 6.0.48; Boersma & Weenink, 2019). Ten voiced cycles prior to and after the voiceless consonant were identified using Praat's autocorrelation algorithm for pitch tracking, and the period and the instantaneous f_o for each cycle were calculated. RFF for each cycle was calculated in semitones (ST) from the equation: ST= 39.86 × $\log_{10}(f_o/\text{reference } f_o)$, in which the reference f_o was offset cycle 1 for offset cycles and onset cycle 10 for onset cycles. Mean RFF was higher in comfortable voice samples than in maximum vocal effort samples.

Increases in self-modulated vocal effort accompanied with increases in listener-perceived strain

We also chose recordings in which the strain increased as the self-modulated vocal effort level increased. Strain of the recordings was evaluated by a voice-experienced speech-language pathologist. Since speakers could increase their vocal effort without actual perceptible increases in strain by listeners, this criterion ensured that the comfortable voice and maximum vocal effort samples had actual differences in strain. The speech-language pathologist rated the strain of each recording on a 100-mm visual analog scale from the CAPE-V form (Kempster et al., 2009). Mean strain was higher in maximum vocal effort samples than in comfortable voice samples.

Minimal listener-perceived breathiness regardless of vocal effort level

Speakers with minimal listener-perceived breathiness were favored because of the study aim to evaluate the effect of mid-to-high frequency noise on perceiving strain and RFF. Since we planned to compare samples with and without added noise, recordings with minimally breathy voices would be ideal to precisely control for noise. Minimal breathiness was determined from both the auditory-perceptual evaluation by a voice-experienced speech-language pathologist and CPPS values. The speech-language pathologist rated the breathiness of each recording on a 100-mm visual analog scale from the CAPE-V form (Kempster et al., 2009) after listening to three /ifi/ samples per speaker. CPPS represents the strength of the cepstral peak compared to the cepstral background noise in acoustic signals, which reflects the degree of the periodicity of the signal. CPPS has shown a strong negative correlation with perceived breathiness (Hillenbrand et al., 1994). CPPS was obtained from the /i/ portions of /ifi/ recordings with the commands and parameters described in Watts, Awan, and Maryn (2017). Mean breathiness and CPPS values were similar to those of 20 young female adults with healthy voices in our previous study (Park, Perkell, Matthies, & Stepp, 2019) and did not differ between the two vocal effort conditions.

Stimuli Synthesis

In order to synthetically modify the selected recordings, we used the Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) algorithm. The STRAIGHT is based on a sophisticated channel VOCODER system, which separates the spectral and source information (Kawahara, 2006). The algorithm incorporates speech analysis, modification, and synthesis. During the analysis, the algorithm extracts information about the spectral envelope, the instantaneous f_o contour, and the aperiodic component of a sound sample. These three components can be modified separately and then synthesized back together to generate a modified voice sample. This method was developed to provide flexible modification of the three components as naturally as possible and was adapted to MATLAB (Ver. R2018a, MathWorks, Natick, MA). The 16 original recordings were analyzed, and the STRAIGHT components were saved for sample modifications. Instead of including the original recordings as a part of the perceptual stimuli, we used STRAIGHT-synthesized versions of the original recordings. The STRAIGHT components from the original recordings were resynthesized back without any modification to be included in the experiment as *unmodified* samples. This ensured that possible perceptual differences between the original samples and RFF modified samples would not be due to being synthesized from the STRAIGHT, although the original recordings and the STRAIGHT-synthesized versions are known to be perceptually identical (Kawahara, 2006). Each sample consisted of three /ifi/s in the same vocal effort condition from the same participant with 300-ms periods between each /ifi/.

Modifying RFF

RFF of the unmodified samples was modified in order to test the hypothesis that the modification of RFF alone would alter the strain. Modifying RFF of the samples and comparing the RFF-modified and -unmodified samples allowed precise evaluation of RFF in relation to strain, since RFF was the only acoustic feature that was different between them. RFF of the comfortable voice sample from each participant was lowered to the RFF values of the same participant's maximum vocal effort sample for all 20 RFF cycles. RFF of the maximum vocal effort sample from each participant was raised to the RFF values of the same participant's comfortable vocal effort sample for all 20 RFF cycles. In order to modify RFF, we modified the f_o contours of the samples, as estimated by the STRAIGHT. Figure 3.1 illustrates the procedure for modifying f_o contours. The modified f_o contours and other STRAIGHT components of the same sample were combined together to synthesize an RFF-modified sample. We performed the RFF modification on all unmodified samples and confirmed that the modified RFF values matched the goal RFF values very closely. The mean difference in RFF between comfortable voice samples with RFF modification and maximum vocal effort samples without RFF modification was 0.14 ST; the mean difference in RFF between maximum vocal effort samples with RFF modification and comfortable voice samples without RFF modification was 0.01 ST. Because the spectral envelopes of the samples were not modified during the process, the formant values of RFF-modified samples were not altered.

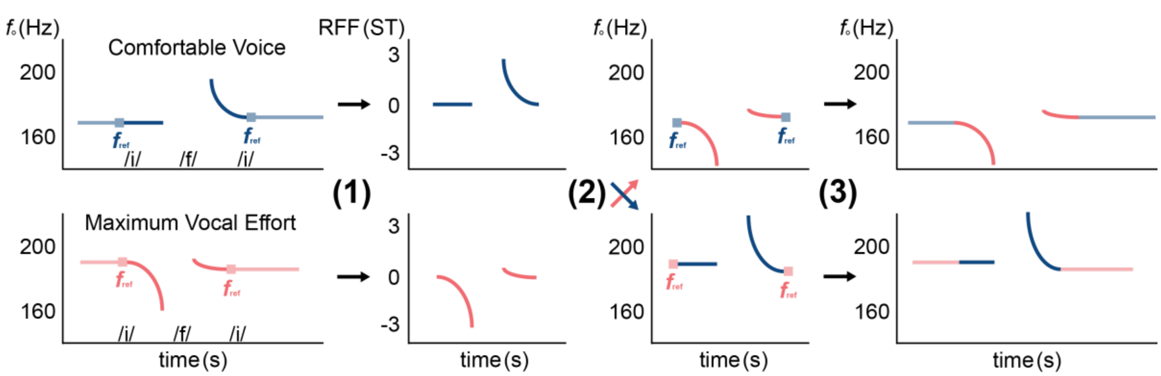

Figure 3.1: The procedures for modifying fundamental frequency (f_o) contours of the comfortable and maximum vocal effort samples to exchange their RFF values. The upper panels present schematic plots of f_o (in Hertz; Hz) and RFF (in semitones; ST) as a function of time (in seconds; s). Blue represents a comfortable voice sample and pink represents a maximum vocal effort sample. RFF portions of the f_o contour contain bolder colors than non-

RFF portions (abbreviation: $f_{ref} = f_o$ of the reference cycle). Following steps outline the details of the procedures:

(1) **RFF portions were selected from the f_o contours of the comfortable voice and maximum vocal effort samples. The selected f_o contour of each sample was normalized by the f_{ref} values of each sample (RFF [ST] = $39.84 \times \log_{10}[f_o/f_{ref}]$).**

(2) **The RFF contours of the comfortable and maximum vocal effort voice samples were exchanged and transformed back to f_o contours that fit their respective counterpart's f_{ref} values (f_o [Hz] = f_{ref} × 10^[RFF/39.84]).**

(3) **The converted f_o contours of the comfortable and maximum vocal effort samples replaced the RFF portions of their respective counterpart's f_o contours. During this process, durations of the converted f_o contours were adjusted, so that the original durations of RFF portions remained the same after RFF modification.**

Adding mid-to-high frequency noise

Versions of samples with and without RFF modification with added mid-to-high frequency noise were created via the 'breathiness' function in the Praat Vocal Toolkit (Corretge, 2019). This function uses linear predictive coding (LPC) to estimate the spectral envelope of the original signal. The function then creates a "whispered" version of the signal by applying the estimated spectral envelope to white noise, decreasing the spectral energy of the low frequencies under 250 Hz, and increasing the spectral energy of the mid-to-high frequencies centered around 2000 Hz. The whispered version of the signal is then added to the original signal in a quantity determined by the user's input. In order to synthesize the samples to be breathy while retaining a natural quality, we decreased the harmonics-to-noise ratio (HNR) values of the samples by -7 dB using the 'breathiness' function. The mean HNR of the samples without added noise was 19.3 dB and the mean HNR of the samples with added noise was 12.3 dB. A speech-language pathologist evaluated the breathiness of all samples; the breathiness ratings increased by an average of 30.6 mm on the 100-mm scale in the samples with added noise.

Total number of stimuli

The number of the STRAIGHT-synthesized, unmodified samples was 16 (8 participants × 2 vocal effort conditions). RFF was modified in all unmodified samples and noise was added in both RFF-modified and -unmodified samples, resulting in a total of 64 samples for the perceptual tasks [16 unmodified samples + 16 RFF-modified samples + 32 samples with added noise (unmodified and RFF-modified)].

Listeners

Twenty healthy participants (10 females and 10 males; mean age = 22.0 years; range = 18 – 34 years) were recruited as listeners from college job posting sites and paper flyers and were paid for their participation. The number of participants was determined by evaluation of the average absolute deviations of strain from mean strain ratings obtained using a visual sort-and-rate (VSR) task from 20 listeners in a previous study (McKenna & Stepp, 2018). Mean strain rated by 18 listeners differed from mean strain rated by 20 participants by 1 mm on a 100 mm scale. We recruited 20 participants to attain similar precision. Participants reported no prior history of speech, language, and hearing disorders or previous participation in any auditory-perceptual study. Participants all scored within normal ranges for the Voice-Related Quality of Life (Hogikyan & Sethuraman, 1999). All but one participant passed a hearing screening with 25 dB HL pure tones at 125, 250, 500, 1000, 2000, 4000, and 8000 Hz (American Speech-Language-Hearing Association, 2005) in a sound-treated room (one participant passed at 30 dB HL at 4000 Hz in his left ear). Inexperienced listeners were recruited, as previous studies did not find differences in interrater reliability values between expert and inexperienced listeners (Eadie, Kapsner, Rosenzweig, Waugh, Hillel, & Merati, 2010).

Perceptual Tasks

Visual sort-and-rate (VSR) task for strain

The participants completed VSR training and experimental tasks in a sound-treated room. The VSR task was chosen because of its higher reliability compared to other auditory-perceptual tasks (Granqvist, 2003). Participants were provided the CAPE-V definition of strain, “perception of excessive vocal effort (hyperfunction)” (Kempster et al., 2009), and the definition of vocal effort, “perceived exertion in producing voice” (Verdolini, Titze, & Fennell, 1994).

First, they were trained to use the VSR module on a desktop computer and familiarized themselves with a wide range of strain. The training module included eight voice samples, each containing three /ifi/s with 300-ms breaks, similar to experimental stimuli. The eight training samples were chosen from the same database of original recordings as the experimental stimuli. The eight training samples were selected to contain a wide range of strain based on strain ratings from three voice-experienced speech-language pathologists. Strain ratings of the training set ranged from 2.2 to 83.4 mm, spread evenly throughout the 100-mm range. At the start of the training, icons for the samples were located horizontally at the middle of the vertical axis, which ranged from 0 mm (no strain) to 100 mm (the most strain). When the participants clicked each icon on the screen, the corresponding sample was presented at 75 dB SPL through a pair of Sennheiser HD-290 headphones. Participants were allowed to listen to the samples as many times as they wished. They were asked to first listen to each stimulus and rate the strain by moving icons vertically on the strain scale. After finishing the initial listening and rating the samples of the training set, participants were asked to re-listen to each sample and adjust their ratings by comparing the samples that were located near each other vertically. When the participants finished rating the training set, they were given the experts' scores of the training samples as feedback, so that they could familiarize themselves with the experts' ratings on these training samples. This familiarization with the experts' ratings was aimed at improving the interrater reliability of the task, as poor interrater reliability of strain ratings has been previously reported (Webb et al., 2004; Zraick et al., 2011).

After the training module, listeners were asked to complete the experimental VSR module, which contained the same screen set-up as the training module. A total of 80 stimuli, the 64 stimuli and 16 randomly chosen stimuli from the 64 stimuli for intra-rater reliability, were divided into 10 sets of 8 stimuli. Each set was designed to contain only one stimulus from each speaker so that the samples from the same speaker would not be compared to each other within a set. Each set was also designed to contain one stimulus from each modification type (e.g., RFF-modified comfortable voice samples with added noise, RFF-unmodified maximum vocal effort samples with added noise) so that every set would contain samples with a wide range of strain. Each set also contained at least one repeated stimulus from the other sets for intra-rater reliability assessment. The 10 sets of stimuli were constructed specifically for each listener to reduce the effects of stimuli order and set composition. The experimental VSR task took approximately 15 minutes to complete.

Same or different task

Participants completed an AX (same or different) task in a sound-treated room to evaluate whether listeners could differentiate between the samples that differed only in their RFF. Each /ifi/ in the RFF-unmodified samples was paired with its own RFF-modified version with a 300-ms interstimulus interval. A total of 96 pairs of RFF-modified and -unmodified samples were possible from the three /ifi/s in our 32 RFF-modified and 32 RFF-unmodified samples. Within each pair, the order of RFF-modified and -unmodified samples were randomly determined. In order to balance the number of same and different trials in the task, 96 stimuli pairs with the same /ifi/s were randomly chosen from the stimuli set and included in the task. After hearing each pair of stimuli, listeners judged whether the two stimuli were same or different in a forced-choice paradigm. They were asked to listen very carefully and were informed that the difference in the two samples could be very small, but they were not given any information about the basis of any differences. In total, 192 trials were performed by each listener, taking approximately 20 minutes to complete.

Data Analysis

Strain ratings for each stimulus obtained from the VSR tasks were averaged across the listening participants. The number of correct "different" responses of each participant was obtained from the AX task, and the correct response rate was calculated for each of the four stimulus conditions: comfortable voice samples with and without noise and maximum vocal effort samples with and without noise. The number of wrong "different" response was also obtained, and the false-alarm rate was calculated for each stimulus condition. The sensitivity index, *d'*, was calculated from the equation below presented in

Macmillan and Creelman (2004) for each stimulus condition of each listener: *d' = z(correct response rate) – z(false-alarm rate),* where *z* is the inverse of the normal distribution. When either rate was 0 (which inhibits the calculation of *d'*), we used 1/(2 × the number of trials in a stimulus condition [24]) instead. A high, positive *d'* value would indicate high discriminability between RFF-modified and -unmodified samples, whereas a zero *d'* value would indicate chance level performance (Macmillan & Creelman, 2004).

Statistical Analysis

Statistical analysis was performed in SPSS (Ver. 24.0, IBM Corp., Armonk, NY). A three-way repeated measures ANOVA on the mean strain of the stimuli was performed with the factors: vocal effort level (comfortable voice or maximum vocal effort), RFF modification (unmodified or modified), noise (no or added noise), and the interactions between the factors. We hypothesized that the interaction between vocal effort level and RFF modification would be statistically significant, which would support the contribution of RFF to strain. We did not hypothesize a main effect of RFF modification on strain because the direction of RFF modification depended on vocal effort level: The comfortable voice samples would have increased strain due to their lowered RFF, whereas the maximum vocal effort samples would have decreased strain due to their raised RFF. We also hypothesized that either noise or the interaction between noise and vocal effort level would be statistically significant because mid-to-high frequency noise was expected to increase strain. Finally, we hypothesized that there would be a statistically significant interaction between noise, vocal effort level, and RFF modification because noise may affect the listeners' acuity of the percept caused by RFF. We also performed a two-way repeated measures ANOVA on *d'* values from the AX task with vocal effort level and noise as factors to evaluate how adding the noise would affect the ability of listeners to notice differences in RFF. Effect sizes were calculated as a partial eta squared (η_p^2) and *post hoc* tests were performed when statistically significant interactions were observed.

Intra-rater reliability of the ratings of strain was assessed using Pearson's correlations from 16 repeated samples. Strain showed intra-rater reliability (Pearson's *r*) above 0.7 in 18 of 20 listeners (median = 0.85, range = 0.42 – 0.99). Interrater reliability was represented as intraclass correlation coefficient (ICC; two-way mixed effects, consistency, single measure) calculated with the ratings of all 64 stimuli from all listeners. ICC below 0.5 has been considered as poor reliability, 0.5 – 0.75 as moderate reliability, 0.75 – 0.9 as good reliability, and above 0.9 as excellent reliability (Portney & Watkins, 2000). Ratings of strain showed moderate interrater reliability (ICC = 0.61, 95% CI = 0.53 – 0.70).

We additionally evaluated intra-rater and interrater reliability separately for samples with and without noise to examine our hypothesis that mid-to-high frequency noise may decrease the reliability of strain rating. Among the 16 repeated samples, half of them were samples with added noise and the other half was without noise. In order to evaluate our hypothesis that mid-to-high frequency noise would decrease intra-rater reliability of strain, mean absolute differences in strain between the actual and repeated samples were calculated separately for samples with and without noise from each listener. An independent *t*-test was performed on the mean absolute differences in strain to evaluate whether the samples with noise resulted in a larger mean absolute difference than samples without noise, which would suggest lower intra-rater reliability. To evaluate our hypothesis that noise would decrease interrater reliability, we calculated mean absolute deviation by obtaining absolute deviations of an individual listener's strain rating of a sample from the sample's mean strain rating by the 20 listeners and averaging absolute deviations within each sample. A paired *t*-test was performed between the mean absolute deviations of the samples with and without added noise to examine if adding noise increased mean absolute deviation, which would suggest decreased interrater reliability. A predetermined level of statistical significance ($\alpha = 0.05$) was used for all statistical tests.

Results

RFF Modification

The three-way ANOVA on mean strain showed a statistically significant effect of the interaction between vocal effort level and RFF modification with a large effect size (p = 0.003, η_p^2 = 0.74). This interaction indicates that RFF modification changed the perceived strain of the samples, but that the effect differed based on the vocal effort level. The *post hoc* paired *t*-test between the comfortable voice samples with and without RFF modification revealed that synthetically lowering RFF values in the comfortable voice samples resulted in increases in strain (t = -5.4, p < 0.001; Figure 3.2), as hypothesized. The *post hoc* paired *t*-test between the maximum vocal effort samples with and without RFF modification revealed that synthetically raising RFF values in the maximum vocal effort samples resulted in decreases in strain (t = 3.5, p = 0.003; Figure 3.2). The mean *d'*, which represents the listeners' performance discriminating between RFF-modified and -unmodified samples on the AX task, ranged from 0.12 to 0.40 (mean = 0.29, 95% confidence interval = 0.17 – 0.40) in the four stimulus conditions, all above 0 in the scale, in which 0 indicates chance-level performance (Figure 3.3).

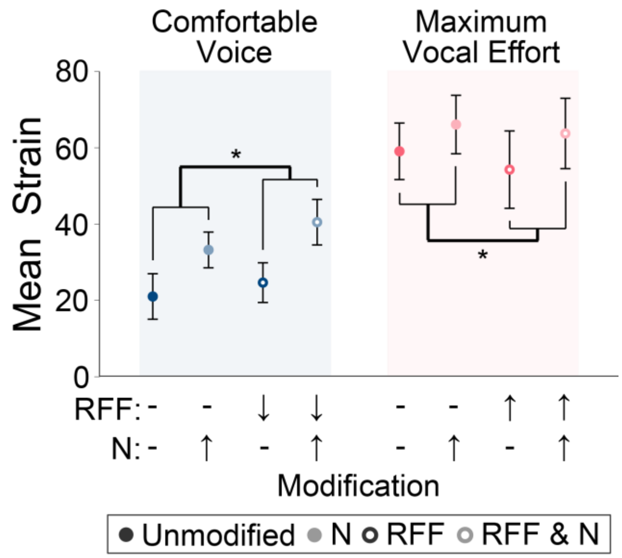

Figure 3.2: Mean strain ratings of comfortable and maximum effort samples as a function of modification condition. Error bars indicate 95% confidence intervals, and bolded brackets and asterisks indicate statistically significant differences between RFF-modified and -unmodified samples within comfortable voice ($p < 0.001$) and maximum vocal effort ($p = 0.003$) conditions. (abbreviations: RFF = relative fundamental frequency, N = noise, - = unmodified, ↑ = increase, ↓ = decrease).

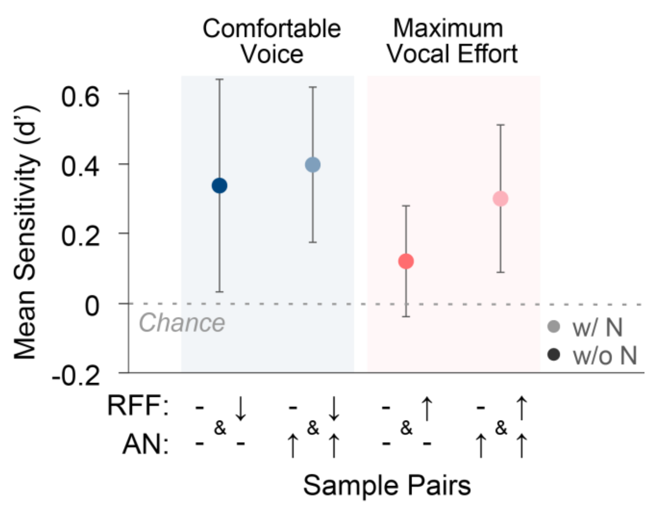

Figure 3.3: Mean sensitivity of discriminating RFF-modified and -unmodified samples from the AX task as a function of paired condition in comfortable voice and maximum vocal effort samples. The addition of noise was not a statistically significant factor on mean *d'* ($p > 0.05$). A dotted line indicates chance-level discrimination. Error bars indicate 95% confidence intervals. (abbreviations: RFF = relative fundamental frequency, w/ N = with noise, w/o N = without noise, - = unmodified, ↑ = increase, ↓ = decrease)

Mid-to-High Frequency Noise

The effect of mid-to-high frequency noise on strain was statistically significant and showed a large effect size ($p < 0.001$, $\eta_p^2 = 0.97$). The samples with added noise had increased strain (Figure 2). There was no statistically significant effect on the interaction between noise, vocal effort, and RFF ($p = 0.83$). The addition of noise was not a statistically significant factor in the one-way ANOVA on mean *d'* ($p = 0.18$; Figure 3).

The mean absolute difference in strain between the samples and their repetitions was statistically greater ($t = 2.45$, $p = 0.01$) in the samples with noise (mean = 12.2, 95% confidence interval = 11.4 – 13.1) relative to the samples without noise (mean = 8.5, 95% confidence interval = 7.9 – 9.0), suggesting lower intrarater reliability of strain ratings in the samples with noise than without noise. The mean absolute deviation was also statistically greater ($t = 2.20$, $p = 0.035$, mean difference = 1.4, 95% confidence interval = 0.1 – 2.7) in the samples with noise than samples without noise, suggesting slightly lower interrater reliability of strain ratings in the samples with noise than without noise.

Discussion

In this study we performed auditory-perceptual experiments with synthetically modified voice samples to evaluate direct, causal contributions of RFF and mid-to-high frequency noise to the perception of strain. We hypothesized that synthetically lowering RFF in voice samples would increase strain, whereas raising RFF in voice samples would decrease strain. We also hypothesized that adding mid-to-high frequency noise in voice samples would both increase strain and decrease intra- and inter-rater reliability of strain ratings.

RFF

The statistically significant interaction between vocal effort level and RFF modification supports the role of RFF as an acoustic contributor to strain. The mean *d'* value greater than 0 from the AX tasks also supports that differences in RFF can be noticed by listeners, although with difficulty (low *d'* values). Our finding is consistent with previously observed correlations between RFF and strain in speakers with healthy voices who modulated their vocal effort (Lien et al., 2015b; McKenna & Stepp, 2018), speakers with healthy voices and vocal hyperfunction (Stepp et al., 2012), and speakers with laryngeal dystonia (Eadie & Stepp, 2013). Our findings further support the contribution of RFF to strain by showing that strain changed when only RFF was modified in the acoustic samples while other acoustic features remained constant.

A potential reason for the relationship between RFF and perceived strain may be the high prevalence of decreased RFF in individuals with increased laryngeal tension and vocal effort in their voice production. In our study, decreasing RFF values in the comfortable voice samples resulted in increases in strain. This decreased RFF pattern has been observed in individuals with increased laryngeal tension and vocal effort during their voice production in previous studies (Eadie & Stepp, 2013; Heller Murray et al., 2017; Lien et al., 2015b; Roy et al., 2016; Stepp, 2013; Stepp et al., 2010a; Stepp et al., 2011b). Specifically, these individuals showed decreasing offset RFF and slightly increased and then decreasing onset RFF, whereas individuals speaking with typical voices showed stable or slightly decreasing offset RFF and substantially increased and then decreasing onset RFF. Stepp et al. (2011b) and Roy et al. (2016) also observed that successful voice therapy sessions normalized this decreased RFF pattern (although Roy et al. (2016) only observed this finding in onset RFF). Individuals with increased laryngeal tension and vocal effort are known to have strained voice quality (Kempster et al., 2009; Roy, 2008), which is suggested to be multidimensional (Kent, 1996; Lowell et al., 2012a). Thus, we may frequently encounter the decreased RFF pattern concurrently present with other acoustic features of strain in individuals with increased laryngeal tension and may associate it with increased strain.

Although the contribution of RFF to perceived strain is supported in this study, RFF is probably a small factor in the overall construct of strain due to its short duration and linguistic constraints. Although we observed a statistically significant and large effect of the interaction between vocal effort level and RFF modification, the average change in strain due to RFF modification was small, less than 10 mm on the 100-mm scale. These small changes in strain may have been due to the fact that the RFF measure only spans a small proportion of each utterance, whereas other acoustic features such as increased spectral energies at higher harmonic frequencies and mid-to-high frequency noise can span entire utterances. The short duration of RFF cycles may also explain why it was challenging for the listeners to consistently differentiate between the samples with and without RFF modification in the AX task. The duration of 20 RFF cycles can be estimated from our speakers' average f_os producing /ifi/, which ranged from 105 to 254 Hz. Based on that f_o range, the duration of 20 RFF cycles is estimated to range only from 79 to 190 ms ($1/f_o \times$ 20 cycles), whereas the duration of the entire utterance ranged from approximately 400 to 1000 ms. The proportion of RFF in a sound sample of typical running speech will further decrease, as these are likely not to contain many VCV contexts with voiceless consonants. Thus, we hypothesize that, in running speech stimuli, the contribution of RFF to strain also would be even smaller.

Mid-to-High Frequency Noise

The results of the VSR task for rating strain showed that mid-to-high frequency noise is also a statistically significant contributor to strain with a large effect. The effect of noise was also stronger than of RFF, probably because it was present in much longer durations of the samples than RFF. Our finding is consistent with previous findings that showed that increases in breathiness or aspiration noise were coincident with increased strain (Hirano, 1981; Lowell et al., 2012a). Listeners may associate increased mid-to-high frequency noise with strain because of the high prevalence of aspiration noise in individuals with strained voices (e.g., individuals with glottal insufficiency, vocal nodules, and paralysis) who need to increase their vocal effort in order to phonate. Aspiration noise also may be perceived as increased respiratory effort, which usually accompanies an increased airflow rate (Zhang, 2015).

There was no statistically significant interaction between noise and RFF. We had predicted that noise would affect the listeners' acuity to the percept caused by RFF. There was also no statistically significant effect of noise on the *d'* in the AX task, which suggests that mid-to-high frequency noise may not affect the ability to notice differences in RFF. However, we may not have observed noise reducing *d'* in the AX task because of the inherent difficulty of the task: the task resulted in overall low values of *d'*, suggesting a floor effect.

Our findings also suggest that mid-to-high frequency noise may decrease both intra- and interrater reliability of strain ratings. These findings are likely not due to noise interfering with the perception of RFF, since the effect of noise on discriminability between samples with and without RFF modification was not statistically significant. Instead, mid-to-high frequency noise is likely to interfere with higher harmonic frequencies, located in a similar frequency range. This speculation is consistent with previous findings from Kreiman and Gerratt (2012) that showed that increased noise in samples reduced sensitivity to harmonic frequencies. Thus, listeners may have perceived different amounts of energy at higher harmonic frequencies when noise was added, resulting in different strain ratings.

Implications for Clinical Evaluation of Strain

The decrease in reliability of strain ratings of samples with noise observed in this study suggests that the auditory-perceptual evaluation of strain may be more challenging for individuals with breathy or dysphonic voices than for individuals with voices without breathiness. The effect of noise on strain may have been the reason that previous studies have observed lower reliability of strain than of other voice qualities (Webb et al., 2004; Zraick et al., 2011) for which individuals with voice disorders were evaluated. When rating strain in dysphonic voices, some listeners may base their ratings more on the presence of noise whereas other listeners may base their rating more on spectral energies at higher harmonic frequencies or RFF. This finding is similar to the findings of Kreiman, Gerratt, Precoda, and Berke (1992), which suggested variability in acoustic cues that individuals use to rate voice quality. Intra-rater reliability of strain may have also been low due to noise interacting with other acoustic features of strain. Mid-to-high frequency noise affecting reliability of strain is problematic because many individuals with voice disorders are likely to present increased aspiration noise due to glottal insufficiency. This population needs to be evaluated accurately for effective treatment. Based on our findings, clinicians should be aware that the auditory-perceptual evaluation of strain may not be reliable in individuals with dysphonia and that their strain ratings should be incorporated with care in their clinical practice.

These issues with auditory-perceptual evaluation of strain support call for more research to develop objective measures to assess strain. The findings of this study further support that RFF exhibits potential as an objective measure for assessing increased vocal effort. RFF has consistently differentiated between individuals with healthy voices and vocal hyperfunction (Roy et al., 2016; Stepp et al., 2010a; Stepp et al., 2011b), whereas conventional acoustic measures have been shown mixed results (Belsky et al., In Press; Holmberg et al., 2003; Schindler et al., 2013). H1-H2 and measures of spectral tilt (e.g., low-to-high spectral ratio), which reflect increased energies at higher harmonic frequencies, may fail to reflect increased vocal effort if individuals with increased vocal effort do not completely adduct their vocal folds due to structural lesions or vocal fold paralysis. Previous attempts to examine the effect of CPPS on strain have resulted in mixed findings (Anand et al., 2019; Lowell et al., 2012a; McKenna & Stepp, 2018), possibly due to occurrences of both increased harmonic energy in higher harmonics (which increases CPPS) and mid-to-high frequency noise (which decreases CPPS) in strained voices. In contrast, RFF is a time-based measure, which is not affected by the spectral contents of voice samples. RFF was also observed to detect possible voice changes from an intense voice-use period that auditory-perception ratings did not reflect (Heller Murray et al., 2016), which suggests that RFF may be more sensitive to small changes in vocal function than auditory-perceptual evaluation. Thus, RFF may be a good complement to clinical evaluation of strain.

Although RFF may reflect strain, the multidimensionality of strain suggests that a single acoustic variable may not be sufficient to capture strain. The present study supports assertions about the multidimensional nature of strain, finding statistically significant contributions of RFF and mid-to-high frequency noise to strain in addition to the previously observed effects of increased energies at higher harmonic frequencies (Anand et al., 2019; Bergan et al., 2004). Due to these acoustic features affecting strain, previous studies may have struggled to find a single acoustic measure that correlates strongly with strain (Bhuta et al., 2004) This multidimensional character of strain is also likely to inhibit the recent efforts to develop analogous scales for the perception of voice quality (e.g., *sones* for the loudness scale) from being applied to strain, since developing analogous scales for perception requires a single physical variable that correlates strongly with perception (e.g., noise-to-harmonic ratio for breathiness; Eddins, Anand, Lang, & Shrivastav, 2020).

Instead of a single acoustic measure, multiparametric tools, similar to Acoustic Voice Quality Index (AVQI; Maryn, Corthals, Van Cauwenberge, Roy, & De Bodt, 2010) and cepstral spectral index of dysphonia (CSID; Awan, Roy, Zhang, & Cohen, 2016), may represent strain more adequately. Both AVQI and CISD have been developed to complement clinical evaluation of the overall severity of dysphonia and a primary acoustic component in both of these indices is CPPS. Although there was a previous attempt to build a multiparametric tool for strain using CPPS as a component (Lowell et al., 2012a), CPPS is not likely to specifically represent strain due to both increased energies at higher harmonic and noise frequencies contributing to strain, as previously explained. In order to develop a multiparametric tool for strain, acoustic measures that can independently estimate energies of harmonic and noise frequencies may be required. RFF could also be one of the factors in this tool, with sentences loaded with RFF instances used as speech samples. Future studies should incorporate these acoustic elements and other potential acoustic contributors to strain into a multiparametric tool for strain.

Limitations

Due to our use of synthetically modified samples, listener reactions to any synthetic sound quality in the samples may have affected the results of this study. To determine whether this was the case, we performed an additional VSR of rating synthetic quality, described in the Appendix 1. We did not find statistical differences between RFF-modified and -unmodified samples, but we did find statistical differences between samples with and without added noise with a large effect size. Increased synthetic quality in the samples with added noise may have affected strain ratings in these samples, but the relationship between synthetic quality and strained voice quality is also unknown. We aimed to add mid-to-high frequency noise as naturally as possible using the breathiness function in Praat, which estimates spectral shapes of voice samples to filter white noise and generates whispered versions of the original voice samples. However, because we added synthetic noise to natural voice samples, we could not avoid our samples with added noise sounding more synthetic. Future studies should investigate methods to increase noise levels in voice samples more naturally for future perceptual studies of voice quality.

Another limitation of this study is the small number of expert raters in auditory-perceptual evaluations of the original voice recordings and the training stimuli for the VSR task. Because of the known poor inter-rater reliability of auditory-perceptual evaluation (Webb et al., 2004; Zraick et al., 2011), the scores from these expert raters may have not been reliable. However, these ratings played subsidiary roles of the experiment and thus they are not likely to substantially affect the findings of the study.

Conclusion

Synthetic modification of RFF and addition of mid-to-high frequency noise changed the perceived strain of the modified samples. Lowering RFF resulted in increased strain and raising RFF resulted in decreased strain, consistent with previous findings of the perceptual studies on RFF. Adding mid-to-high frequency noise resulted in increased strain and decreased intra- and interrater reliability of strain. Our findings support the multidimensionality of strain and suggest that future acoustic assessment of strain can be better achieved through multiparametric tools incorporating multiple acoustic features of strain. The decreased intra-rater reliability of strain in the samples with noise indicates that the clinical perceptual evaluation of strain in dysphonic voices can be problematic and further supports the need for objective assessment of strain to complement auditory-perceptual evaluation.

Chapter 4: Test-Retest Reliability of Relative Fundamental Frequency and Conventional Acoustic, Aerodynamic, and Perceptual Measures in Individuals with Healthy Voices

Abstract

Recent studies have shown that an acoustic measure, Relative Fundamental Frequency (RFF), has potential for the assessment of excessive laryngeal tension and vocal effort associated with functional and neurological voice disorders. This study presents an analysis of the test-retest reliability of RFF in individuals with healthy voices and a comparison of reliability between RFF and conventional measures of voice. Acoustic and aerodynamic measurements and Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) were performed on 28 individuals with healthy voices on five consecutive days. Participants produced RFF stimuli, a sustained /ɑ/, and a reading passage to allow for extraction of acoustic measures and CAPE-V ratings; /pa/ trains were produced to allow for extraction of aerodynamic measures. Moderate reliabilities (ICC 0.64 – 0.71) were found for RFF values. Mean vocal fundamental frequency, smoothed cepstral peak prominence, shimmer, harmonics-to-noise ratio, and mean airflow rate exhibited good-to-excellent reliabilities (ICC 0.76 – 0.99). ICCs for jitter and phonation threshold pressure were moderately reliable (ICC 0.67 – 0.74). ICCs for subglottal pressure estimates and all CAPE-V parameters showed poor reliabilities (ICC 0.31 – 0.58). RFF has comparable reliability to conventional measures of voice. This expands the potential for clinical application of RFF.

Introduction

One of the most common features of voice disorders is vocal hyperfunction (Stemple et al., 2014). Vocal hyperfunction, characterized by strained voice quality, has been defined as “abuse and/or misuse of the vocal mechanism due to excessive and/or ‘imbalanced’ muscular forces” (Hillman et al., 1989). Vocal hyperfunction may accompany voice disorders that change the structure of the vocal folds (e.g., vocal nodules) or it may appear in individuals without organic changes to the larynx, as in muscle tension dysphonia (Hillman et al., 1989). Thus, managing vocal hyperfunction is an important therapeutic strategy to treat a variety of voice disorders. However, clinical assessment currently lacks objective tools to quantify the degree of vocal hyperfunction and evaluate the treatment outcomes (Hillman, Gress, Hargrave, Walsh, & Bunting, 1990a).

Recently, the acoustic measure, Relative Fundamental Frequency (RFF), has been investigated as a possible objective correlate of strained voice quality in vocal hyperfunction (Stepp et al., 2010a). RFF quantifies changes in the fundamental frequency (f_o) of voicing offset and onset during the production of sonorant-voiceless consonant-sonorant constructs. In healthy voices, f_o usually decreases slightly before the voiceless consonant and increases immediately after (Watson, 1998). This f_o change in voicing offset and onset surrounding voiceless consonants is assumed to be related to an increase in vocal fold tension produced by the cricothyroid (CT) muscle, which is thought to aid voicing termination for voiceless consonants (Stevens, 1977). Although the exact cause of the instantaneous f_o change is still unknown, Stepp et al. (2010a) hypothesized that baseline rigidity or tension in the larynx in individuals with vocal hyperfunction would decrease the extent of this f_o change during voiceless consonant production, and thus, the RFF of individuals with vocal hyperfunction would be lower than in those with healthy voices (Stepp et al., 2010a).

Several studies have supported RFF's potential to assess vocal hyperfunction. RFF values were significantly lower in participants with vocal hyperfunction (Heller Murray et al., 2017; Roy et al., 2016; Stepp et al., 2010a; Stepp et al., 2012), Parkinson's disease (Bowen et al., 2013; Goberman & Blomgren, 2008; Stepp, 2013), and adductor spasmodic dysphonia (Eadie & Stepp, 2013) compared to the RFF of individuals with healthy voices. In addition, the RFF of individuals with vocal hyperfunction significantly increased toward the RFF values of typical speakers after successful voice therapy sessions. This finding suggested promise for the usefulness of RFF as an outcome measure for voice therapy (Roy et al., 2016; Stepp et al., 2011b). Studies have also evaluated RFF's ability to assess the degree of baseline laryngeal tension, the findings of which included significant correlations between RFF and both aerodynamic (Lien et al., 2015b) and auditory-perceptual measures (Stepp et al., 2012) of vocal effort, and with a kinematic estimate of laryngeal stiffness (McKenna et al., 2016).

Although RFF continues to show promise as a possible objective marker for vocal hyperfunction, more research is necessary before it can be utilized clinically, such as reliability, sensitivity to change, and diagnostic sensitivity and specificity. This study aimed to examine test-retest reliability of RFF. A few studies have compared RFF values estimated at different times, and no significant group differences in RFF values of healthy individuals were found when measured at 10 weeks apart (Heller Murray et al., 2016) and one-hour apart (Roy et al., 2016). However, very little is known about RFF's reproducibility in *individual speakers* over time. Group effects, while important, can mask whether a measure is a useful indicator at the individual patient level. We examined the test-retest reliability in individuals with healthy voices to minimize any voice changes that could affect the results. We measured the participants' voices every day throughout one work week during which their vocal function was assumed to be relatively stable.

We also measured participants' voices using conventional voice measures throughout the week in order to compare the test-retest reliability of these standard clinical measures with the reliability of RFF. Instrumental measures have provided valuable information to clinicians when they diagnose voice disorders and assess their severity, prognosis, and treatment outcomes (Stemple et al., 2014). Acoustic measures of mean vocal f_0, jitter, shimmer, harmonics-to-noise ratio (HNR) were selected because of their frequent usage in clinic as well as the research evidence that suggests their effectiveness in classifying dysphonia (Desjardins, Halstead, Cooke, & Bonilha, 2017; Eadie & Doyle, 2005; Linder, Albers, Hess, Poppl, & Schonweiler, 2008). Jitter, shimmer, and HNR have been studied due to their hypothesized association with roughness and breathiness (Eskenazi, Childers, & Hicks, 1990; Hillenbrand, 1988). Although the test-retest reliabilities of jitter, shimmer, and NHR have shown mixed findings in previous studies (Bough et al., 1996; Carding et al., 2004; Leong et al., 2013), we included them since they have been commonly used in both clinical and research applications. We also included smoothed cepstral peak prominence (CPPS), a cepstral measure, obtained from the Fourier transform of the power spectrum, which has also shown high accuracy in predicting dysphonia (Heman-Ackah et al., 2003). Aerodynamic measures of mean airflow rate, subglottic pressure, and phonation threshold pressure that have shown effectiveness in detecting vocal changes (Chang & Karnell, 2004; Desjardins et al., 2017; Solomon & DiMattia, 2000) were included as well. Comparing the reliability of these conventional measures with that of RFF could help determine the clinical usefulness of RFF relative to those measures that are already in use in clinical practice. In addition, Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V; American Speech-Language-Hearing Association, 2009) was included since perceptual evaluation is routinely used in clinics (Carding et al., 2009; De Bodt et al., 1996). We hypothesized that RFF would have comparable but slightly lower reliability than the conventional acoustic and aerodynamic measures because most of them are associated with perceived overall dysphonia, often caused by structural changes of the vocal folds, whereas RFF is thought to reflect strain related to laryngeal tension, a functional aspect of vocal production. We hypothesized that RFF would have higher reliability than CAPE-V ratings because of the subjective nature of perceptual evaluation.

We also examined the effect of speaker intensity on RFF reliability. Controlling speaker intensity levels in different recording sessions produces more reliable results in acoustic and aerodynamic measurements (Lee, Stemple, & Kizer, 1999), since different intensity levels between recording sessions can lead to higher variability. Thus, the reliability of RFF stimuli produced in soft, comfortable, and loud voices was compared with the hypothesis that different loudness levels would result in different reliability. Although Park and Stepp (2019) examined the within-subject standard deviation of different loudness levels and found no effect of loudness, the within-subject standard deviations were estimated from nine RFF values obtained from one recording session, not different days. In this study, we examined the reliabilities of RFF mean values produced at different loudness levels over five consecutive days. Since a wider range of loudness may be produced relative to the ranges elicited by instructions for comfortable and soft voices, we hypothesized that using a loud voice would result in lower reliability of RFF due to less consistent sound pressure levels over the experimental week than when using soft and comfortable voices.

Methods

Participants

Thirty-two healthy participants aged 18 to 33 years (16 women, 16 men; M = 22.5, SD = 4.1) were recruited and reported no prior history of speech, language, and hearing disorders. Participants also reported a small-to-medium amount of daily voice use, classified as low voice users. The low voice users were recruited in order to minimize possible voice changes during the week-long course of the study. Participants completed the Voice Handicap Index (VHI) and the Reflux Symptom Index (RSI). VHI and RSI, both self-rating questionnaires, subjectively evaluate the degree of voice handicap and laryngopharyngeal reflux, respectively (Belafsky, Postma, & Koufman, 2002a; Jacobson et al., 1997). Participants completed the questionnaires on their first study visit and scored within normal ranges except for one participant who scored higher than the cut-off score for VHI and was excluded. In addition, all participants passed a hearing screening with 25 dB HL pure tones at 125, 250, 500, 1000, 2000, 4000, and 8000 Hz (American Speech-Language-Hearing Association, 2005). The participants provided written consent prior to participation, in compliance with the Boston University Institutional Review Board. Three additional participants were excluded midway through the experimental week due to sickness. Thus, a total of four participants were excluded during the course of the study resulting in 28 participants.

Experimental Tasks

Participants visited the lab over a period of five consecutive days, Monday through Friday of the same week. They were asked to come at a similar time each day, at least three hours after waking such that their voice conditions would be as consistent as possible during each visit, although there were some cases in which participants had to schedule different times. In order to confirm that their voice conditions were consistent throughout the week, we conducted a detailed voice-use interview and a self-administered vocal rating during each visit.

The chronological order of experimental tasks during each visit is outlined in Table 4.1. RFF stimuli recording was performed after the conventional acoustic measurements, to ensure that the loud phonation in the RFF protocol did not have an impact on participants' voices, and thus affect the results of later recordings. Aerodynamic measurement, which also had a loud voice condition, was placed after the conventional acoustic and RFF recordings for the same reason.

Table 4.1: The timeline of experimental tasks during each visit

Experimental Task	Time
Voice Interview	2 min
Vocal Self-Rating (IPSV)	3 min
Training of RFF stimuli and recording set up	5 min
Conventional acoustic measurement	3 min
RFF stimuli	5 min
Aerodynamic measurement	5 min

Voice interview and vocal self-rating

During every visit, the experimenter interviewed participants with detailed questions about their daily voice use, voice condition, and wake-up time to document any behaviors that might affect their current voice condition. Participants also performed a vocal self-rating task called the Inability to Produce Soft Voice (IPSV), which has shown reliability in tracking teachers' voice changes (Halpern, Spielman, Hunter, & Titze, 2009). IPSV consists of four different tasks, which participants are asked to perform as softly as possible: 1) sustaining /i/ for 5 seconds on a comfortable pitch, 2) gliding on /i/ from a low to a high pitch, 3) saying a train of /i/ production in staccato with a high pitch, and 4) singing a few bars of "Happy Birthday" in a high pitch. After these tasks, participants rated their own score on a scale of 1 (no problem) to 10 (extreme problem).

Conventional acoustic measurements

Participants were equipped with a head-mounted microphone (Shure WH20) in a sound-treated booth. Their voice was recorded with SONAR Artist (Cakewalk, Chicago, IL) using a 44.1 kHz sampling rate. Participants produced sustained /ɑ/ vowels for 3 – 5 seconds in one exhalation with a constant pitch and loudness. We asked participants to produce the sustained utterance nine times to match the sample numbers with RFF productions to more accurately compare the variabilities of conventional measures and RFF. Participants were also asked to read the first paragraph of the "Rainbow Passage" (Fairbanks, 1960).

RFF stimuli recording

RFF stimuli were recorded with the same recording equipment. The RFF short utterance stimulus, /əfə/, was selected because it resulted in lower RFF within-subject standard deviation compared to other stimuli in previous studies (Lien et al., 2014; Park & Stepp, 2019). In addition, participants were asked to produce RFF stimuli with equal stress using similar pitch and loudness in both vowels since this has also been shown to decrease within-subject standard deviations (Park & Stepp, 2019).

RFF stimuli were produced with comfortable, soft, and loud voices to test the effect of loudness on RFF reliability. The degree of softness and loudness were not assigned a specific sound pressure level, but rather were determined by participants, similar to clinical instructions for loud voice (Patel et al., 2018). However, the sound pressure level was estimated offline by calibrating the acoustic waveforms collected, using an electrolarynx and sound pressure level meter. The participants were asked not to whisper for the soft voice, since accurate estimation of fundamental frequency is difficult with whispered voice. Each stimulus under a given loudness condition was produced nine times, as RFF has previously been shown to correlate better with auditory-perceptual judgments when averaged over at least six productions (Eadie & Stepp, 2013).

The experimenter instructed the participants on how to produce stimuli for both conventional acoustic measures and RFF before each recording session (the instruction included practice of each task except for the reading of "the Rainbow Passage"). Participants also listened to sample recordings of RFF stimuli to learn how to produce RFF stimuli with the equal stress; opposite gender recordings were played to avoid pitch mimicking. During the recording, when participants occasionally pronounced the stimuli with the wrong stress, the experimenter asked them to produce the stimulus again. The experimenter also asked the participants to repeat any stimulus produced with clear glottalization or with extremely short vowel productions, as these do not allow for calculation of RFF.

Aerodynamic measurement

Mean airflow rate (ml/s) and intraoral estimates of subglottic pressure (Psub; cm H_2O) were measured with a Phonatory Aerodynamic System (PAS; Model 6600, PENTAX Medical, Montvale, New Jersey). Participants wore a face mask, which fit over their nose and mouth, and placed a small catheter inside their mouth. They produced six trains of five /pɑ/s with a comfortable loudness level and six trains with a loud level. For phonation threshold pressure (PTP), participants produced continuous /pɑ/s in one exhaling breath with decreasing loudness until their voice stopped. They performed this protocol three times at a comfortable pitch. They performed the same protocol three times each at 3 ST below and above their comfortable f_o (Enflo, Sundberg, Romedahl, & McAllister, 2013). Each individuals' comfortable f_o was determined from each participant's recording of sustained /ɑ/ using Praat acoustic software (Boersma & Weenink, 2016). Sample synthetic voices at each participant's comfortable f_o and 3 ST below and above their comfortable f_o were generated and played with a Madde synthesizer (Granqvist, 2010). We played each sample to participants, and they produced the /pa/s with the f_o they heard. The participants were given instructions and practiced before each task.

Data Analysis

RFF was estimated using an automated RFF estimation algorithm in MATLAB (Ver. R2015b, MathWorks, Natick, MA). The automated RFF algorithm identified the voiced cycles before and after the consonants, estimated the periods and the instantaneous f_o for each cycle, and calculated RFF with the RFF equation ST= 39.86 × $\log_{10}(f_o/\text{reference } f_o)$. The algorithm automatically rejected any recorded stimuli that lack periodic cycles or contain glottalization, which can affect the accuracy of the f_o estimation. We focused on offset cycle 10 and onset cycle 1 RFF (Figure 4.1) values because these two cycles best represent the degree of vocal hyperfunction (Lien et al., 2015b; Stepp et al., 2010a). The mean offset cycle 10 and onset cycle 1 RFF values were calculated with the nine RFF values estimated (less in the case of any rejections) from the nine recordings of each stimulus condition.

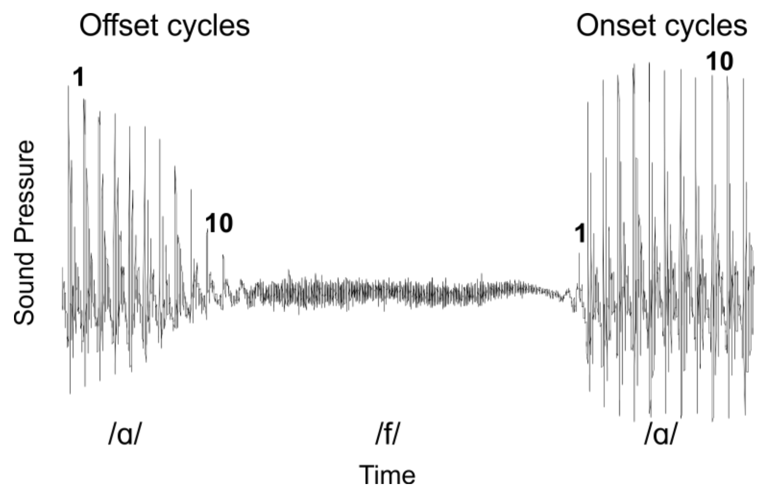

Figure 4.1: An example acoustic waveform of a sonorant-voiceless consonant-sonorant. The offset cycles and onset cycles 1 and 10 are labeled.

Conventional acoustic measures were obtained using Praat acoustic analysis software (Version 6.0.21). Praat's built-in function, "voice report," was used to estimate acoustic measures for a selected stable one-second segment of the recorded sustained vowel /ɑ/ samples. Jitter, shimmer, and HNR were obtained from the nine sustained /ɑ/ productions, and the nine values of each measure on a given experimental day were averaged to obtain the mean value of each individual parameter for the day. CPPS was also obtained from a Praat built-in function developed by Maryn and Weenink (2015), following the protocol in Watts et al. (2017). Nine CPPS values were obtained from each of the nine one-second sustained /ɑ/ recordings and averaged into $CPPS_{vowel}$ for each day. Mean vocal f_o and $CPPS_{sentence}$ were calculated from the recordings of the first and second sentences of the "Rainbow Passage" (Fairbanks, 1960).

Aerodynamic data were analyzed in MATLAB to obtain mean airflow rate and subglottic pressure (see Appendix 2 for the analysis scripts). Airflow rate and intraoral air pressure signals were extracted from the raw aerodynamic data from the PAS system. In order to be considered a valid measurement, the airflow signal had to contain a steady-state, horizontal portion during the vowel and the air pressure signal had to have a flat peak during /p/ stop consonant (Patel et al., 2018). From the six /pɑ/ trains of each airflow rate and air pressure signals, three stable /pɑ/ trains were chosen. In each selected /pɑ/ train, measures were obtained from the middle three syllables, discarding first and last syllables. In airflow rate signals, the horizontal portions during the vowels were selected, and nine selections from three /pa/ trains (three middle vowels × three /pa/ trains) were averaged into mean airflow rate for each day. In intraoral air pressure signals, the interpolation method (Patel et al., 2018) was used to estimate the subglottic pressure over the vowels, and nine estimate subglottic pressures (three middle vowels × three /pa/ trains) were averaged into Psub for each day. Because oral estimates of subglottal pressure are known to be under-estimates if the pressure in the oral cavity is not equalized (Fryd, Van Stan, Hillman, & Mehta, 2016), we decided to eliminate more 'peaky' pressure waveforms by setting a threshold value to the 5% variation of each pressure peak, described in McKenna et al. (2016), and eliminating pressure waveforms with variation above this threshold value from the estimation. We eliminated approximately 10% of all pressure waveforms.

However, we did not see differences in the results based on whether these waveforms were included, consistent with the finding of (Fryd et al., 2016).

Phonation threshold pressure (PTP) was estimated in MATLAB (see Figure 4.2). PTP was estimated as the average of the two intraoral pressure peaks that were surrounding where the acoustic waveform indicated that the voice of the participant stopped or turned into a whisper (Enflo et al., 2013). This selection was performed on all three trials at each f_o, and the selected values were averaged across f_o to represent the PTP for that day.

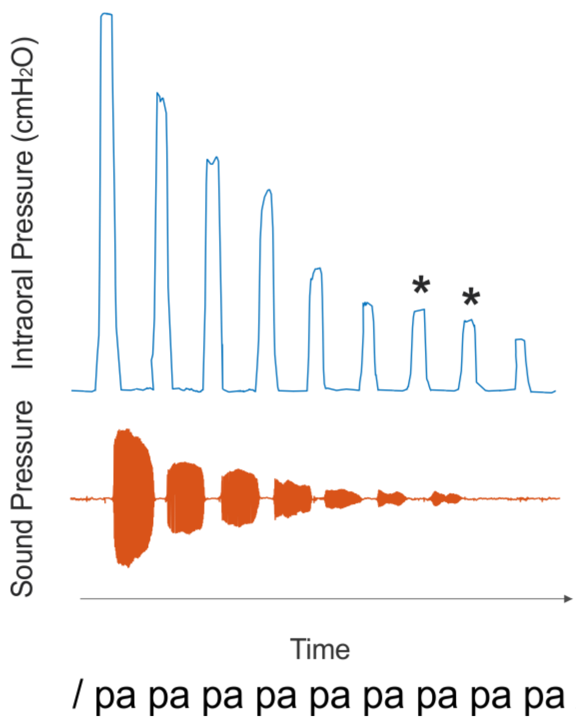

Figure 4.2: Example of phonation threshold pressure estimation. The top panel is the intraoral pressure. The bottom panel is the associated acoustic waveform. Two peaks (marked with asterisks) were selected because they surrounded the point at which the voice stopped (phonation threshold).

CAPE-V was performed by a voice-experienced speech pathologist. The total number of the ratings were 140 (28 participants × 5 days), and the order of the ratings was pseudorandomized across both participants and days. The listening samples consisted of the three 1-s /ɑ/s and the first two sentences of the recording of the "Rainbow Passage" from the acoustic recordings of each participant on each day. After listening to each sample, the rater completed the standardized CAPE-V form that contained 100-mm Visual Analog scales (Kempster et al., 2009). Since all of our participants had healthy voices, we excluded pitch and loudness parameters in the CAPE-V and examined the four parameters of voice quality: overall severity, roughness, breathiness, and strain. Because of the large number of ratings, the rater completed the ratings over five different sessions. For intrarater reliability evaluation, 20% of the ratings were repeated by the same rater at a later date, and for interrater reliability evaluation, another speech-language pathologist with experience in voice also performed 20% of the ratings.

Statistical Analysis

To compare measurement variabilities within healthy speakers, intraclass correlation coefficients (ICCs) were used in previous studies (Awan et al., 2013; Bough et al., 1996; Carding et al., 2004; Leong et al., 2013). Thus, ICC values and their 95% confidence interval for each acoustic and aerodynamic measure and each CAPE-V parameter were calculated using SPSS (Version 24, SPSS Inc., Chicago, IL) based on single measures, absolute-agreement, and the 2-way mixed-effects model (McGraw & Wong, 1996). Although there are no standards to interpret ICCs, ICCs below 0.5 have been suggested as indicative of poor reliability, 0.5 – 0.75 as moderately reliable, 0.75 – 0.9 as having good reliability, and above 0.9 as having excellent reliability (Koo & Li, 2016). In order to assess intrarater reliabilities for the CAPE-V parameters, Pearson's correlation coefficients (*r)* were calculated. Interrater reliabilities for the CAPE-V parameters was calculated with ICCs as the two-way mixed-effects model for the consistency of single measurements for each CAPE-V parameter. Intrarater and interrater reliabilities for each parameter are presented in Table 4.2. Intrarater reliabilities were ≥ 0.64 for all CAPE-V parameters except strain (r = 0.34), and interrater reliabilities were poor ($r \leq 27$) for all parameters.

Table 4.2: Intra- and interrater reliability of CAPE-V parameters

Parameters	Intrarater Pearson's *r*	Interrater ICC
Overall Severity	0.79	0.21
Roughness	0.57	0.21
Breathiness	0.64	0.00
Strain	0.34	0.27

As a post-hoc assessment, we also examined the possible effects of the participants' time awake (time between their wake-up and the recording session) on their voices. Most of the recording sessions were scheduled at a similar time of the day during the week for each participant, but some participants had to schedule at a different time of the day or woke up much earlier or later than the other days. We suspected that being awake much longer and possibly talking more before the recording session may have affected these participants' voices. We chose 10 participants whose ranges of the time awake varied by more than five hours across the experimental week, so that the sample would have sufficient variance in the associated measures to show potential associations. For each of the 10 participants, individual Pearson's correlation analysis was performed between the time awake, mean vocal f_o, offset 10 and onset 1 RFF values, and PTP from the five-day sessions. These instrumental measures were specifically chosen because of their known sensitivities to vocal loading or vocal fatigue (Chang & Karnell, 2004; Kagan & Heaton, 2017; Solomon & DiMattia, 2000; Stemple, Stanley, & Lee, 1995; Welham & Maclagan, 2003). The individual Pearson's correlation coefficients (*r*) from the 10 participants were averaged using Fisher's *z*' transformation.

Results

Mean values of all of the CAPE-V parameters are presented in Table 4.3. All of the parameters in the CAPE-V showed low mean values and thus support that participants had healthy voices. All of our participants had type 1 voices as determined by the first author.

Table 4.3: Mean values of CAPE-V parameters

Parameters	Mean (SD)
Overall Severity	4.6 (2.7)
Roughness	2.0 (2.3)
Breathiness	1.8 (1.9)
Strain	1.2 (1.6)

We obtained ICCs for offset 10 RFF and onset 1 RFF as well as for conventional acoustic, aerodynamic, and perceptual measures from the recordings of five consecutive days in order to assess the test-retest reliability of RFF. The results are presented in Figure 4.3. Both offset 10 and onset 1 RFF had moderate reliability. Excellent reliability was observed for mean vocal f_o, shimmer, and HNR, and good reliability was seen for both $CPPS_{vowel}$ and $CPPS_{speech}$. For the aerodynamic measures, both mean airflow rates measured in both comfortable and loud voice showed good reliability, PTP, moderate-to-good reliability, and both $Psub_{comfortable}$ and $Psub_{loud}$, poor-to-moderate reliability. All of the CAPE-V parameters exhibited poor reliability.

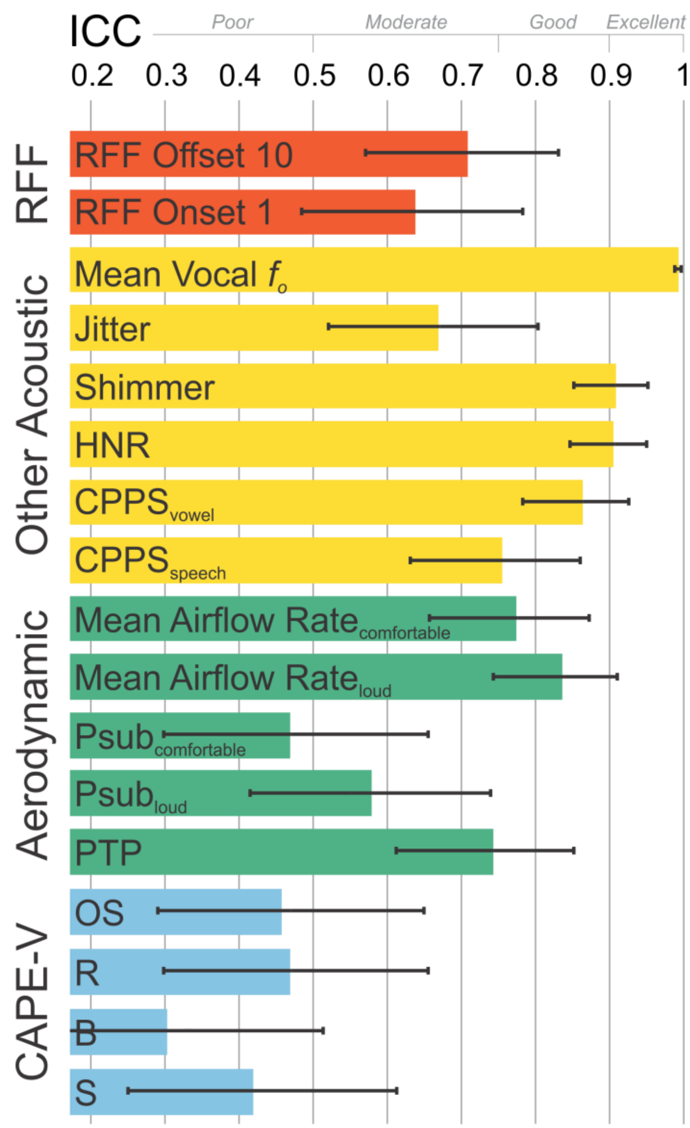

Figure 4.3: Intraclass correlation coefficient (ICC) values obtained with five measurements over five consecutive days (error bars indicates the 95% confidence intervals). ICCs below 0.5 are considered poor, 0.5-0.75 moderately reliable, 0.75 to 0.9 good reliability, and above 0.9 excellent reliability. ICCs of relative fundamental frequency (RFF) are shown as red, other acoustic measures as yellow, aerodynamic measures as green, and CAPE-V parameters as blue (HNR: harmonic-to-noise ratio, CPPS: smoothed cepstral peak prominence, Psub: subglottic pressure estimate, PTP: phonation threshold pressure, OS: overall severity, R: roughness, B: Breathiness, and S: Strain).

We also obtained the ICCs for RFF from different loudness levels (Figure 4.4) in order to compare the effects of loudness on the test-retest reliability. Analysis of soft voice recordings produced the highest ICC for onset 1 RFF, raising the ICC to ‘good’ reliability. Loud voice recordings had good reliability in offset 10 values, but poor reliability in onset 1 values.

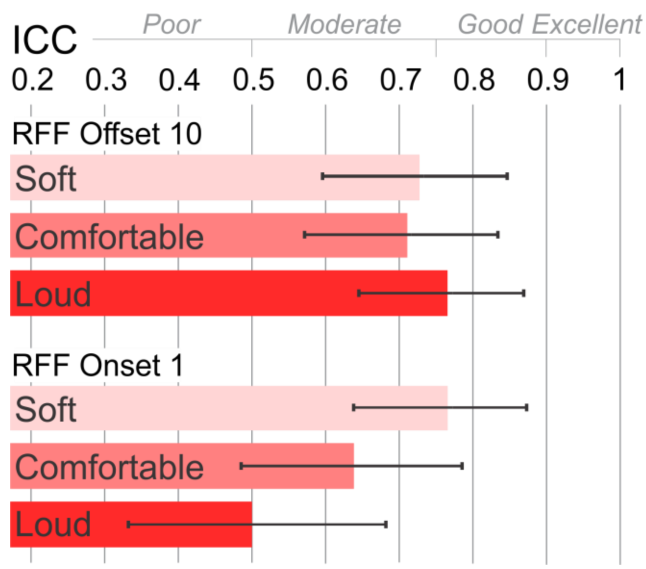

Figure 4.4: Intraclass correlation coefficients (ICCs; error bars indicate 95% confidence intervals) of relative fundamental frequency (RFF) produced with different loudness levels.

The results of the voice-related daily questionnaires, as well as the daily interviews and IPSV scores, suggest that participants’ voices did not change during the experimental week. Participants had normal VHI scores (M = 8.2, SD = 7.3) and RSI scores (M = 3.3, SD = 3.2). Participants did not report any significant voice use with the exception of two participants who reported yelling during a sporting event. Most participants also did not report any discomfort in the throat, with the exception of a few participants reporting slight laryngeal discomfort during some of the recording sessions. The mean IPSV value collected from participants was 2.3, and the mean within-subject standard deviation of IPSV was 0.6.

The log of participants' wake-up times and recording session times indicated that some participants had substantial differences in their wake-up or recording session times over five days. The average range of the time awake before the recording sessions (duration between wake-up and recording session) over the five days was 3.5 hr, and 10 participants had ranges of their time awake that were greater than 5 hours. The possible effects of this variability in time awake before the recording session on participants' voices were evaluated using averaged individual Pearson's correlation coefficients (Table 4.4). The only measures with an averaged correlation coefficient greater than ±0.5 (moderate correlation) with the time awake were onset 1 RFF (-0.50) and mean vocal f_o (0.55).

Table 4.4: Averaged individual correlation coefficients (*r*) and standard errors between the time awake, offset 10 and onset 1 relative fundamental frequency (RFF), mean vocal f_o, and phonation threshold pressure (PTP) among 10 participants whose range of the time awake during the experimental week was over 5 hours

Measures	Time awake	Offset 10 RFF	Onset 1 RFF	Mean vocal f_o	PTP
Time awake	1.00	-0.44 (0.21)	-0.50 (0.16)	0.55 (0.22)	-0.09 (0.21)
Offset 10 RFF		1.00	0.39 (0.18)	-0.44 (0.21)	-0.21 (0.21)
Onset 1 RFF			1.00	-0.41 (0.17)	0.09 (0.19)
Mean vocal f_o				1.00	-0.11 (0.22)
PTP					1.00

Discussion

ICC for RFF versus Other Measures

The test-retest reliabilities of RFF and conventional acoustic, aerodynamic, and perceptual measures were assessed using ICC values (Figure 4.3). Both offset 10 and onset 1 RFF were found to be moderately reliable, which were lower than the reliabilities of mean vocal f_o, shimmer, HNR, $\text{CPPS}_{\text{vowel}}$, mean airflow $\text{rate}_{\text{comfortable}}$, and mean airflow $\text{rate}_{\text{loud}}$. The lower ICCs for RFF compared to ICCs for these conventional measures were expected because RFF measures are thought to correlate with laryngeal tension, whereas the most of the measures that showed higher reliabilities than RFF, except mean vocal $f_{o,}$ correlate with overall dysphonia severity; thus, RFF is likely to be more sensitive to day-to-day functional variation than these measures. The ICCs for RFF were similar to the ICC for PTP possibly because PTP, also reflects day-to-day variations in vocal fold vibratory characteristics and vocal fatigue (Chan & Titze, 2006; Solomon & DiMattia, 2000).

Another potential reason that the test-retest reliability of RFF was lower than of many conventional instrumental measures could be the difference in the speech samples. RFF is measured in a continuous speech context, but the most of the conventional measures are measured during sustained phonation. Sustained phonation may result in better test-retest reliability than continuous speech because it is free from fluctuations in frequency and amplitude due to prosody and is not affected by speech rate (Maryn et al., 2010). Leong et al. (2013) observed higher ICC values in voice quality measures obtained from sustained vowels compared to the measures obtained from sentence stimuli. Similarly, in our study, $\text{CPPS}_{\text{speech}}$, measured from sentence stimuli, had lower test-retest reliability than $\text{CPPS}_{\text{vowel}}$, which was measured from a sustained /ɑ/ phonation. The ICC for $CPPS_{speech}$, in fact, was similar to the ICC value for RFF. The lower test-retest reliability of RFF compared to other conventional instrumental measures may be, in part, the result of differences in stimuli.

The test-retest reliability of RFF was higher than the test-retest reliabilities of CAPE-V parameters, as expected. The test-retest reliabilities of CAPE-V parameters were poor (below 0.5), and these results might be due to low intra- and interrater reliabilities. Intrarater reliabilities for roughness and breathiness (Table 4.2) were below the averaged intrarater reliabities (roughness: r = 0.77, breathiness: r = 0.82) obtained from 21 voice-trained speech-language pathologists (Zraick et al., 2011). In addition, both intrarater reliabilities of strain in our study (Table 4.2) and in the previous study (r = 0.35; Zraick et al., 2011) were poor-to-moderate, despite that the speech-language pathologist was an expert in voice with experience in administering the CAPE-V, which has shown to increase intrarater reliability (Eadie & Baylor, 2006). Auditory fatigue may be one of the reasons for poor intrarater reliabilities, although we aimed to minimize the potential for fatigue by dividing the rating sessions into five different days. Having only one rater might have also resulted in poor test-retest reliabilities (five-day data), since a previous study with seven raters resulted in good test-retest reliabilities for a similar perceptual rating, the GRBAS scale (Webb et al., 2004). However, in typical clinical settings, only one speech-language pathologist is likely to perform the CAPE-V, and thus, our results might be a more accurate reflection of the actual test-retest reliability of CAPE-V in practice. Nevertheless, the test-retest reliabilities of CAPE-V performed by one rater is likely to be heavily dependent on who the rater is, as suggested by the poor interrater reliabilities (Table 4.2). The interrater reliabilities of all CAPE-V parameters in our study were lower than published values with 21 raters (Zraick et al., 2011). These results confirm the need for more objective measures, in addition to perceptual measures, in clinical settings (Hillman et al., 1990a). The higher reliabilities of RFF compared to the reliabilities of CAPE-V, especially the strain parameter, support the potential clinical utility of RFF, since perceptual evaluation is widely used for assessing hyperfunctional voice disorders and evaluating treatment outcomes (Carding et al., 2009).

Soft Voice in RFF Test-Retest Reliability

We included different loudness levels in RFF stimuli recordings to see if loudness would affect the test-retest reliability of RFF. We found that loudness did not affect the reliability of offset 10 values, but the use of loud voice decreased the reliability for onset 1 RFF, whereas the soft voice increased the reliability for onset 1 values (Figure 4.4). This is somewhat surprising because soft voice has been associated with both increased values and variabilities of jitter and shimmer, which may reflect increased variability in vocal fold vibratory characteristics (Brockmann et al., 2008). Soft voice has been also associated with vocal fatigue as it is used in tasks for IPSV and PTP (Chan & Titze, 2006; Hunter, 2011; Solomon & DiMattia, 2000). We also previously found that RFF values produced with a soft voice had high between-subject variability(Park & Stepp, 2019). One possible explanation for the higher ICC value for soft voice in the current study may be that participants used more consistent vocal effort to produce soft voice than comfortable and loud voice, and RFF may be sensitive to this vocal effort.

Comparison to the Literature

The observed ICCs for most of the measures in our study were generally higher compared to those reported in previous studies (Table 4.5), possibly suggesting that the participants' voices were less varied during the experiment in our study. Four previous studies have documented ICC values for the test-retest reliabilities of the measures included in the current study. Bough et al. (1996) examined the test-retest reliabilities of mean vocal f_o, jitter, shimmer, and HNR over 15 test sessions at consistent times of the day. Leong et al. (2013) evaluated the test-retest reliabilities of mean vocal f_o, jitter, shimmer, and CPPS over 10 sessions. Carding et al. (2004) evaluated the test-retest reliabilities of jitter, shimmer, and NHR in 45 participants over 2 hours. Awan et al. (2013) examined the test-retest reliabilities of aerodynamic measures over 2 sessions. The results from the previous studies were within the 95% confidence intervals of the results from the current study for jitter, HNR, $CPPS_{speech}$, and mean airflow rate (bolded; Table 4.5). For acoustic measures in general, we observed higher reliability in our study, and we suspect that having sessions over consecutive days may have resulted in more consistent vocal conditions between the recording sessions compared to other studies. Although there are also other factors that could have affected the reliabilities, including recording environment (Deliyski, Shaw, Evans, & Vesselinov, 2006), gender distributions, the number of participants and sessions, the higher reliabilities of the acoustic measures in our study may suggest that the participants' voices were more consistent during the experimental week, which may be a better environment to assess reliabilities of instrumental measures.

Table 4.5: Comparison of test-retest reliability ICC values to the literature

Category	Measures	Current Study ICC	Previous Studies' ICCs
Acoustic Measures	Mean vocal f_o	0.99	0.32 (female), 0.60 (male)[1]
	Jitter	**0.67**	0.50 (female), 0.91 (male)[1] 0.32[2], **0.73**[3]
	Shimmer	0.91	0.56 (female), 0.53 (male)[1] 0.67[2], 0.55[3]
	HNR	0.91	0.23 (female), 0.05 (male)[1, *] **0.93**[2], 0.68[3]
	$CPPS_{speech}$	0.76	0.45 (female), **0.80 (male)**[1]
Aerodynamic Measures	Mean Airflow Rate	0.78 (comfortable) 0.84 (loud)	**0.67 (comfortable)**[4]
	Subglottic Pressure	0.47 (comfortable) 0.58 (loud)	0.74 (comfortable)[4]

Bolded are the ICC values from the previous studies that were within 95% confidence intervals of the ICC values from the current study. [1]Leong et al. (2013); [2]Bough et al. (1996); [3]Carding et al. (2004), [4]Awan et al. (2013); *Noise-to-Harmonic Ratio

On the other hand, the ICC for Psub was generally lower in our study. We suspect that low ICCs of Psub may have been from varied intensity producing Psub tasks and thus normalizied Psub with dB SPL as described in (Espinoza et al., 2017) and re-calculated ICCs; however, we obtained similar ICCs using normalization (comfortable: 0.51; loud: 0.61) suggesting that the low ICCs of Psub are not due to intensity variability. Another possible reason for the low reliability may be that the Psub measurement contained 'peaky' pressure waveforms which are known to result in underestimation (Holmberg et al., 1984). However, as mentioned in the Methods, we eliminated some particularly 'peaky' pressure waveforms and found no differences in ICC values; there is still a possibility that our data include ‘peaky’ pressure waveforms that could have resulted in underestimation. Finally, we found that many of our participants produced /pa/ trains at a slower rate than the recommended rate of 1.5 – 2 /pa/s per second, which might have led to more ‘peaky’ pressure waveforms (Holmberg et al., 1984).

Time Awake vs. the Outcome Measures

We examined the times between the participants’ wake-up and the recording sessions because we suspected that this time difference may affect the voice condition during the experimental week. We found that this time awake had moderate correlations with onset 1 RFF (r = -0.50) and mean vocal f_o (r = 0.55). We hypothesize that, when the participants were awake longer before the session, they were likely to have talked more prior to the session, and the increased vocalization prior to the session might have increased their baseline laryngeal tension; both RFF and mean vocal f_o may reflect this change. This finding is similar to the finding of Garrett and Healey (1987), who measured participants’ voices three times during a day and found that male participants showed a significant increase in their mean vocal f_o at the later times of the day. This increase in mean vocal f_o was consistent with the findings of Stemple et al. (1995), who observed significant increases in mean vocal f_o in both sustained vowel and reading samples after vocal loading tasks (Stemple et al., 1995). However, previous studies have mixed findings about the effect of vocal loading tasks on RFF (Fujiki, Chapleau, Sundarrajan, McKenna, & Sivasankar, 2017; Kagan & Heaton, 2017), and PTP, a sensitive measure to vocal loading tasks (Chang & Karnell, 2004; Solomon & DiMattia, 2000), was not correlated with the time awake (r = -0.09) in the current study. Thus, the time awake may not modulate only the amount of vocalization before the sessions, but it may have influenced other possible factors that could have affected RFF and mean vocal f_o, but not PTP. Because the time awake was shown to be correlated with RFF, the difference in the range of the time awake during the experimental week may have resulted in the moderate test-retest reliabilities of RFF. In contrast, the test-retest reliability of mean vocal f_o was excellent. Mean vocal f_o may be less sensitive to vocal change due to the time awake, thus RFF, which is more likely to be related to the baseline laryngeal tension (Lien et al., 2015b; McKenna et al., 2016).

Limitations and Future Directions

We examined the test-retest reliability of clinical voice outcome measures, recruiting individuals who reported no history of voice disorders. However, they were not examined by a laryngologist. If they had some degree of vocal hyperfunction or other voice disorder, ICC results and the acoustic measures would not represent solely individuals with healthy voices. The test-retest reliabilities of acoustic measures have shown to be less reliable in dysphonic voices compared to healthy voices (Carding et al., 2004). Individuals with dysphonic voices would have more irregular voice conditions than individuals with healthy voices; thus, test-retest reliabilities of the measures among dysphonic voices should be examined in the future. In addition, although we asked the participants about their voice discomfort and usages, we did not ask them about changes in their emotional stress, which might have affected their voices and influenced the results of this study (Helou, Rosen, Wang, & Verdolini Abbott, 2018).

Although the moderate reliability of RFF may have been related to actual changes in laryngeal function, it may also reflect the actual reliability of RFF and its current estimation process. The automated RFF algorithm used in the current study was developed as an alternative to the manual RFF estimation process, which is subjective and time-consuming. However, the current algorithm shows small differences in estimated RFF values compared to manual estimation (Lien et al., 2017). Improvements in automated RFF estimation to better detect offset and onset cycles may enhance the reliability of RFF for clinical use in the future. Another possibility to increase the reliability of RFF may be online-monitoring of sound pressure level while producing RFF stimuli, since controlling intensity has been shown to increase the reliability of acoustic and aerodynamic measures (Lee et al., 1999). However, our previous work did not indicate that mean RFF values were significantly impacted by sound pressure level (Park & Stepp, 2019).

Conclusion

From recording individuals with healthy voices for five consecutive days, we found that RFF exhibited moderate test-retest reliability, which was slightly lower or comparable to commonly used acoustic and aerodynamic measures. We suspect that our finding of moderate reliability may reflect, to some degree, actual changes in individuals' vocal function or tension, since RFF was affected by the time awake before the recording sessions. RFF was found to be more reliable than CAPE-V parameters, as assessed performed by a voice-trained speech pathologist. In addition, RFF measured from soft voice recordings showed better reliability than those measured from comfortable voice. For future studies, sensitivity-to-change and minimal clinically important differences should be studied to further evaluate the appropriateness of RFF for clinical use.

Chapter 5: Discussion

Summary and Implications of the Dissertation

This dissertation aimed to evaluate the translational potential of relative fundamental frequency (RFF). The three studies in this dissertation successfully addressed and filled some of the gaps in previous research on RFF (Figure 5.1) and thus further supported the potential of RFF as a clinical measure for assessing vocal hyperfunction. The following paragraphs summarize the most important findings and implications of the three studies presented in this dissertation.

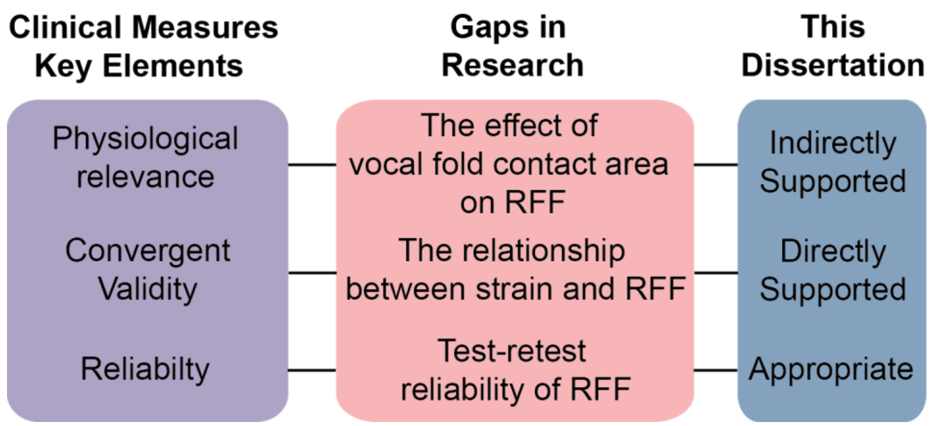

Figure 5.1: An organizational chart that summarizes the gaps in research for RFF to be applied clinically and the results of this dissertation.

The Effect of the Degree of Adduction on Relative Fundamental Frequency

While evaluating the previous hypotheses on possible physiological mechanisms underlying RFF, the high-speed videoendoscopy (HSV) study in Chapter 2 indirectly supported the effect of the degree of adduction on RFF. We recorded individuals producing intervocalic voiceless fricatives and stops with HSV to observe if possible differences in vocal fold kinematics between voiceless fricatives and stops would be related to differences in RFF. We hypothesized that onset RFF would be lower for stops relative to fricatives due to the combined effects of less adducted and stiffer vocal folds in stop production. Our hypothesis was supported by the lower onset RFF, the greater glottal angle at voicing onset, and the greater kinematic estimate of laryngeal stiffness obtained from stops than those obtained from fricatives. Offset RFF was hypothesized and observed as not different due to the combined effect of stiffer and less abducted vocal folds during stop production. This potential positive relationship between the degree of adduction and RFF is in line with the hypothesis from Watson (1998) on the negative effect of decreased vocal fold contact area on RFF.

If the decreased vocal fold contact area actually decreases RFF, we suggest expanding previous hypotheses on physiological factors for RFF in healthy voices and vocal hyperfunction (Stepp et al., 2011b) to include an additional factor that was also suggested by Heller Murray et al. (2017), *adduction*. Although abduction was one of the factors in previous hypotheses, along with increased cricothyroid (CT) muscle activation and airflow rate surrounding voiceless consonants, no hypothesis was formulated on the effect of adduction on RFF (Stepp et al., 2011b; Watson, 1998). The offset RFF pattern was thought to be the result of the combined effects of increased CT muscle activity and abduction, both suggested to be devoicing mechanisms. The onset RFF pattern was thought to be the result of the combined effects of increased CT muscle activity and increased airflow rate from obstruent production. In addition to these factors, we suggest the degree of adduction should be included as an additional physiological factor for RFF (See Figure 5.2; the aerodynamic factor was not included due to its conflicting hypothesis with the hypothesis on adduction, which is explained later) based on the observation that onset RFF was lower along with less adducted vocal folds for stops relative to fricatives. The degree of adduction is hypothesized to have a positive effect on RFF; less adducted vocal folds would decrease RFF, because of the negative effect of decreased vocal fold contact area on RFF, hypothesized by Watson (1998). However, the effect of adduction on RFF is expected to be much smaller than the effect of abduction since adduction is usually completed in a shorter duration than abduction, which would allow less time for decreased vocal fold contact area to affect RFF.

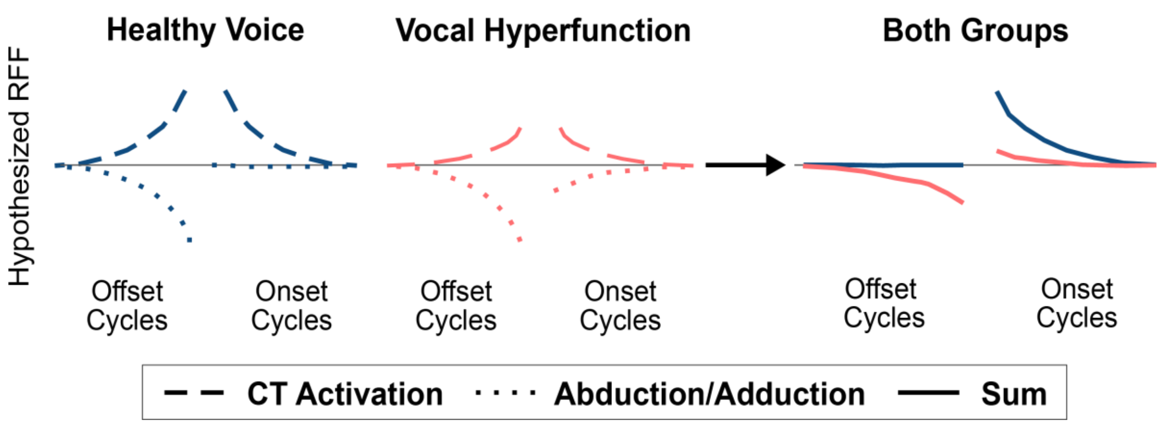

Figure 5.2 Schematics of hypothesized physiological mechanisms for RFF in individuals with healthy voices (blue) and vocal hyperfunction (pink). The left panels display the hypothesized effect of each factor on RFF separately for each group, whereas the right panel display summed RFF values for both groups.

If adduction is a physiological factor for RFF, we speculate that the degree of adduction also contributes to RFF patterns unique to individuals with vocal hyperfunction (See Figure 5.2). The lower offset and onset RFF values in individuals with vocal hyperfunction relative to typical speakers have only been attributed to the decreased effects of CT muscle activity on the vocal folds due to a ceiling effect from already increased baseline laryngeal tension in individuals with vocal hyperfunction. However, individuals with vocal hyperfunction may also have less adducted vocal folds at voicing onset than individuals with typical voices. Previous findings suggest that there may be possible increases in glottal angles and durations of voicing onset due to glottal insufficiency in individuals with both phonotraumatic and non-phonotraumatic vocal hyperfunction (Galindo et al., 2017; Woo, 2017). Individuals with non-phonotraumatic, also known as non-adducted, vocal hyperfunction usually do not achieve complete vocal fold closure in voice production (Hillman et al., 1989). Individuals with phonotraumatic vocal hyperfunction may also have less adducted vocal folds at voicing onset, although they may eventually achieve complete vocal fold closure from compensation for glottal insufficiency. A computational modeling study demonstrated that phonotraumatic vocal hyperfunction could be modeled as a compensation mechanism applied to the vocal fold model with incomplete glottal closure (Galindo et al., 2017). Galindo et al. (2017) observed that in order for less adducted vocal folds to achieve the same level of vocal output as more adducted vocal folds, vocal fold contact force would increase, which increases the risk of phonotrauma. Woo (2017) found that individuals with voice disorders related to vocal hyperfunction required a longer time to achieve steady voicing after voicing onset from HSV images. Thus, less adducted vocal folds at voicing onset in individuals with vocal hyperfunction, in addition to decreasing the effect of the CT muscles, may decrease their onset RFF.

The hypothesized positive effect of adduction on RFF is in contrast with the previous hypotheses on the effects of increased airflow rate and adduction on RFF (Heller Murray et al., 2017; Stepp et al., 2011b). Increased airflow rate was hypothesized to increase RFF due to the increased Bernoulli's force, which may increase the speed of vocal fold closure in a vibratory cycle (Ladefoged, 1967; Stepp et al., 2011b). The increased airflow rate, on the other hand, may result in the decreased degree of adduction at voicing onset, which was suspected to result in lower onset RFF in stops than in fricatives in Chapter 2. Because the decreased vocal fold adduction is hypothesized to decrease onset RFF in this dissertation, our hypothesis suggests a negative relationship between airflow rate, which may decrease the degree of adduction at voicing onset, and RFF. This hypothesis on the negative effect of airflow rate on RFF is in contrast with the previous hypothesis on the positive effect of airflow on RFF. Our hypothesis of the effect of adduction on RFF is also in contrast with the hypothesis of the effect of adduction on RFF suggested by Heller Murray et al. (2017). The authors suspected that their findings of higher onset 1 RFF values in non-phonotraumatic vocal hyperfunction relative to phonotraumatic vocal hyperfunction may be due to less adducted vocal folds in non-phonotraumatic vocal hyperfunction. Less adducted vocal folds were suspected to result in higher airflow rate, and higher airflow rate was suspected to result in higher onset RFF in non-phonotraumatic vocal hyperfunction relative to phonotraumatic vocal hyperfunction (Heller Murray et al., 2017). However, Heller Murray et al. (2017) did not measure vocal fold kinematics during the production of RFF stimuli, so it is inconclusive whether the degree of adduction, especially at voicing onset, would be different between phonotraumatic and non-phonotraumatic vocal hyperfunction. Precisely controlled modeling experiments may elucidate the exact effect of the airflow and adduction on RFF.

The effect of adduction on RFF may be related to the differences in RFF depending on the participants' time spent awake before the recording session observed in Chapter 4.

In this test-retest reliability study, we aimed to keep participants voices consistent as much as possible, so that the assessed reliability would not be affected by actual changes in participants' voices. To achieve this aim, we recorded participants throughout five consecutive workdays. However, we observed that ten participants had more variable time spent awake before the recording sessions than the rest of the participants, and these ten participants displayed negatively correlated time awake and RFF values. One speculation about this finding was increased vocal fatigue affecting RFF. We can also speculate that, based on the findings of Chapter 2, increased fatigue in adductor laryngeal muscles may reduce the degree of glottal closure at voicing onset, which may have resulted in lower RFF recorded after spending more time awake.

Another finding of Chapter 4, the strongest test-retest reliability found in RFF obtained from soft voice, may also be attributed to the degree of adduction. The production of RFF stimuli in soft voice may have resulted in a more constant vocal fold configuration at voicing onset compared to the production of RFF stimuli in other loudness conditions since it is likely to be easier to achieve a loudness goal in soft voice production relative to other loudness conditions. Speakers control their loudness by using both laryngeal and respiratory mechanisms (Zhang, 2016b). Producing soft voice may reduce those mechanisms to control loudness and thus results in a constant vocal fold configuration at voicing onset, which may have increased test-retest reliability.

RFF as an Acoustic Factor in Strain

The perceptual study in Chapter 3 supports the convergent validity of RFF by showing a direct contribution of RFF to strain. Because RFF has been found to be lower in individuals with increased vocal effort, previous studies examined possible correlations between RFF and strain in individuals with healthy voices and vocal hyperfunction (Stepp et al., 2012), with laryngeal dystonia (Eadie & Stepp, 2013), and with healthy voices self-modulating their vocal effort (Lien et al., 2015b; McKenna & Stepp, 2018). These studies have shown weak to moderate correlations between RFF and strain. However, we suspected that other acoustic features in voice samples, in addition to RFF, might have affect these correlation results because these acoustic features were likely to differ between the samples with and without increased vocal effort. Thus, we synthetically modified RFF alone in voice samples to explore if the changes in RFF in the acoustic samples would result in changes in strain in those samples. We observed that synthetically lowering RFF in the comfortable voice samples resulted in increased strain, whereas raising RFF in the maximum vocal effort samples resulted in decreased strain.

One of the reasons for the contribution of RFF to strain may be the high prevalence of decreased RFF patterns in individuals with increased laryngeal tension and vocal effort. We may have often heard the decreased RFF pattern concurrent with other acoustic characteristics of strain, and we may associate this pattern with strain as well. This rationale implies that decreased RFF patterns are frequently present in individuals with increased laryngeal tension and vocal effort. Previous studies have shown that increases in self-modulated vocal effort cause a decrease in RFF (Lien et al., 2015b; McKenna & Stepp, 2018). Both the laryngeal and respiratory systems are thought to increase vocal effort (Hunter et al., 2020). Increased laryngeal muscle activity may decrease RFF based on the hypothesis that increased laryngeal tension may reduce the effect of CT muscle activity on

RFF (Stepp et al., 2011b). Increased activity in the respiratory system may increase both subglottic pressure and airflow rate, which may start vocal fold vibrations in a less adducted state than typical voice production, similar to the observation of less adducted vocal folds after stop productions relative to fricative production in Chapter 2. The decreased degree of vocal fold adduction may decrease RFF based on the hypothesis of this dissertation and Watson (1998). In summary, decreased RFF patterns, often present in voices with increased laryngeal tension and vocal effort, were observed to have a direct contribution to strain.

Test-Retest Reliability of RFF

The last research question of this dissertation was whether the test-retest reliability of RFF would be appropriate for clinical application. We recorded the voices of individuals with healthy voices on five consecutive workdays and found a moderate test-retest reliability of RFF, similar to some conventional voice measures, such as CPPS obtained from speech and mean airflow rate. RFF reliability was lower than the reliability of many acoustic measures such as mean vocal f_o, shimmer, HNR, and CPPS obtained from vowel, but it was higher than the subglottic pressure estimates and perceptual measures from CAPE-V. We concluded that the test-retest reliability of RFF is appropriate and comparable to those of conventional measures.

Based on the findings of the study, we also suggested that slightly lower RFF reliability than that of conventional acoustic measures may reflect the actual variation in vocal function that other acoustic and auditory-perceptual measures may not be able to indicate. Although we attempted to record each participant at a similar time of the day during the experimental week, half of the participants were recorded at more variable times than the other half and had more variable time spent awake before the recording sessions. Among the half of the participants who were recorded at more variable times, their RFF values of the five consecutive days resulted in moderate to large negative correlations with their time-awake. This negative correlation between RFF and time-awake suggests that the participants might have had greater baseline laryngeal tension or vocal effort during recording sessions at later times of the day relative to earlier times. These small variations in RFF also may also not be perceivable by listeners, inferred from the findings of Chapter 3. Although we observed a significant contribution of RFF to strain from the perceptual experiment with synthetically modified RFF, the average RFF values modified in the samples were around 3 ST, which was the difference in RFF values between comfortable and maximum vocal effort samples. The resulting changes in strain after RFF modification were around 10 on a 100 mm scale. Thus, differences in RFF from variations in vocal effort during the day, relative to differences in RFF between comfortable and maximum vocal effort, may not be noticeable. This idea is in line with the findings of Heller Murray et al. (2016), which showed that after a period of long-term high voice use, individuals displayed decreased RFF, whereas their perceived voice quality remained the same.

Clinical Applications of RFF

A part of the Diagnosis of Vocal Hyperfunction

This dissertation supports RFF's potential for assessing voice disorders related to vocal hyperfunction, along with previous studies (Eadie & Stepp, 2013; Roy et al., 2016; Stepp, 2013; Stepp et al., 2010a; Stepp et al., 2011b). The results of this dissertation further support RFF's role as an acoustic measure that can be a part of the diagnosis of vocal hyperfunction by showing a possible connection between RFF and the degree of vocal fold closure at voicing onset, strain perception, and appropriate test-retest reliability (Figure 5.1).

Many conventional measures have resulted in inconsistent findings when they were compared between individuals with healthy voices and vocal hyperfunction (Belsky et al., In Press; Mehta et al., 2015; Van Stan et al., 2020). One possible reason for these findings may be that most of the conventional measures usually indicate decreased periodicity in voice, often due to glottal insufficiency, which can be caused by both structural and functional pathologies. However, within individuals with vocal hyperfunction, the degree of glottal insufficiency can vary depending on the types of vocal hyperfunction and varying compensating mechanisms and outcomes as well (Hillman et al., 1989; Hillman et al., 1990b). Thus, any acoustic measures reflecting glottal insufficiency are likely to vary depending on the degree of vocal fold closure, and these measures may not capture vocal hyperfunctional behavior specifically. Aerodynamic measures were also observed to show different patterns among individuals with vocal hyperfunction (Gillespie et al., 2013; Hillman et al., 1989; Holmberg et al., 2003).

RFF, on the other hand, quantifies an acoustic feature on a much shorter scale than most of the conventional measures, and thus may only be affected by glottal insufficiency at voicing onset, hypothesized from the findings of Chapter 2. The glottal insufficiency at voicing onset is likely to be common in both phonotraumatic and non-phonotraumatic vocal hyperfunction (Galindo et al., 2017; Hillman et al., 1989), although individuals with phonotraumatic vocal hyperfunction may compensate and close the glottal gap completely after voicing onset. Thus, the reasons that RFF has been consistently different between individuals with healthy voices and vocal hyperfunction may include RFF possibly reflecting glottal insufficiency only at voicing onset, along with increased baseline laryngeal tension (Stepp et al., 2011b).

The results from Chapters 3 and 4 further support RFF's clinical potential for diagnosing vocal hyperfunction. The results of chapter three support decreased RFF patterns relating to increases in strain. RFF has also shown a comparable test-retest reliability with conventional measures and better test-retest reliability than perceptual ratings.

Tracking Changes in Vocal Function

In addition to RFF's utility in diagnosis, RFF may also be a good measure of voice therapy outcomes and vocal function monitoring within individuals. Previous studies have already observed significant increases in RFF after successful voice therapy sessions (Roy et al., 2016; Stepp et al., 2011b). RFF has also reflected changes in vocal effort within individuals when they modulated their vocal effort. RFF was also found to be different after a heavy voice-use period, whereas auditory-perceptual measures did not change

(Heller Murray et al., 2016). In Chapter 4, we observed that individuals showed lower RFF values when they were recorded later in the day compared to earlier in the day of the experimental week. This finding also suggests that RFF may be able to monitor small changes in vocal function within individuals.

RFF may also perform better in monitoring vocal function than other acoustic, aerodynamic, and perceptual measures. Individuals with healthy voices may keep their overall voice quality consistent throughout the day despite possible vocal fatigue and small functional changes during the day. Thus, most voice measures that reflect overall voice quality may not be able to detect underlying functional changes, whereas RFF may be able to detect those changes, possibly due to the hypothesized physiological factors for RFF. Specifically, the effect of CT muscle activation and degree of vocal fold closure at voicing onset may be affected by these small changes and will, therefore, be reflected in RFF. In addition, auditory-perceptual measures are unlikely to be sensitive to small functional changes in voice because the previous study reported no changes in perceptual measures after a high voice-use period, whereas RFF showed statistical changes (Heller Murray et al., 2016). Although we observed that RFF modification resulted in changes in the perception of strain, the degree to which RFF changes within an individual throughout a day is likely to be much smaller than the degree of RFF modification made in Chapter 3, and thus a small variation in RFF is likely to be perceptually unnoticeable.

The potential ability of RFF to track changes in vocal function may be very useful for implementing vocal hygiene as well. Decreased RFF values in individuals with healthy voices may indicate that they are using more laryngeal tension than they usually do or that their voices may be more fatigued. These indications can inform speakers to be aware of their vocal condition and exercise proper care for their voices. These practices can help keep healthy voices healthy because vocal hyperfunction can become a habit when not treated appropriately, possibly resulting in damage to the vocal folds (Hillman et al., 1989). The practice of vocal hygiene with RFF may be especially useful for heavy voice users such as performers, actors, singers, and teachers who are at high risk for developing voice disorders. Recent research has sought to develop an ambulatory monitoring device using an accelerometer to track voices throughout a day (Hillman, Heaton, Masaki, Zeitels, & Cheyne, 2006; Hunter, 2012). Ambulatory monitoring using an accelerometer can utilize RFF as one of the measures since RFF values measured from accelerometer signals have been observed to be similar to those from microphone signals (Lien, Calabrese, Michener, Murray, Van Stan, Mehta, Hillman, Noordzij, & Stepp, 2015a).

RFF may also be useful for tracking changes in vocal function in aging voices. Although we did not observe differences in RFF and vocal fold kinematics between younger and older adult groups in Chapter 2, we suspected that our older group was not old enough to observe meaningful differences. Previous studies have shown aging-related changes in laryngeal anatomy and physiology. Videolaryngoscopic examinations demonstrate that larynges show increased cases of atrophy as they age (Honjo & Isshiki, 1980). Histological exams show a decrease in vocal fold mucosa thickness over time (Rodeno et al., 1993). Morphologic changes of elastic fibers and collagen fibers in superficial lamina propria have been observed as well (Sato & Hirano, 1997; Sato et al., 2002). These changes are likely to decrease the pliability of the vocal folds. Aged laryngeal systems have also shown decreased type II muscle fibers (which contract faster than type I fibers) in the TA muscles (Rodeno et al., 1993). Additionally, these systems have less myelinated fibers in the recurrent laryngeal nerves (Tiago, Pontes, & do Brasil, 2007). These changes in the laryngeal system in aging voices are likely to affect older speakers gradually as they age, and an acoustic measure that could capture initially small changes in their voices would be a beneficial tool for older speakers. RFF may be more sensitive than conventional measures to changes related to aging, because RFF may be able to detect small functional changes that occur prior to actual changes in voice quality. RFF was also found to be lower for older adults (Watson, 1998). RFF can be developed as a tracking tool for older adults to provide them with targeted voice training to maintain their vocal health.

RFF Measurement Protocol

In order for the most effective use of RFF in the applications described above, RFF must be measured reliably. In order to achieve a test-retest reliability of RFF that is similar to the one observed in Chapter 4, we recommend the protocols from Lien et al. (2014) and Park and Stepp (2019). These studies suggested the use of uniform utterances (e.g.,/afa/) produced with a comfortable voice with equal stress on both vowels. Our study in Chapter 3 confirmed that the test-retest reliability of RFF obtained with this protocol would be appropriate. However, RFF obtained with different protocols is likely to result in different, possibly worse, reliability. For example, RFF stimuli produced with different stress types may result in lower reliability since equal stress type showed the lowest within-participant variation (Park & Stepp, 2019). In addition, since stress type has a significant effect on RFF values, recordings of RFF stimuli with different stress types within an individual will not be able to accurately track changes in RFF over time. Sentence stimuli may have more ecological validity, but its test-retest reliability would be lower than the reliability observed in Chapter 4 since RFF obtained from sentence stimuli was observed to have a higher within-participant standard deviation (Lien et al., 2014).

RFF-Loaded Sentences for the Auditory-Perceptual Evaluation of Strain

Although sentence stimuli may result in more variable RFF than uniform utterances, sentence stimuli including RFF instances may be useful in the auditory-perceptual evaluation of strain. Since we observed RFF contributing to strain in Chapter 3, sentences loaded with RFF instances may provide more acoustic cues to assess strain than other sentences and thus may increase the reliability of the evaluation. RFF-loaded sentences may also be better at perceptually differentiating between individuals with healthy and increased vocal effort than other sentences. Thus, RFF-loaded sentences can be designed specifically for assessing strain perceptually, similar to CAPE-V sentences that were designed to assess each element related to vocal function and voice quality (e.g., nasality; Kempster et al., 2009). Less aspirated sounds would be preferred for RFF-loaded sentences to increase reliability because adding aspiration noise was shown to decrease intra-rater reliability of strain in Chapter 3. RFF-loaded sentences for auditory-perceptual evaluation would be a simple way to incorporate RFF in clinics because auditory-perceptual evaluation can be administered easily and is still considered the gold-standard (Oates, 2009).

Future Research

Computational Modeling

Future research using computational modeling can further provide evidence for proposed physiological mechanisms for RFF. As we attempted to gain more insights on physiological mechanisms for RFF in this dissertation, Chapter 2 suggested the possible effect of the degree of vocal fold closure on RFF. However, the direct causal relationship could not be evaluated due to the impossibility of controlling a parameter of interest in isolation from other parameters in speakers' voice production. Mathematical models have provided insight into self-oscillating vibrations of the vocal folds in order to understand phonation and its connection with the biomechanical properties of the vocal folds (Story & Titze, 1995; Zañartu, Mongeau, & Wodicka, 2007). However, understanding physiological mechanisms for RFF may require more complex models than those currently available. These models should be able to elicit at least a VCV utterance and reflect intrinsic laryngeal muscular control during the production. Manríquez, Peterson, Prado, Orio, Galindo, and Zañartu (2019) also suggested that current computational models lack appropriate representation of actual neurophysiological muscle activation. The representation of actual neural control would be important in understanding RFF because one of the hypothesized physiological mechanisms for RFF is the activation of the CT muscles in conjunction with other intrinsic laryngeal muscles (Stepp et al., 2010a; Stepp et al., 2011b). In addition, as RFF stimuli contain obstruents, the effect of supraglottal pressure on the vocal folds may need to be incorporated in order to accurately understand the effect of aerodynamics on RFF.

Real-Time RFF Measurement

Future research can continue to develop more reliable and effective RFF recording and estimation methods. Since controlling stress type was emphasized in RFF protocol (Park & Stepp, 2019) and averaging at least 6 RFF values was suggested (Eadie & Stepp, 2013), real-time software that can check for appropriate stress type and usability of recorded RFF instances should be developed. If unequal stress was produced or the production contained glottalization, voiced consonants, or vowel portions that were too short, the program can notify speakers to produce more instances until six usable RFF instances would be recorded. Then, this program can be paired with an existing semi-automated RFF algorithm (Lien et al., 2017; Vojtech et al., 2019) to provide real-time RFF values. This real-time software for RFF would be useful both in clinics and as individual tracking tools.

Ecological Validity of Currently Recommended RFF Stimuli

Although RFF values obtained from equal stress on both vowels showed smaller within-participant variations than RFF values from other stress types, RFF values produced with equal stress have not been compared between individuals with healthy voices and vocal hyperfunction. Thus, it is not yet certain that RFF values produced with equal stress would effectively reflect vocal hyperfunction. Previous studies have shown differences in RFF values obtained from uniform utterances when individuals modulated their vocal effort (McKenna & Stepp, 2018) and between individuals with healthy voices and vocal hyperfunction (Heller Murray et al., 2017). However, stress type was not controlled, and controlling stress types may decrease ecological validity. Therefore, before the currently recommended RFF protocol is applied to clinics, the protocol must be evaluated in both individuals with healthy voices and vocal hyperfunction.

Normative RFF Values

Future studies should also publish the normative RFF values of both individuals with healthy voices and vocal hyperfunction. Since RFF was found to differ significantly by the stimulus context and stress type, we would need to control for these factors and recruit a large number of individuals with healthy voices and vocal hyperfunction. One of the challenges in obtaining normative data is that individuals with healthy voices may present some degree of vocal hyperfunction that may not be noticeable in their voice quality and their self-report. Mild cases of structural lesions may also not be noticeable by speakers and listeners. Videolaryngoscopy is encouraged to confirm that individuals with healthy voices do not have any structural lesions or strong evidence of vocal hyperfunction. Different age groups and sex are important to consider since previous studies have suggested possible effects of age (Watson, 1998) and sex (Stepp, 2013).

Normative data will provide an important reference in the clinical evaluation of vocal hyperfunction. From the RFF values obtained from large groups of speakers with healthy and disorder voices, a cut-off score can be determined, and the cut-off score's diagnostic strength can be evaluated with a receiver operating characteristic curve using sensitivity and specificity with their diagnosis. The cut-off score will be useful in a clinical setting to provide objective evidence for diagnosis.

Sensitivity to Change

The minimal clinically important difference of RFF should be examined for RFF to be used as an outcome measure or a tracking tool for assessing vocal function. The minimal clinically important difference has been defined as "the smallest difference in score in the domain of interest which patients perceive as beneficial" (Jaeschke, Singer, & Guyatt, 1989). The minimal clinically important difference of RFF will aid in objectively evaluating outcomes of voice therapy targeted for vocal hyperfunction. In addition, this value can be used for a tracking tool to notify individuals of changes in their RFF values when the change is near this clinically important difference, which may indicate a potential worsening of their vocal function. A large group of individuals with vocal hyperfunction will be required for the purpose of evaluating the minimal clinically important difference, and they will have to be successfully treated with voice therapy sessions. RFF values can be recorded longitudinally as participants undergo therapy sessions, and the minimal clinically important difference can be determined by changes in RFF that would reflect actual improvements in voice production. The management of vocal hyperfunction is a major interest in clinical management of many voice disorders. Thus, the continuous development of RFF, leading to its clinical application, will improve the quality of voice care that seeks to preserve an essential element of human life.

Appendices

1. VSR Task for Synthetic Quality

Additional VSR training and experimental tasks were performed at the end of the session to determine how synthetic the synthesized stimuli were perceived. This task was included to ensure that potential differences in strain ratings between RFF-modified and -unmodified samples or between samples with and without noise were not due to the modified samples sounding synthetic. Participants completed a VSR training module before the actual VSR task for rating synthetic quality. We presented the same eight samples that contained a wide range of strain and were used in the training module before the VSR task for rating strain. We placed all of them at 0 on a synthetic quality scale, and listeners were informed that these samples were rated 0 for synthetic quality because all of them were produced naturally (i.e., objectively not synthetic) and were not synthesized, despite having various levels of strain.

The same protocol from the VSR task for strain ratings was provided to the listeners to complete the experimental VSR module for rating synthetic quality. Each set was designed specifically for each listener as described in the methods and a total of 10 sets containing 80 items (64 stimuli + 16 stimuli for intra-rater reliability) were completed. The VSR task for synthetic ratings took approximately 15 minutes to complete. Synthetic quality ratings for each stimulus obtained from the VSR tasks were averaged across the listeners. A three-way repeated measures ANOVA was performed on mean synthetic quality ratings. Synthetic quality showed intra-rater reliability (Pearson's r) above 0.7 in only 11 of 20 listeners (median = 0.71, range = -0.12 – 0.91) and poor inter-rater reliability

(ICC = 0.18, 95% CI = 0.13 – 0.26).

The effect of RFF modification in the three-way ANOVA on synthetic quality was not statistically significant (p = 0.67), which indicates that the synthetic quality ratings between the samples with and without RFF modification were not statistically different (Figure S1). Thus, the statistically significant effect of the interaction between vocal effort level and RFF modification on strain is unlikely to have resulted from the RFF-modified samples possibly sounding synthetic. However, the effect of noise on synthetic quality was statistically significant and had a large effect size (p = 0.001, η_p^2 = 0.81). The samples with added noise had increased synthetic quality ratings (Figure), which could suggest that increased synthetic quality in the samples with noise might have contributed to their increased strain ratings.

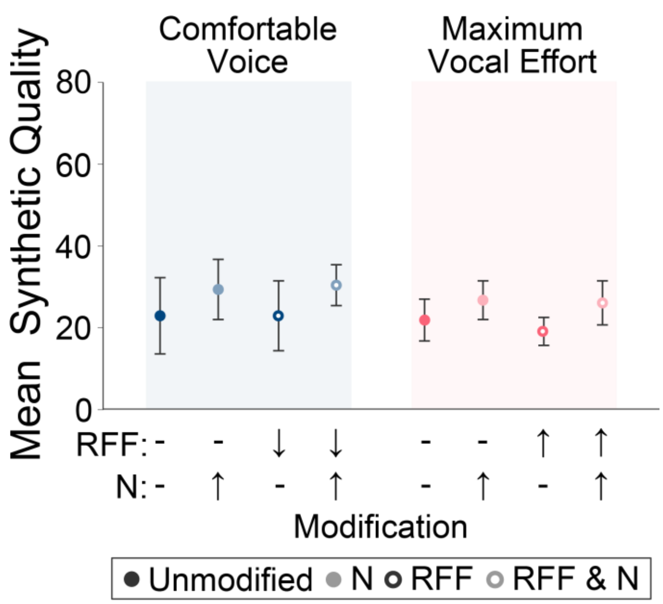

Figure A.1: Mean synthetic quality ratings of comfortable and maximum effort samples as a function of modification condition. There were no statistically significant differences among conditions (*p* > 0.05). Error bars indicate 95% confidence intervals. (abbreviations: RFF = relative fundamental frequency, N = noise, - = unmodified, ↑ = increase, ↓ = decrease)

2. MATLAB Scripts for Obtaining Mean Airflow and Subglottic Pressure Estimates

PASdata.m

```
% This script is to extract data from txt output from PAS and convert them to MATLAB vectors.
fid=fopen('a.txt');     %open the txt file from PAS
for i=1:17              %obtaining the data from the file
   aline=fgetl(fid);
end
nfid=fopen('na.txt','w');%write a temporary text file
while feof(fid)==0
   aline=fgetl(fid);
   fprintf(nfid, '%s\n', aline)
end
fclose('all');
fid=fopen('na.txt');
%obtaining the data as MATLAB vectors
pasmat=fscanf(fid,'    %f    %f    %f    %f    %f',[5, inf]);
fclose('all');
t=pasmat(1,:); % time vector
A=pasmat(4,:); % Airflow vector
S=pasmat(5,:); % Pressure vector
t=t';
A=A';
S=S';
clear fid ans i nfid pasmat aline
```

Airflow.m

```
%This script is to hand-select the stable middle portion of the airflow signal.
%To select,
%1. first zoom in to the desired pa train and press enter.
%2. Only middle three /pa/s will be used.
%3. Click the start and end of the stable portions of all three /pa/s (total of 6 clicks).
%4. When you are done selecting within this train, press enter.
%5. Repeat 1-4 twice to obtain total of 9 pas in three trains.
plot (t,A);
hold on
zoom on
pause ();
[x1,y1]=ginput;
zoom out
zoom on
pause ();
[x2,y2]=ginput;
zoom out
zoom on
pause ();
[x3,y3]=ginput;
zoom out
[Amean Astd]=AirflowMean(x1, x2, x3, t, A); %calculates mean airflow
clear x1 x2 x3 y1 y2 y3
```

AirflowMean.m

```
%This function is to calculate mean airflow from selected data points.
```

```
function [Amean Astd]=AirflowMeanfast(x1, x2, x3, t, A)
x=[x1' x2' x3']; %Combining selected points into one vector
for i=1:numel(x) %To obtain time points of each selected points
   val= x(i);
   [~,na(i)] = min(abs(t-val));
end
%Separating data points into each /pa/ train for better scripting
n(1,:)=na(1:6);   %First /pa/ train
n(2,:)=na(7:12);  %Second /pa/ train
n(3,:)=na(13:18); %Third /pa/ train

for j=1:3   %looping for three /pa/ train
   for i=1:3   %looping for four /pa/s
      Aval=A(n(j,2*i-1):n(j,2*i));  %Obtaining airflow value during /p/
      Atrim=Aval((length(Aval)/8):(length(Aval)*7/8));%Trim each 1/8 end
      Axmean=mean(Atrim); %averaging mean
      Ax(i)=Axmean;
   end
   Aall(1:3,j)=Ax';
end
%Averaging values all together to represent mean airflow
Amean=mean(mean(Aall));
Astd=std(std(Aall));
```

Subglottic.m

```
%This script is to hand-select the stable middle peak of the air pressure signal.
%To select,
%1. first zoom in to the desired pa train and press enter.
%2. Only last four /pa/s will be used (no first /pa/).
%3. Click the start and end of the stable peaks of four /pa/s (total of 8 clicks).
%4. When you are done selecting within this train, press enter.
%5. Repeat 1-4 twice to obtain total of 12 pas in three trains.
plot (t,S);
hold on
zoom on
pause ();
[xs1,ys1]=ginput;
zoom out
zoom on
pause ();
[xs2,ys2]=ginput;
zoom out
zoom on
pause ();
[xs3,ys3]=ginput;
zoom out

%calculates mean subglottic pressure
[Pall Pest]=SubglotticMean(xs1,xs2,xs3,t,S) close all
clear x1 x2 x3 y1 y2 y3
```

SubglotticMean.m

```
%This function is to estimate mean subglottic pressure from data points.
function [Pall Pest]=SubglotticMean(xs1,xs2,xs3,t,S)
xs=[xs1' xs2' xs3']; %Combining selected points into one vector
for i=1:numel(xs) %To obtain time points of each selected points
   val= xs(i);
   [~,nsa(i)] = min(abs(t-val));
end
%Separating data points into each /pa/ train for better scripting
ns(1,:)=nsa(1:8);   %First /pa/ train
ns(2,:)=nsa(9:16);  %Second /pa/ train
ns(3,:)=nsa(17:24); %Third /pa/ train

for j=1:3   %looping for three /pa/ train
   for i=1:4   %looping for four /pa/s
      Pval=S(ns(j,2*i-1):ns(j,2*i));%Obtaining airpressure value during /p/
      Pmax=max(Pval); % finding the maximum value
      VP=1.96*std(Pval(find(Pval>0.95*Pmax))); %To calculated 5% variation
      Psub(i,:)=[Pmax VP];
   end
   Pall((j+3*(j-1)):4*j,:)=Psub;
end
%This step is to average two adjacent /p/s to interpolate subglottic pressure during the vowel in between.
%Four /pa/ peak values estimate three subglottic pressure values for middle three vowels in one train.
for k=1:length(Pall)-1
   if k<4
      Psubest(k)=mean(Pall(k:k+1,1));
   elseif k>4&&k<8
      Psubest(k)=mean(Pall(k:k+1,1));
   elseif k>8&&k<12
      Psubest(k)=mean(Pall(k:k+1,1));
   else
      Psubest(k)=[0];
   end
end
%averaging all the estimates to represent mean subglottic pressure
Pest=mean([mean(Psubest(1:3)) mean(Psubest(5:7)) mean(Psubest(9:11))]);
```

Bibliography

Altman, K. W., Atkinson, C., & Lazarus, C. (2005). Current and emerging concepts in muscle tension dysphonia: a 30-month review. *Journal of Voice, 19*(2), 261–267.

American Speech-Language-Hearing Association. (1993). *Definitions of comunication disorders and variations*. Relevant Paper. Retrieved from http://www.asha.org/policy/RP1993-00208/

American Speech-Language-Hearing Association. (2005). *Guidelines for manual pure-tone threshold audiometry* [Guideline]. Available from https://www.asha.org/policy

American Speech-Language-Hearing Association. (2009). *Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V): ASHA Special Interest Group 3, Voice and Voice Disorders*. Available from https://www.asha.org/uploadedFiles/ASHA/SIG/03/CAPE-V-Procedures-and-Form.pdf

Anand, S., Kopf, L. M., Shrivastav, R., & Eddins, D. A. (2019). Objective Indices of Perceived Vocal Strain. *Journal of Voice, 33*(6), 838–845.

Angsuwarangsee, T., & Morrison, M. (2002). Extrinsic laryngeal muscular tension in patients with voice disorders. *Journal of Voice, 16*(3), 333–343.

Awan, S. N., & Awan, J. A. (2020). A two-stage cepstral analysis procedure for the classification of rough voices. *Journal of Voice, 34*(1), 9–19.

Awan, S. N., Novaleski, C. K., & Yingling, J. R. (2013). Test-retest reliability for aerodynamic measures of voice. *Journal of Voice, 27*(6), 674–684.

Awan, S. N., & 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.

Awan, S. N., Roy, N., Zhang, D., & Cohen, S. M. (2016). Validation of the cepstral spectral index of dysphonia (CSID) as a screening tool for voice disorders: Development of clinical cutoff scores. *Journal of Voice, 30*(2), 130–144.

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology, 8*, 456.

Belafsky, P. C., Postma, G. N., & Koufman, J. A. (2002a). Validity and reliability of the reflux symptom index (RSI). *Journal of Voice, 16*(2), 274–277.

Belafsky, P. C., Postma, G. N., Reulbach, T. R., Holland, B. W., & Koufman, J. A. (2002b). Muscle tension dysphonia as a sign of underlying glottal insufficiency. *Otolaryngology—Head and Neck Surgery, 127*(5), 448–451.

Belsky, M. A., Rothenberger, S. D., Gillespie, A. I., & Gartner-Schmidt, J. L. (In Press). Do phonatory aerodynamic and acoustic measures in connected speech differ between vocally healthy adults and patients diagnosed with muscle tension dysphonia? *Journal of Voice*.

Bergan, C. C., Titze, I. R., & Story, B. (2004). The perception of two vocal qualities in a synthesized vocal utterance: ring and pressed voice. *Journal of Voice, 18*(3), 305–317.

Bhattacharyya, N. (2014). The prevalence of voice problems among adults in the United States. *Laryngoscope, 124*(10), 2359–2362.

Bhuta, T., Patrick, L., & Garnett, J. D. (2004). Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of Voice, 18*(3), 299–304.

Boersma, P., & Weenink, D. (2016). Praat: doing phonetics by computer (Version 6.0.21) [Computer Program]. Retrieved from http://www.praat.org/

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer (Version 6.0.48). Retrieved from http://www.praat.org

Bough, I. D., Jr., Heuer, R. J., Sataloff, R. T., Hills, J. R., & Cater, J. R. (1996). Intrasubject variability of objective voice measures. *Journal of Voice, 10*(2), 166–174.

Bouwers, F., & Dikkers, F. G. (2009). A retrospective study concerning the psychosocial impact of voice disorders: Voice Handicap Index change in patients with benign voice disorders after treatment (measured with the Dutch version of the VHI). *Journal of Voice, 23*(2), 218–224.

Bowen, L. K., Hands, G. L., Pradhan, S., & Stepp, C. E. (2013). Effects of Parkinson's disease on fundamental frequency variability in running speech. *Journal of Medical Speech-Language Pathology, 21*(3), 235–244.

Brinca, L. F., Batista, A. P., Tavares, A. I., Goncalves, I. C., & Moreno, M. L. (2014). Use of cepstral analyses for differentiating normal from dysphonic voices: a comparative study of connected speech versus sustained vowel in European Portuguese female speakers. *Journal of Voice, 28*(3), 282–286.

Brockmann-Bauser, M., Bohlender, J. E., & Mehta, D. D. (2018). Acoustic perturbation measures improve with increasing vocal intensity in individuals with and without voice disorders. *Journal of Voice, 32*(2), 162–168.

Brockmann, M., Storck, C., Carding, P. N., & Drinnan, M. J. (2008). Voice loudness and gender effects on jitter and shimmer in healthy adults. *Journal of Speech, Language, and Hearing Research, 51*(5), 1152–1160.

Brodnitz, F. S. (1966). Rehabilitation of the human voice. *Bulletin of the New York Academy of Medicine*, *42*(3), 231–240.

Carding, P. N., Steen, I. N., Webb, A., MacKenzie, K., Deary, I. J., & Wilson, J. A. (2004). The reliability and sensitivity to change of acoustic measures of voice quality. *Clinical Otolaryngology and Alied Sciences*, *29*(5), 538–544.

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

Carroll, L., Rooney, A., Ow, T. J., & Tan, M. (2018). Pressure and flow comparisons across vocal pathologies. *Journal of Voice*, *32*(1), 95–100.

Chan, R. W., & Titze, I. R. (2006). Dependence of phonation threshold pressure on vocal tract acoustics and vocal fold tissue mechanics. *Journal of the Acoustical Society of America*, *119*(4), 2351–2362.

Chang, A., & Karnell, M. P. (2004). Perceived phonatory effort and phonation threshold pressure across a prolonged voice loading task: a study of vocal fatigue. *Journal of Voice*, *18*(4), 454–466.

Chen, S. H., Chiang, S. C., Chung, Y. M., Hsiao, L. C., & Hsiao, T. Y. (2010). Risk factors and effects of voice problems for teachers. *Journal of Voice*, *24*(2), 183–190, quiz 191–182.

Chhetri, D. K., Neubauer, J., & Berry, D. A. (2012). Neuromuscular control of fundamental frequency and glottal posture at phonation onset. *The Journal of Acoustical Society of America*, *131*(2), 1401–1412.

Collier, R., Lisker, L., Hirose, H., & Ushijima, T. (1979). Voicing in intervocalic stops and fricatives in Dutch. *Journal of Phonetics*, *7*, 357–373.

Cooke, A., Ludlow, C. L., Hallett, N., & Selbie, W. S. (1997). Characteristics of vocal fold adduction related to voice onset. *Journal of Voice, 11*(1), 12–22.

Cooke, J. D. (1980). The organization of simple, skilled movements. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 199–212). Amsterdam: North-Holland Publishing Company.

Corretge, R. (2019). *Praat Vocal Toolkit*. Available from http://www.praatvocaltoolkit.com

Dabirmoghaddam, P., Aghajanzadeh, M., Erfanian, R., Aghazadeh, K., Sohrabpour, S., Firouzifar, M., Maroufizadeh, S., & Nikravesh, M. (In Press). Comparative study of increased supraglottic activity in normal individuals and those with muscle tension dysphonia (MTD). *Journal of Voice*.

Dailey, S. H., Kobler, J. B., Hillman, R. E., Tangrom, K., Thananart, E., Mauri, M., & Zeitels, S. M. (2005). Endoscopic measurement of vocal fold movement during adduction and abduction. *Laryngoscope, 115*(1), 178–183.

De Bodt, M. S., Van de Heyning, P. H., Wuyts, F. L., & Lambrechts, L. (1996). The perceptual evaluation of voice disorders. *Acta Oto-rhino-laryngologica Belgica*, *50*(4), 283–291.

Dehqan, A., Scherer, R. C., Dashti, G., Ansari-Moghaddam, A., & Fanaie, S. (2012). The effects of aging on acoustic parameters of voice. *Folia Phoniatrica et Logopaedica*, *64*(6), 265–270.

Deliyski, D. D., Shaw, H. S., Evans, M. K., & Vesselinov, R. (2006). Regression tree approach to studying factors influencing acoustic voice analysis. *Folia Phoniatrica et Logopaedica*, *58*(4), 274–288.

Desjardins, M., Halstead, L., Cooke, M., & Bonilha, H. S. (2017). A systematic review of voice therapy: what “effectiveness" really implies. *Journal of Voice, 31*(3), 392 e313–392 e332.

Dietrich, M., & Verdolini Abbott, K. (2012). Vocal function in introverts and extraverts during a psychological stress reactivity protocol. *Journal of Speech, Language, and Hearing Research, 55*(3), 973–987.

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.

Eadie, T. L., & Doyle, P. C. (2005). Classification of dysphonic voice: acoustic and auditory-perceptual measures. *Journal of Voice, 19*(1), 1–14.

Eadie, T. L., Kapsner, M., Rosenzweig, J., Waugh, P., Hillel, A., & Merati, A. (2010). The role of experience on judgments of dysphonia. *Journal of Voice, 24*(5), 564–573.

Eadie, T. L., & Stepp, C. E. (2013). Acoustic correlate of vocal effort in spasmodic dysphonia. *The Annals of Otology, Rhinology, and Laryngology*, *122*(3), 169–176.

Eddins, D. A., Anand, S., Lang, A., & Shrivastav, R. (In Press). Developing Clinically Relevant Scales of Breathy and Rough Voice Quality. *Journal of Voice*.

Enflo, L., Sundberg, J., Romedahl, C., & McAllister, A. (2013). Effects on vocal fold collision and phonation threshold pressure of resonance tube phonation with tube end in water. *Journal of Speech, Language, and Hearing Research, 56*(5), 1530–1538.

Eskenazi, L., Childers, D. G., & Hicks, D. M. (1990). Acoustic correlates of vocal quality. *Journal of Speech and Hearing Research, 33*(2), 298–306.

Espinoza, V. M., Zanartu, M., Van Stan, J. H., Mehta, D. D., & Hillman, R. E. (2017). Glottal Aerodynamic Measures in Women With Phonotraumatic and Nonphonotraumatic Vocal Hyperfunction. *Journal of Speech, Language, and Hearing Research, 60*(8), 2159–2169.

Fairbanks, G. (1960). *Voice and articulation drillbook, 2nd edn.* New York: Harper & Row.

Feldman, A. G. (1980). Superposition of motor programs – I. Rhythmic forearm movements in man. *Neuroscience, 5*(1), 81–90.

Fex, B., Fex, S., Shiromoto, O., & Hirano, M. (1994). Acoustic analysis of functional dysphonia: before and after voice therapy (accent method). *Journal of Voice, 8*(2), 163–167.

Fryd, A. S., Van Stan, J. H., Hillman, R. E., & Mehta, D. D. (2016). Estimating Subglottal Pressure From Neck-Surface Acceleration During Normal Voice Production. *Journal of Speech, Language, and Hearing Research, 59*(6), 1335–1345.

Fujiki, R. B., Chapleau, A., Sundarrajan, A., McKenna, V., & Sivasankar, M. P. (2017). The Interaction of Surface Hydration and Vocal Loading on Voice Measures. *Journal of Voice, 31*(2), 211–217.

Fukui, N., & Hirose, H. (1983). Laryngeal adjustments in Danish voiceless obstruent production. *Annual Bulletin Research Institute of Logpedics and Phoniatrics, 17*, 61–71.

Galindo, G. E., Peterson, S. D., Erath, B. D., Castro, C., Hillman, R. E., & Zanartu, M. (2017). Modeling the pathophysiology of phonotraumatic vocal hyperfunction with a triangular glottal model of the vocal folds. *Journal of Speech, Language, and Hearing Research, 60*(9), 2452–2471.

Gallena, S., Smith, P. J., Zeffiro, T., & Ludlow, C. L. (2001). Effects of levodopa on laryngeal muscle activity for voice onset and offset in Parkinson disease. *Journal of Speech, Language, and Hearing Research, 44*(6), 1284–1299.

Garrett, K. L., & Healey, E. C. (1987). An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day. *The Journal of the Acoustical Society of America, 82*(1), 58–62.

Gillespie, A. I., Dastolfo, C., Magid, N., & Gartner-Schmidt, J. (2014). Acoustic analysis of four common voice diagnoses: moving toward disorder-specific assessment. *Journal of Voice, 28*(5), 582–588.

Gillespie, A. I., Gartner-Schmidt, J., Rubinstein, E. N., & Abbott, K. V. (2013). Aerodynamic profiles of women with muscle tension dysphonia/aphonia. *Journal of Speech, Language, and Hearing Research, 56*(2), 481–488.

Goberman, A. M., & Blomgren, M. (2008). Fundamental frequency change during offset and onset of voicing in individuals with Parkinson disease. *Journal of Voice, 22*(2), 178–191.

Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication, 40*, 189–212.

Granqvist, S. (2003). The visual sort and rate method for perceptual evaluation in listening tests. *Logopedics Phoniatrics Vocology, 28*(3), 109–116.

Granqvist, S. (2010). Madde (Version 3.0.0.2) [Computer Program]: Tolvan Data. Retrieved from http://www.tolvan.com/

Halpern, A. E., Spielman, J. L., Hunter, E. J., & Titze, I. R. (2009). The inability to produce soft voice (IPSV): a tool to detect vocal change in school-teachers. *Logopedics Phoniatrics Vocology, 34*(3), 117–127.

Heller Murray, E. S., Hands, G. L., Calabrese, C. R., & Stepp, C. E. (2016). Effects of Adventitious Acute Vocal Trauma: Relative Fundamental Frequency and Listener Perception. *Journal of Voice, 30*(2), 177–185.

Heller Murray, E. S., Lien, Y. S., Van Stan, J. H., Mehta, D. D., Hillman, R. E., Pieter Noordzij, J., & Stepp, C. E. (2017). Relative Fundamental Frequency Distinguishes Between Phonotraumatic and Non-Phonotraumatic Vocal Hyperfunction. *Journal of Speech, Language, and Hearing Research, 60*(6), 1507–1515.

Helou, L. B., Rosen, C. A., Wang, W., & Verdolini Abbott, K. (2018). Intrinsic Laryngeal Muscle Response to a Public Speech Preparation Stressor. *Journal of Speech, Language, and Hearing Research, 61*(7), 1525–1543.

Helou, L. B., Wang, W., Ashmore, R. C., Rosen, C. A., & Abbott, K. V. (2013). Intrinsic laryngeal muscle activity in response to autonomic nervous system activation. *Laryngoscope*, *123*(11), 2756–2765.

Heman-Ackah, Y. D., Heuer, R. J., Michael, D. D., Ostrowski, R., Horman, M., Baroody, M. M., Hillenbrand, J., & Sataloff, R. T. (2003). Cepstral peak prominence: a more reliable measure of dysphonia. *The Annals of Otology, Rhinology, and Laryngology*, *112*(4), 324–333.

Hillenbrand, J. (1988). Perception of aperiodicities in synthetically generated voices. *The Journal of the Acoustical Society of America*, *83*(6), 2361–2371.

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.

Hillman, R. E., Gress, C., Hargrave, J., Walsh, M., & Bunting, G. (1990a). The efficacy of speech-language pathology intervention: Voice disorders. *Seminars Speech Language*, *11*, 297–310.

Hillman, R. E., Heaton, J. T., Masaki, A., Zeitels, S. M., & Cheyne, H. A. (2006). Ambulatory monitoring of disordered voices. *The Annals of Otology, Rhinology, and Laryngology*, *115*(11), 795–801.

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.

Hillman, R. E., Holmberg, E. B., Perkell, J. S., Walsh, M., & Vaughan, C. (1990b). Phonatory function associated with hyperfunctionally related vocal fold lesions. *Journal of Voice, 4*(1), 52–63.

Hirano, M. (1981). Psycho-acoustic evaluation of voice: GRBAS scale for evaluating the hoarse voice. In *Clinicain Examination of Voice*. New York: Springer-Verlag.

Hogikyan, N. D., & Sethuraman, G. (1999). Validation of an instrument to measure voice-related quality of life (V-RQOL). *Journal of Voice, 13*(4), 557–569.

Holmberg, E. B., Doyle, P., Perkell, J. S., Hammarberg, B., & Hillman, R. E. (2003). Aerodynamic and acoustic voice measurements of patients with vocal nodules: variation in baseline and changes across voice therapy. *Journal of Voice, 17*(3), 269–282.

Holmberg, E. B., Hillman, R. E., Perkell, J. S., & Gress, C. (1994). Relationships between intra-speaker variation in aerodynamic measures of voice production and variation in SPL across repeated recordings. *Journal of Speech and Hearing Research, 37*(3), 484–495.

Holmberg, E. B., Perkell, J. S., & Hillman, R. E. (1984). Methods for using a noninvasive technique for estimating glottal functions from oral measurements. *The Journal of the Acoustical Society of America, 75*, S7.

Honda, K., Hirai, H., Masaki, S., & Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in F0 control. *Language and Speech, 42 (Pt 4)*, 401–411.

Honjo, I., & Isshiki, N. (1980). Laryngoscopic and voice characteristics of aged persons. *Archives of Otolaryngology, 106*(3), 149–150.

Hottinger, D. G., Tao, C., & Jiang, J. J. (2007). Comparing phonation threshold flow and pressure by abducting excised larynges. *Laryngoscope, 117*(9), 1695–1699.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America, 25*, 105–113.

Hsiao, T. Y., Liu, C. M., Hsu, C. J., Lee, S. Y., & Lin, K. N. (2001). Vocal fold abnormalities in laryngeal tension-fatigue syndrome. *Journal of the Formosan Medical Association, 100*(12), 837–840.

Hunter, E. J. (2011). *General Statistics of the NCVS Self-Administered Vocal Rating (SAVRa). The National Center for Voice and Speech Online Technical Memo.* Available from www.ncvs.org/research_techbriefs.html

Hunter, E. J. (2012). Teacher response to ambulatory monitoring of voice. *Logopedics Phoniatrics Vocology, 37*(3), 133–135.

Hunter, E. J., Cantor-Cutiva, L. C., van Leer, E., van Mersbergen, M., Nanjundeswaran, C. D., Bottalico, P., Sandage, M. J., & Whitling, S. (2020). Toward a consensus description of vocal Effort, vocal Load, vocal Loading, and vocal fatigue. *Journal of Speech, Language, and Hearing Research, 63*(2), 509–532.

Jacobson, B. H., Johnson, A., Grywalski, C., Silbergleit, A. K., Jacobson, G. P., Benninger, M. S., & Newman, C. W. (1997). The Voice Handicap Index (VHI): Development and Validation. *American Journal of Speech-Language Pathology, 6*(3), 66–69.

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials, 10*(4), 407–415.

Jafari, N., Salehi, A., Izadi, F., Talebian Moghadam, S., Ebadi, A., Dabirmoghadam, P., Faham, M., & Shahbazi, M. (2017). Vocal Function Exercises for Muscle Tension Dysphonia: Auditory-Perceptual Evaluation and Self-Assessment Rating. *Journal of Voice, 31*(4), 506 e525–506 e531.

Jafari, N., Salehi, A., Meerschman, I., Izadi, F., Ebadi, A., Talebian, S., Khoddami, S. M., Dabirmoghadam, P., Drinnan, M., Jordens, K., D'Haeseleer, E., & Van Lierde, K. (2020). A novel laryngeal palpatory scale (LPS) in patients with muscle tension dysphonia. *Journal of Voice, 34*(3), 488 e489–488 e427.

Johns, M. M. (2003). Update on the etiology, diagnosis, and treatment of vocal fold nodules, polyps, and cysts. *Current Opinion in Otolaryngology & Head and Neck Surgery, 11*(6), 456–461.

Ju, Y. H., Jung, K. Y., Kwon, S. Y., Woo, J. S., Cho, J. G., Park, M. W., Park, E. H., & Baek, S. K. (2013). Effect of voice therapy after phonomicrosurgery for vocal polyps: a prospective, historically controlled, clinical study. *The Journal of Laryngology & Otology, 127*(11), 1134–1138.

Kagan, L. S., & Heaton, J. T. (2017). The Effectiveness of Low-Level Light Therapy in Attenuating Vocal Fatigue. *Journal of Voice, 31*(3), 384.315–384.323.

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and*

Technology, 27(6), 349–353.

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

Kent, R. D. (1996). Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology, 5*(3), 7–23.

Khoddami, S. M., Ansari, N. N., & Jalaie, S. (2015). Review on Laryngeal Palpation Methods in Muscle Tension Dysphonia: Validity and Reliability Issues. *Journal of Voice, 29*(4), 459–468.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of Acoustical Society of America, 87*(2), 820–857.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Koufman, J. A., & Blalock, P. D. (1991). Functional voice disorders. *Otolaryngologic Clinics of North America, 24*(5), 1059–1073.

Kreiman, J., & Gerratt, B. R. (2012). Perceptual interaction of the harmonic source and noise in voice. *The Journal of Acoustical Society of America, 131*(1), 492–500.

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

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

Kreiman, J., Shue, Y. L., Chen, G., Iseli, M., Gerratt, B. R., Neubauer, J., & Alwan, A. (2012). Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *The Journal of Acoustical Society of America, 132*(4), 2625–2632.

Ladefoged, P. (1967). Three ares of experimental phonetics. In R. Mackin & P. D. Strevens (Eds.), *Language and Language Learning* (pp. 33). London: Oxford University Press.

Lee, L., Stemple, J. C., & Kizer, M. (1999). Consistency of acoustic and aerodynamic measures of voice production over 28 days under various testing conditions. *Journal of Voice, 13*(4), 477–483.

Leong, K., Hawkshaw, M. J., Dentchev, D., Gupta, R., Lurie, D., & Sataloff, R. T. (2013). Reliability of objective voice measures of normal speaking voices. *Journal of Voice, 27*(2), 170–176.

Lien, Y. S., Calabrese, C. R., Michener, C. M., Murray, E. H., Van Stan, J. H., Mehta, D. D., Hillman, R. E., Noordzij, J. P., & Stepp, C. E. (2015a). Voice relative fundamental frequency via neck-skin acceleration in individuals with voice disorders. *Journal of Speech, Language, and Hearing Research, 58*(5), 1482–1487.

Lien, Y. S., Gattuccio, C. I., & Stepp, C. E. (2014). Effects of phonetic context on relative fundamental frequency. *Journal of Speech, Language, and Hearing Research, 57*(4), 1259–1267.

Lien, Y. S., Heller Murray, E. S., Calabrese, C. R., Michener, C. M., Van Stan, J. H., Mehta, D. D., Hillman, R. E., Noordzij, J. P., & Stepp, C. E. (2017). Validation of an Algorithm for Semi-automated Estimation of Voice Relative Fundamental Frequency. *The Annals of Otology, Rhinology, and Laryngology, 126*(10), 712–716.

Lien, Y. S., Michener, C. M., Eadie, T. L., & Stepp, C. E. (2015b). Individual monitoring of vocal effort with relative fundamental frequency: Relationships with aerodynamics and listener perception. *Journal of Speech, Language, and Hearing Research, 58*(3), 566–575.

Lin, L., Sun, N., Yang, Q., Zhang, Y., Shen, J., Shi, L., Fang, Q., & Sun, G. (2014). Effect of voice training in the voice rehabilitation of patients with vocal cord polyps after surgery. *Experimental and Therapeutic Medicine, 7*(4), 877–880.

Linder, R., Albers, A. E., Hess, M., Poppl, S. J., & Schonweiler, R. (2008). Artificial neural network-based classification to screen for dysphonia using psychoacoustic scaling of acoustic voice features. *Journal of Voice, 22*(2), 155–163.

Löfqvist, A., Baer, T., McGarr, N. S., & Story, R. S. (1989). The cricothyroid muscle in voicing control. *The Journal of Acoustical Society of America, 85*(3), 1314–1321.

Löfqvist, A., Koenig, L. L., & McGowan, R. S. (1995). Vocal-tract aerodynamics in /aCa/ utterances: Measurements. *Speech Communication, 16*(1), 49–66.

Lowell, S. Y., Kelley, R. T., Awan, S. N., Colton, R. H., & Chan, N. H. (2012a). Spectral- and cepstral-based acoustic features of dysphonic, strained voice quality. *The Annals of Otology, Rhinology, and Laryngology, 121*(8), 539–548.

Lowell, S. Y., Kelley, R. T., Colton, R. H., Smith, P. B., & Portnoy, J. E. (2012b). Position of the hyoid and larynx in people with muscle tension dysphonia. *Laryngoscope, 122*(2), 370–377.

Macmillan, N. A., & Creelman, D. (2004). *Detection theory: A user's guide* (2 ed.): Lawrence Erlbaum.

Manríquez, R., Peterson, S. D., Prado, P., Orio, P., Galindo, G. E., & Zañartu, M. (2019). Neurophysiological muscle activation scheme for controlling vocal fold models. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 27*(5), 1043–1052.

Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N., & De Bodt, M. (2010). Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. *Journal of Voice, 24*(5), 540–555.

Maryn, Y., & Weenink, D. (2015). Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index. *Journal of Voice, 29*(1), 35–43.

McGarr, N. S., & Löfqvist, A. (1988). Laryngeal kinematics in voiceless obstruents produced by hearing-impaired speakers. *Journal of Speech and Hearing Research, 31*(2), 234–239.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

McKenna, V. S., Diaz-Cadiz, M. E., Shembel, A. C., Enos, N. M., & Stepp, C. E. (2019). The relationship between physiological mechanisms and the self-perception of vocal effort. *Journal of Speech, Language, and Hearing Research, 62*(4), 815–834.

McKenna, V. S., Heller Murray, E. S., Lien, Y. S., & Stepp, C. E. (2016). The relationship between relative fundamental frequency and a kinematic estimate of laryngeal stiffness in healthy adults. *Journal of Speech, Language, and Hearing Research, 59*(6), 1283–1294.

McKenna, V. S., & Stepp, C. E. (2018). The relationship between acoustical and perceptual measures of vocal effort. *The Journal of Acoustical Society of America, 144*(3), 1643.

Mehta, D. D., & Hillman, R. E. (2008). Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. *Current Opinion in Otolaryngology & Head and Neck Surgery, 16*(3), 211–215.

Mehta, D. D., Van Stan, J. H., Zanartu, M., Ghassemi, M., Guttag, J. V., Espinoza, V. M., Cortes, J. P., Cheyne, H. A., 2nd, & Hillman, R. E. (2015). Using ambulatory voice monitoring to investigate common voice disorders: Research update. *Frontiers in Bioengineering and Biotechnology, 3*, 155.

Morris, M. A., Meier, S. K., Griffin, J. M., Branda, M. E., & Phelan, S. M. (2016). Prevalence and etiologies of adult communication disabilities in the United States: Results from the 2012 National Health Interview Survey. *Disability and Health Journal, 9*(1), 140–144.

Morrison, M. D. (1997). Pattern recognition in muscle misuse voice disorders: how I do it. *Journal of Voice, 11*(1), 108–114.

Morrison, M. D., & Rammage, L. A. (1993). Muscle misuse voice disorders: description and classification. *Acta Oto-Laryngologica, 113*(3), 428–434.

Nawka, T., & Konerding, U. (2012). The interrater reliability of stroboscopy evaluations. *Journal of Voice, 26*(6), 812.811–812.810.

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

Olthoff, A., Woywod, C., & Kruse, E. (2007). Stroboscopy versus high-speed glottography: a comparative study. *Laryngoscope, 117*(6), 1123–1126.

Park, Y., Perkell, J. S., Matthies, M. L., & Stepp, C. E. (2019). Categorization in the perception of breathy voice quality and its relation to voice production in healthy speakers. *Journal of Speech, Language, and Hearing Research, 62*(10), 3655–3666.

Park, Y., & Stepp, C. E. (2019). The effects of stress type, vowel identity, baseline f_0, and loudness on the relative fundamental frequency of individuals with healthy voices. *Journal of Voice, 33*(5), 603–610.

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., Paul, D., Svec, J. G., & Hillman, R. (2018). Recommended Protocols for Instrumental Assessment of Voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. *American Journal of Speech-Language Pathology, 27*(3), 887–905.

Poburka, B. J., Patel, R. R., & Bless, D. M. (2017). Voice-vibratory assessment with laryngeal imaging (VALI) form: reliability of rating stroboscopy and high-speed videoendoscopy. *Journal of Voice, 31*(4), 513.511–513.514.

Portney, L. G., & Watkins, M. P. (2000). *Foundations of clinical research: applications to practice*. New Jersey: Prentice Hall.

Ramig, L. O., & Verdolini, K. (1998). Treatment efficacy: voice disorders. *Journal of Speech, Language, and Hearing Research, 41*(1), S101–116.

Rodeno, M. T., Sanchez-Fernandez, J. M., & Rivera-Pomar, J. M. (1993). Histochemical and morphometrical ageing changes in human vocal cord muscles. *Acta Oto-Laryngologica, 113*(3), 445–449.

Rousseau, B., Kojima, T., Novaleski, C. K., Kimball, E. E., Valenzuela, C. V., Mizuta, M., Daniero, J. J., Garrett, C. G., & Sivasankar, M. P. (2017). Recovery of Vocal Fold Epithelium after Acute Phonotrauma. *Cells Tissues Organs, 204*(2), 93–104.

Roy, N. (2003). Functional dysphonia. *Current Opinion in Otolaryngology & Head and Neck Surgery, 11*(3), 144–148.

Roy, N. (2008). Assessment and treatment of musculoskeletal tension in hyperfunctional voice disorders. *International Journal of Speech-Language Pathology, 10*(4), 195–209.

Roy, N., Barkmeier-Kraemer, J., Eadie, T., Sivasankar, M. P., Mehta, D., Paul, D., & Hillman, R. (2013). Evidence-based clinical voice assessment: a systematic review. *American Journal of Speech-Language Pathology, 22*(2), 212–226.

Roy, N., Bless, D. M., & Heisey, D. (2000). Personality and voice disorders: a multitrait-multidisorder analysis. *Journal of Voice, 14*(4), 521–548.

Roy, N., Fetrow, R. A., Merrill, R. M., & Dromey, C. (2016). Exploring the clinical utility of relative fundamental frequency as an objective measure of vocal hyperfunction. *Journal of Speech, Language, and Hearing Research, 59*(5), 1002–1017.

Roy, N., & Leeper, H. A. (1993). Effects of the manual laryngeal musculoskeletal tension reduction technique as a treatment for functional voice disorders: Perceptual and acoustic measures. *Journal of Voice, 7*(3), 242–249.

Roy, N., Merrill, R. M., Gray, S. D., & Smith, E. M. (2005). Voice disorders in the general population: prevalence, risk factors, and occupational impact. *Laryngoscope, 115*(11), 1988–1995.

Sato, K., & Hirano, M. (1997). Age-related changes of elastic fibers in the superficial layer of the lamina propria of vocal folds. *The Annals of Otology, Rhinology, and Laryngology, 106*(1), 44–48.

Sato, K., Hirano, M., & Nakashima, T. (2002). Age-related changes of collagenous fibers in the human vocal fold mucosa. *The Annals of Otology, Rhinology, and Laryngology, 111*(1), 15–20.

Scherer, K. R. (1978). Personality inference from voice quality: the loud voice of extroversion. *European Journal of Social Psychology, 8*, 467–487.

Schindler, A., Mozzanica, F., Maruzzi, P., Atac, M., De Cristofaro, V., & Ottaviani, F. (2013). Multidimensional assessment of vocal changes in benign vocal fold lesions after voice therapy. *Auris Nasus Larynx*, *40*(3), 291–297.

Shim, H., Jung, H., Koul, R., & Ko, D. (2016). Spectral and cepstral based acoustic features of voices with muscle tension dysphonia. *Clinical Archives of Communication Disorders*, *1*(1), 41–47.

Smith, A. B., & Robb, M. P. (2013). Factors underlying short-term fundamental frequency variation during vocal onset and offset. *Speech, Language, and Hearing, 16*(4), 208–214.

Smith, E., Verdolini, K., Gray, S., Nichols, S., Lemke, J., Barkmeier, J., Dove, H., & Hoffman, H. (1994). Effect of Voice Disorders on Quality of LIfe. *NCVS Status and Progress Report*, *7*, 1–17.

Smith, E., Verdolini, K., Gray, S., Nichols, S., Lemke, J., Barkmeier, J., Dove, H., & Hoffman, H. (1996). Effect of voice disorders on quality of life. *Journal of Medical Speech-Language Pathology, 4*(4), 223–244.

Solomon, N. P., & DiMattia, M. S. (2000). Effects of a vocally fatiguing task and systemic hydration on phonation threshold pressure. *Journal of Voice, 14*(3), 341–362.

Stemple, J. C. (2019). Principles of successful voice thearpy. In *Voice Therapy: Clinical Case Studies* (5 ed.). San Diego, CA: Plural Publishing.

Stemple, J. C., Roy, N., & Klaben, B. K. (2014). *Clinical voice pathology* (5th ed.). San Diego, CA: Plural Publishing, Inc.

Stemple, J. C., Stanley, J., & Lee, L. (1995). Objective measures of voice production in normal subjects following prolonged voice use. *Journal of Voice, 9*(2), 127–133.

Stepp, C. E. (2013). Relative fundamental frequency during vocal onset and offset in older speakers with and without Parkinson's disease. *The Journal of Acoustical Society of America*, *133*(3), 1637–1643.

Stepp, C. E., Heaton, J. T., Braden, M. N., Jette, M. E., Stadelman-Cohen, T. K., & Hillman, R. E. (2011a). Comparison of neck tension palpation rating systems with surface electromyographic and acoustic measures in vocal hyperfunction. *Journal of Voice*, *25*(1), 67–75.

Stepp, C. E., Hillman, R. E., & Heaton, J. T. (2010a). The impact of vocal hyperfunction on relative fundamental frequency during voicing offset and onset. *Journal of Speech, Language, and Hearing Research, 53*(5), 1220–1226.

Stepp, C. E., Hillman, R. E., & Heaton, J. T. (2010b). A virtual trajectory model predicts differences in vocal fold kinematics in individuals with vocal hyperfunction. *The Journal of Acoustical Society of America*, *127*(5), 3166–3176.

Stepp, C. E., Merchant, G. R., Heaton, J. T., & Hillman, R. E. (2011b). 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.

Stepp, C. E., Sawin, D. E., & Eadie, T. L. (2012). The relationship between perception of vocal effort and relative fundamental frequency during voicing offset and onset. *Journal of Speech, Language, and Hearing Research*, *55*(6), 1887–1896.

Stevens, K. N. (1977). Physics of laryngeal behavior and larynx modes. *Phonetica*, *34*(4), 264–279.

Story, B. H., & Titze, I. R. (1995). Voice simulation with a body-cover model of the vocal folds. *The Journal of the Acoustical Society of America*, *97*(2), 1249–1260.

Sundberg, J., & Gauffin, J. (1978). Waveform and spectrum of the glottal voice source. *Dept. for Speech Music and Hearing Quarterly Progress Status Report*, *19*(2–3), 35–50.

Tiago, R., Pontes, P., & do Brasil, O. C. (2007). Age-related changes in human laryngeal nerves. *Otolaryngology Head and Neck Surgery*, *136*(5), 747–751.

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.

Van Stan, J. H., Mehta, D. D., Ortiz, A. J., Burns, J. A., Toles, L. E., Marks, K. L., Vangel, M., Hron, T., Zeitels, S., & Hillman, R. E. (2020). Differences in weeklong ambulatory vocal behavior between female patients with phonotraumatic lesions and matched controls. *Journal of Speech, Language, and Hearing Research*, 1–13.

Verdolini, K., Titze, I. R., & Fennell, A. (1994). Dependence of phonatory effort on hydration level. *Journal of Speech and Hearing Research*, *37*(5), 1001–1007.

Vojtech, J. M., & Heller Murray, E. (2019). *Tutorial for Manual Relative Fundamental Frequency (RFF) Estimation using Praat*. Tutorial. Boston University. Stepp Lab.

Vojtech, J. M., Segina, R. K., Buckley, D. P., Kolin, K. R., Tardif, M. C., Noordzij, J. P., & Stepp, C. E. (2019). Refining algorithmic estimation of relative fundamental frequency: Accounting for sample characteristics and fundamental frequency estimation method. *The Journal of the Acoustical Society of America*, *146*(5), 3184.

Watson, B. C. (1998). Fundamental frequency during phonetically governed devoicing in normal young and aged speakers. *The Journal of Acoustical Society of America, 103*(6), 3642–3647.

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.

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.

Welham, N. V., & Maclagan, M. A. (2003). Vocal fatigue: current knowledge and future directions. *Journal of Voice, 17*(1), 21–30.

Woo, P. (2017). High-speed Imaging of Vocal Fold Vibration Onset Delay: Normal Versus Abnormal. *Journal of Voice*, *31*(3), 307–312.

Xue, S. A., & Deliyski, D. (2001). Effects of aging on selected acoustic voice parameters: preliminary normative data and educational implications. *Educational Gerontology, 27*(2), 159–168.

Yin, J., & Zhang, Z. (2013). The influence of thyroarytenoid and cricothyroid muscle activation on vocal fold stiffness and eigenfrequencies. *The Journal of Acoustical Society of America, 133*(5), 2972–2983.

Yoshioka, H., Löfqvist, A., & Collier, R. (1982). Laryngeal adjustments in Dutch voiceless obstruent production. *Annual Bulletin Research Institute of Logpedics and Phoniatrics, 16*, 27–35.

Zañartu, M., Mongeau, L., & Wodicka, G. R. (2007). Influence of acoustic loading on an effective single mass model of the vocal folds. *The Journal of the Acoustical Society of America, 121*(2), 1119–1129.

Zhang, Z. (2015). Regulation of glottal closure and airflow in a three-dimensional phonation model: Implications for vocal intensity control. *The Journal of Acoustical Society of America, 137*(2), 898–910.

Zhang, Z. (2016a). Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model. *The Journal of Acoustical Society of America, 139*(4), 1493.

Zhang, Z. (2016b). Mechanics of human voice production and control. *The Journal of Acoustical Society of America, 140*(4), 2614.

Zheng, Y. Q., Zhang, B. R., Su, W. Y., Gong, J., Yuan, M. Q., Ding, Y. L., & Rao, S. Q. (2012). Laryngeal aerodynamic analysis in assisting with the diagnosis of muscle tension dysphonia. *Journal of Voice, 26*(2), 177–181.

Zraick, R. I., Kempster, G. B., Connor, N. P., Thibeault, S., Klaben, B. K., Bursac, Z., Thrush, C. R., & 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.

Curriculum Vitae

[redacted]

[redacted]

[redacted]

[redacted]