6-22-2021 2:00 PM

Modelling loudness: Acoustic and perceptual correlates in the context of hypophonia in Parkinson's disease

Daryn Amory Cushnie-Sparrow, *The University of Western Ontario*

Supervisor: Adams, Scott G., *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Health and Rehabilitation Sciences

Abstract

Hypophonia (quiet speech) is a common speech symptom associated with Parkinson's disease (PD), and is associated with reduced intelligibility, communicative effectiveness, and communicative participation. Studies of hypophonia commonly employ average speech intensity as the primary dependent measure, which may not entirely capture loudness deficits. Loudness may also be affected by the frequency components of speech (i.e. spectral balance) and speech level variability. The present investigation examined relationships between perceived loudness and intelligibility with acoustic measures of loudness, speech intensity, and spectral distribution in individuals with hypophonia secondary to Parkinson's disease (IWPDs) and neurologically healthy older adults (HOAs).

Samples of sentence reading and conversational speech from 56 IWPDs and 46 HOAs were presented to listeners for ratings of perceived loudness and intelligibility. Listeners provided ratings of loudness using visual analogue scales (VAS) and direct magnitude estimation (DME). Acoustic measures of speech level (e.g. mean intensity), spectral balance (e.g. spectral tilt), and speech level variability (e.g. standard deviation of intensity) were obtained for comparison with perceived characteristics. In a spectral manipulation experiment, a gain adjustment altered the spectral balance of sentence samples while maintaining equal mean intensity. Listeners provided VAS ratings of perceived loudness of these manipulated samples.

IWPDs were quieter, less intelligible, and had a relatively greater concentration of low-frequency energy than HOAs. Speech samples with weaker contributions of mid- (2-5 kHz) and high-frequency (5-8 kHz) energy were perceived as quieter. Results of the spectral manipulation experiment indicated that increases in the relative contribution of 2-10 kHz energy were associated with increases in perceived loudness. The acoustic time-varying loudness model (TVL) demonstrated stronger associations with perceived loudness and larger differences between IWPDs and HOAs, and successfully identified differences in loudness in the

spectral manipulation experiment. Loudness ratings provided with VAS and DME were consistent, both providing excellent reliability.

Findings of this investigation indicate that perceived loudness, acoustic loudness, and spectral balance are important components of hypophonia evaluation. Incorporating spectral manipulation in amplification by increasing high-frequency energy may improve efficacy of amplification devices for hypophonia management.

Keywords: Parkinson's disease, hypophonia, loudness, intelligibility, speech acoustics, speech intensity, spectral balance, amplification, visual analogue scales, direct magnitude estimation, loudness model

Lay Summary

Most individuals with Parkinson's disease (PD) experience changes in their speech and voice. Quiet speech (hypophonia) is a common speech symptom associated with PD. Hypophonia interferes with the ability of individuals with PD (IWPDs) to effectively communicate because they may not be heard or understood, and some IWPDs may avoid communicating in situations they previously enjoyed. Effective assessment and evaluation of hypophonia is important in research and clinical settings to understand the condition and provide strategies to reduce the impacts of this condition on the lives of IWPDs. This study investigated several measures that can be used to assess hypophonia in order to identify components of effective assessment.

IWPDs and neurologically healthy older adults (HOAs) were recorded while reading sentences aloud and while participating in a conversation. Recordings of their speech were played for listeners, who rated the loudness and intelligibility (how much of their speech they could understand) of each sample. Acoustic measures were obtained from the speech recordings to compare how the sound characteristics of their speech related to the listeners' perceptions. Acoustic algorithms designed to estimate perceived loudness were also included for comparison with perceived loudness. In a second experiment, listeners heard manipulated samples of speech. Frequency characteristics of speech were altered to investigate how the loudness would change.

IWPDs were quieter, less intelligible, and had disrupted spectral balance (frequency characteristics of their speech). Speech samples with relatively weaker high-frequency energy sounded quieter. Time-varying loudness (TVL; acoustic algorithm estimating loudness) provided effective measurement of loudness in both IWPDs and HOAs. Effective assessment of hypophonia may include listener judgments of loudness, acoustic calculations of loudness, and descriptions of spectral balance. Some IWPDs use amplification devices, similar to the microphone and loudspeaker used by a speaker in a large auditorium. Findings of this study suggest that incorporating a high-frequency boost to these amplifiers might further improve the loudness which would be a more effective tool for IWPDs with hypophonia.

Acknowledgments

While conducting this investigation, I have lived and worked on the traditional lands of the Anishinaabek, Haudenosaunee, Lūnaapéewak, Attawandaron, Huron-Wendat, and Ojibway/Chippewa peoples, on the Treaty Lands and the Territory of the Mississaugas of the Credit, and on lands connected with the London Township and Sombra Treaties of 1796 and the Dish with One Spoon Covenant Wampum. I am grateful to have the opportunity to work on this land and give my respect to its first inhabitants. I respect the longstanding relationships that Indigenous Nations have to this land, as they are the original caretakers. I acknowledge historical and ongoing injustices that Indigenous Peoples (First Nations, Métis and Inuit) endure in Canada, and accept responsibility to contribute toward revealing and correcting miseducation as well as renewing respectful relationships with Indigenous communities.

Thank you to my supervisor, Dr. Scott Adams. From the very beginning, you have supported me and my interests, fostered my curiosity and creativity, and were a wonderful guide throughout this process. I feel incredibly fortunate to have had access to your guidance, your feedback, and your expertise.

Thank you to my advisory committee for your support and contributions. To Dr. Allyson Page, for always providing kindness and new perspectives. To Dr. Vijay Parsa, for helping me to navigate the tricky, complex world of hearing science and signal processing, and doing so in a way that always felt supportive and patient. To Dr. Mandar Jog, for your commitment to movement disorders and your enthusiastic feedback and insights. Thank you also to all members of the Movement Disorders Centre for your collaborations and conversations.

Thank you to my examiners. Your commitment to scholarship and your generosity with your time are greatly appreciated. To Dr. Julie Theurer, for being an inspiring example of a clinician-researcher and a kind and friendly face. To Dr. François Poiré, for being a wonderful reading course advisor many moons ago, and for supporting me at the beginning and the end of this journey. To Dr. Christopher Dromey, for being an inspiration in this field and for contributing your expertise.

Thank you to all of our participants, members of the Parkinson's community, and the Parkinson Society of Southwestern Society for your gracious support. This work would not have been possible without you generously donating your time and energy.

Thank you to the faculty and staff of Health and Rehabilitation Sciences and the School of Communication Sciences and Disorders. I am very grateful for each of your contributions to my learning.

Thank you to my lab mates, combined program colleagues, and my clinical cohort. To Anita, Cynthia, and Thea for making me feel so welcome and supported as I started on this journey, and I will always cherish the times we've spent exploring, celebrating, and commiserating! Special thanks to Thea, for being such a wonderful friend and mentor to me, always being ready for a rant, or a desperate code problem, or a read-over. I feel so grateful to have met you in this process. To Caitlin, for being at my side throughout these five years - your friendship has been such a blessing.

Thank you to all of my friends that I've met along the way. Each of you has provided me with support and with joy that I've needed to complete this work. Whether we've been friends for months or years or seen each other frequently or rarely, I have enjoyed your time and your company. To Becca, for joining me in many rants, sighs, and cheers.

Thank you to Kristi, for being the best sister I could have asked for. I am so grateful for you as my chosen family and am excited to keep sharing life's journeys with you.

Thank you to my family, for always enthusiastically cheering me on. To Mum, for all of your contributions to my life and my growth. I would not have been able to achieve this or experience many of life's blessings without the time and energy you've given me. To my parents, for being the wind beneath my wings. I am so blessed to have parents who support me in every imaginable way, allowing me to reach higher and farther. I am glad to call you my parents, my family, and my friends.

Thank you to Abdullah, for being my partner throughout this journey. I have learned so much from you and am grateful to have your support and your companionship. Thank you for always helping me to find balance, peace, and joy when I've needed it most.

It really does take a village. Thank you to everyone who has supported my life. To the people who grow my food, sew my clothes, deliver things around the world, design and produce the technology I love, provide me with the entertainment I need to stay balanced, and many more. I am grateful for the contributions of every person who has touched my life, near and far. I will make every effort to live my life with gratitude for the lifestyle we build for one another and will always seek to contribute to our social fabric in the ways I can.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

AGC: Automatic gain control

AIC: Akaike information criterion

ANOVA/s: Analysis/analyses of variance

CLS: Category loudness scaling

dB: Decibel

dB FS: Decibels relative to full scale

dB SPL: Decibels sound pressure level

DBS: Deep brain stimulation

DBS-STN: Deep brain stimulation of the subthalamic nucleus

DME: Direct magnitude estimation

EMM: Estimated marginal means

ERB: Equivalent rectangular bandwidth

HOA/s: Neurologically healthy older adult/s

HSD: Tukey's Honestly Significant Difference

Hz: Hertz

ICC: Intra-class correlation coefficient

kHz: Kilohertz

IWPD/s: Individual/s with Parkinson's disease

LDEL: Level difference at equivalent loudness

LKFS: Loudness (K-Weighted) relative to full scale

LMER: Linear mixed effects regression

LRT: Likelihood ratio test

LSVT: Lee Silverman Voice Therapy

LTAS: Long-term average spectrum

ME: Magnitude estimation

MoCA: Montreal Cognitive Assessment

MP: Magnitude Production

PD: Parkinson's disease

SIT: Sentence Intelligibility Test

SLM: Sound level meter

SNR: Speech-to-noise ratio

TVL: Time-varying loudness

UPDRS: Unified Parkinson's Disease Rating Scale

VAS: Visual analogue scales

VIF: Variable inflation factor

1 Thesis Overview

1.1 Objective

The objective of this investigation was to examine the relationships between perceived loudness and intelligibility with acoustic measures of loudness, speech intensity, and spectral distribution in individuals with hypophonia secondary to Parkinson's disease (IWPDs) and neurologically healthy older adults (HOAs). Quiet speech of IWPDs with hypophonia affects their communicative effectiveness and intelligibility. In the literature, hypophonia is often captured using vocal sound pressure level (SPL) or mean intensity. However, as a measure of physical strength that does not take listener factors into account, it is possible that acoustic measures of loudness may be more sensitive to hypophonia. Acoustic models of loudness have been designed to incorporate these listener factors, but application to speech has been limited. Consequently, the relationships between acoustic measures of loudness and perceived loudness and intelligibility of speech are not well understood. In IWPDs with hypophonia, features of hypokinetic dysarthria may further contribute to loudness deficits not captured by sound pressure measures alone. Disrupted spectral balance has been identified in IWPDs relative to HOAs, such that IWPDs may demonstrate a greater concentration of energy in lower frequency and weak contribution of higher frequencies. Spectral tilt has been shown to have a relationship with loudness, in that weak high-frequency energy can be associated with lower perceived loudness. These kinds of differences between IWPDs and HOAs may lead to acoustic loudness measures being of particular benefit in the measurement of hypophonia. Additionally, prosodic differences that may be associated with hypokinetic dysarthria, such as monoloudness and excessive loudness decay, may also influence overall perceived loudness. Relationships between spectral characteristics and perceived loudness may be a potential avenue into management of hypophonia via enhanced amplification devices, and an aim of this investigation is to provide evidence for further exploration of these avenues.

The present investigation involved three groups of participants. Sentence reading and conversational monologues were elicited from IWPDs and HOAs. Ten listener participants provided perceptual ratings of intelligibility using a visual analogue scale, and ratings of loudness using both visual analogue scaling and direct magnitude estimation. Listeners also provided perceptual ratings of loudness using a visual analogue scale from samples of speech that were manipulated to have a greater or lesser proportion of mid-high frequency energy (2-10 kHz). This spectral manipulation experiment was intended to specifically examine the relationship between spectral balance and perceived loudness in the context of equal mean intensity. Acoustic measures of loudness, speech level, spectral balance, and variability were investigated with regard to perceived loudness and intelligibility.

1.2 Organization of dissertation

Chapter 2 provides an overview of Parkinson's disease, hypophonia, and speech characteristics associated with Parkinson's disease (hypokinetic dysarthria), as well as a review of the many hypotheses and explanations for hypophonia. Loudness and its measurement are discussed, including characteristics of speech that may affect it. Finally, research questions and hypotheses of this investigation are presented. Chapter 3 describes the methodology of this investigation. The study is reported as two experiments:

1. **Experiment 1**: Natural speech
2. **Experiment 2**: Manipulated speech

Chapter 4 discusses the results of each experiment. In Chapter 5, the results are interpreted relative to the research questions, the literature, and the implications of the findings.

2 Introduction

2.1 Overview of PD and Hypophonia

Parkinson's disease (PD) is a neurological disorder characterized primarily by degeneration of dopaminergic neurons in the basal ganglia and connected brain regions. Based on the Mapping Connections report by Neurological Health Charities Canada, Public Health Agency of Canada, Health Canada and the Canadian Institutes of Health Research (2014), the estimated worldwide prevalence of PD is 428.5 per 100,000 people between the ages of 60 and 69, and 1903 per 100,000 people over the age of 80. In Canada, survey data suggests that 390 per 100,000 Canadians over the age of 45 have PD, rising to 1420 per 100,000 people over the age of 80. PD affects more men than women at about a 1.5:1 ratio.

The basal ganglia are a group of structures, including the striatum, the external and internal globus pallidi, subthalamic nucleus, and substantia nigra pars compacta and pars reticulata (Sapir, 2014). Dopamine fine-tunes neuronal excitability in the basal ganglia, and depletion results in physiologic imbalances which manifest as a variety of motor and non-motor symptoms (Obeso et al., 2010). Cardinal symptoms of PD include bradykinesia, rigidity, resting tremor, gait abnormalities, and postural instability. Additional symptoms associated with PD include dysphagia, anosmia, sleep disorders, cognitive abnormalities and a speech disorder known as hypokinetic dysarthria (Goldman, Williams-Gray, Barker, Duda, & Galvin, 2014; Jankovic, 2008).

No cure exists for PD, and treatment involves management of symptoms. Levodopa, a pharmacological dopamine replacement therapy, is widely considered the gold-standard for treatment of PD's motor symptoms (Fahn & Poewe, 2015). Deep-brain stimulation of the subthalamic nucleus (DBS-STN) is a treatment that is becoming increasingly prevalent, particularly later in the disease process when levodopa becomes associated with dyskinesias, unintended movements (Fahn et al., 2004). Both levodopa and DBS-STN have been associated with success in treatment of cardinal motor impairments, but effects of these treatments on speech and voice are less clear (Cushnie-Sparrow et al., 2018; Hammer, Barlow, Lyons, & Pahwa, 2011; Knowles et al., 2018; Spencer, Morgan, & Blond, 2009).

As many as 70-90% of individuals with PD (IWPDs) may develop speech and voice abnormalities, known as hypokinetic dysarthria, at some point in the disease process (Logemann, Fisher, Boshes, & Blonsky, 1978). In the seminal Mayo Clinic studies of dysarthria (Darley, Aronson, & Brown, 1969; Darley, Aronson, & Brown, 1969a), hypokinetic dysarthria was most closely associated with imprecise consonants, variable rate, short rushes of speech, reduced stress, monopitch, monoloudness, short rushes of speech, and variable rate. Additional characteristics of hypokinetic dysarthria include hypophonia (quiet speech) and abnormal voice quality (Adams & Dykstra, 2009).

Various acoustic abnormalities have been identified in hypokinetic dysarthria related to these perceived characteristics. IWPDs have been found to have reduced vowel and consonant distinctiveness on the basis of vowel space, formant transitions, spectral means, and voice onset time (Bunton & Weismer, 2002; Cushnie-Sparrow, Adams, Knowles, Leszcz, & Jog, 2016; Lam & Tjaden, 2016; Lansford & Liss, 2014; Rusz et al., 2013; Sapir, Spielman, Ramig, Story, & Fox, 2007; Tjaden, Lam, & Wilding, 2013). Additionally, there is evidence that IWPDs have a relatively greater concentration of energy in the lower frequencies than neurologically healthy older adults (HOAs). Dromey (2003) found that in sustained vowels, reading passages, and conversational monologues, IWPDs demonstrated lower spectral mean, lower spectral standard deviation and higher skewness. These characteristics all highlight a relatively greater concentration of energy in lower frequencies, and similar findings were reported by Smith and Goberman (2014).

Hypophonia may be the most common speech symptom associated with PD (Johnson & Adams, 2006) and may be most apparent during conversation (Adams, Dykstra, Jenkins, & Jog, 2008). On average, speech intensity of IWPDs is estimated to be 3-5 dB SPL quieter than HOAs (Adams et al., 2006; Adams, Haralabous, Dykstra, Abrams, & Jog, 2005; Adams, Winnell, & Jog, 2010; Adams et al., 2006a; Fox & Ramig, 1997; Matheron, Stathopoulos, Huber, & Sussman, 2017).

2.2 Noise and Distance

Noise and interlocutor distance are important factors to consider in the investigation of hypophonia, as both represent common adverse communication contexts that present particular barriers to communication for individuals with speech disorders. Increasing noise and increasing interlocutor distance can both require an increase in speech loudness to maintain effective communication. The effect of noise on speech was first described by Lombard (1911), consequently called the Lombard effect. The Lombard effect indicates that with an increase in background noise, individuals automatically speak more loudly (Lane & Tranel, 1971). This has important implications for understanding hypophonia's etiology with connections to sensorimotor integration, self-monitoring, and cuing, and for developing treatments, including Lombard-based treatments designed to trigger automatic increases in speech loudness.

Inconsistent findings have been reported in the literature regarding the Lombard effect and response of interlocutor distance in IWPDs, with some reports of attenuated Lombard effects and a limited response to increased distance in IWPDs (Ho, Iansek, & Bradshaw, 1999a). However, a majority of studies support that IWPDs demonstrate a response to noise and distance that is similar to HOAs, despite lower speech intensity across conditions (Adams et al., 2006; Adams et al., 2005; Adams & Lang, 1992; Adams et al., 2006a; Stathopoulos et al., 2014).

Even in the context of similar increases in intensity between IWPDs and HOAs, IWPDs may still be less intelligible in the presence of noise due to the contribution of speech-to-noise ratio (SNR). Adams et al. (2008) investigated SNR and intelligibility in IWPDs and HOAs speaking conversationally in multiple levels of background noise. In 70 dB SPL of background noise, approximately the level of a moderately busy cafeteria, IWPDs had only 1.4 dB SNR and 45% intelligibility. A SNR of 5-7 dB was deemed to provide approximately 80% intelligibility for both HOAs and IWPDs, indicating that the SNR achieved by IWPDs was not sufficient for them to be intelligible. Similarly, Dykstra, Adams, and Jog (2013) found a non-parallel effect of noise on intelligibility, in that IWPDs' intelligibility was more affected by increased noise. HOAs were able to maintain normal intelligibility across noise levels. While IWPDs may have a similar response to noise as HOAs, the intelligibility consequences may be amplified in IWPDs with hypophonia since their generally lower speech intensity leads to lower SNR.

In the present investigation, noise and interlocutor distance were not examined, allowing a narrower focus on speech produced in what might be considered optimal conditions (minimal noise, small interlocutor distance). However, it is possible that the underlying characteristics of hypophonic speech investigated in the present study would have greater effects on loudness and intelligibility at higher levels of background noise, in highly reverberant spaces, or at larger interlocutor distances.

2.3 Underpinnings of Hypophonia

Many causes and contributors have been suggested as underpinnings of hypophonia, and it is likely that these contributors combine in a multifactorial way to produce hypophonia. Hypothesized contributors include physiological deficits, sensory and somatosensory deficits, abnormal sensorimotor integration, deficits in loudness perception and autophonic loudness perception, abnormal perception of effort, inappropriate scaling, and deficits in cuing.

2.3.1 Sensory and Somatosensory Deficits

The basal ganglia have connections and relationships with many brain regions, and as a result, dopamine dysregulation in PD affects much more than the basal ganglia. Differences in activity and connectivity of various brain regions have been identified (Cao, Xu, Zhao, Long, & Zhang, 2011), in part due to extensive intermingling of sensory and motor activity in the striatum (Conte, Khan, Defazio, Rothwell, & Berardelli, 2013). IWPDs have been found to show reduced sensitivity to proprioceptive information (Conte et al., 2013), and abnormal patterns of activity in cortical and cerebellar areas (Cao et al., 2011; Rascol et al., 1992).

Sensory deficits specific to laryngeal structures may be even more closely related to hypophonia. Hammer and Barlow (2010) found that the laryngeal mucosa of IWPDs was less sensitive to air bursts, which the authors suggested was likely to affect respiratory and phonatory control, thereby contributing to speech deficits. Specifically, abnormal sensitivity of laryngeal mucosal mechanoreceptors could generate a false sense of effort when speaking. It is possible that these sensory deficits contribute in a bottom-up fashion, in that poorer sensation leads to poorer sensory feedback which then further disrupts the integration of this sensory feedback with motor plans. The authors speculate that this could be explained by an increase in sensory gating. Sensory gating is the process through which irrelevant or unhelpful information is filtered out by a sensory system. Hammer and Barlow (2010) hypothesized that increased gating at the laryngeal somatosensory level leads to compensation, manifesting in an increased sensitivity to auditory feedback.

This increased sensitivity would then result in reduced loudness. However, that study did not directly examine sensory gating, as somatosensory levels during laryngeal movements were not measured. Other perspectives on the effects of sensory gating in IWPDs exist. Conte et al. (2017) investigated the relationship between tactile perception and motor actions in the index fingers of individuals with IWPDs and HOAs. Their findings identified abnormal reduction in sensory gating (loss of sensory selectivity) during movement in PD and found that this gating deficit was related to the severity of bradykinesia. In the auditory realm, Gulberti et al. (2015) provided support for a reduction in auditory sensory gating in PD via electroencephalography data collected while participants vigilantly listened to rhythmic clicks. These authors suggest that the pedunculopontine nucleus of the pons may play a key role here. The pedunculopontine nucleus has close links with the basal ganglia's subthalamic nucleus. Their findings indicated that DBS-STN may improve this gating deficit, further suggesting that abnormal basal ganglia function in PD underpins this deficit. However, the influence of the process of speech production on such a gating deficit is currently not well understood. Arnold, Gehrig, Gispert, Seifried, and Kell (2014) found increased activation of the auditory cortex during speech in IWPDs relative to HOAs. These findings may further support the notion of decreased gating of auditory information during speech, in that more auditory-sensory information related to self-intensity is passed through to the auditory cortex. Additionally, task effects have been suggested to be an important consideration in studies of sensory gating (Lei, Ozdemir, & Perez, 2018), which may make it challenging to predict the nature and results of abnormal sensory gating as it relates to loudness in PD.

Primary auditory deficits have also been identified in IWPDs. Higher prevalence of peripheral sensorineural hearing impairments has been observed in IWPDs (Vitale et al., 2012). IWPDs have also demonstrated deficits in central auditory processing and have reported greater difficulty hearing spoken words (Folmer, Vachhani, Theodoroff, Ellinger, & Riggins, 2017). Chen and Watson (2017) investigated the relationship between tactile and auditory sensation related to /s/ and /sh/ production and perception in IWPDs and healthy older adults. Their findings indicated that IWPDs were less sensitive to auditory and tactile stimuli, with higher just-noticeable-difference and area of uncertainty, consistent with a flatter psychophysical function. These findings of reduced sensitivity to tactile stimuli in the vocal tract and reduced sensitivity to speech-relevant auditory stimuli, as well as the observed differences in cortical and cerebellar areas, are examples of abnormal sensory processes that could contribute to hypophonia by undermining IWPDs' ability to perceive their productions via auditory and proprioceptive routes.

Given this evidence of sensory abnormalities in IWPDs, it follows that IWPDs may have difficulty combining sensory information with motor plans to produce efficient and accurate movements. General sensorimotor integration deficits have been identified in PD, including abnormal facial reflexes (Caligiuri & Abbs, 1987; Schneider, Diamond, & Markham, 1986) and abnormal integration of feedback in finger musculature (Tamburin et al., 2003). Within the speech domain, previous investigations of sensorimotor adaptation have suggested that abnormalities may exist in the feedback and feedforward systems of IWPDs (Abur et al., 2018; Ho, Bradshaw, Iansek, & Alfredson, 1999; Liu, Wang, Metman, & Larson, 2012; Mollaei, Shiller, & Gracco, 2013; Senthinathan, Adams, Page, & Jog, 2021). Mollaei et al. (2013) and Abur et al. (2018) employed pitch perturbation using sensorimotor adaptation paradigms to investigate the responses of IWPDs and HOAs to altered auditory feedback. Both sets of findings demonstrated that IWPDs showed reduced compensation to perturbation relative to HOAs, which Abur et al. (2018) hypothesized could mean that IWPDs are over-relying on the feedback system due to an impaired feedforward system. This feedforward deficit aligns well with the identified sensory deficits discussed previously. However, Liu et al. (2012) found an increased magnitude of response to perturbations in pitch and intensity among IWPDs. Senthinathan et al. (2021) investigated the response of IWPDs and HOAs to altered intensity feedback, but using long-term alterations rather than short-term perturbations, and found reduced compensation to altered intensity in IWPDs similar to findings of Mollaei et al. (2013) and Abur et al. (2018). It is possible that the method of altering feedback explains the inconsistencies in these findings. Across these investigations, it is clear that there are anomalies in the sensorimotor integration of IWPDs, although specific details continue to emerge in this line of research.

2.3.2 Loudness Perception

Auditory deficits (Chen & Watson, 2017; Folmer et al., 2017; Troche, Troche, Berkowitz, Grossman, & Reilly, 2012; Vitale et al., 2012) are likely to affect the way IWPDs perceive loudness. Dromey and Adams (2000) asked IWPDs and HOAs to provide direct magnitude estimation (DME) ratings of the loudness of pure tones and identified no group differences between IWPDs and HOAs on this task of external loudness perception. Similarly, Abur, Lupiani, Hickox, Shinn-Cunningham, and Stepp (2018a) did not find group differences in perception of pure tones between IWPDs and HOAs. Clark, Adams, Dykstra, Moodie, and Jog (2014) presented IWPDs and HOAs with a standard pre-recorded speech sample at varying levels of intensity and asked participants to rate loudness using magnitude estimation. No significant group differences were identified, but a trend indicated a flatter psychophysical function for loudness and restricted range of loudness ratings, based on IWPDs overestimating loudness of quieter stimuli and underestimating loudness of louder stimuli. It is possible that differences in pure tone and speech loudness perception are responsible for this identified trend. Richardson and Sussman (2019) investigated intensity discrimination of vowels in IWPDs, HOAs, and young adult controls. In two experiments of differing complexity, participants were asked to identify which samples differed in intensity of presentation, with samples differing in 1 dB increments from 1-8 dB above the 70 dB SPL standard. IWPDs demonstrated poorer intensity discrimination than controls and a shallower discrimination function slope as the intensity contrast widened. IWPDs required a larger intensity contrast to achieve their best discrimination performance (7-8 dB) compared to controls (4-5 dB). In a loudness rating task, both IWPDs and HOAs demonstrated a flatter psychophysical function of loudness compared to younger controls. This absence of a group difference between IWPDs and HOA is consistent with Abur et al. (2018a), but not with the trend of Clark et al. (2014), which may suggest that perception of vowels is more analogous to pure tone perception than perception of connected speech. More research is needed to clearly understand loudness discrimination and loudness perception of pure tones, vowels, and connected speech in IWPDs and HOAs.

The studies described above investigated perception of external stimuli, but it is possible that hypokinetic dysarthric deficits may further interfere with how IWPDs use judgments of their own loudness (autophonic loudness) to regulate their speech intensity. Ho et al. (1999a) studied self-loudness perception of IWPDs and HOAs using a loudness-matching paradigm in which participants adjusted the volume knob on a tape player until their player sounded as loud as a second player at either 1 m or 8 m. While HOAs and IWPDs provided similar loudness ratings at the 1 m distance, IWPDs showed a smaller difference in their judgments of loudness between these near and far distances. The authors interpreted this result as a flatter psychophysical function of loudness, leading to soft sounds seeming louder and loud sounds seeming softer. In a further investigation, Ho, Bradshaw, & Iansek (2000) asked 15 IWPDs and 15 HOAs to read aloud a standard passage, estimate their own loudness immediately (autophonic), and then estimate their loudness based on a playback of their voice using the same volume knob procedure. Participants completed this series of tasks at their habitual loudness, during quiet speech, and during loud speech. IWPDs again estimated their quiet speech as louder than it was, supporting this hypothesis of a flatter psychophysical function of loudness in IWPDs. Clark et al. (2014) also included magnitude production and imitation tasks in their investigation of loudness perception in IWPDs, providing insight into self-loudness perception. Consistent with Ho et al. (1999a) and Ho et al. (2000), IWPDs overestimated quieter stimuli and underestimated louder stimuli. Similar findings were also identified by Keyser et al. (2016). However, Brajot, Shiller and Gracco's (2016) investigation of 12 IWPDs and 12 HOAs did not identify differences in autophonic loudness perception between groups.

It is possible that findings regarding loudness perception of speech are complicated by stimuli being self-generated or external, as autophonic loudness perception may be differentially affected in PD. However, controversy continues, as IWPDs in studies with both external speech stimuli (Clark et al., 2014) and self-generated speech (Clark et al., 2014; Ho et al., 1999a, 2000; Keyser et al., 2016) have demonstrated a trend of a flatter psychophysical function, despite other studies having failed to find significant group effects (Brajot et al., 2016; Dromey & Adams, 2000). Clark et al. (2014) and Ho et al. (1999a; 2000) recruited solely IWPDs with hypophonia to participate in speech-based loudness perception tasks, and it is possible that greater consistency between these investigations may have aided in clarifying results. Imitation tasks may be particularly helpful in revealing loudness perception deficits (Clark et al., 2014). Abur et al. (2018a) concluded that it is unlikely that abnormal loudness perception plays a primary role in hypophonia; however, this is based on an investigation of pure tone loudness perception. Future investigations should seek to clarify the relationships between loudness perception of pure tones and externally-generated and self-generated speech stimuli specifically in IWPDs with hypophonia.

2.3.3 Effort Perception and Cuing

IWPDs may perceive their effort differently than HOAs, which could affect their ability to regulate speech intensity and contribute to hypophonia. Vocal effort is complex, and can be defined as "perceived exertion of a vocalist to a perceived communication scenario," encompassing a combination of physiological effort, the experience of effort, psychological effort, effort as a speech production level (intentionally speaking with more or less effort), and effort as affected by communication environment (interlocutor distance, background noise, time in vocal use; Hunter et al., 2020). Solomon and Robin (2005) investigated IWPDs' and HOAs' task-related, generalized, and speaking effort ratings. IWPDs provided greater overall ratings of effort, but moment-to-moment effort levels were not significantly different between groups. IWPDs may have an inflated sense of overall effort, which could contribute to hypophonia by leading individuals to reduce their speech intensity to align with what feels like habitual effort. However, the current evidence for the role of effort perception in hypophonia is very limited, and it is difficult to support effort and calibration as having a primary role in hypophonia in the absence of additional investigations in this area. However, effort is a critical component of the popular Lee Silverman Voice Therapy (LSVT) LOUD voice therapy for IWPDs. LSVT is an intensive voice treatment, focusing on a high number of repetitions per task, a focus of increased movement amplitude directed mostly at respiratory and laryngeal components, and frequent, specific feedback about speech intensity. In LSVT, modeling is used as the primary method of instruction, with the rationale that this avoids excessive cognitive burden (Sapir, Ramig, & Fox, 2011). Additionally, IWPDs are encouraged to 'recalibrate' their effort and loudness perception, learning to recognize that they have been speaking too quietly and that their target voice is not too loud. This relates to the loudness perception literature reviewed in Section 2.3.2, as IWPDs may overestimate the loudness of quiet sounds and underestimate the loudness of loud sounds. Literature support for the efficacy of LSVT is strong, particularly for short-term results, though many studies are related to the original research group (Constantinescu et al., 2011; Fox, Morrison, Ramig, & Sapir, 2002; Fox & Ramig, 1997; Howell, Tripoliti, & Pring, 2009; Ramig, Halpern, Spielman, Fox, & Freeman, 2018; Ramig, Countryman, O'Brien, Hoehn, & Thompson, 1996; Ramig et al., 1995; Ramig et al., 2001; Sapir et al., 2007; Theodoros et al., 2006; Tindall, Huebner, Stemple, & Kleinert, 2008). Long-term maintenance of these treatment benefits is less clear, with inconsistent findings in the literature (Watts, 2016; Wight & Miller, 2015). It may also be difficult for some clients to apply the knowledge they learn in a quiet, controlled clinical environment to real-life communication environments with noise, reverberation, higher interlocutor distance, and greater cognitive demand.

It is also possible that IWPDs have deficits in internal cuing, resulting in difficulty adjusting their speech intensity. Explicit cues are externally generated, specific feedback about the action that needs to be taken, such as increasing speech intensity. For example, when an IWPD is asked to speak louder or is asked to repeat something they have said, this is an explicit cue to speak louder. Implicit cues are based on external information, but without an explicitly defined action. For example, noise and distance may represent implicit cues to speak louder, as they are known to negatively affect intelligibility. Internal cues are self-generated, such as an IWPD attempting to remember to increase their speech intensity in general, knowing that they are generally too quiet. Explicit cues have been found to mitigate hypokinetic deficits of IWPDs in gait (Ford, Malone, Nyikos, Yelisetty, & Bickel, 2010) and writing (Oliveira, Gurd, Nixon, & Marshall, 1997), and have also improved vocal loudness and speech clarity in IWPDs (Tjaden et al., 2013; Tjaden & Wilding, 2004). These studies indicate that explicit cuing improves the ability of IWPDs to generate appropriate scaling in the context of writing, walking and regulating speech intensity. However, the effects of implicit cues (such as interlocutor distance and noise) are less clear. If IWPDs possess implicit cuing deficits that contribute to hypophonia, it would be expected that IWPDs would be able to increase speech intensity when receiving explicit cues (being asked to speak louder), but not when receiving implicit cues (distance and noise). Findings of a parallel Lombard effect discussed previously do not support this, as IWPDs and HOAs both increased their speech intensity in response to noise and distance cues (Adams et al., 2006; Adams et al., 2005; Adams & Lang, 1992; Adams et al., 2006a; Stathopoulos et al., 2014).

It is possible that IWPDs are differentially impaired in internal cuing (self-generated cues), and that responses to implicit and explicit cues remain intact. An internal cuing deficit could manifest as an inability to self-cue to maintain appropriate speech intensity, but an ability to increase speech intensity when requested or when prompted by environmental factors. Even in studies in which IWPDs responded to these implicit cues by increasing speech intensity, IWPDs with hypophonia were still 2-5 dB quieter on average than HOAs. This may represent a scaling deficit, rather than a cuing deficit, to be further discussed in Section 2.3.4. IWPDs might be seen as responding to the cues, but not to an appropriate extent due to scaling limitations. Additionally, differences in loudness perception and sensorimotor integration, discussed previously, may interact with the way IWPDs perceive cues, thereby affecting their responsiveness and providing important clinical implications to management of hypophonia. As discussed above, the relationships between effort, cuing, sensorimotor integration, and loudness perceptions are fundamental underpinnings of LSVT, which may provide further support for the need to explore these relationships in IWPDs. It is also possible that cuing deficits do not exist in IWPDs with hypophonia, and that difficulties are better explained by loudness perception deficits discussed previously. However, based on the inconsistency in the literature surrounding loudness perception of speech and the more consistently negative findings regarding loudness perception of pure tones, it may not be possible to rule out the existence or role of cuing deficits in PD.

2.3.4 Scaling

As introduced above, it is possible that hypokinesia, a downscaling of movement amplitude, is an underlying feature of hypophonia given its role as an important feature of PD as a whole. Hypokinesia in PD has been hypothesized as stemming from the basal ganglia's role in movement planning amplitude. Investigations of the use of visual cuing in scaling of hand movements have indicated that hypokinesia in IWPDs specifically affects scaling of amplitude, as opposed to scaling of direction (Desmurget, Grafton, Vindras, Grea, & Turner, 2004). Downscaling of movement amplitude in oral musculature of IWPDs was also identified through lip and jaw kinematic analyses of Walsh and Smith (2012). Reduced habitual intensity is a key characteristic of hypophonia, and has been discussed above. Scaling deficits may not be particularly influential in this overall reduction of intensity. However, it is possible that this downscaling of amplitude leads to a decreased range of available vocal amplitudes, as demonstrated by changes in maximum intensity and changes in the responses of IWPDs to noise and distance effects. In the domain of speech intensity, reduced maximum intensity has been observed in IWPDs relative to HOAs (Adams et al., 2006a). Ho, Iansek and Bradshaw (1999a) investigated loudness perception of IWPDs at distances between 1 m and 8 m, and conversation and counting samples were covertly collected to examine their response to distance cues. IWPDs did not adjust to increasing distances as much as HOAs, which the authors also interpreted as possible evidence of a decreased motor set related to speech intensity. As discussed above, however, the majority of studies of hypophonia and distance have not identified differences in how IWPDs and HOAs respond to noise and distance. Future investigations of the relationships between movement scaling deficits and loudness scaling deficits are needed.

2.3.5 Physiological Contributors to Hypophonia

Physiological explanations for hypophonia also exist, such as vocal fold bowing, laryngeal rigidity, and respiratory deficits. Vocal fold bowing, in which the glottal folds do not fully approximate during phonation due a bowed shape, has been identified in a majority of IWPDs (Hanson, Gerratt, & Ward, 1984). Vocal fold bowing influences hypophonia because an inability to achieve sufficient medial compression of the vocal folds reduces the ability to achieve adequate speech intensity. Laryngeal electromyography findings have identified two possible explanations for hypophonia: rigidity and hypokinesia. Both are primary features of PD as a whole. Baker, Ramig, Luschei, and Smith's (1998) findings were more consistent with hypokinesia, and Gallena, Smith, Zeffiro, and Ludlow's (2001) results among de novo IWPDs were more consistent with rigidity. A possible explanation of this discrepancy is that prolonged levodopa use may change the activity patterns of the thyroarytenoid musculature in IWPDs, consistent with rigidity playing a larger role in early stages of the disease. Ho, Bradshaw, and Iansek (2008) described hypophonia as a possible laryngeal analogue of limb hypokinesia. The authors connect laryngeal aspects of hypophonia to the role of the basal ganglia in regulation of force, similar to Desmurget et al. (2004). IWPDs have also demonstrated lower subglottal pressure and glottal closed time, contributing to hypophonia, as increasing subglottal pressure is an important physiological adjustment needed to increase speech intensity (Isshiki, 1964; Matheron et al., 2017).

In addition to these identified abnormalities in laryngeal physiology, the respiratory system has also been implicated in PD, including reduced forced vital capacity, forced expiratory volume, lower lung volume initiations and terminations, larger abdominal volume initiations, smaller rib cage volume initiations, abnormal rib cage excursions, and more variability in respiratory movements than HOAs (Huber & Darling, 2011; Sadagopan & Huber, 2007; Solomon & Hixon, 1993). Reduced respiratory driving pressure can also contribute to reduced subglottal pressure (Hammer & Barlow, 2010).

As introduced in the context of the potential role of scaling deficits in hypophonia, Walsh and Smith (2012) investigated oral hypokinesia in 16 IWPDs using lip and jaw kinematics as well as with acoustic measures of speech intensity, and vowel formants. IWPDs spoke with reduced lower lip and jaw movement amplitudes and velocities, decreased vocal intensity and shallower formant slopes. Similarly, small lip, jaw, and tongue movements and small articulatory working space have been identified in IWPDs (Dromey, 2000; Kearney et al., 2017; Yunusova et al., 2017). These studies on speech movement hypokinesia focused on the effects of speech movement size on articulation and intelligibility. However, it is also possible that small oral aperture could reduce speech intensity and loudness. Adams, Dykstra, and Jog (2012) examined the speech intensity of IWPDs and HOAs speaking in multiple levels of background noise, simultaneously recorded using a throat microphone and a head-mounted microphone 8 cm from the mouth. Findings indicated that the speech intensity of HOAs was more stable between the throat and the mouth. IWPDs demonstrated lower speech intensity at the head-mounted microphone than at the throat microphone, whereas HOAs demonstrated no significant difference. The authors hypothesized that IWPDs show an abnormality in the use of mouth opening to modulate speech intensity, and findings support the contribution of supraglottic factors in the presentation of hypophonia. Small oral aperture of IWPDs may also alter normal resonance. At the extreme, this could simulate a cul-de-sac resonance, contributing to muffled-sounding speech (Kummer, 2020). More research is needed to clearly understand the effects that oral aperture might have on resonance, speech intensity and loudness in IWPDs with hypophonia. Insights into the effects of small oral aperture on loudness and intelligibility would create new treatment avenues.

A challenge in the investigation of the role of physiological deficits in hypophonia is the problem of correlation versus causality. Some relationships have been identified between hypophonia and abnormal physiology; however, it is unclear whether these physiologic differences are causes and contributors to hypophonia or are epiphenomena of hypophonia. For example, if hypophonia is primarily driven by sensory or sensorimotor contributions, IWPDs with hypophonia speaking at a reduced speech intensity could demonstrate changes in physiological measures such as vocal fold closure and respiratory volumes as a result of the changes in speech intensity.

2.4 Measurement of Hypophonia

Hypophonia is often measured acoustically or perceptually, and physiological factors hypothesized to contribute to hypophonia such as laryngeal, articulatory and respiratory deficits may be measured with glottography, electromyography, kinematics or aerodynamics. While acoustic measurements of hypophonia are common in the literature, hypophonia can also be judged perceptually by clinicians, IWPDs, partners of IWPDs, or naïve listeners. Judgments can be provided with pre-existing tools such as the Voice Handicap Index (Jacobson et al., 1997) which uses a visual analogue scale, allowing raters to provide a continuous rating along a fixed line. Other rating tools provide a discrete scale requiring a rater to quantify their judgment (i.e., on a scale from 1-10) or categorize the individual's speech (i.e., not impaired, mildly impaired, moderately impaired). Ratings can also be provided informally, such as the rater's overall opinion of their own speech or of an IWPD's speech.

Speech intensity is a very common measure of hypophonia, as it is thought to represent the acoustic correlate of perceived loudness. This is not entirely true, as will be further discussed in Section 2.5. Speech intensity may be more accurately known as speech sound level or speech sound pressure level, as discussed by Švec and Granqvist (2018). Despite this, the term speech intensity (or voice intensity) is very commonly used in the hypophonia literature to describe speech sound level, and the term mean intensity is used in this investigation. Sound pressure level (dB

SPL) is a decibel-scaled expression of sound pressure (in Pascals) and provides an estimate of sound power.

Consistency in methodology is particularly important in acoustics. Throughout their tutorial on the measurement of voice sound pressure level (SPL), Švec and Granqvist (2018) emphasize the need for strong and consistent methodology. In addition to these recommendations, an expert panel assembled by the American Speech-Language and Hearing Association (ASHA) sought to assemble a collection of best practice guidelines for instrumental voice assessment (Patel et al., 2018). Rusz, Tykalova, Ramig, and Tripoliti (2021) recently published a set of guidelines extending these recommendations more specifically to dysarthrias of movement disorders. These guidelines provide a well-rounded framework for assessment, though manipulations in noise and interlocutor distance are not discussed which may improve ecological validity of assessment for this population. In this framework, however, mean intensity is presented as the acoustic measure representing the dimension of speech loudness without a caveat as to the gaps this may introduce. It is hoped that the present investigation can provide additional insights into the effectiveness of mean intensity in representing loudness.

When measuring and reporting speech intensity, it is critical that calibrated equipment is used to ensure that the levels reported are correctly referenced to the reference pressure of the dB SPL scale, as outlined by Švec and Granqvist (2018). Calibration involves a sound level meter (SLM) and is complicated by factors like distance and SLM settings (time-weighting, frequency-weighting), and significant differences in values can occur with variations in these settings. Detailed description of calibration procedures should be included in manuscripts reporting acoustic measures of speech, as methodological differences complicate knowledge synthesis across studies and across research groups. A scoping review was recently conducted by this author to characterize the methodological variability of hypophonia studies (Cushnie-Sparrow, Adams, Page, and Parsa, 2018 [in prep]). Findings indicated that in many ways, methodologies of studies of hypophonia have been consistent with recent best practice guidelines (Patel et al., 2018; Rusz et al., 2021) in terms of measures and tasks employed. An area of improvement would be better application and reporting of appropriate calibration procedures. Specifically, room conditions, microphone and SLM configurations and distance, SLM frequency weighting, calibration procedure and digital calibration adjustment method, software analysis methods and contour-averaging methods should be consistently reported in future studies of hypophonia. This would improve the clarity and specificity of studies of hypophonia and allow for improved knowledge syntheses and meta-analyses in the future. Additionally, it was found that loudness measures have not yet been applied to deepen our understanding of the nature of hypophonia, a driving factor in the development of the present investigation.

2.5 Loudness

Loudness has been defined as the subjective intensity of a sound, and is often described as the psychophysical correlate of sound level (Florentine, 2011). Prior studies investigating the speech intensity of IWPDs with hypophonia relative to HOAs mostly employed intensity (dB SPL) as a proxy for loudness. However, intensity in dB SPL is not designed to represent perceived loudness, but rather to convey sound pressure on an appropriate scale. Several metho ds have been employed to quantify loudness in ways that take listener characteristics into account, which sound pressure level does not.

Two metrics used to express loudness are loudness level, in phons, and loudness, in sones. A loudness level of *N* phons is as loud as a 1 kHz tone at *N* dB SPL (International Standards Organization [ISO] 226, 2003; Marks & Florentine, 2011). Conversion from sound pressure level to loudness level requires equal-loudness contours developed using equal-loudness matching, in which listeners adjust a variable stimulus to match a standard stimulus. A more detailed exploration of equal-loudness matching is provided in Section 2.5.1. Equal-loudness curves centered around 1 kHz were originally reported by Fletcher and Munson (1933) based on loudness matching data. Additional studies have sought to replicate and refine these curves, with generally similar results across investigations. A synthesis by Suzuki and Takeshima (2004) sought to compile many of these replication efforts into a single set of curves. Jesteadt and Leibold (2011) highlighted the similarity between Fletcher and Munson's (1933) loudness power functions and the more recent, complex, and robust calculations.

Regardless of the equal-loudness curve selected for use, a loudness level of 40 phons corresponds to 40 dB SPL at 1 kHz, and other frequencies are quantified relative to the 40 phons equal-loudness curve. Loudness level in phons provides both a nominal and ordinal indicant of loudness; nominal, in that all acoustic signals equal in loudness are equal in loudness level, and ordinal, in that relative loudness of signals can be ranked (Marks & Florentine, 2011). Jesteadt and Joshi (2011)'s Figure 5.1 provides a clear depiction of equal-loudness contours as outlined by the American National Standards Institute [ANSI] and Acoustical Society of America [ASA] S3.4 (2007) standard, as modified by Glasberg and Moore (2006).

Sones are a unit for loudness in which 1 sone is the loudness of a 1 kHz tone presented in a free field at 40 dB SPL. A sound with a loudness of 2 sones is twice as loud as a sound with a loudness of 1 sone, a property of true ratio scales that many procedures used to measure loudness do not have (Jesteadt & Joshi, 2013). The original sone function came from the Fletcher and Munson (1933) study introduced above. Matching data was converted to ratios by assuming that a tone presented binaurally would be twice as loud as the same tone presented monaurally, and also by assuming that a tone complex consisting of *n* equally loud tones with wide spacing in frequency would be *n* times as loud as a single tone. Sones were obtained by dividing values by the value for a 40-dB tone. Jesteadt and Joshi (2013) provide an overview of the progression from this original function for sones to the more recent ANSI/ASA S3.4 (2007) standard. Since the writing of Marks and Florentine (2011), the ANSI/ASA S3.4 (2007) model has been further expanded by Moore and Glasberg, leading to the current ISO 532-2 (2017) standard. This model of time-varying loudness (TVL) will be reviewed in greater detail in Section 2.5.2.3.

2.5.1 Perceived Loudness

Perceptual measures of loudness can be used in isolation as reliable and valid measures of loudness. Marks and Florentine (2011) highlighted two key characteristics of acceptable loudness measurement: measures must be internally consistent, such that greater intensity of signal A relative to signal B must correspond to greater loudness of signal A, and be transitive, in that if signal A is louder than signal B, and signal B is louder than signal C, signal A must be louder than signal C. While logical, these broad, fundamental characteristics leave considerable room for diversity in measurement of loudness. Some of the current methods available to subjectively quantify loudness include scaling methods, such as category loudness scaling, visual analogue scaling, and magnitude estimation, as well as equal-loudness matching.

Equal-loudness matching was introduced above as the method used in Fletcher and Munson (1933). Loudness matching can employ a simple adjustment method or more advanced, adaptive methods. In adjustment paradigms, listeners are presented with two sounds, a standard tone and a tone that varies in frequency and level across trials. Listeners have direct control over the variable tone, and are asked to adjust it until its loudness matches the standard tone. While simple, this paradigm can be affected by systematic errors including time-order bias, systematic preference for the first or last stimulus depending on the interstimulus interval, and a bias to comfortable listening levels leading to overestimation of low levels and underestimation of high levels. In adaptive methods, the listener is again presented with two stimuli, but instead of controlling the tone, they identify which tone is louder. Their response determines the presentation level of the next trial (e.g. using an up-down procedure). The amount of change in level reduces as a point of subjective equality is approached, and results of loudness-matching are often described by the level difference at equal loudness (LDEL). Adaptive methods are affected by different error patterns than the adjustment method. Listeners can become aware of which stimulus is varied and may try to adjust the level by perseverating or changing responses. They may also compare stimuli to their memory of past items, rather than the current item. Critically, a fine balance must be sought between variability of responses and the number of trials subjects complete. Variability decreases with a higher number of trials; however, large numbers of trials may produce fatigue and increase variability over time (Marks & Florentine, 2011).

While loudness level provides a specific loudness equivalent, or a rank order of loudness, it does not specifically express the extent to which the loudness of one sound exceeds that of another (Marks & Florentine, 2011). Loudness scaling seeks to fill this gap. One form of scaling is categorical loudness scaling (CLS). Listeners are presented with a sound and provide a rating on a discrete category scale. For computational purposes it is often assumed that equal distance between category labels exists, but it should be noted that this is not a property of interval scales (Stevens, 1946). A key advantage of CLS is that it is simple and easy to administer with minimal training (Rasetshwane et al., 2015). However, judgments on rating scales are relativistic, and the lowest and highest stimuli levels may serve as anchors. This can lead to reduced variability in extreme categories relative to responses in between, referred to as an edge resolution effect. Concerns about the internal consistency of judgments made on categorical scales have also been raised (Marks & Florentine, 2011). Hellman (1999) also discusses that that CLS uses arbitrary units that do not relate to standard units of loudness and have been found to underestimate the slope of loudness growth. Additionally, in order to increase the amount of information transmitted and mutual discriminability among stimuli, a large number of response categories is needed (Marks, 1968). ISO 16832 (2006) outlines the standard for loudness scaling by categories based on work by Appell (2002) and Brand and Hohmann (2002).

Increased information transmission and discriminability are among the reasons that visual analogue scales (VAS) are becoming more popular. In VAS, listeners rate each stimulus using a line segment of set length, responding by making a mark at the point on the line that corresponds to their perceived loudness (i.e. a cross further toward the left typically means a weaker stimulus). VAS is a bounded, continuous scale presented as line segments, described as an effective method of scaling because individuals can easily use spatial length and position as metaphors for perceived strength (Lakoff & Johnson, 1980). Like CLS, VAS is easy to administer with minimal training, has been found to be more reliable than equal appearing interval scales in the context of voice quality (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993), and greatly improves resolution (Karnell et al., 2007). Marks and Florentine (2011) suggest that VAS is likely to avoid many of the pitfalls of CLS. While VAS has not been used as frequently as CLS in the loudness literature, it has precedent in the measurement of loudness within the hypophonia literature (Ma, Whitehill, & Cheung, 2010; Ramig et al., 1995; Wight & Miller, 2015).

Magnitude estimation (ME) is an unbounded, continuous scaling method for quantification of loudness and other perceptual parameters. In ME, the listener is presented with a series of stimulus levels and then is asked to respond with a number that matches its number. ME can include a fixed modulus, whereby listeners are asked to rate a target stimulus in relative quantities of the modulus. Over time, some studies have shifted away from a modulus entirely. In absolute ME, instructions avoid any reference to ratio relations between stimuli, and listeners are instead encouraged to assign any numeral to the stimulus to match the perceived magnitudes of the sensation (Marks & Florentine, 2011). Similar to ME, in magnitude production (MP), the subject hears a variable stimulus and is instructed to adjust its loudness to match a target number (Marks & Florentine, 2011). In the context of speech, MP paradigms have also been used by requesting speakers to adjust their loudness by a given ratio relative to their habitual loudness or relative to a target stimulus (Clark et al., 2014). Like other scaling methods, ME and MP are easy to administer with minimal training. Similar to VAS, ME and MP provide ratio-based data, avoiding the statistical concerns associated with interval scales such as CLS (Gescheider, 1997). Additionally, by measuring across a wide range of stimulus intensities, experimenters are able to generate overall psychophysical magnitude functions. However, ME and MP are still subject to systematic context biases (McRobert, Bryan, & Tempest, 1965). It is possible that in absolute ME, these context effects are particularly pronounced, as the absence of a defined scale means that listeners are relying more heavily on their own frame of reference.

ME has a history of usage in hearing research, championed by Stevens (1955, 1956) and subsequently used in many studies investigating loudness of tones (Epstein & Florentine, 2006; Marks & Florentine, 2011; McRobert et al., 1965). ME has also been used within the motor speech literature to scale various parameters including loudness (Clark et al., 2014; Dromey & Adams, 2000; Ma, Schneider, Hoffmann, & Storch, 2015) and intelligibility (Tjaden & Wilding, 2011; Walshe, Miller, Leahy, & Murray, 2008; Weismer & Laures, 2002). Jesteadt and Joshi (2013) noted that CLS is more popular than ME in hearing aid research and demonstrates strong reliability and practicality. This led them to compare loudness ratings provided via ME, MP and CLS based on a range of 1 kHz tones. Results of their comparison indicated that CLS was more reproducible and more robust than MP and ME, with ME and MP highly influenced by test order. ME was also found to be affected by participants' experience with CLS. Practically, this could be an issue when listeners have different experience levels using rating scales.

Equal-loudness matching is useful for obtaining fine-grained evaluations of loudness at specific levels and specific frequencies. It is likely that equal-loudness matching is less suitable for longer, complex, and variable sounds, including speech.

Marks and Florentine (2011) suggest that the use of the method of continuous judgment by category may be more suitable to long, dynamic stimuli. For dynamic stimuli of moderate length, such as judgments of a single sentence of speech, scaling methods like ME, MP, CLS, and VAS may be appropriate. Given the drawbacks of CLS, VAS may be preferable between the two in order to capitalize on higher resolution of responses and avoid the problems of interval-based data. Despite its strengths, VAS has not been employed as frequently in the loudness literature. ME has a long history of use and is well-respected as a robust measure, but can be impractical for use outside of a research setting, as it requires an experimental setup and multiple listeners for results to be informative. VAS, like categorical rating scales, is practical as a clinical measure used for clinicians, patients, and communicative partners to provide ratings of their loudness and other speech characteristics (Ramig et al., 1995a).

2.5.2 Acoustic Loudness

Acoustic models of loudness differ in complexity, but all seek to use acoustic characteristics to accurately predict perceived loudness by capturing spectral and variability characteristics.

2.5.2.1 LKFS

Integrated loudness in LKFS, Loudness Units Relative to Full Scale (K-weighted), is a measure of loudness developed for broadcast material. The LKFS scale is designed to quantify loudness and loudness range for regulating the loudness of broadcast programs, and LKFS units are designed to be proportional to decibels relative to full scale (dB FS). Because it is designed for such dynamic stimuli, it is possible that LFKS would be an appropriate measure of the loudness of speech. LKFS is implemented in MATLAB (MathWorks, 2020) as part of the Audio Toolbox. The algorithm for calculation of loudness is detailed in ITU-R BS.1770-4 (2015), and loudness range through EBU R-128 (2014).

There are two defining characteristics of this model: K-weighting, and gating. The signal is first weighted using a K-weighted filter. The filtering has two phases: the first phase accounts for acoustic effects of the shape of the head, and the second phase applies the revised low-frequency B-curve (RLB) filter. This filter, sloping with a cut-off of around 14 kHz to emphasize higher frequencies, was designed based on empirical results of subjective testing in which 97 listeners participated in a loudness-matching paradigm (Soulodre, 2004). Listeners were asked to adjust the level of a sample of broadcast material until it matched the reference signal, a sample of English speech at 60 dBA SPL. Of several candidate loudness meters investigated in Soulodre (2004), the resulting RLB filter was ranked highest in nearly all performance metrics, leading to its use in the K-weighting of this model.

After applying the K-weighting, momentary power and loudness (as per the formula in ITU-R BS.1770-4, 2015) values are obtained from 400 ms blocks with 300 ms overlap, which are then gated in two steps. The first yields a set of blocks in which loudness is above absolute threshold and calculates loudness with the same formula as momentary loudness. The second yields a further subset of these blocks that are greater than the relative threshold (-10 LKFS less than the gated loudness estimate obtained from the first step) and again calculates loudness with the same formula. This result is the integrated loudness of the sample, which will be referred to as LKFS throughout this investigation to clearly delineate it from other loudness metrics investigated.

Loudness range is based on the statistical distribution of measured loudness and is designed so that a short but very loud event would not affect the loudness range of a longer segment. The range of distribution of loudness levels is determined by estimating the difference between the lower (10th) and higher percentiles (95th) of the distribution. Loudness range is calculated in the same way as integrated loudness but over a larger window with greater overlap (3 second window with 2.9 seconds of overlap). The power and loudness components of loudness range are called short-term power and short-term loudness.

2.5.2.2 Active Speech Level

Unlike LKFS, active speech level is not specifically designed to describe perceived loudness. However, it is included in this investigation as it is an attempt to transform a measure of physical strength to be more applicable to speech measurement. Active speech level is outlined in ITU-T P.56 (2011) and is the speech level over the time during which speech is present, excluding low intensity segments like pauses. This measure can be obtained via the Voicebox: Speech Processing Toolbox for MATLAB (Brookes, 2020).

The determination of whether speech is active relies on an adaptive threshold applied to the signal, with a default margin of 15.9 dB between speech and noise. Following band-pass filtering, instantaneous power estimates are obtained. Active speech level is calculated by integrating instantaneous power estimates aggregated over the active time, expressed proportional to total energy divided by active time. The output value is expressed in dB FS.

In the context of hypokinetic dysarthria, active speech level may be an effective measure because it can efficiently reduce the effect of pauses on overall intensity. This may be particularly helpful for measurement of spontaneous speech over longer periods of conversation, where pauses may be more frequent and less predictable. Active speech level might be of particular interest to studies involving long-term, remote collection of speech via portable voice accumulators or voice dosimeters (Schalling, Gustafsson, Ternström, Wilén, & Södersten, 2013; Szabo & Hammarberg, 2013; Titze, Hunter, & Švec, 2007).

2.5.2.3 TVL

The time-varying loudness (TVL) model has been developed by hearing scientists Moore and Glasberg over the course of more than 30 years of research. This model can be seen as an expansion from Zwicker's model, to be discussed below, using a similar approach but with different assumptions and different features taken into account. It is far more complex and robust than the LKFS model. MATLAB code for running the current version of TVL was last updated in 2018 as per Moore, Jervis, Harries, and Schlittenlacher (2018) and is freely available on the authors' departmental website.

The model begins by applying a transfer filter to simulate the sound's travel through the middle ear depending on its method of presentation. This signal is then converted to a running short-term spectrum through 6 fast Fourier transforms run for every millisecond of the sample, based on Hann-windowed segments of various lengths centered around that millisecond. These running spectra are then converted to an excitation pattern, the effective spectrum reaching the cochlea, defined as a pattern of outputs from the auditory filters as a function of filter center frequency, based on the rounded-exponent function (Patterson, Nimmo-Smith, Weber, & Milroy, 1982). Excitation is then converted to specific loudness, a form of loudness density representing the loudness evoked over a 1-Cam wide range of centre frequencies (where 1 Cam is 1 number on the ERB_N scale). Early versions of these conversions are thoroughly described in Glasberg and Moore (1990) and Moore, Glasberg, and Baer (1997). The parameters of this conversion have been empirically designed and adjusted throughout the model's life. They are designed to account for the shape of the auditory filter and extent of cochlear gain at different frequencies and different sound levels.

The specific loudness pattern obtained for a single short-term spectral estimate is called the instantaneous specific loudness pattern, which is then smoothed over time by calculating a running average of instantaneous specific loudness, separately for each center frequency. The result is called the short-term specific loudness pattern. This smoothing employs a circuit similar to automatic gain control (AGC) with greater attack time than release time, meaning that short-term specific loudness can increase relatively quickly but takes longer to decay. Short-term specific loudness is then binaurally inhibited and smoothed, such that the signal at each ear is inhibited (reduced) by the signal's presence at the right ear (Moore et al., 1997). This broad tuning is implemented by smearing each ear's specific loudness pattern with a Gaussian weighting function. Inhibition is then implemented by reducing the loudness evoked at the left ear proportionally to the signal at the right ear, and vice versa. In cases where the sound is diotic, the signal in each ear has been identical to this point. For diotic sounds equal in short-term specific loudness at each ear, a diotic sound is predicted to be 1.5 times as loud as the same sound if presented monoaurally (Moore, Glasberg, Varathanathan, & Schlittenlacher, 2016).

Short-term loudness for each ear is then calculated by summing the inhibited short-term specific loudness values over each Cam value on the ERB_N scale from 1.75 to 39. Overall binaural short-term loudness is obtained by summing each ear's short-term loudness. Long-term loudness for each ear is then calculated by averaging each ear's short-term loudness using a similar AGC-style smoothing, and the overall long-term loudness is calculated by summing the long-term loudness values for each ear. The overall loudness estimate returned by the model is the maximum obtained value of long-term loudness, as this has been found to be slightly more accurate than mean of long-term loudness for transient sounds and speech (Marshall & David, 2007; Moore et al., 2016; Zorilă, Stylianou, Flanagan, & Moore, 2016). Individual parameters and components of this model have been empirically tested and refined over time. Moore et al. (2018) conducted such testing and refining on the model described in Moore et al. (2016), leading to the most recent refinements in time constants and dramatically improving its predictive performance for some signals. In loudness matching experiments of Moore et al. (2018), mean LDEL was small, indicating that the model's predictions were quite close to listener's perceptions.

2.5.2.4 Zwicker

As discussed above, the TVL model is built upon the principles of the Zwicker model. An important difference between TVL and the Zwicker model is that TVL uses ERB_N and the Zwicker model uses critical-bands and the Bark scale (Zwicker & Scharf, 1965). Additionally, through its improvements over time, TVL has incorporated binaural inhibition, an important consideration for sounds presented in free-field, diffuse-field, and naturalistic listening environments. The Zwicker model is detailed in its current standard (ISO-531:2017, Part 1).

Rennies, Holube, and Verhey (2013) applied the TVL and Zwicker models to signals along a continuum of real speech to speech-like noise. Thirteen listeners rated the loudness of these signals using categorical loudness scaling. Results indicated that TVL yielded better predictions, and also indicated that TVL estimates were particularly affected by high-frequency components. This may be important in the context of IWPDs, who may demonstrate disrupted spectral balance, including lower energy in high-frequency ranges. Due to the findings of Rennies et al. (2013) and difficulty accessing code for the Zwicker model's implementation, it was not selected for inclusion in the present investigation.

None of the acoustic methods described above has been thoroughly investigated in speech research. While LKFS is designed for broadcast material, including speech, its perceptual model is not as comprehensive and may not be as suitable for research purposes. Similarly, active speech level is designed for application to speech, but is not directly intended to describe loudness. TVL is a very comprehensive model of loudness, but application to speech has been limited. TVL may not, then, be sensitive enough to clinical differences in speech characteristics. Additionally, TVL is computationally intensive, making it prohibitively slow to apply to longer samples of speech.

2.6 Loudness, Intelligibility, and Hypophonia

Due to the complex nature of hypophonia and of the speech system, hypophonia and hypokinetic dysarthria may have an interconnected influence on the perceived loudness and intelligibility of IWPDs. Several areas of speech may interact here, including voice quality, glottal closure patterns, articulation and other supraglottic contributions, and prosody. Of particular interest to this investigation are contributions of spectral balance and speech level variability related to monoloudness and loudness decay.

2.6.1 Spectral Balance

Equal-loudness contours demonstrate that the perceived loudness of two pure tones with the same intensity can differ depending on their frequencies. This has important implications for speech. Spectral balance may be seen as an overarching term describing the distribution of energy across the frequency spectrum. Numerous spectral balance measures exist for the description of speech, such as spectral tilt, spectral slope, alpha, low-high spectral ratio, spectral moments, parabolic spectral parameter, and spectral emphasis (Alharbi, Cannito, Buder, & Awan, 2019; Corcoran, Hensman, & Kirkpatrick, 2019; Dromey, 2003; Hammarberg, Fritzell, Gaufin, Sundberg, & Wedin, 1980; Smith & Goberman, 2014; Titze, 2020; Titze & Palaparthi, 2020; Tjaden, Sussman, Liu, & Wilding, 2010; Watts & Awan, 2011; Weingartová & Volín, 2014). These measures vary in their calculations and interpretations, but it is notable that spectral tilt and low-high spectral ratio are the same measure in that both express the difference in dB between low and high frequency energy. The frequency cut-off separating these two bands varies across studies and should be considered in the interpretation of results.

Flatter spectral tilt, with a greater proportion of high frequency energy, has been associated with greater perceived loudness in synthetic vowels (Duvvuru &

Erickson, 2013), disproportionate loudness increases compared to intensity increases (Titze, 2020; Titze & Palaparthi, 2020), and vowel prominence (Sluijter & Heuven, 1996). Steep spectral tilt has also been associated with breathiness and dysphonia (Alharbi et al., 2019; Hillenbrand & Houde, 1996).

Deviations in spectral balance have been identified in the speech of IWPDs, including a reduction of energy in the high-frequency range as characterized by lower spectral mean, lower spectral standard deviation, higher skewness, and higher kurtosis (Dromey, 2003). Corcoran et al. (2019) found that the parabolic spectral parameter of sustained vowels was successful in distinguishing IWPDs from healthy adults. Parabolic spectral parameter is a method of fitting a parabola to lower frequencies of the glottal source spectrum to measure spectral decay, and these findings support the contribution of spectral tilt to the voice differences of IWPDs. Tjaden et al. (2010) also found a positive relationship between skewness, kurtosis, and perceived voice severity of IWPDs.

2.6.2 Spectral Balance and Vocal Effort

Flatter spectral tilt has also been associated with effortful speech. As discussed in Section 2.3.3, vocal effort encompasses physiological effort, the experience of effort, psychological effort, effort as a speech production level, and effort as affected by communication environment (Hunter et al., 2020). Speakers may use both somatosensory feedback and auditory feedback when rating their own effort, and some speakers may have sensory preferences, such as a bias to auditory feedback (Lane, Catania, & Stevens, 1961; McKenna & Stepp, 2018). Lane et al. (1961) investigated autophonic scale with auditory masking and stated that "under extensive changes in the auditory feedback that a speaker receives from his own voice, the scale of vocal effort remains relatively invariant in form and slope" (pg. 164). This is consistent with overall vocal effort being a complex phenomenon with many inputs. Listeners rating vocal effort may rely on a combination of mean intensity and spectral balance (Brandt, Ruder, & Shipp, 1969; McKenna & Stepp, 2018; Sluijter, Heuven, & Pacilly, 1997), but do not have access to components like somatosensory feedback or psychological effort. As a result, self-reported effort is thought to be the most accurate since the speaker can account for all these modalities (Rosenthal, Lowell, & Colton, 2014).

In the literature, vocal effort has been studied in a number of ways, including directly requesting different effort levels (Brandt et al., 1969; Glave & Rietveld, 1975; McKenna & Stepp, 2018), altering interlocutor distance (Liénard & Benedetto, 1999), and requesting different loudness levels (Gauffin & Sundberg, 1989; Lane et al., 1961). There are some challenges involved with studying vocal effort. Given the complex nature of vocal effort, it is multidisciplinary and requires input from several fields (McKenna & Stepp, 2018). Additionally, inconsistencies in definitions and conceptualizations make comparisons even more difficult, which Hunter et al. (2020) sought to mitigate with their review. Studies also vary in the ways high and low effort states are elicited, as stated above, which can complicate knowledge synthesis. In many studies investigating effort, effort is conflated with speaking loudly. While louder speech tends to require greater effort, not all effort is intended to increase loudness. Effort may involve speaking with greater clarity, speaking slower, speaking in a different mode (i.e., a whisper), or intentionally altering laryngeal tension without a goal of increased loudness, as requested by McKenna and Stepp (2018). Providing clear, specific operational definitions of effort within each experiment is important to clarify findings across studies and better investigate the relationships between effort and other parameters of speech.

Several voice changes are associated with high vocal effort. Physiological changes associated with effort manifest in acoustic changes. These may include increased subglottal pressure (Hunter et al., 2020; McKenna, Diaz-Cadiz, Shembel, Enos, & Stepp, 2019; Rosenthal et al., 2014), greater lung volume initiations and terminations (Dromey & Ramig, 1998), increased cervical muscle tension and laryngeal tension (Hunter et al., 2020; McKenna et al., 2019), larger displacement and higher peak velocities of lip movements (Dromey & Ramig, 1998; Dromey, 2000), increased mean intensity (Dromey & Ramig, 1998; Hunter et al., 2020; McKenna & Stepp, 2018), increased proportion of high-frequency energy (Eriksson

& Traunmuller, 2002; Gauffin & Sundberg, 1989; Liénard & Benedetto, 1999; McKenna & Stepp, 2018), increased fundamental frequency, standard deviation of fundamental frequency, and first formant frequency (Dromey & Ramig, 1998; Hunter et al., 2020; Liénard & Benedetto, 1999), and shorter glottal closing phase (Gauffin & Sundberg, 1989; Glave & Rietveld, 1975; Sluijter & Heuven, 1996). Shorter glottal closing phase affects spectral balance because the steeper glottal pulse shifts intensity over the spectrum, leading to the additional intensity gained with the increased effort being added to the high-frequency range instead of a flat increase across frequencies (Sluijter & Heuven, 1996). Low vocal effort has received less investigation than high vocal effort, because of the importance of high vocal effort in understanding hyperfunctional voice disorders. Rosenthal et al. (2014) found that low effort speech was associated with decreased laryngeal resistance and decreased subglottal pressure.

This intersection of spectral balance and vocal effort may be important in the discussion of hypophonia for a number of reasons. As discussed in Section 2.3.3, effort perception has been implicated as a possible contributor to hypophonia, though evidence is limited (Solomon & Robin, 2005), and is an important component of LSVT LOUD voice therapy for IWPDs (Sapir et al., 2011). LSVT techniques encourage the use of higher vocal effort. Dromey, Ramig, and Johnson (1995) investigated phonatory and articulatory changes in IWPDs before and after LSVT. Among the observed changes were a relatively greater proportion of high-frequency energy post-treatment as measured by harmonic spectral slope. These findings support that spectral balance of IWPDs is sensitive to changes in their effort level, at least as it pertains to loud, effortful speech.

Neel (2009) also investigated the relationship between loud, effortful speech and amplification on intelligibility in 5 IWPDs that had previously completed LSVT in the 1-2 years prior to the testing session. Speakers produced sentences and words at habitual effort and with loud speech, and were regularly cued to use LSVT techniques. Louder speech, compared to habitual speech, was associated with higher spectral mean, higher spectral standard deviation, lower spectral skewness and lower spectral kurtosis, consistent with an increase in high-frequency energy and in the reverse direction of tendencies of IWPD speech identified by Dromey (2003). These changes were associated with an increase in intelligibility. Both loud speech and amplified habitual speech were associated with a significant increase in intelligibility, but loud speech was found to be more effective than amplification alone. The authors stated that the increase of speech-to-noise ratio accounted for up to half of the observed increase in intelligibility with loud speech, and hypothesized that glottic and supraglottic changes must be responsible for the remainder. While this study did not evaluate perceived loudness and included a small sample of IWPDs and no control group, their findings suggest that ongoing evaluation of supraglottic and glottic contributions to intelligibility is needed. Additionally, these findings further support that amplification of speech could be improved with the use of filters that adjust spectral balance.

It is notable that some patterns of high effort speech observed in the normal system are seen to be flipped in IWPDs. Specifically, a relatively greater proportion of high-frequency energy is associated with effortful speech, and a weaker proportion of high-frequency energy is associated with hypokinetic dysarthria. It is possible that low effort speech produced by a normal speech system may be analogous to the hypophonic system of IWPDs, such that normal effort speech produced by a hypofunctional system mimics low effort speech produced by a normal system. IWPDs may need to speak at a higher effort level in order to compensate for this hypofunction. Findings of laryngeal abnormalities in IWPDs have long been identified, and have been differentially associated to hypokinesia, rigidity, and respiratory influences. It has been seen that normal speakers can intentionally produce similar acoustic manifestations of breathiness as breathy dysphonic speakers (Hillenbrand & Houde, 1996). Similarly, relative fundamental frequency patterns of normal speakers using increased vocal effort are similar to individuals with hyperfunctional voice disorders and spasmodic dysphonia (McKenna, Murray, Lien, & Stepp, 2016). If normal speakers using high vocal effort can mimic vocal hyperfunction, perhaps it is possible that normal speakers using low vocal effort could mimic laryngeal hypofunction. This relationship might be used to infer that hypofunction is an important contributor to the overall presentation of hypophonia. This hypofunction may be due to laryngeal and/or respiratory influences and may stem from hypokinetic and/or rigid mechanisms. Support for this hypothesis may also come from findings of Watts and Awan (2011). This investigation studied 16 hypofunctional speakers with glottic incompetence from a number of disease populations including Parkinson's disease and unilateral recurrent laryngeal nerve paralysis/paresis, as well as 16 matched controls. Their findings indicated that low-high spectral ratio was successful in distinguishing hypofunctional speakers, with specificity of 88% and sensitivity of 69%. This study did not focus on hypophonia or on loudness, and modest sensitivity may be due to the heterogeneous patient population studied. However, findings still provide support for the relationship between spectral balance and laryngeal hypofunction (Watts & Awan, 2011). This hypothesis may provide future directions for investigation of the specific effects of modulating vocal effort on individual systems of speech and voice in IWPDs, extending the findings of Neel (2009).

2.6.3 Prosodic Influences

Prosodic characteristics of Parkinsonian speech, such as monoloudness, may also affect perceptions of loudness. Monoloudness is a pronounced perceptual feature of hypokinetic dysarthria identified in seminal dysarthria literature (Darley et al., 1969, 1969a). IWPDs may also demonstrate higher loudness decay, such that loudness abnormally decreases over the course of the utterance (Clark, 2012; Ho, Iansek, & Bradshaw, 2001; Matheron et al., 2017; Rosen, Kent, & Duffy, 2005).

It is not well-understood how these features contribute to the overall perception of loudness. For example, the peaks of the intensity contour are flatter in monoloud speech, and it is possible that this is a key consideration for listeners judging the sample. Similarly, intensity declination and loudness decay may be key contributors to perceptions of average loudness over the course of a longer speech sample.

2.6.4 Intelligibility

Given that hypophonia may be most apparent in conversational speech (Adams, Dykstra, et al., 2006; Fox & Ramig, 1997; Ho et al., 1999a), hypophonia can be expected to significantly influence the speech activities of IWPDs. Lower intelligibility has been identified among IWPDs (Chiu, Neel, & Loux, 2020; Miller et al., 2007; Tjaden, Sussman, & Wilding, 2014; Weismer, Jeng, Laures, Kent, & Kent, 2001). Intelligibility may be particularly affected by hypophonic deficits, due to previously discussed influences of SNR (Adams et al., 2008; Dykstra et al., 2013) and spectral balance (Tjaden et al., 2010) on intelligibility.

Loudness and intelligibility are important measures of hypophonic speech. Characteristics of hypokinetic dysarthria may affect both loudness and intelligibility, but perhaps in different ways. For example, intelligibility may be particularly affected by articulatory deficits, which might be expected to have a smaller effect on perceived loudness. Intelligibility is an important component of speech assessment, representing an ecologically valid evaluation of an individual's ability to make their speech understood. However, some features of speech contribute more than others to intelligibility. It cannot be directly inferred that a characteristic or intervention that affects loudness would equally affect intelligibility.

2.7 Research Questions and Hypotheses

The overall purpose of this investigation was to examine the relationships between perceived loudness and acoustic measures of loudness, speech level, spectral balance, and variability in individuals with hypophonia secondary to Parkinson's disease (IWPDs) and neurologically healthy older adults (HOAs).

RQ1: Do group differences exist between IWPDs with hypophonia and HOAs in perceived loudness, mean intensity, acoustic loudness, intelligibility, spectral balance, or speech level variability?

Hypotheses:

- IWPDs will be perceived as quieter and less intelligible than HOAs.
- IWPDs will be quieter than HOAs as measured by speech level measures of intensity and acoustic loudness.
- IWPDs will show differences in spectral composition compared to HOAs.
- IWPDs will show differences in speech level variability compared to HOAs.

RQ2: Are acoustic models of loudness more predictive of perceived loudness than mean intensity?

Hypothesis: Acoustic models of loudness will be more predictive of perceived loudness than mean intensity, as they have been incorporate listener factors.

RQ3: Can perceived loudness be predicted by speech level, spectral balance, or speech level variability?

Hypotheses:

- Speech level will be predictive of perceived loudness such that lower speech level is predictive of lower perceived loudness.
- Spectral balance will be predictive of perceived loudness such that a relatively greater concentration of energy in low frequencies (e.g. steep tilt) is predictive of lower perceived loudness.
- Speech level variability will be predictive of perceived loudness such that low standard deviation and high decay are predictive of lower perceived loudness.

RQ4: Can differences in perceived loudness between IWPDs and HOAs be explained by acoustic characteristics of their speech?

Hypotheses:

- Speech level deficits of IWPDs will be associated with lower perceived loudness.
- Spectral balance deficits of IWPDs will be associated with lower perceived loudness.
- Speech level variability deficits of IWPDs such as low speech level standard deviation and high speech level decay will be associated with lower perceived loudness.

RQ5: Are loudness ratings collecting using visual analogue scales and direct magnitude estimation consistent and reliable?

Hypothesis: Loudness ratings collected using visual analogue scales and direct magnitude estimation will be consistent and offer similar reliability.

RQ6: Do acoustic measures that predict loudness also predict intelligibility?

Hypothesis: Measures predicting loudness may also contribute to intelligibility, but perceived loudness and intelligibility will differ enough that intelligibility ratings could not be considered to encompass loudness.

RQ7: Do manipulations of spectral composition predict perceived loudness ratings?

Hypothesis: Increases and decreases in the gain of mid- and high-frequency energy will increase and decrease loudness, respectively. A relatively greater proportion of energy in the higher frequencies (i.e. flatter tilt) will be associated with greater perceived loudness.

Experiments 1 and 2 may both inform some of these research questions. RQ5 and RQ6 will be answered through the results of Experiment 1 (natural speech). RQ1, RQ2, RQ3, and RQ4 will be primarily answered through the results of Experiment 1, with contributions from Experiment 2 (spectral manipulation). RQ7 will be answered through the results of Experiment 2. Methodology and results of each experiment will be described separately, and findings will be integrated in Chapter 5 in the interpretation and discussion.

3 Methods

3.1 Data Collection and Preparation

3.1.1 Data Sources

Audio data for this investigation was pooled from archived audio of previous investigations of hypophonia in IWPDs. Combining data across studies was possible because of methodological similarity in the collection of the data in terms of speech tasks, recordings and calibration. Creation of this pooled dataset for analysis and presentation to listeners was approved by the Western University Health Sciences Research Ethics Board (ID: 115159).

A total of 152 candidate speaker participants (97 IWPDs, 55 HOAs) were available using this pooled data. All participants provided written consent to participate in the respective study in which data was collected. Within each of these previous investigations, IWPDs were selected as individuals between the ages of 50-90 with idiopathic PD and with hypophonia as their primary speech concern noted by their neurologist. All participants with PD had been diagnosed at least 6 months prior to the study session and were on a stable dopaminergic medication for the previous 6 months. All participants were diagnosed by and receiving regular treatment from an experienced movement disorders neurologist (M. Jog) at the Movement Disorders Centre of London Health Sciences Centre, London, Ontario. IWPDs were excluded if they had a history of speech, language, or neurological conditions other than PD. None of the IWPDs had a history of speech therapy within the year prior to the study session. HOAs served as the control group, and were individuals between the ages of 50-90 without a history of speech, language, or neurological conditions. Participants were required to speak, read, and write English to the extent necessary to participate in speech testing. Participants were also required to pass a 40 dB HL hearing screen at 0.5, 1, and 2 kHz in at least one ear. While this allows for the inclusion of individuals with unilateral hearing loss, prevalence of hearing deficits is higher in older adults, and excluding individuals with any form of hearing deficit may result in a non-representative sample. The presence of hearing deficits could pose a greater problem for studies involving background noise, as participants might hear the noise at different levels, affecting the observed Lombard effect. As background noise was not included in this investigation, it was deemed to be acceptable for some speaker participants to have hearing deficits.

Prior to inclusion in the present investigation, audio data was screened to ensure a high-quality pooled dataset. For inclusion, speakers needed to speak independently and fluently enough to not compromise intelligibility (i.e., repeating large portions of sentences or requiring additional prompts). Accented speakers were removed from the analysis if their accent was deemed to affect their intelligibility. These choices regarding intelligibility were intended to achieve greater consistency within the dataset. Data was also removed if any unacceptable noise or distortion was present in the recording due to the interest in spectral characteristics. The majority of removed candidate participants were removed due to the presence of noise and distortion in the audio recordings. Following this screening, 102 speaker participants (56 IWPDs, 46 HOAs) were selected.

3.1.2 Speaker Participants

Limited demographic information is available for the speaker participants selected for inclusion in this study. Sex and age were recorded for all participants, except one participant whose age was not available. Basic demographic information about the speaker participants is presented in Table 3.1. The higher proportion of males among IWPDs is consistent with the greater prevalence of the disease among men, as per the Mapping Connections report by Neurological Health Charities Canada, Public Health Agency of Canada, Health Canada and the Canadian Institutes of Health Research (2014).

Group	*Age (Mean)*	*Age (SD)*	*Age (Range)*	*N*	*Sex*	*Proportion Male*
HOA	71	7.88	55-86	46	F = 27, M = 19	41%
IWPD	69	7.07	54-88	56	F = 14, M = 42	75%

Table 3.1: Demographic characteristics of speaker participants.

Detailed characteristics of IWPDs including years since diagnosis, dosage, and disease severity were not available for all participants. Table 3.2 presents the available characteristics. The Unified Parkinson's Disease Rating Scale (UPDRS) is an assessment of overall PD severity (Goetz et al., 2008). The Montreal Cognitive Assessment (MoCA) evaluates cognitive performance and is frequently used as a screening criterion (Chou et al., 2010; Nasreddine et al., 2005). 7 IWPDs in this study had previously undergone deep brain stimulation surgery, implanting an electrode to stimulate the subthalamic nucleus (DBS-STN). DBS-STN can reduce the required dosage of dopaminergic medication, which can help to reduce medication-related dyskinesias and side effects developed by some IWPDs following long-term medication use (Okun, 2012). The considerable variability in cognitive ability, disease severity, and years since diagnosis of the IWPDs in this study is reflective of the heterogeneity of the PD population.

	Mean	*SD*	*Range*	*N*
Years Since Diagnosis	10.27	6.68	0.5-31	50
UPDRS	36.83	11.52	17-66	35
MoCA	24.00	3.73	14-29	36

Table 3.2: Detailed characteristics of individuals with PD. Characteristics were not available for all participants given the retrospective nature of this study. Each parameter was summarized from all participants for whom the data was available. The number of participants with available data on each parameter is presented alongside the statistics. Higher UPDRS scores reflect greater disease severity, and lower MoCA scores reflect greater cognitive impairment. MoCA scores above 26 may be considered to reflect normal cognitive function (Nasreddine et al., 2005). Among HOAs, an average UDPRS score of 1.4 was reported by Zitser et al. (2021).

3.1.3 Speech Recordings

All audio data included in this study was recorded in either a quiet room or a sound-treated booth (Industrial Acoustic Company) with a headset microphone (AKG c520) placed 6 cm from the speaker's mouth at a 30-45 degree angle. Audio was digitally recorded using either a DAT recorder (Tascam DA-P1) or USB audio interface (M-Audio Mobile Pre USB MKII). Sustained vowel calibration was performed with a sound level meter (Quest 215) placed 15 cm from the mouth using A-frequency weighting. Prior to all analyses and listener presentation, each sample was calibrated to accurate sound pressure level values based on the sustained vowel calibration, resampled to a sampling frequency of 22.05 kHz as some files were originally sampled at 44.1 kHz, and band-pass filtered from 70 Hz to 10 kHz to remove noise.

3.1.4 Speech Tasks

Sentence reading and conversation samples were obtained from each speaker participant. Each participant read aloud 11 sentences from the Sentence Intelligibility Test (SIT) varying in length from 5 to 15 words (Yorkston, Beukelman, & Tice, 1996). Participants were provided with the full word list and were instructed to read each sentence at a comfortable rate, pitch, and loudness. Conversational monologues were obtained by asking participants biographical questions about their life, career, interests, or vacations. In Praat (Boersma & Weenink, 2020), speech samples were extracted using manually annotated TextGrids and custom scripts. All SIT sentences were extracted. Samples of conversation were selected by identifying 3 complete utterances 4-8 seconds in length. Variability in sample length was required in order to obtain utterances expressing a complete thought.

3.2 Perceptual Analyses

3.2.1 Listener Participants

Listener participants were recruited from clinical communication sciences graduate students halfway through the speech-language pathology program at Western University. All listener participants had received education in auditory-perceptual evaluation of speech and voice, but with limited practical experience. Listener participants were required to be between the ages of 18-35, speak English as their first language, read and write in English, and pass a 25 dB HL hearing screening at 0.5, 1, 2, and 4 kHz in both ears. Listener participants were excluded if they had a history of a speech, language, or neurological disorder or if they had extensive research or clinical experience with individuals with PD. Extensive experience was defined as working directly with the population of interest for longer than a short-term volunteer position (e.g. 10 hours), or having been directly involved in research studies of people with Parkinson's disease that involved listening to or analyzing their speech. These requirements reduced the variability in experience of the raters. 10 listeners were recruited for this investigation, and demographic information describing the listener participants is presented in Table 3.3.

Age (Mean)	*Age (SD)*	*Age (Range)*	*N*	*Gender*
24	1.62	22-28	10	F = 9, M = 1

Table 3.3: Demographic characteristics of listener participants.

3.2.2 Listening Experiment Setup

Listener participants completed all ratings in a sound-treated booth (Industrial Acoustic Company). Listeners were seated 1.5 m from a loudspeaker (Yamaha HS8 Audio Monitor). Prior to each listening session, the loudspeaker was calibrated using a 1 kHz tone, calibrated to 70 dB SPL at the position of the listener's head (1.5 m from the speaker and 1 m from the ground) with a sound level meter (Quest 215). When combined with the calibration of each file in Praat, this loudspeaker calibration ensures that the sound pressure level of each audio sample is consistent with the level at which it was spoken by the speaker participant. Listeners provided all ratings using a digital interface presented on a laptop computer via custom scripts written in Praat. Further details of each listening task are presented below with regard to each experiment.

3.2.3 Experiment 1: Natural Speech

Within Experiment 1, 4 samples per participant were presented to listeners. The 8- and 10-word SIT sentences and 2 conversational samples were presented to listeners, for a total of 408 samples. The 8- and 10-word SIT sentences were selected for their moderate length. Other samples were retained for acoustic analyses (N = 1424).

3.2.3.1 Listening Tasks

Loudness ratings were collected via direct magnitude estimation (DME) and visual analogue scaling (VAS). Intelligibility ratings were collected via VAS. Ratings were provided by listeners using custom Praat scripts. Samples were provided in a fully randomized order within each rating task. A random 10% of samples were duplicated for reliability calculation and randomly mixed into the presentation order. Listeners completed all ratings within each rating task in a single session lasting 60-90 minutes. Listeners were able to take breaks at any time to reduce effects of fatigue. Listeners heard each sample only once before providing their rating, and could only confirm their rating once they had heard the full sample. They were instructed only to repeat the sample in the rare event that they were unable to hear the sample the first time, rather than to verify their rating. In order to reduce bias, DME loudness ratings were always completed before exposure to VAS

(Jesteadt & Joshi, 2013). Loudness and intelligibility VAS blocks were counterbalanced such that half of participants rated intelligibility before loudness.

3.2.3.1.1 DME Ratings

DME was performed with a standard modulus assigned a value of 100. The modulus was selected by evaluating the 9-word SIT samples to find a sample with moderate mean intensity, moderate loudness and good intelligibility based on subjective estimation and preliminary loudness ratings by this author. 9-word SIT samples by IWPDs and HOAs were considered for selection. The selected modulus was spoken by an IWPD. The modulus was presented every 5 samples as well as after any break of longer than 30 seconds between samples.

Listeners were instructed to assign the standard modulus a value of 100 and provide all ratings relative to the modulus such that higher numbers reflected louder samples and smaller numbers reflected quieter samples. No upper or lower limit was imposed on their ratings. Listeners were instructed to use any increment and any scale for their ratings, including decimals or negative values if they felt it necessary. A screen capture of the interface used by listeners to provide their ratings is displayed in Figure 3.1. Listeners typed their numerical response into the box on-screen after hearing the sample.

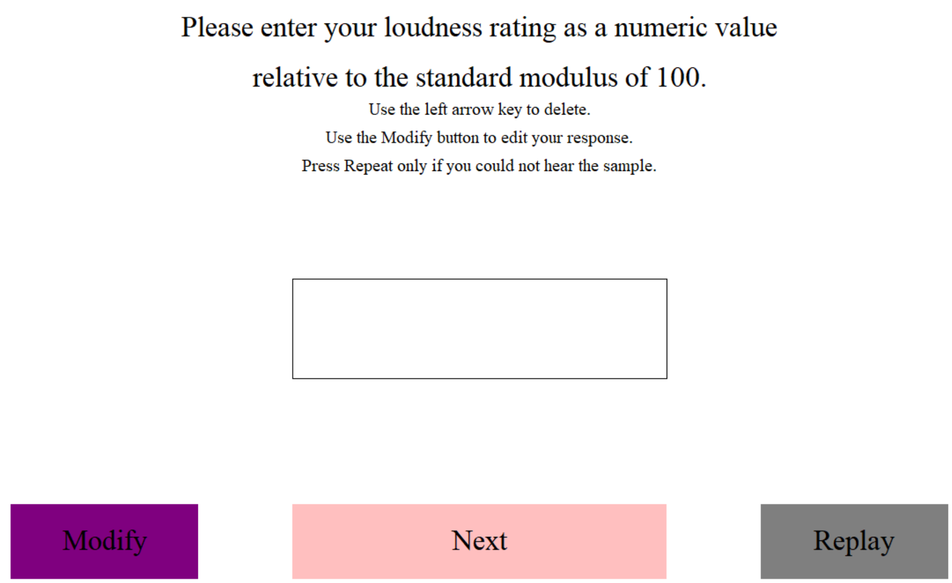


Figure 3.1: Screen capture of the listener rating interface for direct magnitude estimation of loudness.

3.2.3.1.2 VAS Ratings

Listeners provided ratings of loudness and intelligibility via VAS. Listeners provided their rating by clicking along the line on-screen. This rating could be adjusted after the initial click before confirming their rating. Ratings were saved as a percentage of total line length (15 cm). For intelligibility, the anchors were “Low intelligibility” and “High intelligibility.” For loudness, the anchors were “Low loudness” and “High loudness.” These anchors allow for a task-specific rating of loudness. Previous studies using VAS for loudness ratings have used a generalized rating, such as “Always loud enough” and “Never loud enough” in the LSVT assessment questionnaire (Wight & Miller, 2015). It is difficult to translate this type of anchor to a task-specific loudness rating needed in this investigation. Wilson, Page, and Adams (2020) included task-specific VAS ratings of perceived loudness of IWPDs with anchors of “normal” to “severely impaired or abnormal.” Due to the inclusion of HOAs in this investigation, neutral anchors were deemed to be preferable to a severity-based anchor.

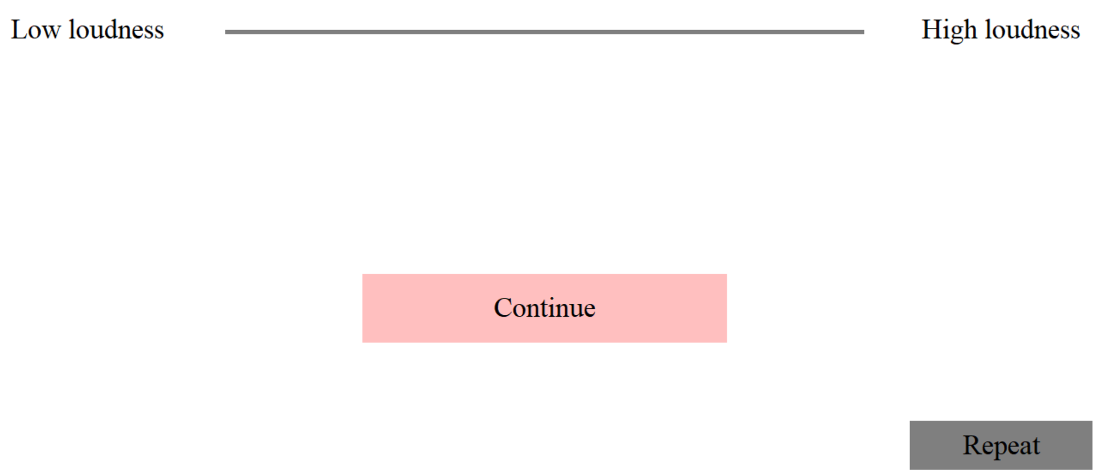


Figure 3.2: Screen capture of the listener rating interface for visual analogue scaling of loudness.

3.2.3.2 Averaged Perceived Values

For analysis of perceptual ratings, values from each listener were averaged such that a single value was obtained for each sample on the each measure (DME loudness, VAS loudness, and VAS intelligibility). VAS ratings were averaged using arithmetic means across participants. DME ratings were averaged via geometric mean and percentage averaging. Geometric mean is consistent with uses of DME ratings in the literature (Constantinescu et al., 2011; Walshe et al., 2008; Weismer & Laures, 2002). Percentage averaging was attempted to simplify analysis of DME ratings for future investigations. Each listener's ratings were converted to a percentage bounded by their smallest and largest ratings. Percentage scores were then averaged across participants via arithmetic means. Percentage averaging allows for comparison of scores between participants and simpler calculation of inter-rater reliability. It was included alongside geometric means to verify the consistency between these averaging methods.

3.2.4 Experiment 2: Spectral Manipulation

For the spectral manipulation experiment, the 5-word SIT sentence from each participant was selected for presentation. Including the 4 spectral manipulations, this resulted in 5 samples per participant presented to listeners, for a total of 510 samples within Experiment 2.

3.2.4.1 Spectral Manipulation

The spectral manipulations employed in this investigation were simple spectral balance adjustments performed by a custom MATLAB script. The script applied interpolated gains to the spectrum within target frequencies. Frequencies between 0-1 kHz were unaltered. Above 2 kHz, a flat gain was applied of +5, +10, -5, or -10 dB. A gradual transition was applied between 1-2 kHz to achieve less distortion and a more natural adjustment of tilt. Finally, the output amplitude was normalized to the input amplitude such that the manipulation would not affect the overall mean intensity of the sample. This normalization isolates the effect of spectral balance on perceived loudness, without the contribution of mean intensity. Examples of the long-term average spectra (LTAS) resulting from this manipulation are presented in Figure 3.3. Despite the gradual transition, some distortion was audible in the resulting files; however, it was deemed that this did not prevent listeners from rating the loudness effectively.

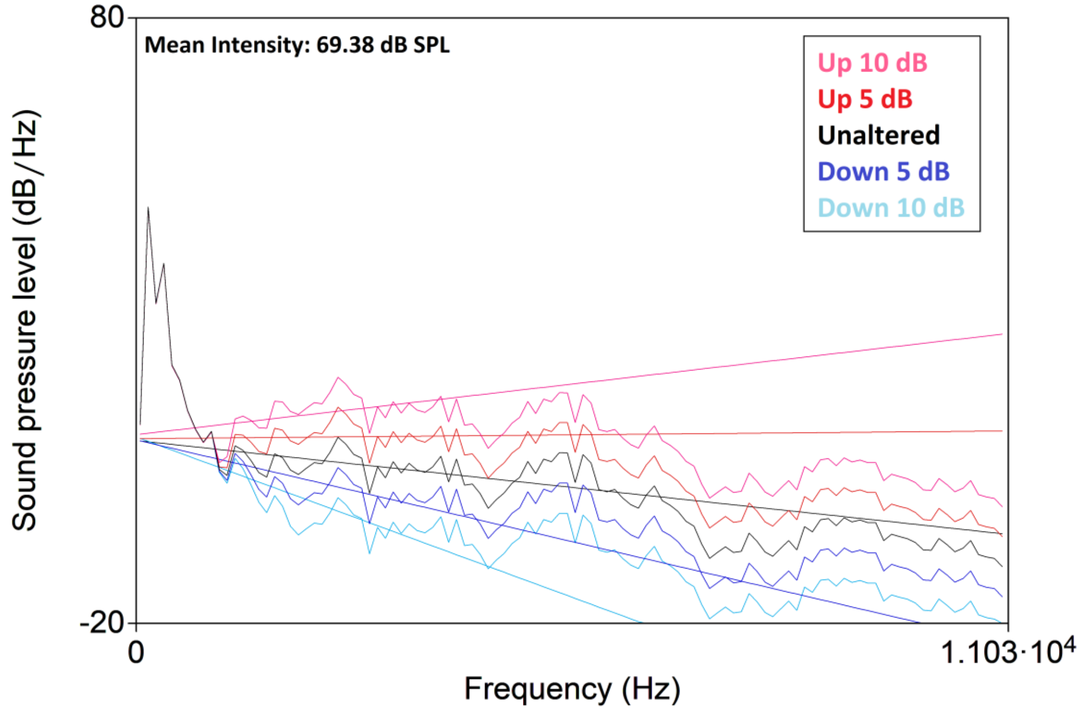


Figure 3.3: Long-term average spectra representing the effect of the spectral manipulation, with a trendline based on the 1-5 kHz range to demonstrate the shift in tilt. All 5 files have the same mean intensity (69.38 dB SPL). Frequencies below 1 kHz are unaltered in all 5 files, gain is gradually increased between 1-2 kHz, and a flat gain is applied above 2 kHz.

3.2.4.2 Listening Task

The rating procedures for Experiment 2 were the same as the VAS loudness rating procedures of Experiment 1. The same script and scale anchors were employed. Samples were fully randomized and listeners completed their ratings in one session. Perceptual ratings of Experiment 2 were always the last rating task completed by participants. As with VAS loudness in Experiment 1, perceptual ratings were averaged across participants via arithmetic mean.

3.3 Acoustic Analyses

Acoustic analyses were conducted in Praat and MATLAB using custom scripts. Acoustic measures were clustered into three conceptual groups to aid in interpretation, comparison of similar measures, and in the stepwise regression models. These clusters included measures of speech level, spectral balance, and speech level variability.

3.3.1.1 Speech Level

The term ‘speech level’ was used for cohesion to refer generally to sound level or acoustic loudness. Speech level measures in this investigation were mean, median, and maximum intensity, TVL and TVL mean, LKFS, and active speech level.

- **Intensity:** Mean, median, and maximum intensity (dB SPL) were obtained in Praat. Mean intensity is the most common measure reported in the hypophonia literature. However, it is possible that maximum intensity relates more closely to loudness and it was included as an alternative. Median intensity may better account for variability in the intensity contour.
- **Time-Varying Loudness (TVL):** TVL was obtained in MATLAB using code available on the creators’ departmental website, last updated in 2018 as per Moore et al. (2018). TVL is the maximum of the long-term loudness calculated by the model, and is the default output of the model. TVL mean is the mean of the long-term loudness, included in this investigation to compare the effectiveness of the long-term maximum and mean in the context of connected speech. Details of TVL’s calculation were discussed in Section 2.5.2.3. TVL and TVL mean are expressed in sones.
- **Loudness (K-weighted) Relative to Full Scale (LKFS):** Integrated loudness (in LKFS) was obtained via the function `integratedLoudness`, available through MATLAB’s Audio Toolbox, which implements the algorithm outlined

by ITU-R BS. 1770-4 (2015). Loudness range was not included in this investigation as the majority of speech samples were too short for its calculation. Details of the calculation of LKFS were discussed in Section 2.5.2.1. Integrated loudness is expressed in LKFS units, proportional to decibels relative to full scale (dB FS). In this investigation, integrated loudness is described as LKFS to clearly delineate it from other measures.

- **Active Speech Level:** Active speech level was obtained via the function `v_activelev`, available through VOICEBOX: Speech Processing Toolbox for MATLAB, which implements the algorithm outlined in ITU-T P.56 (2011). Details of the calculation of active speech level were discussed in Section 2.5.2.3. Active speech level is expressed in decibels relative to full scale (dB FS).

3.3.1.2 Spectral Balance

Spectral balance measures described the distribution of energy across the frequency spectrum. Spectral balance measures in this investigation were tilt, voiced tilt, tilt ratio, LTAS skewness and kurtosis, mid-ratio, and high-ratio.

- **Tilt:** Spectral tilt was calculated as the difference in energy between the 0-1 kHz range and 1-10 kHz range of the long-term average spectrum (LTAS), obtained in Praat. Tilt is expressed in dB.
- **Voiced Tilt:** Voiced segments of speech were obtained and concatenated, obtained in Praat using a script adapted from the AVQI (Maryn & Weenink, 2015). Tilt was calculated from these voiced segments in the same way as outlined above. Voiced tilt is expressed in dB.
- **Tilt Ratio:** Tilt ratio was calculated as the ratio between the tilt in voiced-only segments and the overall tilt.
- **Spectral Moments:** Skewness and kurtosis were obtained from the LTAS in Praat, each describing the distribution of energy across the LTAS. Kurtosis describes the concentration of energy, and skewness describes the relative emphasis of low-frequency energy.

- **Mid-Ratio:** The proportion of the mean energy in the 2-5 kHz range relative to the overall mean intensity (2-5 kHz mean / mean intensity).
- **High-Ratio:** The proportion of the mean energy in the 5-8 kHz range relative to the overall mean intensity (5-8 kHz mean / mean intensity).

Mid-ratio and high-ratio were calculated as proportions, dividing the power spectral density (dB/Hz) mean of the target range by the overall mean intensity. A proportion was used to express the relative concentration of energy in the target range on a clearer scale, avoiding the complications of negative power spectral density estimates that occurred in many individuals. Appropriate ratio characteristics of this proportion were observed relative to the uncorrected power spectral densities and the mean intensity, guided by the discussion of ratio measures by Curran-Everett (2013).

3.3.1.3 Variability

Variability measures characterized speech level variability. Standard deviation and decay were calculated for both intensity and TVL. Excessive intensity declination and monoloudness are both features that have been associated with hypokinetic dysarthria, and it was of interest the extent to which these characteristics affected overall judgments of loudness.

- **Intensity Variability:** Standard deviation of intensity was obtained in Praat. Intensity decay was obtained in Praat and R, expressed as the slope of a linear regression of intensity values across the sample in 8 ms intervals.
- **TVL Variability:** TVL decay was calculated as the slope of the linear regression of TVL's short-term loudness estimates across the sample in 1 ms intervals. Standard deviation of TVL was calculated as the standard deviation of TVL's long-term loudness estimates across the sample in 1 ms intervals.

3.4 Statistical Analyses

All statistical analyses were conducted in R (R Core Team, 2021). Detailed package citations are presented in Appendix B. Correlations, tests of group differences, linear mixed effects regression (LMER) models, logistic regression, analyses of variance (ANOVA), and classification decision trees were employed to answer research questions. Specific details of each analysis and related research questions are presented below, separately for each experiment.

3.4.1 Reliability

Reliability of perceptual ratings was calculated using intraclass correlation (ICC) as per Koo and Li (2016). Intra-rater reliability for each listener was calculated using ICC 3 (two-way mixed effects, single rater, consistency) based on the randomly repeated 10% of samples within each condition. Average inter-rater reliability across listeners was calculated using ICC 3k (two-way mixed effects, multiple raters, consistency).

3.4.2 Experiment 1: Natural Speech

Correlations and tests of group differences were performed using values averaged within each participant to maintain independence of observations. When tests were run within each task (each SIT sentence and conversational sample), value distributions and results within tasks were consistent with the averaged values. As a result, averaging provided a simple and robust means of analyzing the overall results. Non-normality was observed based on visual inspection and Shapiro-Wilk tests of normality. However, parametric tests were maintained, as the large sample size in this experiment means that parametric tests are likely to be robust to these deviations (Lumley, Diehr, Emerson, & Chen, 2002). Appendix D provides Shapiro-Wilk tests of normality for all measures.

Pearson correlations between VAS loudness and DME loudness ratings were used to inform RQ5 (consistency between VAS and DME). Correlations between acoustic and perceptual measures provided starting points for RQ2 (acoustic models of loudness) and RQ3 (acoustic measures as predictors of loudness) by identifying measures that correlated most strongly with loudness. Correlations between perceived loudness, intelligibility, and acoustic measures provided insight into RQ6 (loudness and intelligibility). Separate correlations were obtained among IWPDs and HOAs to support RQ1 (IWPD-HOA differences). Corrections for multiple comparisons were not employed with correlation analyses due to the exploratory focus of these correlations.

Tests of group differences were employed to inform RQ1 (IWPD-HOA differences) and RQ4 (interaction of IWPD-HOA differences on perceived loudness). Welch *t*-tests were used to evaluate group differences, as heteroscedasticity was observed based on visual inspection and Levene's tests of equality of variances. Appendix D provides Levene's tests for all measures. The Holm-Bonferroni method was used to correct *p*-values for multiple comparisons, providing a balance between Type-I and Type-II error. Cohen's *d* was used as an effect size estimate for these differences, generally interpreted as d = 0.2 associated with small effects, d = 0.5 with medium effects, and d = 0.8 with large effects.

Linear mixed effects regression (LMER) models investigating interactions between group (PD status) and acoustic measures in their prediction of perceived loudness were used to inform RQ4 (interaction of IWPD-HOA differences on perceived loudness). LMER allows for modelling of hierarchical or repeated measures data by including multiple predictors (fixed effects) and random effects to control for variation across repeated measures. This approach combines the benefits of multiple linear regression and repeated measures ANOVA. Sonderegger, Wagner, and Torreira (2018) provides a detailed, online tutorial and reference for the use of LMER in linguistic research, which offers many helpful considerations for use in clinical speech research. LMER has seen limited application in clinical speech research until recent years. A recent tutorial by Gordon (2019) provides a useful example of the application of this statistical approach within this field. Given the observed consistency and similar reliability between loudness ratings obtained using DME and VAS, either could be used as the outcome measure of LMER models. VAS loudness was selected over DME as the outcome measure as it is simple and practical to apply in a clinical setting. Simple models were used to investigate the group interactions while accounting for within- and between-speaker variability. The formula for each of these models was defined as:

$$Loudness \sim Measure * Group + (1|Participant) + (1|Task)$$

This formula states that VAS loudness was predicted by the measure of interest, group, the interaction of the measure and group, and by-participant and by-task random intercepts. LMER models include fixed effects and random effects. Fixed effects are analogous to the effects of a multiple linear regression and tend to be effects that are of primary interest to the investigation. Random effects account for additional variances, providing more robust prediction. In the model formula above, Measure and Group are included as fixed effects, as well as their interaction, represented by the * operator. Random intercepts allow the intercept of the linear regression to randomly vary. By-participant random intercepts vary for each participant, capturing between-participant variability. By-task random intercepts vary for each task (SIT sentences and conversation), capturing between-task variability which in this context contributes to within-speaker variability. Random slopes allow a fixed effect to randomly vary across participant or task, further refining the relationship between that fixed effect and the outcome. Random slopes were not included in the LMER group interaction models, but were included as candidate components of the maximal LMER models, described in detail in Section 3.4.4. The use of LMER does not require normality or homoscedasticity between groups or contrast levels in the underlying data. Appropriate use of LMER requires that the residuals are normally distributed and display homoscedasticity, which was observed in all models reported in this investigation. Effect sizes of LMER models are an area of active research in the statistical field. For this investigation, LMER effect sizes were calculated as Cohen's *d* analogues as per Westfall, Kenny, and Judd (2014), referred to throughout this investigation as delta (δ). These effect sizes were calculated as the regression coefficient divided by the square root of the sum of all variances (residual, participant, and task variances for fixed and random effects). Use of these effect sizes is further discussed in a tutorial and review by Brysbaert and Stevens (2018). It is currently unknown if these delta effect sizes can be interpreted on the same scale as classical Cohen's *d*. Brysbaert and Stevens (2018) notes that effect size estimates of Westfall et al. (2014) may be 'optimistic.' Within this investigation, these effect sizes are used consistently across LMER models, and it will be assumed that these effect sizes are roughly analogous to Cohen's *d* with the caveat that they may be inflated. The primary goal of these effect sizes is to compare effects of different measures and between models, while removing effects of scale and controlling for variability. Direct comparison of Cohen's *d* and δ effects is not within the scope of this investigation.

Maximal LMER models were built using a stepwise approach based on the conceptual grouping of acoustic measures outlined in Section 3.3. LMER was used to identify combinations of acoustic predictors that provide the best prediction of loudness and intelligibility, respectively, while taking into account the effects of within- and between-speaker variability. The model building process is described further in Section 3.4.4. Maximal LMER models predicting loudness informed RQ2 (acoustic models of loudness), RQ3 (acoustic measures as predictors of loudness), and RQ4 (interaction of IWPD-HOA differences on perceived loudness).

Maximal LMER models predicting intelligibility informed RQ6 (loudness and intelligibility), as measures that successfully predict loudness may not predict intelligibility. The same stepwise approach was used, described in Section 3.4.4. The underlying distribution of intelligibility is negatively skewed. Transformations or alternative methods of modeling such as generalized linear mixed models could be considered to achieve a more robust model of intelligibility. Transforms were explored to correct the skew of intelligibility to improve residual normality, including log and logit transformations. Because of the secondary role of intelligibility in this study and the exploratory goal of this model, it was decided that the benefits of more direct comparison in interpretation between the intelligibility and loudness models was more in line with the goals of this investigation, and it was decided that intelligibility models would be built with untransformed data using LMER. Visual assessment of residuals indicated acceptable adherence to assumptions despite underlying skew.

Classification decision trees were used to provide additional insight into group differences between IWPDs and HOAs, informing RQ1 (IWPD-HOA differences). Kuhn and Johnson (2018) provide a detailed overview of the use of classification trees as predictive models. Advanced classification methods like support vector machines and neural networks can provide more robust predictive performance at the cost of interpretability. The choice of decision trees aligns with the exploratory goals of this investigation due to their simplicity and ease of interpretation. However, because of the instability of decision trees, 10 trees predicting group were run on different random 80-20 splits of the data into train-test sets. The full acoustic data (N = 1424) was used for these splits. Suiting the exploratory nature of this investigation, these 10 trees were described and compared to observe the trends and tendencies. The choice to evaluate 10 simple trees was arbitrary, as this was an exploratory and descriptive exercise. A bagged (bootstrap aggregated) tree was also conducted with k-fold cross-validation (k = 10) to evaluate classification performance in a more stable tree and identify important variables. Bagged trees are also described in Kuhn and Johnson (2018). Variable importance was used to compare the results of the bagged tree to the observed trends among simpler decision trees. Variable importance represents the impact of the measure in classifying IWPDs and HOAs, not the magnitude of effect of that measure. In this context, impact refers to the way the trees make decisions at each branch. A measure with high importance is likely to be seen in high-level splits and in splits that cause a large number of participants to be classified as belonging to a particular group.

A logistic regression model predicting group (PD status) was built using a similar model building approach as the LMER model building process described in Section 3.4.4. While linear regression models the predicted value of the outcome, logistic regression models the probability of a binary outcome. The logistic regression model informed RQ1 (IWPD-HOA differences) and RQ3 (acoustic measures and loudness) by providing another perspective on the differences between predicting loudness (RQ3) and predicting PD (RQ1). It was expected that there would be characteristics that effectively identified IWPDs while not being strong predictors of loudness. Averaged values were used for this model to maintain independence of observations, similar to the correlation analyses and tests of group differences.

Based on the considerable heterogeneity observed among IWPDs, IWPDs were divided into two subgroups on the basis of their perceived loudness. This subgrouping provided further investigation into group differences between IWPDs and HOAs (RQ1 and RQ4). IWPDs with an average VAS loudness more than 2 standard deviations below the average VAS loudness of HOAs were deemed to have low loudness. Differences between these 3 groups (HOAs, low loudness IWPDs, normal loudness IWPDs) were investigated using LMER. The advantage of using LMER to investigate these subgroups, rather than ANOVA, is that LMER manages unequal grouping and subsequent heteroscedasticity more effectively and allows for inclusion of repeated measures data. Estimated marginal means (Searle, Speed, & Milliken, 1980) were used to summarize these LMER models and calculate pairwise *t*-tests analogous to post-hoc comparisons using the `emmeans` package in R. Estimated marginal means are determined from the model predictions, rather from the underlying data. This provides the benefit of managing heteroscedasticity and incorporating repeated measures data to these summary statistics.

3.4.3 Experiment 2: Spectral Manipulation

Results of Experiment 2 primarily informed RQ7 (spectral manipulation and loudness), while also providing insight into RQ2 (acoustic models of loudness) and RQ4 (interaction of IWPD-HOA differences on perceived loudness).

ANOVAs were used to evaluate the effects of group (PD status), spectral manipulation, and their interaction on perceived loudness and on acoustic measures. Homoscedasticity of variance was observed across manipulation conditions for the majority of measures. Parametric tests were deemed sufficiently robust to deviations in normality and heteroscedasticity for use in this investigation due to the large sample size and similar sample sizes between IWPDs and HOAs. Shapiro-Wilk tests of normality and Levene's tests of equality of variance within Experiment 2 are presented in Appendix D.

Loudness amplification was defined as the difference score of perceived loudness between the unaltered speech and the positive 10 dB manipulation condition. Pearson correlations between loudness amplification and acoustic measures were obtained to investigate the relationships between underlying acoustic characteristics and effectiveness of spectral manipulation.

3.4.4 LMER Model Building Process

Maximal LMER models were built using a manual stepwise approach based on the conceptual grouping of measures described in Section 3.3. Akaike Information Criterion (AIC) and likelihood ratio tests (LRT) were used for model selection between nested models, models between which only one term differs. AIC is a measure of model fit based on information loss, identifying the models which provide a better fit to the data (Burnham & Anderson, 2004). Lower AIC values represent better fit when comparing two nested models. The absolute values of AIC can vary considerably, which is why it is used specifically for selecting between nested models. Similarly, LRT is a statistical test comparing the likelihood ratios via chi-square tests of two nested candidate models, assessing goodness-of-fit and providing hypothesis testing. Using these model selection criteria reduces overparameterization. Significant LRT indicates that the addition of a model term significantly improves model performance.

Predictors were added into the model in the order of the categories: speech level, spectral balance, variability. Speech level predictors were added first, then spectral balance predictors, and then variability predictors. This order was based on the expected magnitude of contribution of each category to the overall perceived loudness. Within a category, each candidate predictor was added individually and each candidate model was compared to the previous (or baseline) model using LRT and to the other candidates using AIC. Combinations of predictors were then attempted, so long as variance inflation factor (VIF) remained below 5 to manage collinearity. VIF measures the collinearity of terms within a particular model by evaluating the effect of correlation between predictors on the variance of the regression coefficients (Akinwande, Dikko, & Samson, 2015). VIF values above 5 represent high correlations that are likely to influence model results, while a value of 1 would indicate no correlation between predictors. As terms were added to the model, they were maintained if they contributed to model performance as demonstrated by LRT $p < .05$ and reduced AIC. Fixed effects were identified first, then interactions between each fixed effect, then random slopes for each fixed effect. Only significant interactions were maintained in the final maximal models.

The baseline model for each outcome (perceived loudness or intelligibility) was defined as:

$$Outcome \sim Group + (1|Participant) + (1|Task)$$

A fixed effect of group was included in all models based on conceptual expectations and due to consistent group differences identified by t-tests and visual inspection of distributions. By-participant and by-task random intercepts are included to account for the variation within- and between-speakers. Full details of the model building process for each model, including the intermediate tables from each stage of the selection process, are presented in Appendix G.

3.5 Sample Size and Power Analysis

As the present investigation is based on archived audio data, the sample size was determined by the available data. Power analysis was conducted in G*Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) to confirm that the available sample size provided sufficient power. Power analyses specifically focused on LMER models were not conducted. As discussed in Section 3.4.2, power analyses and effect sizes are active areas of research with regard to LMER; consequently, obtaining a confident estimate of power is challenging. The focuses of the power analyses outlined below were to estimate the required power to detect 1) group differences between IWPDs and HOAs, 2) relationships between speech level and perceived loudness, and 3) an effect of spectral manipulation on perceived loudness.

To determine the required sample size of IWPDs and HOAs to detect group differences related to PD status, one-tailed t-tests were conducted as the directions of effect were expected to be consistent. IWPDs were expected to be quieter as measured by mean intensity and perceived loudness, less intelligible, and show a relatively greater concentration of low-frequency energy in their spectral distributons. Perceived loudness has not been included in many investigations comparing IWPDs and HOAs. Ludlow and Bassich (1984) found an effect size of d = 1.20 between the 12 IWPDs and 12 HOAs in their investigation. Large effect sizes of intelligibility have also been reported between IWPDs and HOAs, varying from d = 0.90 to d = 1.02 (Miller et al., 2007; Weismer et al., 2001). Assuming large effect sizes of d = 1.0, 56 IWPDs and 46 HOAs were estimated to provide a power of 0.99 to detect group differences in perceived loudness and intelligibility. With regard to mean intensity, an effect size of d = 0.87 was expected based on several studies of hypophonia that reported means and standard deviations of mean intensity for both groups (Brajot et al., 2016; Huber & Darling, 2011; McCaig, Adams, Dykstra, & Jog, 2016; Sapir et al., 2007; Tjaden & Martel-Sauvageau, 2017; Tjaden & Wilding, 2004). With 56 IWPDs and 46 HOAs, it was expected that power of 0.99 would be obtained to detect group differences in mean intensity. Large effect sizes have been reported in studies of spectral balance of IWPDs and HOAs, including spectral mean, skewness, and kurtosis with effect size estimates on these parameters ranging from d = 1.0 to d = 1.76 (Dromey, 2003; Smith & Goberman, 2014). Assuming a large effect size of d = 1.0, 56 IWPDs and 46 HOAs were estimated to provide a power of 0.99 to detect group differences in spectral balance. Overall, it was deemed that inclusion of 56 IWPDs and 46 HOAs would provide adequate power to detect group differences in mean intensity, perceived loudness, intelligibility, and spectral balance.

It was more challenging to estimate the power required to detect relationships between acoustic measures and perceived loudness, as this area of the present investigation is relatively novel. Ludlow and Bassich (1984) found a Pearson correlation of $r = 0.36$ between mean intensity and loudness with 24 participants. With 102 participants (56 IWPDs and 46 HOAs), it was estimated that power of 0.99 would be provided to detect the correlation between mean intensity and perceived loudness.

Estimating power required to detect an effect of spectral manipulation on perceived loudness was particularly difficult. While previous findings suggest the contribution of high-frequency energy to loudness, the experimental designs are considerably different from the present investigation (Duvvuru & Erickson, 2013; Titze, 2020). Titze (2020) investigated the effects of single-harmonic spectral manipulations of loudness with a computational paradigm, but loudness was acoustically determined based on equal-loudness contours, rather than based on experimental perceived loudness findings. Duvvuru and Erickson (2013) investigated spectral slope and loudness in synthesized vocal stimuli, with loudness ratings obtained from 15 listeners using a loudness-matching paradigm. The spectral slope of each synthesized stimulus was modified by 3 dB/octave and 6 dB/octave, and the differences in perceived loudness between conditions were reported. A 3 dB/octave adjustment yielded an effect size of $d = 0.73$, and the 6 dB/octave adjustment yielded an effect size of $d = 1.04$. The spectral manipulation in the present investigation is a targeted gain adjustment, rather than a slope adjustment. As a result, it is difficult to directly translate these effect sizes to the design of the present study. Assuming a moderately-large effect size of *d* = 0.7 and 102 participants (56 IWPDs and 46 HOAs), it was expected that power of 0.99 would be obtained.

Overall, it is believed that this investigation was sufficiently powered with the available data from 56 IWPDs and 46 HOAs as all power estimates significantly exceeded the 0.80 recommendation of Cohen (1992). Due to the use of archived audio data, it was possible to investigate a larger group of IWPDs than would usually available, as the balance between participant time and adequate power is an important consideration in determining how many individuals will be recruited. By incorporating across several studies, this balance was not required for the present investigation and it was possible to include more individuals than would be suggested by power analysis. This investigation will also provide clearer expectations for the power required for future studies exploring these dimensions in greater detail.

4 Results

Results for each experiment are provided separately. Findings will be integrated and interpreted with respect to the research questions in Chapter 5.

4.1 Reliability

Reliability of perceptual ratings was calculated as described in Chapter 3. ICC values between 0.75 and 0.90 are deemed to represent 'good' reliability. Based on the results of reliability analyses, 2 listeners were removed from further analyses due to poor reliability. Demographic characteristics of the final group of listeners is presented in Table 4.1.

Age (Mean)	*Age (SD)*	*Age (Range)*	*N*	*Gender*
24	1.77	22-28	8	F = 8

Table 4.1: Demographic characteristics of listener participants included in the analysis.

Results of reliability analyses from the selected 8 listeners are presented in Table 4.2. Inter-rater reliability was higher in general than intra-rater reliability. This may reflect the contribution of perceptual drift, such that listeners' ratings were affected by the neighbouring samples. Randomized presentation order is intended to mitigate this problem, but may not entirely remove it. Inter-rater reliability for VAS was higher than for DME, whereas the reverse was true for intra-rater reliability, showing higher intra-rater reliability in DME than VAS. This may be because DME is naturally more idiosyncratic as a method, as each listener chooses their scale and increment. Overall, excellent intra-rater and inter-rater reliability was observed for intelligibility and for loudness across measurement techniques and experiments.

Measure	*ICC Type*	*ICC*	*95% CI*	*p*
Intra-Rater, Spec. Manip. Loudness	ICC3k	**0.920**	0.906 – 0.932	< 0.001
Intra-Rater, VAS Loudness	ICC3k	**0.905**	0.886 – 0.921	< 0.001
Intra-Rater, DME Loudness	ICC3k	**0.929**	0.915 – 0.941	< 0.001
Intra-Rater, Intelligibility	ICC3k	**0.932**	0.919 – 0.944	< 0.001
Inter-Rater, Spec. Manip. Loudness	ICC2k	**0.958**	0.951 – 0.963	< 0.001
Inter-Rater, VAS Loudness	ICC2k	**0.960**	0.954 – 0.965	< 0.001
Inter-Rater, DME Loudness	ICC2k	**0.949**	0.931 – 0.962	< 0.001
Inter-Rater, Intelligibility	ICC2k	**0.945**	0.918 – 0.961	< 0.001

Table 4.2: Inter-rater and intra-rater reliability results calculated via ICC.

4.2 Experiment 1: Unaltered Speech

4.2.1 Descriptive Statistics

Means and standard deviations of all measures are presented in Table 4.3.

Measure	*HOA*	*IWPD*
VAS Loudness (%)	69.54 (11.13)	50.23 (20.34)
DME Percent (%)	62.94 (11.06)	44.75 (17.81)
DME Geometric	92.06 (16.85)	64.16 (27.33)
Intelligibility (%)	86.44 (7.88)	59.64 (23.76)
Mean Intensity (dB SPL)	69.44 (3.17)	66.56 (4.66)
Median Intensity (dB SPL)	66.72 (3.80)	63.12 (7.63)
Max Intensity (dB SPL)	76.90 (3.31)	73.56 (4.43)
TVL (sones)	19.22 (4.00)	15.27 (5.55)
TVL Mean (sones)	13.75 (2.76)	10.71 (4.18)
LKFS	-25.12 (3.18)	-27.88 (4.57)
Active Speech Level (dB FS)	-25.60 (3.25)	-29.32 (5.14)
Tilt (dB)	-24.97 (3.52)	-28.61 (4.95)
Voiced Tilt (dB)	-26.09 (3.67)	-29.27 (5.16)
Tilt Ratio	1.05 (0.06)	1.02 (0.06)
Mid-Ratio	0.17 (0.06)	0.08 (0.11)
High-Ratio	0.11 (0.08)	-0.02 (0.12)
Skewness	9.19 (3.15)	14.11 (8.86)
Kurtosis	139.83 (112.28)	397.51 (544.93)
SD Intensity (dB)	12.24 (1.98)	12.96 (2.70)
SD TVL (sones)	3.68 (0.96)	3.12 (1.21)
*Intensity Decay **	-22.33 (25.61)	-18.92 (27.09)
*TVL Decay **	-1.36 (1.39)	-0.90 (1.23)

Table 4.3: Means and standard deviations for each perceptual and acoustic measure. Perceptual measures (VAS Loudness, DME Percent, DME Geometric, Intelligibility) are calculated from the 4 speech tasks presented to listeners (N = 408). Acoustic measures are calculated from all 14 speech tasks (N = 1424). * Mean and standard deviation values for intensity decay and TVL decay are expressed in scientific notation (x 10^3).

4.2.2 T-Tests

4.2.2.1 Perceptual

Group differences between IWPDs and HOAs were evaluated using t-tests based on values averaged across the 4 presented speech tasks. Results of these tests, as well as Cohen's d effect sizes, are presented in Table 4.4. p-values were adjusted for multiple comparisons with the Holm-Bonferroni method.

Measure	*Mean Difference*	*HOA Mean*	*IWPD Mean*	*t*	*95% CI*	*p*	*d*
VAS Loudness (%)	19.32	69.54	50.23	6.68	13.57 – 25.07	< 0.001	1.15
DME Percent (%)	18.19	62.94	44.75	7.04	13.05 – 23.34	< 0.001	1.20
DME Geometric	27.90	92.06	64.16	7.07	20.06 – 35.75	< 0.001	1.21
Intelligibility (%)	26.80	86.44	59.64	8.71	20.65 – 32.94	< 0.001	1.46

Table 4.4: Welch t-tests evaluating group differences between IWPDs and HOAs on perceptual measures, averaged within each participant across the 4 tasks provided to listeners. p-values were adjusted for multiple comparisons using Holm-Bonferroni correction.

For the first measure listed in Table 4.4, VAS loudness, the results of the Welch t-test indicated that HOAs demonstrated significantly greater perceived loudness (M = 69.54) than IWPDs (M = 50.23; $t(81)$ = 6.68, p < .001, d = -1.15). Results for the other perceptual variables are presented in Table 4.4. Significant group differences were observed between IWPDs and HOAs on all perceptual measures, with large effect sizes. Figure 4.1 presents violin plots of each measure, making clear the considerable difference in distributions between measures. Overall, IWPDs were found to be significantly quieter and less intelligible than HOAs. Greater variability existed among IWPDs than among HOAs. The distribution of intelligibility is particularly skewed in HOAs. As no background noise was used in the presentation of speech samples, HOAs were expected to be intelligible to listeners, consistent with the observed skew.

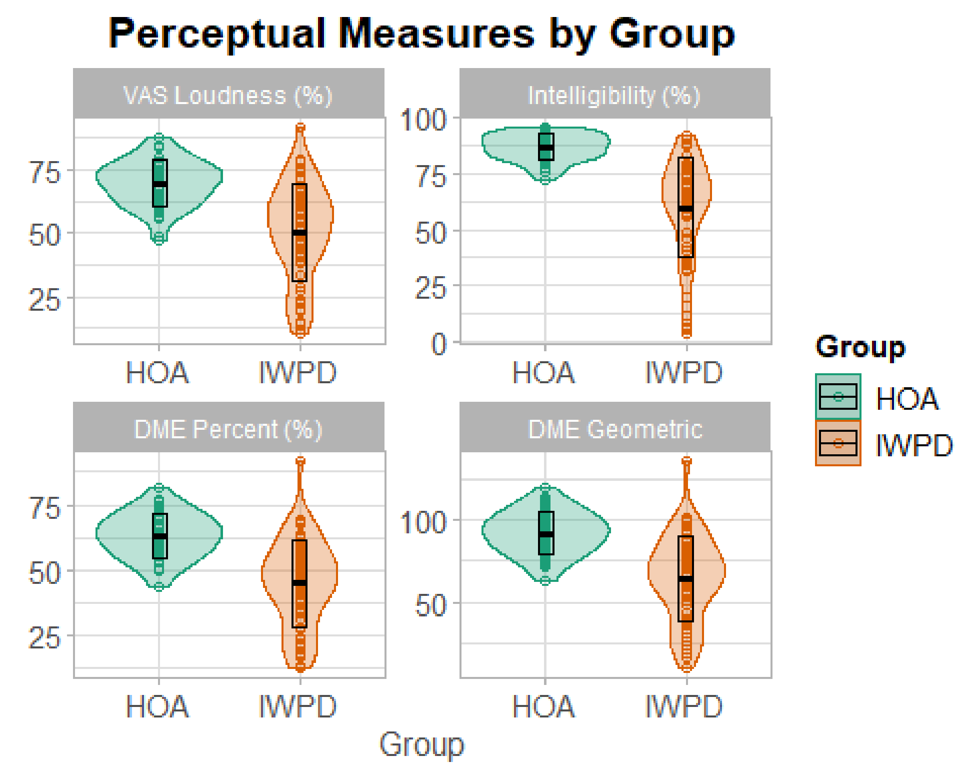


Figure 4.1: Violin plots visualizing differences in distributions of perceptual measures between IWPDs and HOAs. Crossbars within each violin plot present the mean ± 1 SD.

4.2.2.2 Acoustic

Group differences between IWPDs and HOAs were evaluated using t-tests calculated based on values averaged across all 14 speech tasks. Results of these tests, as well as Cohen's *d* effect sizes, are presented in Table 4.5. *p*-values were adjusted for multiple comparisons with the Holm-Bonferroni method.

Measure	*Mean Difference*	*HOA Mean*	*IWPD Mean*	*t*	*95% CI*	*p*	*d*
Mean Intensity (dB SPL)	2.90	69.43	66.54	4.15	1.51 – 4.28	< 0.001	0.71
Median Intensity (dB SPL)	3.63	66.71	63.08	4.00	1.83 – 5.44	0.001	0.58
Max Intensity (dB SPL)	3.36	76.89	73.54	4.88	1.99 – 4.72	< 0.001	0.84
TVL (sones)	3.98	19.22	15.24	4.74	2.31 – 5.65	< 0.001	0.81
TVL Mean (sones)	3.06	13.74	10.68	4.98	1.84 – 4.28	< 0.001	0.84
LKFS	2.77	-25.13	-27.90	4.00	1.40 – 4.15	0.001	0.69
Active Speech Level (dB FS)	3.74	-25.60	-29.34	5.00	2.25 – 5.22	< 0.001	0.85
Tilt (dB)	3.66	-24.97	-28.64	4.85	2.16 – 5.16	< 0.001	0.84
Voiced Tilt (dB)	3.20	-26.09	-29.30	4.07	1.64 – 4.77	< 0.001	0.70
Tilt Ratio	0.02	1.05	1.02	2.67	0.01 – 0.04	0.027	0.36
Mid-Ratio	0.08	0.17	0.08	5.40	0.05 – 0.11	< 0.001	0.92
High-Ratio	0.13	0.11	-0.02	7.40	0.09 – 0.16	< 0.001	1.22
Skewness	-4.95	9.19	14.15	-4.34	-7.23 – -2.68	< 0.001	0.71
Kurtosis	-259.52	139.87	399.40	-3.97	-390.14 – -128.90	0.001	0.63
SD Intensity (dB)	-0.72	12.24	12.96	-2.57	-1.28 – -0.17	0.027	0.30
SD TVL (sones)	0.56	3.68	3.12	3.35	0.23 – 0.90	0.005	0.51
*Intensity Decay**	-3.35	-22.32	-18.97	-1.73	-7.19 – 0.48	0.086	0.13
*TVL Decay**	-0.46	-1.36	-0.89	-3.87	-0.70 – -0.23	0.001	0.36

Table 4.5: Welch t-tests evaluating group differences between IWPDs and HOAs on acoustic measures, averaged within each participant across all 14 speech tasks. *p*-values were adjusted for multiple comparisons using Holm-Bonferroni correction.
* Means, mean difference, and confidence interval values for intensity decay and TVL decay are expressed in scientific notation (x 10^3).

For the first measure listed in Table 4.5, mean intensity, the results of the Welch t-test indicated that HOAs demonstrated significantly greater mean intensity (M = 69.43 dB SPL) than IWPDs (M = 66.54 dB SPL; $t(94) = 4.15$, $p < .001$, $d = -0.71$). Results for the other acoustic variables are presented in Table 4.5. Most acoustic measures demonstrated significant group differences between IWPDs and HOAs. Notably, effect sizes were smaller for most acoustic measures compared to perceptual measures, though effect sizes were still medium to large for several measures. Larger effect sizes were observed for TVL measures compared to intensity measures. A large effect size was also observed for active speech level. Particularly large effect sizes were observed for mid-ratio and high-ratio, highlighting the considerable difference in mid- and high-frequency energy between IWPDs and HOAs. Figure 4.2 presents violin plots of each acoustic measure. As with perceptual measures, greater variability among IWPDs compared to HOAs was observed for most acoustic measures, reflecting the heterogeneity of IWPDs as a population.

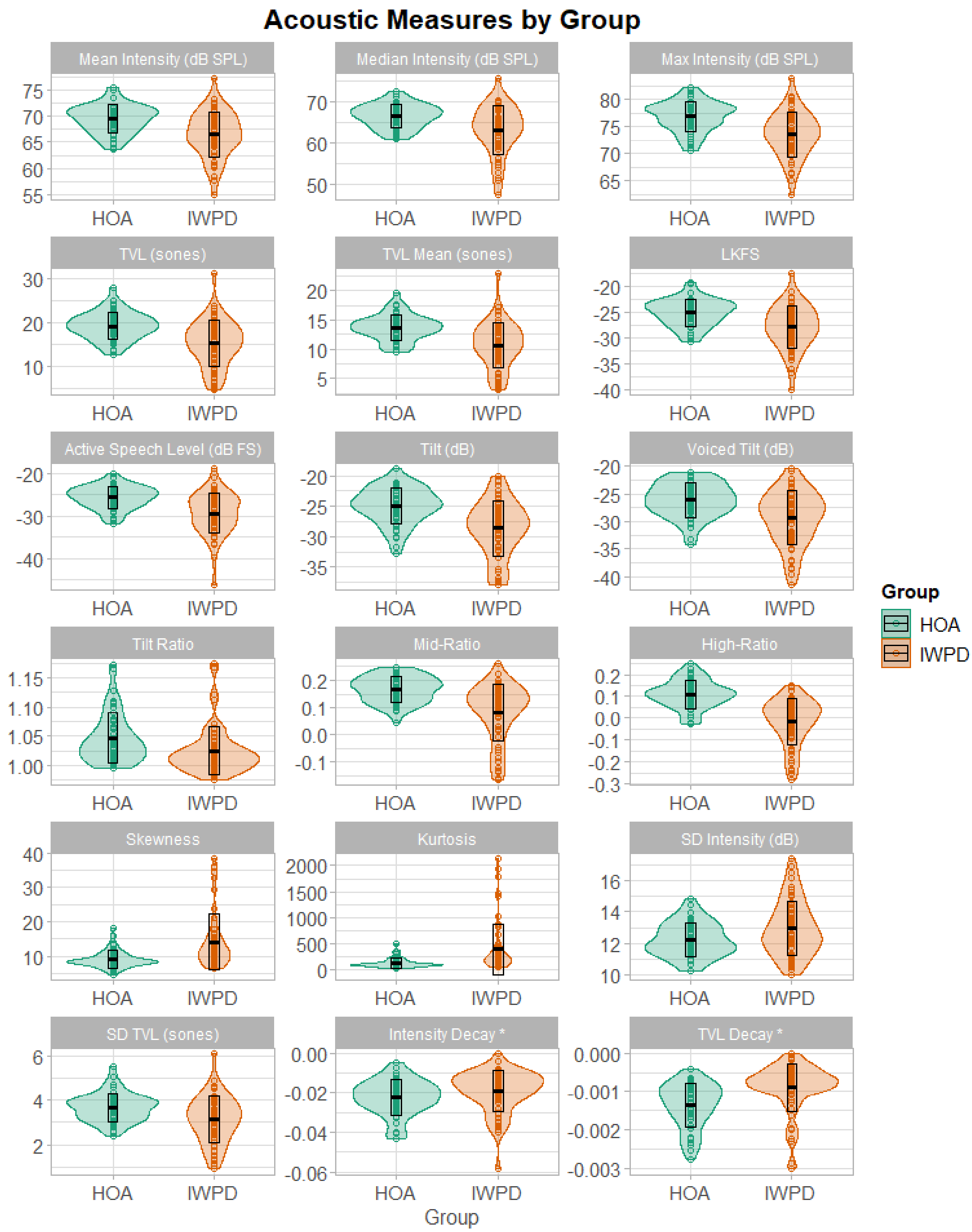


Figure 4.2: Violin plots visualizing differences in distributions of acoustic measures, averaged within participant, between IWPDs and HOAs. Crossbars within each violin plot present the mean ± 1 SD.

4.2.3 Correlations

Pearson correlations of participant-averaged values are presented below. Correlations involving perceptual measures were based on values averaged across the 4 presented speech tasks. Correlations between acoustic measures were based on values averaged across all 14 speech tasks.

4.2.3.1 Perceptual

Correlations between perceptual measures are presented in Table 4.6.

		r	*N*	*p*
DME Percent (%)	*VAS Loudness (%)*	0.987	102	< 0.001
DME Geometric	*VAS Loudness (%)*	0.986	102	< 0.001
DME Percent (%)	*DME Geometric*	0.999	102	< 0.001
Intelligibility (%)	*VAS Loudness (%)*	0.749	102	< 0.001
Intelligibility (%)	*DME Percent (%)*	0.767	102	< 0.001
Intelligibility (%)	*DME Geometric*	0.774	102	< 0.001

Table 4.6: Pearson correlations between perceptual measures, averaged within each participant across the 4 tasks provided to listeners.

The correlation between percent-averaged and geometric-averaged DME loudness ratings approached unity ($r(100) = 0.999$, p < .001), supporting the use of either method of averaging. Very strong positive correlations were identified between loudness ratings provided using VAS and DME rating methods, indicating consistency between these tools. Moderately strong positive correlations were identified between intelligibility and loudness ratings.

4.2.3.2 Acoustic

A correlation plot and table is presented in Figure 4.3 to efficiently present a large number of correlations. Full correlation tables are presented in Appendix E.

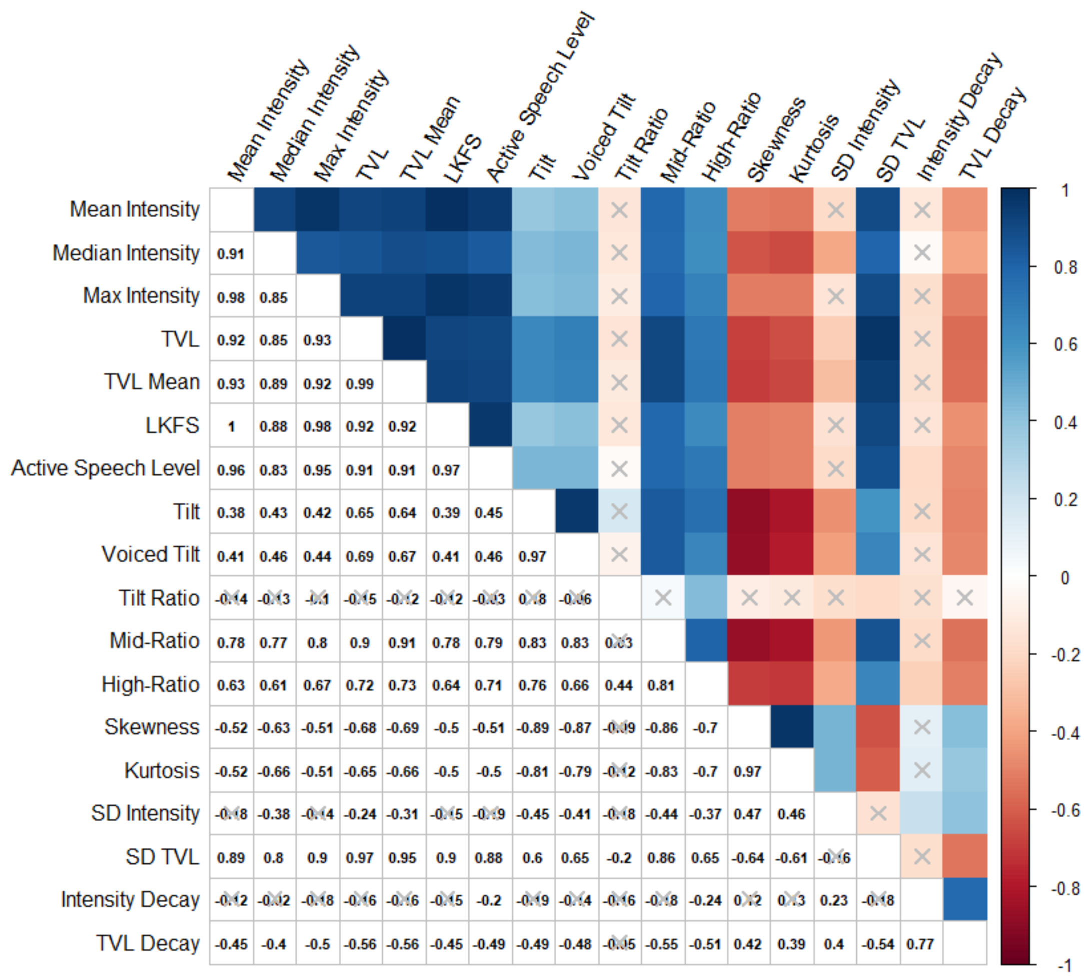


Figure 4.3: Correlation plot presenting Pearson correlations between acoustic measures, averaged within each participant (N = 102). Darker squares represent stronger correlations. Positive correlations are coloured in blue, and negative correlations in red. Correlations that were not significant at p < .05 are represented by an X. All correlations were significant at p < .05.

Strong positive correlations were observed between the speech level measures. Measures of spectral balance tended to correlate more strongly with TVL and TVL mean than with other speech level measures, consistent with the contribution of spectral distribution within TVL's algorithm. Strong negative correlations were observed between skewness and kurtosis and other spectral balance measures, showing stronger correlations with tilt and voiced tilt than with mid-ratio or high-ratio. Tilt ratio demonstrated weak correlations with most measures, but a moderate negative correlation with high-ratio. Higher tilt ratio reflects a greater similarity between voiced tilt and overall tilt. A wider discrepancy between voiced tilt and overall tilt would be reflected by a lower tilt ratio, and could be explained by high-frequency turbulent energy of voiceless sibilants and stop consonants included in the overall tilt. A negative correlation between high-ratio and tilt ratio suggests that a lower proportion of high-frequency energy is associated with a greater similarity between voiced tilt and overall tilt, consistent with high-frequency deficits being driven by weak high-frequency harmonic energy. Decay demonstrated weak correlations with all measures except TVL decay. TVL decay's correlations were moderate with several measures of speech level and spectral balance. Similarly, correlations for SD TVL were much stronger than SD intensity, particularly with measures of speech level and with mid-ratio. This may reflect the additional smoothing of TVL's long-term loudness compared to the intensity contour.

4.2.3.3 Perceptual-Acoustic

Correlations between acoustic and perceptual measures are presented in the figures below to allow for efficient presentation of several correlations. Figure 4.4 presents the correlations between acoustic measures and both VAS loudness and percent-averaged DME loudness. Full correlation tables are presented in Appendix E.

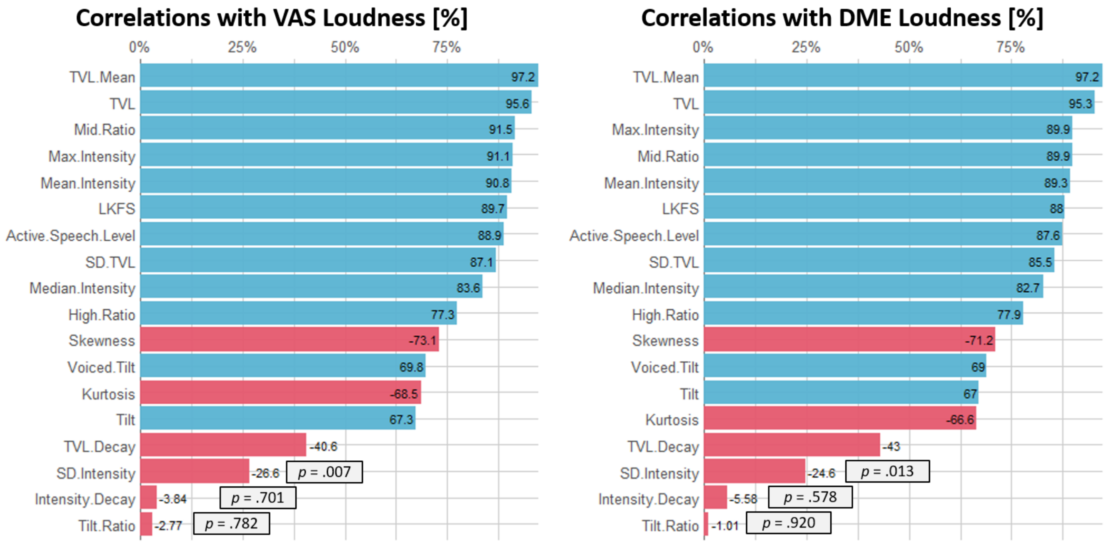


Figure 4.4: Correlation plot presenting Pearson correlations between acoustic measures and perceived loudness measures, averaged within each participant (N = 102). Correlations are presented as percentages to allow both positive and negative correlations to be visualized on the same scale. Positive correlations are coloured in blue, and negative correlations in red. All correlations were significant at $p < .001$, except where noted.

Perceived loudness collected via VAS and DME demonstrated similar patterns of correlations. All speech level measures correlated positively with loudness. Very strong correlations were observed between TVL and loudness, with slightly higher correlations for TVL mean (VAS: $r(100) = 0.97$, $p < .001$; DME: $r(100) = 0.97$, $p < .001$) than for TVL (VAS: $r(100) = 0.96$, $p < .001$; DME: $r(100) = 0.95$, $p < .001$). Correlations between loudness and mean intensity were weaker (VAS: $r(100) = 0.91$, $p < .001$; DME: $r(100) = 0.89$, $p < .001$). Moderate positive correlations with loudness were observed for tilt, voiced tilt, skewness and kurtosis, but correlations were stronger for mid-ratio (VAS: $r(100) = 0.92$, $p < .001$; DME: $r(100) = 0.90$, $p < .001$) and high-ratio (VAS: $r(100) = 0.77$, $p < .001$; DME: $r(100) = 0.78$, $p < .001$). In particular, mid-ratio's correlations with loudness were similar in strength to mean intensity. This supports the particular importance of mid-frequency energy to perceived loudness. Intensity decay did not significantly correlate with loudness (VAS: $r(100) = -0.04$, $p = 0.701$; DME: $r(100) = -0.06$, $p = 0.578$), but TVL decay demonstrated significant but weak correlations with loudness (VAS: $r(100) = 0.23$, $p = 0.020$; DME: $r(100) = 0.21$, $p = 0.037$). SD intensity demonstrated significant but weak correlations with loudness (VAS: $r(100) = -0.27$, $p = 0.007$; DME: $r(100) = -0.25$, $p = 0.013$). Among variability measures, the strongest correlation was between SD TVL and loudness (VAS: $r(100) = 0.87$, $p < .001$; DME: $r(100) = 0.86$, $p < .001$), which may reflect the effects of TVL's long-term loudness smoothing.

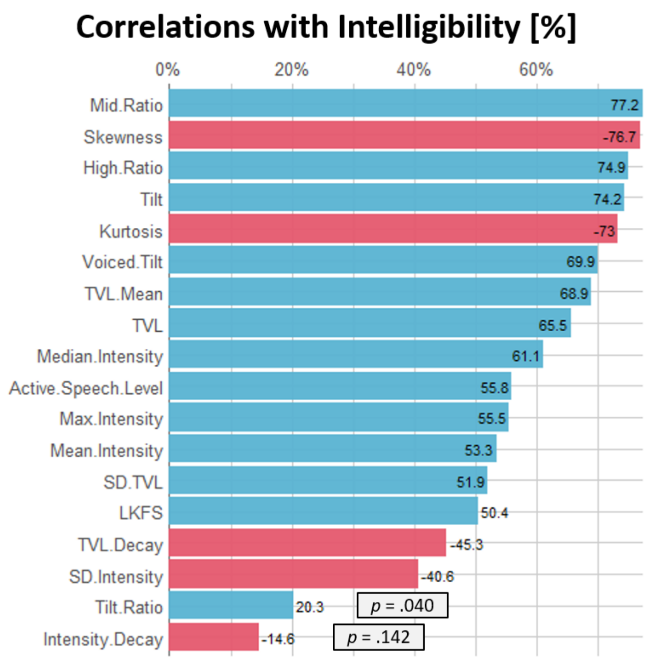


Figure 4.5: Correlation plot presenting Pearson correlations between acoustic measures and intelligibility, averaged within each participant ($N = 102$). Correlations are presented as percentages to allow both positive and negative correlations to be visualized on the same scale. Positive correlations are coloured in blue, and negative correlations in red. All correlations were significant at $p < .001$, except where noted.

Figure 4.5 presents the correlations between acoustic measures and intelligibility.

Correlations with intelligibility were more modest. Intensity measures demonstrated only weak-moderate correlations with intelligibility, such as for mean intensity ($r(100) = 0.53$, $p < .001$), whereas TVL mean ($r(100) = 0.69$, $p < .001$) and TVL ($r(100) = 0.65$, $p < .001$) correlated moderately strongly. The strongest correlations with intelligibility were observed for spectral balance measures. Strong positive correlations with intelligibility were observed for mid-ratio ($r(100) = 0.77$, $p < .001$) and high-ratio ($r(100) = 0.75$, $p < .001$). A larger discrepancy was observed in the correlations of mid-ratio and high-ratio with perceived loudness, which demonstrated a stronger association between mid-ratio and perceived loudness. In the context of intelligibility, similar correlations were observed for both mid-ratio and high-ratio, suggesting the importance of both mid- and high-frequency energy in the perception of intelligibility. Strong negative correlations between intelligibility and both skewness and kurtosis indicated that a concentration of energy in the lower frequencies was associated with lower intelligibility, consistent with the findings of high-ratio.

4.2.3.3.1 IWPDs vs. HOAs

Correlations were also obtained within each group, as the relationships between measures may vary based on the different speech characteristics of IWPDs and HOAs. Correlations between VAS loudness and acoustic measures are presented in Figure 4.6. Full correlation tables are presented in Appendix E.

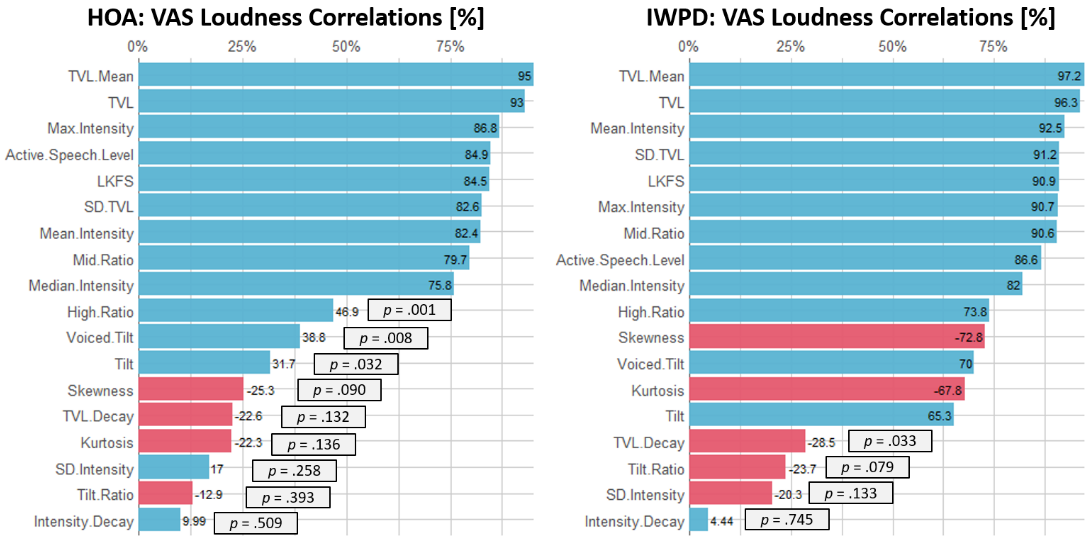


Figure 4.6: Correlation plot presenting Pearson correlations between acoustic measures and VAS loudness, averaged within each participant (N = 46 for HOA, N = 56 for IWPD). Correlations are presented as percentages to allow both positive and negative correlations to be visualized on the same scale. Positive correlations are coloured in blue, and negative correlations in red. All correlations were significant at $p < .001$, except where noted.

In general, correlations between perceived loudness and acoustic measures were stronger in IWPDs, though similar patterns in strength and direction of association were observed between IWPDs and HOAs. This may reflect the effects of a greater range of perceived loudness values among IWPDs. Additionally, weaker and often insignificant correlations were observed between perceived loudness and spectral balance measures in HOAs, whereas the majority of spectral balance measures showed significant, moderate correlations with perceived loudness in IWPDs. For example, tilt's correlations with VAS loudness were stronger among IWPDs ($r(54)$ = 0.65, $p < .001$) than among HOAs ($r(44)$ = 0.32, p = 0.032). Similarly, the correlation between SD TVL and perceived loudness was much stronger in IWPDs than among HOAs.

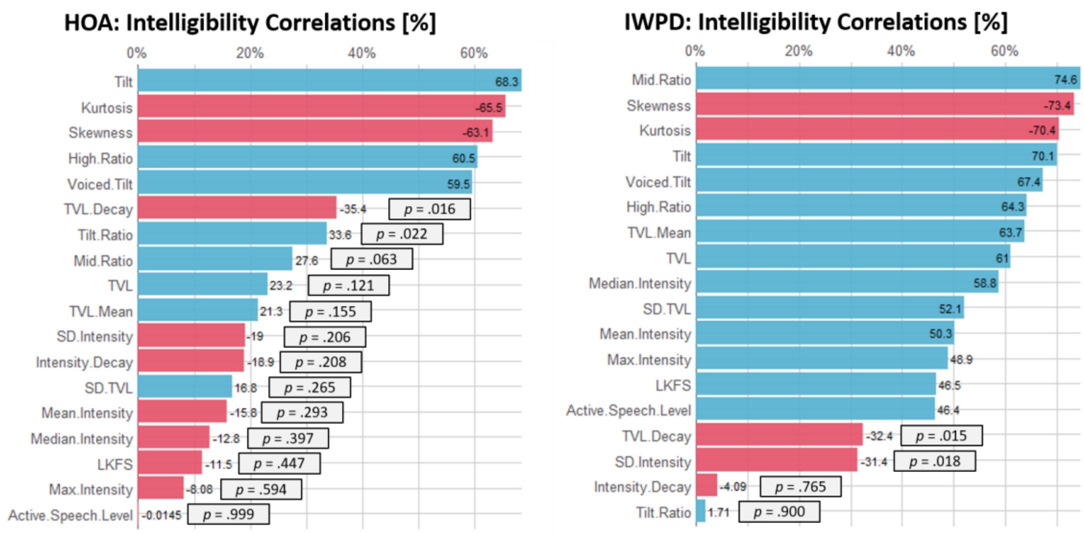


Figure 4.7: Correlation plots presenting Pearson correlations between acoustic measures and intelligibility, averaged within each participant ($N = 46$ for HOA, $N = 56$ for IWPD). Correlations are presented as percentages to allow both positive and negative correlations to be visualized on the same scale. Positive correlations are coloured in blue, and negative correlations in red. All correlations were significant at $p < .001$, except where noted.

Correlations between intelligibility and acoustic measures, presented in Figure 4.7, were even weaker in HOAs than the correlations with loudness. The distribution of intelligibility was particularly skewed for HOAs, which may reflect a restricted range. Notably, correlations between kurtosis, skewness, tilt, and intelligibility were similar in strength for HOAs and IWPDs. As HOAs demonstrated a broader energy distribution across the frequency spectrum compared to the low-frequency concentration of IWPDs, overall measures of spectral shape may be more effective descriptors of HOA spectral characteristics than finer measures like mid-ratio and high-ratio. In IWPDs, a moderately strong correlation between mid-ratio and intelligibility was observed ($r(54) = 0.75$, $p < .001$), whereas this correlation was weak in HOAs ($r(44) = 0.28$, $p = 0.063$). Correlations between intelligibility and speech level measures were weak and insignificant in HOAs but moderate in IWPDs, such as mean intensity (HOA: $r(44) = -0.16$, $p = 0.293$; IWPD: $r(54) = 0.50$, $p < .001$).

4.2.4 LMER: Group Interactions on Loudness

LMER models were used to identify measures which showed an interaction with group on loudness. Figure 4.8 presents interaction plots of each of these models.

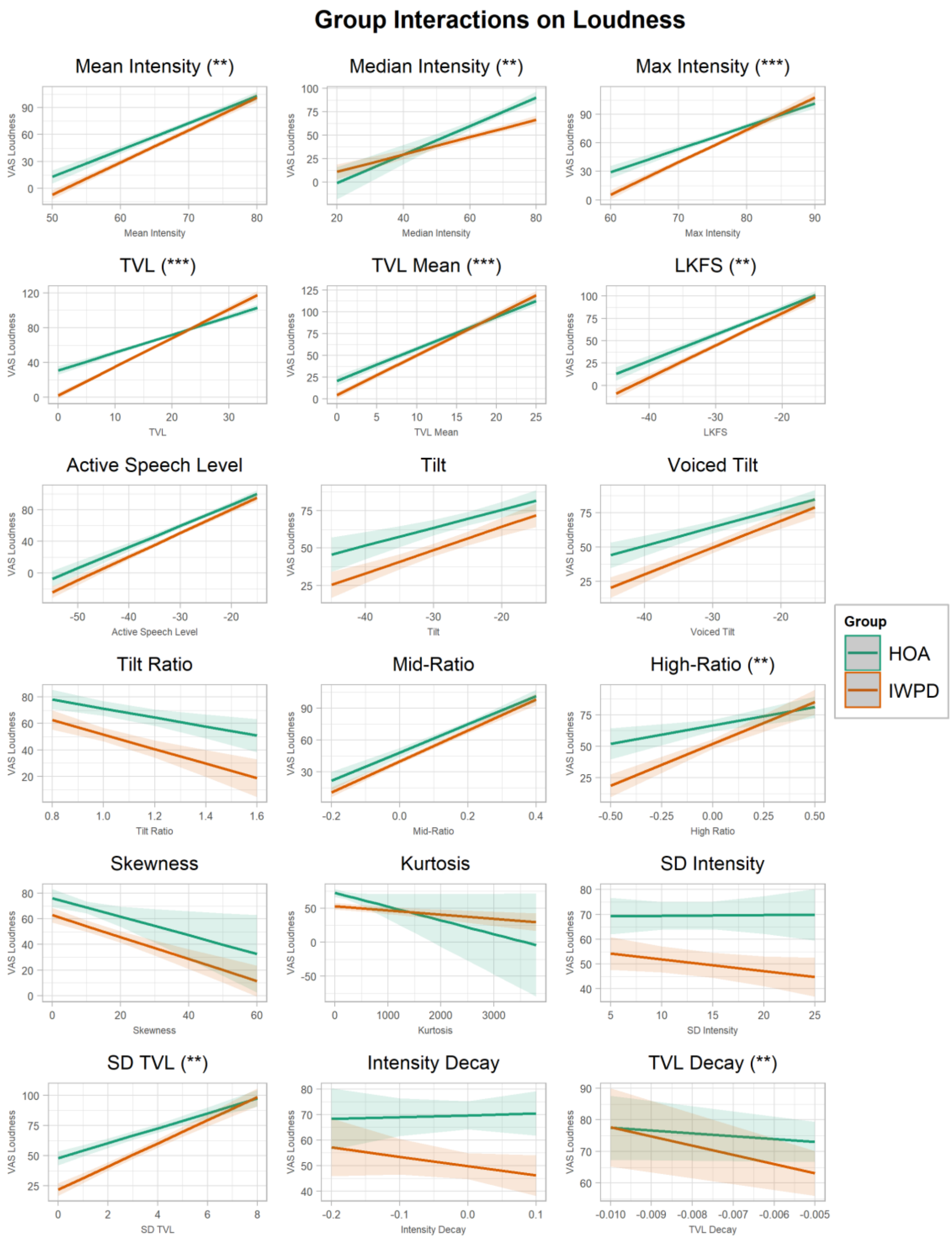


Figure 4.8: Interaction plots representing the group-predictor interactions from each LMER. Significance indicators in the title represent the p-value: (**) $p < .01$, (***) $p < .001$. Shaded areas around each line represent the 95% CI.

Results of the interaction between mean intensity and group on loudness indicated a positive relationship between mean intensity and loudness in both groups, with a stronger effect among IWPDs than among HOAs (β = 5.19, t = 2.75, p = 0.006, δ = 0.63). Similar details of each LMER model are presented in Appendix F.

Overall, intensity measures, TVL measures and LKFS demonstrated significant interactions with group, all revealing stronger positive relationships between speech level and loudness among IWPDs. This may be related to the broader distribution of loudness and of speech level in IWPDs compared to HOAs. TVL and maximum intensity showed a greater divergence between IWPDs and HOAs at lower values, suggesting that among IWPDs, low values of TVL or maximum intensity were particularly influential on perceived loudness. An interesting reversal of the usual speech level group interactions on loudness was observed with median intensity. A weaker relationship between median intensity and loudness was observed for IWPDs, reversing the observed interactions for mean intensity and maximum intensity. It is possible that this difference in averaging captures characteristics of the intensity contour, such as speech level variability, that are important to the overall perceived loudness. Particular importance of median intensity to loudness in HOAs might reflect their larger intensity modulation, which reduces the median intensity.

Most spectral balance measures did not demonstrate significant interactions with group except for high-ratio. High-ratio also showed a stronger positive relationship with loudness in IWPDs. Measures of intensity variability did not reveal group interactions, but measures of TVL variability did. A stronger positive relationship between SD TVL and loudness was observed in IWPDs, and a stronger negative relationship between TVL decay and loudness was observed in IWPDs. Notably, very high variability was observed for TVL decay given the large confidence interval, suggesting considerable individual variability.

4.2.5 LMER: Loudness

A maximal LMER model predicting VAS loudness was built using the model building approach described in Chapter 3. Full details of the model building process of each model, including intermediate models, is presented in Appendix G. The model building process began with the baseline model, with the formula:

$$Loudness \sim Group + (1|Participant) + (1|Task)$$

All speech level predictors were found to significantly improve model performance, with lowest AIC obtained via TVL mean. Combinations of speech level predictors improved performance, but in models with two speech level predictors, VIF exceeded the threshold as soon as spectral balance predictors were incorporated. As a result, only TVL mean was maintained. All spectral balance predictors except tilt ratio significantly improved model performance, with lowest AIC obtained via mid-ratio. Combinations of spectral balance predictors did not improve model performance, and only mid-ratio was maintained. Both SD TVL and SD intensity significantly improved model performance, but AIC and VIF were lower via SD intensity, and SD intensity was maintained. Interactions between predictors were attempted. A significant improvement in model performance was obtained with the interaction of TVL mean and mid-ratio. Smaller significant improvements in performance (based on a smaller change in AIC) were observed with interactions between mid-ratio and SD intensity, mid-ratio and group, TVL mean and SD intensity, TVL mean and group, and mid-ratio and group. Combinations of interactions did not significantly improve performance over the single interaction of TVL mean and mid-ratio, and that was the only interaction maintained. Only the addition of a by-participant random slope of TVL mean improved model performance. The final model predicting VAS loudness was defined as:

$$Loudness \sim TVL\ Mean * MidRatio + SD\ Intensity + Group +$$
$$(TVL\ Mean \mid Participant) + (1|Task)$$

Acceptable VIF was observed in the final model, presented in Table 4.7.

Predictor	*VIF*
TVL Mean	2.57
Mid-Ratio	2.81
SD Intensity	1.05
Group	1.17
TVL Mean * Mid-Ratio	1.14

Table 4.7: Variance inflation factor values for each predictor in the final model predicting VAS loudness. Values of 1 represent no collinearity between predictors, and values of 5 or greater represent high collinearity.

Results of the final model are presented in Table 4.8. Table 4.9 provides the means and standard deviations for each of the predictors in the final model, allowing for interpretation of coefficients. The coefficient for a predictor represents the predicted change in VAS loudness (on the original scale, 0-100) for a 2 SD change in that predictor. For example, results of this model indicate that a 30.19 unit increase in VAS loudness is expected for a 2 SD (7.83 sone) increase in TVL mean. The coefficient of group represents the predicted difference in perceived loudness between groups at the average of all other predictors. Table 4.10 provides delta (δ) effect sizes for each predictor.

VAS Loudness

Predictors	*Estimates*	*95% CI*	*P*
(Intercept)	63.03	60.71 – 65.35	< 0.001
TVL Mean	30.19	27.64 – 32.73	< 0.001
Mid-Ratio	4.55	1.92 – 7.18	< 0.001
SD Intensity	3.02	1.89 – 4.14	< 0.001
Group (IWPD)	-3.65	-4.98 – -2.31	< 0.001
*TVL Mean * Mid-Ratio*	-8.39	-10.97 – -5.82	< 0.001
SD: Participant Intercept	1.88		
SD: TVL Slope	4.45		
SD: Task Intercept	2.12		
SD: Residual	4.47		
N participant	102		
N task	4		
N observations	408		
AIC	2,495.57		
Conditional R^2	0.947		
Marginal R^2	0.913		

Table 4.8: Results of the maximal LMER model predicting VAS loudness.

Measure	*Mean (SD)*
TVL Mean (sones)	12.08 (3.91)
Mid-Ratio	0.12 (0.10)
SD Intensity (dB)	12.63 (2.43)

Table 4.9: Means and standard deviations for predictors included in the maximal LMER model predicting VAS loudness.

Predictor	δ
TVL Mean	4.37
Mid-Ratio	0.86
SD Intensity	0.57
Group (IWPD)	0.69
*TVL Mean * Mid-Ratio*	1.59

Table 4.10: Delta effect sizes for each predictor in the maximal LMER model predicting VAS loudness, calculated as per Westfall et al. (2014).

Results of this model indicated that TVL mean had a substantial effect on the prediction of loudness (β = 30.19, t = 23.24, p < .001, δ = 4.37). Coefficients of other predictors were more modest. Mid-ratio (β = 4.55, t= 3.40, p < .001, δ = 0.86) and SD intensity (β = 3.02, t = 5.26, p < .001, δ = 0.57) both positively predicted loudness such that a larger proportion of mid-frequency energy or greater intensity variability increased predicted loudness. IWPDs were predicted to be quieter than HOAs, even at the average of other predictors (β = -3.65, t = -5.35, p < .001, δ = 0.69). Both the marginal R^2 (0.91) and conditional R^2 (0.95) values were high, suggesting good prediction with fixed effects alone which was further supported by the random effects. An interaction plot presenting the interaction between TVL mean and mid-ratio on VAS loudness is presented in Figure 4.9. This interaction demonstrated that at lower values of mid-ratio, the association between TVL mean and VAS loudness was stronger, such that a low proportion of mid-frequency energy was particularly attenuative to perceived loudness.

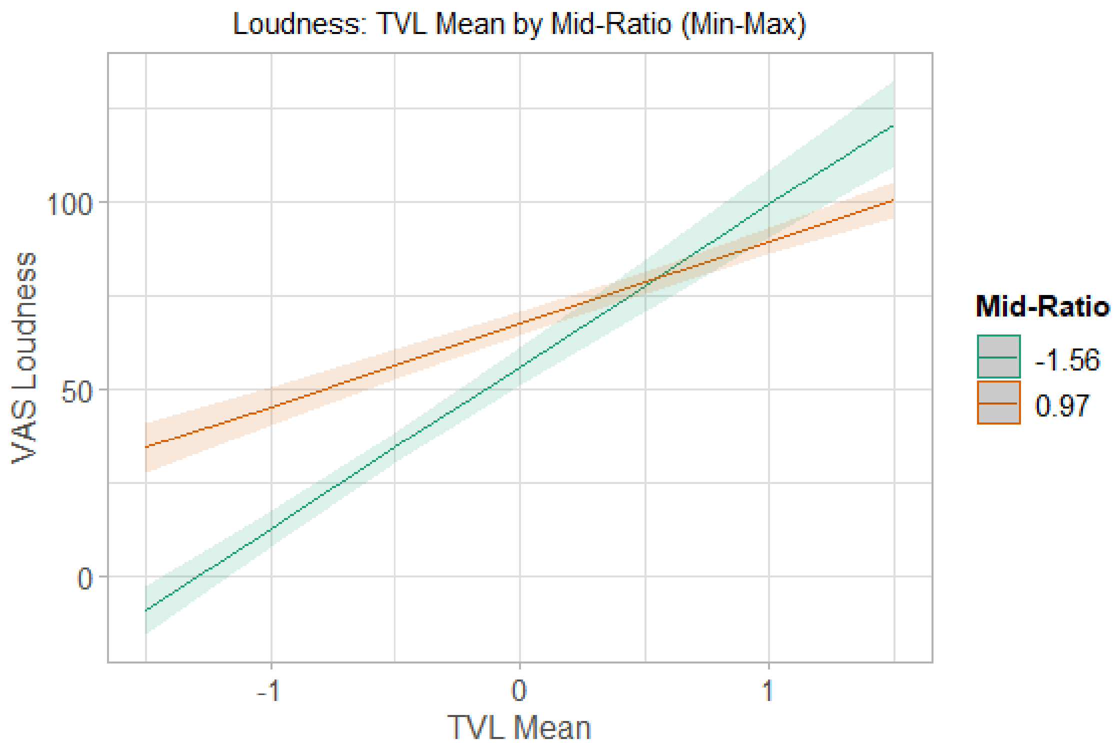


Figure 4.9: Interaction plot presenting the interaction between TVL mean and mid-ratio on VAS loudness. Moderator values for this plot are the minimum and maximum. Similar trends are observed when plotting the mean ± 1 SD. Visualizing the pattern of interaction is clearer with extreme values.

4.2.6 LMER: Intelligibility

The model building process began with the baseline model, with the formula:

$$Intelligibility \sim Group + (1\ |Participant) + (1|Task)$$

All speech level predictors significantly improved model performance, but AIC was lowest for mean intensity and active speech level. Combinations of predictors were attempted, but VIF either exceeded threshold when speech level predictors were combined or when spectral balance was incorporated. Mean intensity was selected as the speech level predictor. All spectral balance predictors significantly improved model performance except for voiced tilt. The lowest AIC was obtained via mid-ratio, followed by high-ratio, skewness, and tilt. Combinations of two spectral balance predictors were attempted, and all 2-predictor models significantly improved model performance. Combinations of three spectral balance predictors (mid-ratio, high-ratio, and skewness; mid-ratio, high-ratio, and tilt) both significantly improved model performance, with lowest AIC from the model with skewness. VIF was found to be acceptable despite the inclusion of multiple predictors from the same conceptual grouping. SD TVL was the only variability predictor that improved performance and VIF was acceptable. Group interactions improved performance for mean intensity, mid-ratio, high-ratio, and skewness, with lowest AIC from the interaction of mid-ratio and group. Among interactions between predictors, only the interaction between mid-ratio and high-ratio significantly improved model performance. Combining both mid-ratio interactions (group and high-ratio) reduced AIC, but the interaction between mid-ratio and high-ratio was no longer significant and was not maintained. Only the interaction between mid-ratio and group was maintained. The addition of a by-participant slope of high-ratio led to a singular fit. The by-participant slopes of mean intensity, mid-ratio, and skewness each improved performance, with lowest AIC obtained via mid-ratio. The combination of both slopes did not improve performance relative to mid-ratio alone. The final model predicting intelligibility was defined as:

$$Intelligibility \sim Mean\ Intensity + Skewness + High\ Ratio + SD\ TVL +$$

$$Mid\ Ratio * Group + (Mid\ Ratio\ |Participant) + (1|Task)$$

Acceptable VIF was observed in the final model, presented in Table 4.11.

Predictor	*VIF*
Mean Intensity	2.41
Mid-Ratio	3.67
Group	1.48
High-Ratio	1.42
Skewness	1.63
SD TVL	1.91
Mid-Ratio * Group	2.57

Table 4.11: Variance inflation factor values for each predictor in the final model predicting intelligibility. Values of 1 represent no collinearity between predictors, and values of 5 or greater represent high collinearity.

Results of the final model are presented in Table 4.12. Table 4.13 provides the means and standard deviations for each of the predictors in the final model, allowing for interpretation of coefficients. The coefficient for a predictor represents the predicted change in intelligibility (on the original scale, 0-100) for a 2 SD change in that predictor. For example, results of this model indicate that a 6.50 unit increase in intelligibility is expected for a 2 SD (0.24) increase in high-ratio. The coefficient of group represents the predicted difference in intelligibility between groups at the average of all other predictors. Table 4.14 provides delta effect sizes for each predictor.

Intelligibility			
Predictors	*Estimates*	*95% CI*	*P*
(Intercept)	83.25	79.37 – 87.14	< 0.001
Mean Intensity	4.27	-0.89 – 9.42	0.105
Mid-Ratio	0.92	-6.72 – 8.57	0.813
Group (IWPD)	-16.94	-21.70 – -12.17	< 0.001
High-Ratio	6.50	2.66 – 10.35	0.001
Skewness	-2.39	-7.28 – 2.51	0.341
SD TVL	-0.83	-4.44 – 2.78	0.654
*Mid-Ratio * Group*	13.71	5.66 – 21.75	0.001
SD: Participant Intercept	9.93		
SD: Mid-Ratio	9.67		
SD: Task Intercept	1.38		
SD: Residual	7.54		
N participant	102		
N task	4		
N observations	408		
AIC	3,031.58		
Conditional R^2	0.871		
Marginal R^2	0.592		

Table 4.12: Results of the maximal LMER model predicting intelligibility.

Measure	*Mean (SD)*
Mean Intensity (dB SPL)	67.86 (4.30)
Mid-Ratio	0.12 (0.10)
High-Ratio	0.04 (0.12)
Skewness	11.89 (7.31)
SD TVL (sones)	3.37 (1.14)

Table 4.13: Means and standard deviations for predictors included in the maximal LMER model predicting intelligibility.

Predictor	δ
Mean Intensity	0.34
Mid-Ratio	0.06
Group (IWPD)	1.35
High-Ratio	0.52
Skewness	0.19
SD TVL	0.07
*Mid-Ratio *Group*	1.09

Table 4.14: Delta effect sizes for each predictor in the maximal LMER model predicting intelligibility, calculated as per Westfall et al. (2014).

Results of this model indicated that most of these predictors offered limited predictive value for intelligibility. The largest coefficient was observed for group (β = -16.94, t = -6.96, p < .001, δ = 1.35), consistent with the large difference between the distributions of intelligibility in IWPDs and HOAs. High-ratio was also found to significantly positively predict intelligibility (β = 6.50, t = 3.31, p = 0.001, δ = 0.52), such that a greater proportion of high-frequency energy was associated with improved intelligibility. Main effects of mid-ratio, mean intensity, skewness and SD TVL were no longer significant in the maximal model. An interaction plot presenting the interaction between mid-ratio and group on intelligibility is presented in Figure

4.10. This interaction indicated that the positive relationship between mid-ratio and loudness was observed only in IWPDs, whereas no effect of mid-ratio on loudness was observed in HOAs. This may be the result of restricted range of intelligibility in HOAs or could reflect the relatively lower importance of mid-frequency energy for intelligibility, rather than for loudness.

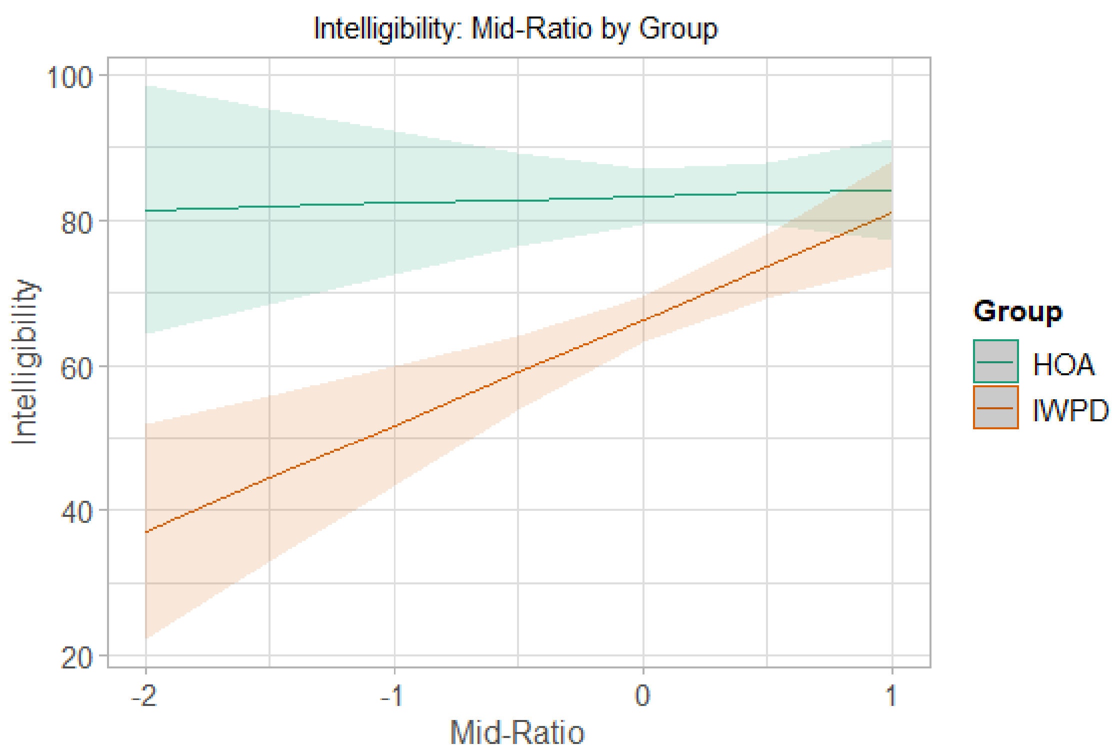


Figure 4.10: Interaction plot presenting the interaction between mid-ratio and group on intelligibility.

4.2.6.1 LMER: Intelligibility in IWPDs

Due to the restricted range of intelligibility among HOAs and the pattern of differences between intelligibility correlations in IWPDs and HOAs, a second maximal model predicting intelligibility was built within only IWPDs. The task intercept was removed, as singular fits were observed in several models while moving through the stepwise progression. It is possible that with the smaller dataset of only IWPDs, the random effect structure was too complex. The model building process began with a baseline model with the formula:

$$Intelligibility \sim (1|Participant)$$

All speech level predictors significantly improved model performance, with lowest AIC from mean intensity, LKFS and active speech level. Combinations of predictors did not significantly improve model performance, and mean intensity was selected as the speech level predictor. All spectral balance predictors significantly improved model performance. Lowest AIC was obtained via mid-ratio, followed by tilt and high-ratio. All combinations of two spectral balance predictors significantly improved model performance, with lowest AIC obtained via the mid-ratio and high-ratio combination. Mid-ratio and tilt yielded a similar AIC, but VIF was higher for mid-ratio when combined with tilt. Mid-ratio and high-ratio were selected as spectral balance predictors. None of the variability predictors or interactions between predictors significantly improved model performance. Random slopes did not significantly improve model performance. In this maximal model, mean intensity was not a significant predictor of intelligibility. Removing it did not significantly decrease model performance, and it was removed from the model. With the fixed effects structure specified, the by-task intercept was re-integrated and no longer led to a singular fit. This intercept did not improve model performance, but it was re-integrated to maximize consistency between models. The final model formula predicting intelligibility among IWPDs was defined as:

$$Intelligibility \sim MidRatio + HighRatio + (1\ |Participant) + (1|Task)$$

Acceptable VIF was observed in the final model, presented in Table 4.15.

Predictor	*VIF*
Mid-Ratio	1.76
High-Ratio	1.76

Table 4.15: Variance inflation factor values for each predictor in the final model predicting intelligibility among IWPDs. Values of 1 represent no collinearity between predictors, and values of 5 or greater represent high collinearity.

Results of the final model are presented in Table 4.16. Table 4.17 provides the means and standard deviations for each of the predictors in the final model among IWPDs, allowing for interpretation of coefficients. For example, results of this model indicate that a 8.42 unit increase in intelligibility is expected for a 2 SD (0.24) increase in high-ratio. Table 4.14 provides delta effect sizes for each predictor.

IWPD Intelligibility

Predictors	*Estimates*	*95% CI*	*P*
(Intercept)	65.60	61.37 – 69.83	< 0.001
Mid-Ratio	19.41	12.76 – 26.06	< 0.001
High-Ratio	8.42	1.57 – 15.27	0.017
SD: Participant Intercept	14.53		
SD: Task Intercept	0.51		
SD: Residual	9.37		
N participant	56		
N task	4		
N observations	224		
AIC	1,768.60		
Conditional R^2	0.820		
Marginal R^2	0.385		

Table 4.16: Results of the maximal LMER model predicting intelligibility among IWPDs.

Measure	*Mean (SD)*
Mid-Ratio	0.08 (0.11)
High-Ratio	-0.02 (0.12)

Table 4.17: Means and standard deviations among IWPDs for predictors included in the maximal LMER model predicting intelligibility among IWPDs.

Predictor	δ
Mid-Ratio	1.12
High-Ratio	0.49

Table 4.18: Delta effect sizes for each predictor in the maximal LMER model predicting intelligibility among IWPDs, calculated as per Westfall et al. (2014).

Similar to results of the model with both IWPDs and HOAs, most of these predictors offered limited predictive value for intelligibility. The large divide between conditional and marginal R^2 is likely the result of a poor fit between fixed effects and the outcome. Considerable individual variability was observed, as reflected by a large standard deviation of participant intercepts (14.53). Both mid-ratio (β = 19.41, t = 5.72, p < .001, δ = 1.12) and high-ratio (β = 8.42, t = 2.41, p = 0.017, δ = 0.49) positively predicted intelligibility, with mid-ratio demonstrating a larger effect. These effects indicate the importance of a robust spectral distribution on perceived intelligibility. The absence of a significant contribution of speech level predictors in both the IWPD-only model and the integrated model of intelligibility suggests that the contribution of audibility to intelligibility is small in the context of sufficient speech-to-noise ratio. The model among IWPDs is considerably simpler, due in part to the choice to remove mean intensity from the final model, but also likely due to the simpler and more normal underlying distribution of the outcome variable.

4.2.7 Classification

Results of the bagged tree model are presented in Table 4.19. The variable importance plot is presented in Figure 4.11. Variable importance represents the impact of the predictor in classifying IWPDs and HOAs, not the magnitude of effect of a predictor.

Metric	*Group*	*Estimate*	*p*
Accuracy		0.818	< 0.001
Sensitivity	HOA	0.791	
Specificity	HOA	0.840	
Balanced Accuracy	HOA	0.815	

Table 4.19: Accuracy metrics of the bagged (bootstrap aggregated) tree model classifying IWPD-HOA on the basis of acoustic measures.

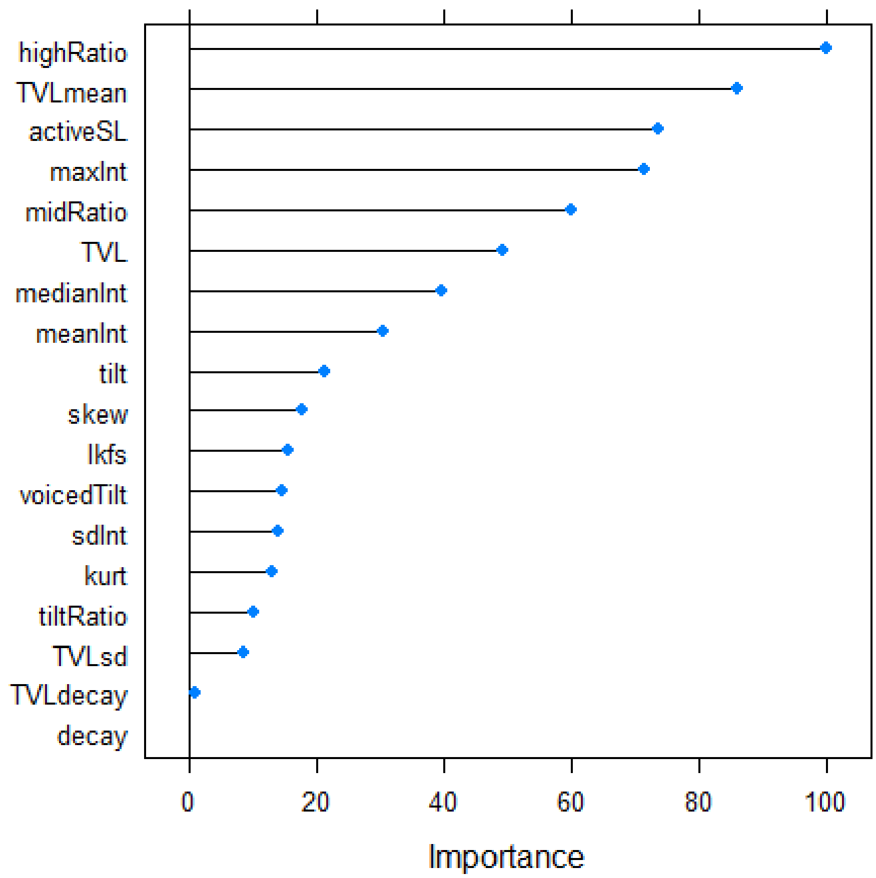


Figure 4.11: Variable importance of each acoustic measure based on the results of the bagged tree model.

Results of the bagged tree model indicated acceptable classification accuracy. Higher specificity than sensitivity indicated that the model was more successful at identifying IWPDs than HOAs. Variable importance was highest for high-ratio, TVL mean, active speech level and maximum intensity. Notably, mean intensity offered only moderate variable importance.

All 10 classification trees are presented in Appendix H. Overall, trends among the decision trees aligned well with variable importance in the bagged tree in terms of the predictors that tended to be selected for high-level nodes. High-ratio was the first node in all trees, explaining its high variable importance. Cut-off values varied from 0.034 to 0.067, with lower values (low proportion of high-frequency energy) more consistent with speech of IWPDs. A second high-ratio split was often present, dividing individuals with even larger high-ratio values (greater than approximately 0.10) as more likely to be HOAs. Maximum intensity was included in all trees as a mid-level node, with maximum intensity values greater than approximately 77 dB being consistent with speech of HOAs.

Lower level nodes varied considerably between trees. It is important to note that part of what makes decision trees unstable is a tendency to overfit, such that some splits are counter-intuitive to the known distributional tendencies between IWPDs and HOAs. This occurs because after high-impact splits such as high-ratio and maximum intensity, relatively few participants remain. Due to high between-participant variability in acoustic measures, the remaining participants' particular traits lead to overfitting. A good example of this was that in one tree, a node split maximum intensity such that values higher than 74 dB were consistent with IWPDs, identifying 5% of samples. Given the stable and consistent high-level node identifying maximum intensity values of 77 dB or above as consistent with HOAs, this node likely reflects overfit. General tendencies of low-level nodes included that high kurtosis and skewness, steep (very negative) tilt, high SD TVL, and low TVL mean were consistent with IWPD. Interesting splits that may reflect overfit included mean intensity, median intensity, and mid-ratio, which all split such that higher values were consistent with IWPDs, despite this being opposite from distributional tendencies. SD intensity often included multiple nodes, such that values greater than 10 dB but less than 15 dB were identified as HOAs. Figure 4.12 presents violin plots of each acoustic measure based on all data, rather than averaged values as per Figure 4.2.

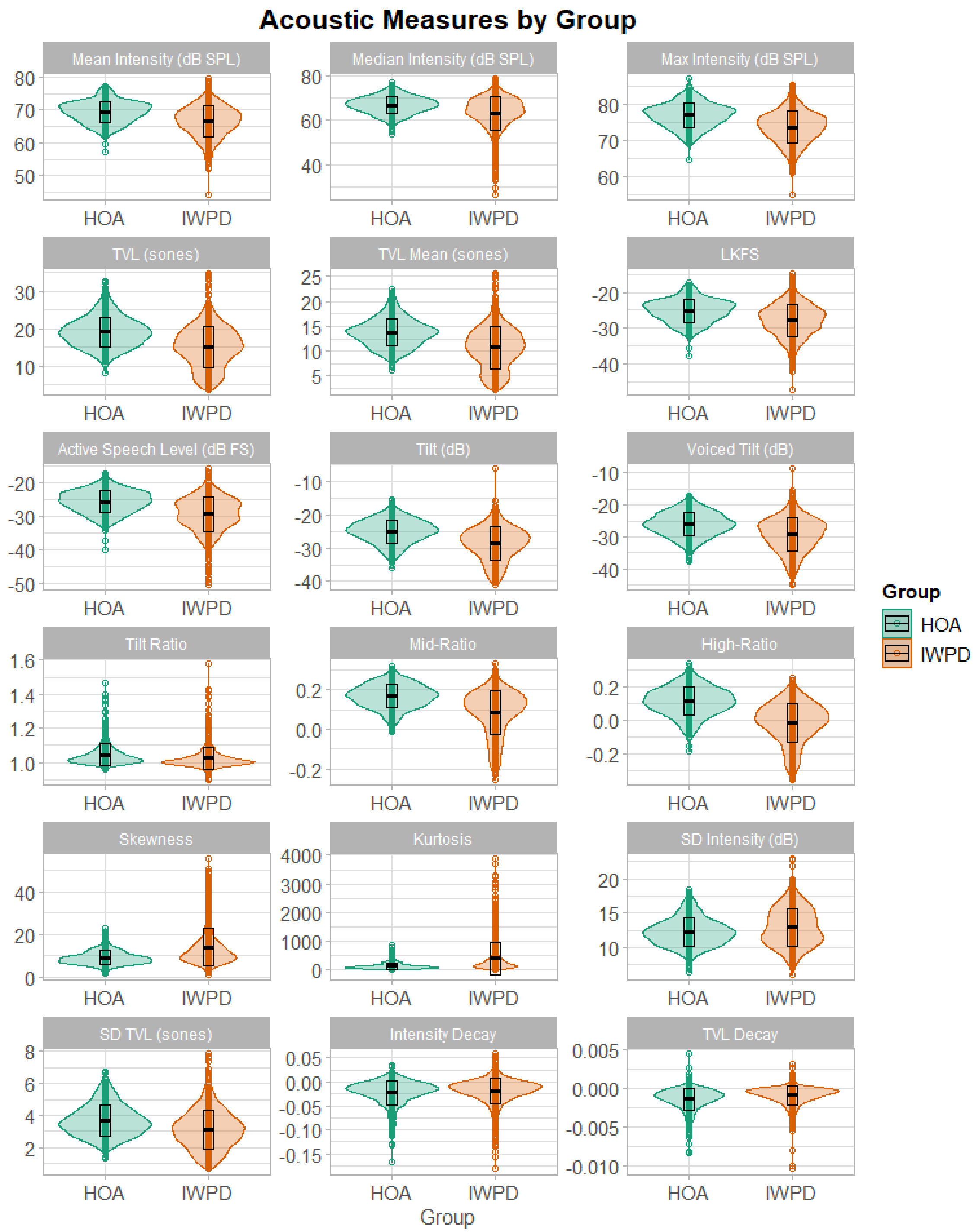


Figure 4.12: Violin plots visualizing differences in distributions of acoustic measures, without participant averaging, between IWPDs and HOAs. Crossbars within each violin plot present the mean ± 1 SD.

Figure 4.13 provides an example of one of these trees. In all trees, the left branch of a node represents the direction more consistent with HOAs than IWPDs. In this tree, high-ratio was first split such that low values were consistent with IWPDs, identifying 48% of samples as IWPDs. The next 2 nodes identified 27% of samples as HOAs based on even higher high-ratio values and higher TVL mean. The remaining 25% of samples were divided by 5 further splits. These splits identified samples with high maximum intensity as HOAs, high skewness as IWPDs, low SD TVL as HOAs, and high SD intensity as HOAs.

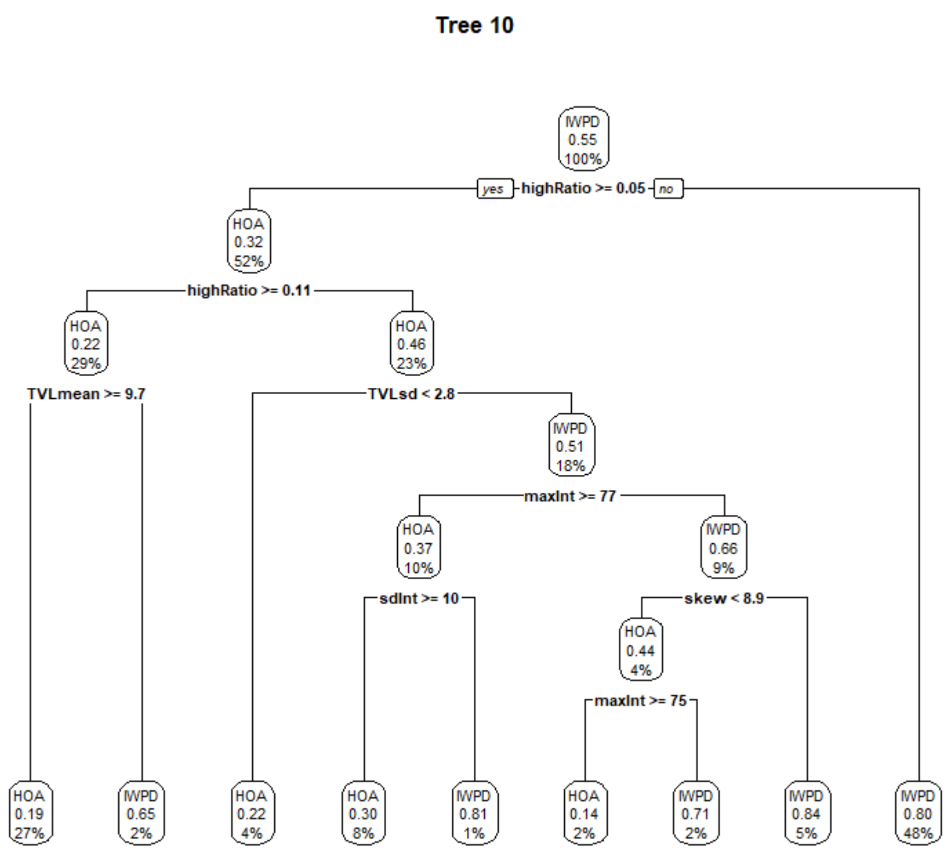


Figure 4.13: One of the ten classification decision trees (Tree 10). All trees are presented in Appendix H. Each tree is equally valid, and the choice of this tree for demonstration purposes is arbitrary.

4.2.8 Logistic Regression: IWPD-HOA

The logistic regression model predicting group (PD status) was based on values averaged within each participant. The model building process began with an intercept model. As this was not a mixed effects model, there were no random intercepts to specify.

All speech level predictors significantly improved model performance over the intercept model. The largest reduction in AIC was observed for active speech level, followed by TVL mean and maximum intensity. None of these combinations improved model performance, and active speech level was selected. All spectral balance predictors significantly improved model performance, but the reduction in AIC was particularly large for high-ratio, followed by kurtosis, tilt ratio, tilt, and skewness. Combinations of spectral balance predictors did not improve model performance. Among variability predictors, only SD TVL improved model performance. None of the interactions between predictors significantly improved model performance. The final model was defined by the formula:

$$Group \sim ActiveSpeechLevel + HighRatio + SDTVL$$

Acceptable VIF was observed in the final model, presented in Table 4.20.

Predictor	*VIF*
Active Speech Level	3.32
High-Ratio	1.10
SD TVL	3.42

Table 4.20: Variance inflation factor values for each predictor in the final logistic regression model predicting group (PD status). Values of 1 represent no collinearity between predictors, and values of 5 or greater represent high collinearity.

Results of the final model are presented in Table 4.21. Table 4.22 provides the means and standard deviations based on the values averaged within participants for each of the predictors in the final model, allowing for interpretation of coefficients. HOA was defined as the baseline group, such that the model is interpreted in terms of probability that the sample in question came from an IWPD. Interpretation of coefficients is such that for a 2 SD increase in a predictor, the log odds that the sample is an IWPD is equal to the coefficient. For example, for a 2 SD (8.72 dB) increase in active speech level, it is 3.39 times less likely that the sample came from an IWPD (due to the negative valence of the coefficient) compared to the likelihood at the mean of all predictors.

	Group (IWPD)		
Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	0.65	0.08 – 1.29	0.0339
Active Speech Level	-3.39	-6.41 – -0.77	0.0172
High-Ratio	-4.56	-6.91 – -2.63	< 0.001
SD TVL	2.69	0.44 – 5.17	0.0242
N observations	102		
AIC	94.05		
Residual Deviance	86.05		
Null Deviance	140.42		

Table 4.21: Results of the logistic regression model predicting group (PD status).

Measure	*Mean (SD)*
Active Speech Level (dB FS)	-27.66 (4.36)
High-Ratio	0.04 (0.11)
SD TVL (sones)	3.37 (0.92)

Table 4.22: Means and standard deviations based on values averaged within participant for predictors included in the logistic regression model predicting group (PD status).

Figure 4.14 presents the distributions of each predictor by group. High-ratio demonstrated the largest effect in the prediction of group (β = -4.56, z = -4.22, p < .001), indicating that a low proportion of high-frequency energy was associated with IWPDs. Similarly, low active speech level was also associated with IWPDs (β = -3.39, z = -2.38, p = 0.017). An interesting difference between the distribution and the coefficient of SD TVL was observed. The differences in means between IWPDs and HOAs on SD TVL were relatively smaller than for high-ratio and active speech level, but SD TVL was generally higher among HOAs. However, the positive coefficient of SD TVL (β = 2.69, z = 2.25, p = 0.024) indicated that in this model, higher SD TVL values were associated with IWPDs. Importantly, this was at the mean of other predictors (high-ratio and active speech level). This result suggests that at moderate high-ratio and active speech level, which may consist of IWPDs with relatively mild speech impairment, high SD TVL was predictive of PD. This may reflect the inability of IWPDs to maintain a high loudness level throughout an utterance, perhaps due to pausing or low intensity of unstressed syllables rather than utterance-level decay.

Figure 4.14: Density plots of each predictor included in the logistic model predicting group (PD status).

4.2.9 PD Subgrouping

Considerable heterogeneity existed among IWPDs. To further investigate differences between IWPDs and HOAs, IWPDS were divided into two subgroups based on their perceived loudness. IWPDs with an average VAS loudness more than 2 SD below the HOA mean were called 'Low' (N = 23), and other IWPDs were called 'Norm' (N = 33). Estimated marginal means (EMM) based on the LMER models investigating subgrouping were used to calculate pairwise *t*-tests comparing these subgroups. Figure 4.15 presents density plots displaying the creation of these subgroups. IWPDs were divided on the basis of their average loudness across all tasks, and the figure presents all data points, which explains the overlap between the 'Low' and 'Norm' distributions in Figure 4.15. Table 4.23 presents means and standard deviations of acoustic measures within these groups.

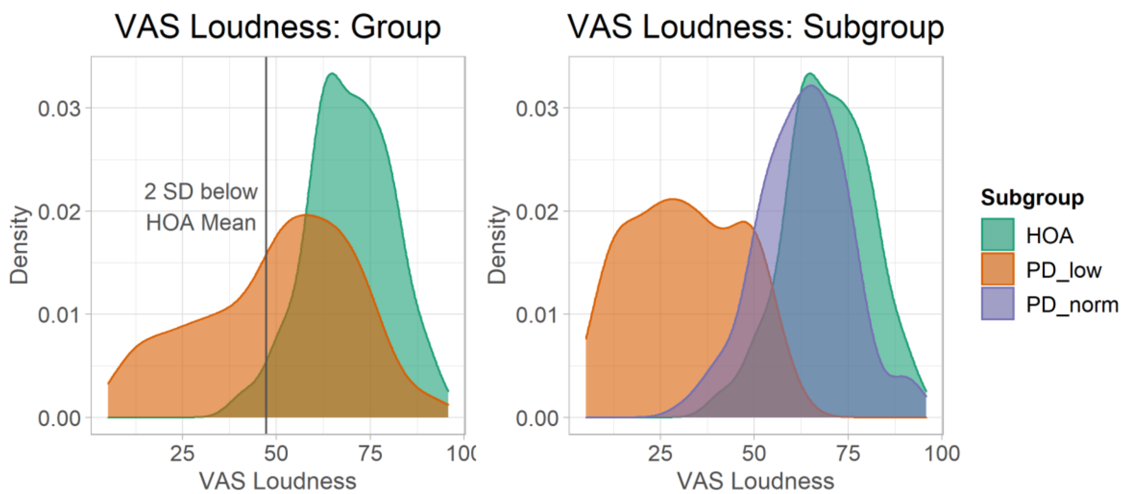


Figure 4.15: Density plots of VAS loudness within the original IWPD-HOA groups and created subgroups. A vertical line on the left plot indicates 2 SD from the HOA mean of VAS loudness, the cut-off used to create the subgroups.

Measure	*HOA*	*IWPD Low*	*IWPD Norm*
Mean Intensity (dB SPL)	68.87 (3.27)	62.16 (3.48)	68.46 (3.45)
Median Intensity (dB SPL)	66.27 (3.83)	57.45 (7.76)	65.43 (6.43)
Max Intensity (dB SPL)	76.68 (3.44)	69.66 (3.47)	75.49 (3.37)
TVL (sones)	18.81 (4.25)	9.92 (3.33)	17.81 (3.99)
TVL Mean (sones)	13.23 (2.82)	6.64 (2.52)	12.31 (3.17)
LKFS	-25.72 (3.28)	-32.20 (3.49)	-25.97 (3.34)
Active Speech Level (dB FS)	-26.27 (3.41)	-34.26 (4.46)	-27.23 (3.54)
Tilt (dB)	-25.05 (3.58)	-31.90 (4.86)	-26.98 (3.33)
Voiced Tilt (dB)	-26.25 (3.73)	-32.94 (4.65)	-27.43 (3.65)
Tilt Ratio	1.05 (0.07)	1.04 (0.08)	1.02 (0.04)
Mid-Ratio	0.16 (0.06)	-0.01 (0.10)	0.13 (0.06)
High-Ratio	0.10 (0.09)	-0.10 (0.12)	0.02 (0.08)
Skewness	9.28 (3.26)	20.32 (10.92)	11.01 (3.65)
Kurtosis	140.30 (106.04)	722.26 (762.45)	211.79 (128.38)
SD Intensity (dB)	12.34 (1.98)	13.68 (2.47)	13.26 (2.73)
SD TVL (sones)	3.52 (1.00)	2.09 (0.75)	3.60 (0.96)
*Intensity Decay **	-16.11 (21.81)	-13.11 (20.43)	-12.45 (21.15)
*TVL Decay **	-1.08 (1.26)	-0.47 (0.59)	-0.70 (1.12)

Table 4.23: Means and standard deviations of acoustic measures within each subgroup; HOA (N = 46), IWPD Low (N = 23), IWPD Norm (N = 33). * Mean difference values for intensity decay and TVL decay are expressed in scientific notation (x 10^3).

Table 4.24 presents the details of each estimated marginal means t-test. Figure 4.16 presents violin plots for each measure, allowing for comparison of distributions to better illustrate the differences between subgroups. Table 4.25 provides effect sizes for each contrast based on the estimated marginal means. Each effect size is calculated as the mean difference divided by the square root of the sum of all variances as an extension of Cohen's d, as per Westfall et al. (2014).

	HOA - Low			HOA - Norm			Low - Norm		
Measure	*EMM Diff.*	*t*	*p*	*EMM Diff.*	*t*	*p*	*EMM Diff.*	*t*	*p*
Mean Intensity (dB SPL)	6.61	9.07	< 0.001	0.32	0.49	0.877	-6.29	-8.12	< 0.001
Median Intensity (dB SPL)	8.37	8.47	< 0.001	0.33	0.38	0.925	-8.04	-7.64	< 0.001
Max Intensity (dB SPL)	6.84	9.33	< 0.001	0.93	1.43	0.33	-5.90	-7.58	< 0.001
TVL (sones)	8.70	10.28	< 0.001	0.70	0.92	0.628	-8.00	-8.89	< 0.001
TVL Mean (sones)	1.48	8.54	< 0.001	-0.07	-0.48	0.882	-1.55	-8.43	< 0.001
LKFS	6.44	8.88	< 0.001	0.22	0.34	0.937	-6.22	-8.06	< 0.001
Active Speech Level (dB FS)	7.86	10.19	< 0.001	0.87	1.27	0.417	-6.99	-8.52	< 0.001
Tilt (dB)	6.67	7.51	< 0.001	1.57	1.97	0.125	-5.11	-5.40	< 0.001
Voiced Tilt (dB)	6.57	7.21	< 0.001	0.86	1.06	0.544	-5.71	-5.89	< 0.001
Tilt Ratio	0.01	0.90	0.64	0.03	3.32	0.004	0.02	1.94	0.134
Mid-Ratio	0.17	10.63	< 0.001	0.02	1.59	0.253	-0.15	-8.65	< 0.001
High-Ratio	0.20	9.95	< 0.001	0.08	4.29	< 0.001	-0.12	-5.76	< 0.001
Skewness	-10.21	-7.52	< 0.001	-1.29	-1.06	0.538	8.92	6.18	< 0.001
Kurtosis	-547.95	-6.86	< 0.001	-58.52	-0.82	0.692	489.44	5.76	< 0.001
SD Intensity (dB)	-0.99	-2.63	0.027	-0.54	-1.61	0.248	0.45	1.12	0.503
SD TVL (sones)	1.48	8.54	< 0.001	-0.07	-0.48	0.882	-1.55	-8.43	< 0.001
*Intensity Decay **	-4.07	-1.62	0.241	-2.94	-1.31	0.392	1.13	0.42	0.906
*TVL Decay **	-0.73	-4.88	< 0.001	-0.28	-2.09	0.097	0.45	2.84	0.015

Table 4.24: Results of pairwise t-tests based on estimated marginal means (EMM) calculated from the subgrouping LMER models. * Mean difference values for intensity decay and TVL decay are expressed in scientific notation (x 10^3). HOA (N = 46), PD Low (N = 23), PD Norm (N = 33).

Measure	*HOA - Low*	*HOA - Norm*	*Low – Norm*
Mean Intensity (dB SPL)	1.94	0.09	1.85
Median Intensity (dB SPL)	1.51	0.06	1.45
Max Intensity (dB SPL)	2.02	0.28	1.75
TVL (sones)	2.17	0.17	2.00
TVL Mean (sones)	1.54	-0.08	1.61
LKFS	1.92	0.07	1.85
Active Speech Level (dB FS)	2.16	0.24	1.92
Tilt (dB)	1.67	0.39	1.28
Voiced Tilt (dB)	1.60	0.21	1.39
Tilt Ratio	0.15	0.50	0.34
Mid-Ratio	2.31	0.31	2.00
High-Ratio	2.10	0.81	1.29
Skewness	1.66	0.21	1.45
Kurtosis	1.46	0.16	1.31
SD Intensity (dB)	0.41	0.22	0.19
SD TVL (sones)	1.54	0.08	1.61
Intensity Decay	0.15	0.11	0.04
TVL Decay	0.56	0.22	0.35

Table 4.25: Effect sizes based on estimated marginal means (EMM) are calculated as the mean difference divided by the square root of the sum of all variances, as an extension to Cohen's *d* as per Westfall et al. (2014).

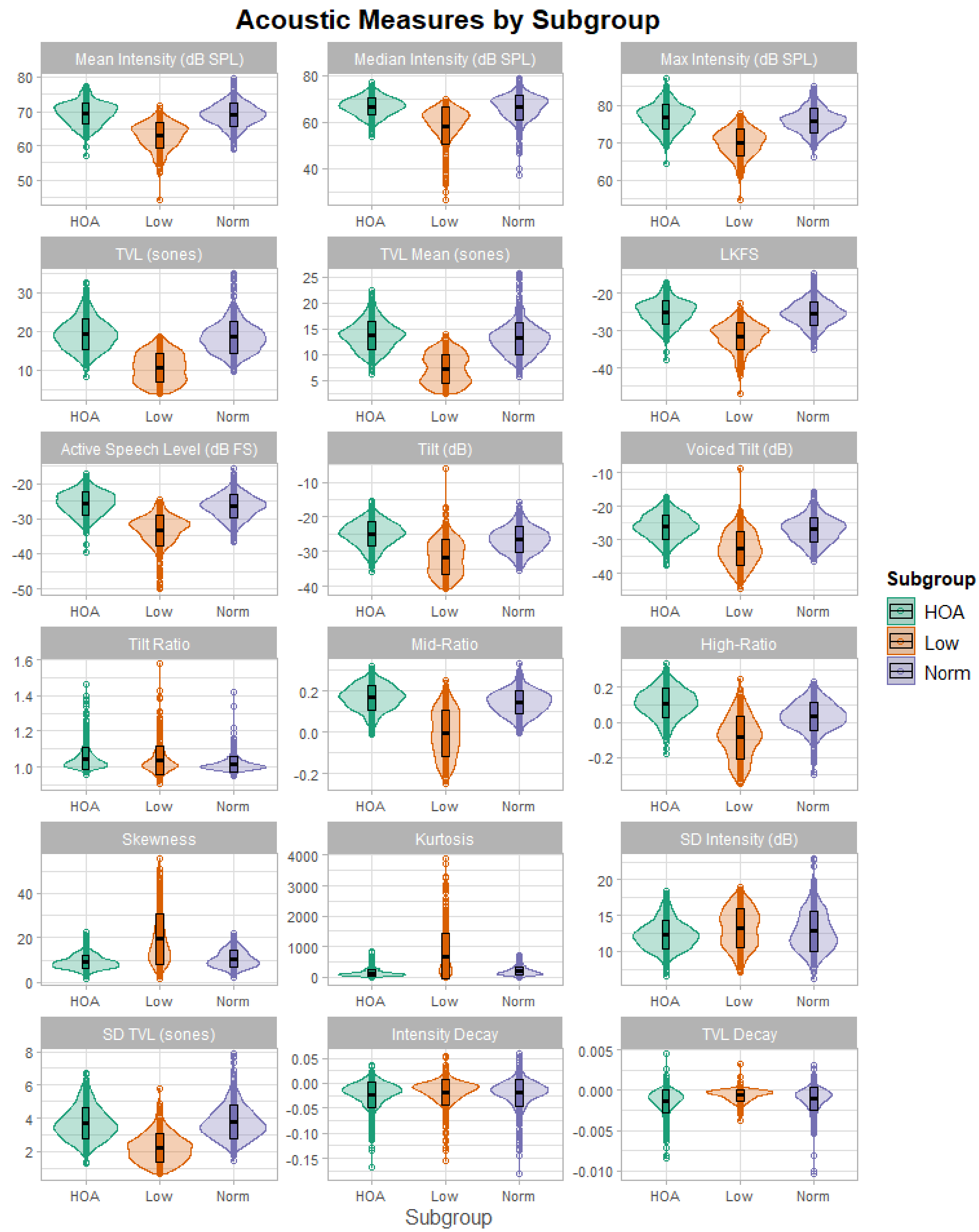


Figure 4.16: Violin plots visualizing differences in distributions of acoustic measures, without participant averaging, between HOAs (N = 46), Low IWPDs (N = 23) and Norm IWPDs (N = 33). IWPDs were split on a cut-off of 2 SDs from the HOA mean of VAS loudness. Crossbars within each violin plot present the mean ± 1 SD.

Across measures, it was notable that the distributions of the IWPD Norm subgroup were more similar to HOAs. These groups did not significantly differ on most measures, with the exceptions of high-ratio and tilt ratio. While significant, the difference between HOA ($M = 1.05$, $SD = 0.07$) and IWPD Norm ($M = 1.04$, $SD = 0.08$) on tilt ratio was small ($t(99) = 3.32$, $p = 0.004$, $\delta = 0.50$), and the distributions were similar across all subgroups. Conversely, high-ratio showed clear contrasts between the HOA ($M = 0.10$, $SD = 0.09$), IWPD Low ($M = -0.10$, $SD = 0.12$), and IWPD Norm ($M = 0.02$, $SD = 0.08$) subgroups (HOA - IWPD Norm: $t(99) = 4.29$, $p < .001$, $\delta = 0.81$; HOA - IWPD Low: $t(99) = 9.95$, $p < .001$, $\delta = 2.10$). This indicates that high-ratio is sensitive to differences between IWPDs and HOAs, despite similarity between the IWPD Norm subgroup and HOA group in speech level, variability, and most spectral balance measures.

Measures with particularly pronounced differences between the IWPD Low subgroup and both the IWPD Norm subgroup and HOA group included TVL, TVL mean, tilt, mid-ratio, high-ratio, skewness, kurtosis, and SD TVL. Specifically, individuals in the IWPD Low subgroup demonstrated lower TVL and TVL mean, consistent with overall low loudness. This reflects the consistency between VAS loudness and TVL measures. High skewness and kurtosis and low tilt, mid-ratio and high-ratio all reflect a relatively weaker contribution of mid- and high-frequency energy among individuals in the IWPD Low subgroup. Patterns of variability measures between subgroups were less clear with the exception of SD TVL. SD TVL among the IWPD Low subgroup was much lower than HOAs and IWPD Norm. This may be consistent with monoloudness among IWPDs with low loudness.

4.3 Experiment 2: Spectral Manipulation

4.3.1 ANOVA

Within the spectral manipulation experiment, the primary focus was on the change in loudness and speech level measures following a positive or negative shift in mid- to high-frequency energy. Table 4.26 presents the means and standard deviations for perceived loudness and acoustic measures within the spectral manipulation experiment. Separate means and standard deviations for HOAs and IWPDs in this experiment are presented in Appendix J.

Measure	*Down 10 dB*	*Down 5 dB*	*Unaltered Speech*	*Up 5 dB*	*Up 10 dB*
VAS Loudness (%)	52.37 (16.18)	56.61 (16.86)	61.96 (16.91)	66.54 (17.08)	72.36 (17.51)
Mean Intensity (dB SPL)	68.97 (4.10)	68.97 (4.10)	68.97 (4.10)	68.97 (4.10)	68.97 (4.10)
Median Intensity (dB SPL)	65.96 (5.68)	65.98 (5.62)	66.06 (5.44)	66.23 (5.20)	66.50 (4.93)
Max Intensity (dB SPL)	75.30 (3.86)	75.30 (3.85)	75.30 (3.86)	75.29 (3.89)	75.43 (3.98)
TVL (sones)	15.07 (4.57)	16.11 (4.89)	17.53 (5.30)	19.36 (5.77)	21.49 (6.23)
TVL Mean (sones)	10.88 (3.33)	11.69 (3.57)	12.78 (3.87)	14.18 (4.22)	15.80 (4.58)
LKFS	-25.58 (4.15)	-25.56 (4.15)	-25.51 (4.16)	-25.35 (4.16)	-25.00 (4.19)
Active Speech Level (dB FS)	-26.63 (4.46)	-26.63 (4.46)	-26.65 (4.46)	-26.67 (4.46)	-26.68 (4.45)
Tilt (dB)	-30.91 (5.94)	-29.00 (5.46)	-26.18 (5.06)	-22.54 (4.84)	-18.30 (4.76)
Voiced Tilt (dB)	-30.34 (5.34)	-28.99 (5.26)	-26.85 (5.22)	-23.71 (5.13)	-20.03 (5.20)
Tilt Ratio	0.98 (0.04)	1.00 (0.04)	1.03 (0.05)	1.05 (0.08)	1.10 (0.12)
Mid-Ratio	0.00 (0.10)	0.08 (0.10)	0.15 (0.09)	0.22 (0.09)	0.28 (0.08)
High-Ratio	-0.09 (0.12)	-0.02 (0.12)	0.05 (0.11)	0.12 (0.11)	0.19 (0.10)
Skewness	10.73 (9.53)	12.09 (8.17)	10.86 (6.21)	7.97 (4.48)	5.10 (2.96)
Kurtosis	435.14 (950.07)	373.61 (588.85)	236.98 (308.97)	113.20 (142.63)	43.80 (56.46)
SD TVL (sones)	3.56 (1.13)	3.79 (1.20)	4.10 (1.29)	4.49 (1.41)	4.95 (1.53)
SD Intensity (dB)	12.89 (2.07)	12.51 (2.07)	12.04 (2.08)	11.49 (2.10)	10.90 (2.10)
*Intensity Decay **	-44.79 (50.67)	-43.41 (48.38)	-41.73 (45.62)	-39.77 (42.54)	-37.59 (39.40)
*TVL Decay **	-1.55 (1.88)	-1.64 (1.95)	-1.74 (2.06)	-1.85 (2.22)	-1.97 (2.42)

Table 4.26: Means and standard deviations for perceived loudness and acoustic measures within the spectral manipulation experiment. 'nat' is unaltered speech. 'Up 5 dB' and 'Up 10 dB' refer to positive 5 dB and 10 dB manipulations. 'Down 5 dB' and 'Down 10 dB' refer to negative 5 dB and 10 dB manipulations. * Mean difference values for intensity decay and TVL decay are expressed in scientific notation (x 10^3).

Table 4.27 presents the results of the ANOVA evaluating perceived loudness by group and manipulation. No significant interaction was observed between PD status and manipulation condition, consistent with a similar effect of the spectral manipulation on perceived loudness in speech of IWPDs and HOAs. ANOVA results for acoustic measures are presented in Appendix J. Table 4.28 presents the effect sizes for group, manipulation, and their interaction for perceived loudness and acoustic measures. A large effect of manipulation was observed for perceived loudness ($F(1,4) = 28.77$, $p < .001$, $\eta^2 = 0.23$), indicating that spectral manipulation was successful in altering loudness despite equal mean intensity. The only speech level measures with large effects of manipulation were TVL ($F(1,4) = 25.63$, $p < .001$, $\eta^2 = 0.10$) and TVL mean ($\eta^2 = 0.05$), with TVL showing a comparable effect size to perceived loudness. Spectral balance measures showed large effect sizes, reflecting the manipulation itself. Among spectral balance measures, effect sizes are highest for mid-ratio ($F(1,4) = 173.57$, $p < .001$, $\eta^2 = 0.58$), high-ratio ($F(1,4) = 137.39$, $p < .001$, $\eta^2 = 0.52$), and tilt ($F(1,4) = 107.96$, $p < .001$, $\eta^2 = 0.12$). Medium-to-large effects of intensity variability were observed for SD TVL ($F(1,4) = 18.82$, $p < .001$, $\eta^2 = 0.05$), with no effects of SD intensity ($F(1,4) = 14.63$, $p < .001$, $\eta^2 = 0.00$), intensity decay ($F(1,4) = 0.40$, $p = 0.807$, $\eta^2 = 0.00$) or TVL decay ($F(1,4) = 0.64$, $p = 0.636$, $\eta^2 = 0.01$).

	df	*Sum of Squares*	*Mean Squares*	*F*	*p*
Manipulation	4	25,465.29	6,366.32	28.77	< 0.001
Group	1	33,799.71	33,799.71	152.75	< 0.001
*Manipulation * Group*	4	27.02	6.75	0.03	0.998
Residuals	500	110,636.79	221.27		

Table 4.27: Results of the ANOVA evaluating the effect of the spectral manipulation on loudness of IWPDs and HOAs.

	Partial η²		
Measure	*Group*	*Manipulation*	*Manipulation * Group*
VAS Loudness (%)	0.23	0.19	0.00
Mean Intensity (dB SPL)	0.08	0.00	0.00
Median Intensity (dB SPL)	0.03	0.00	0.00
Max Intensity (dB SPL)	0.15	0.00	0.00
TVL (sones)	0.10	0.17	0.00
TVL Mean (sones)	0.05	0.13	0.00
LKFS	0.09	0.00	0.00
Active Speech Level (dB FS)	0.15	0.00	0.00
Tilt (dB)	0.12	0.46	0.00
Voiced Tilt (dB)	0.09	0.36	0.00
Tilt Ratio	0.04	0.26	0.01
Mid-Ratio	0.14	0.58	0.00
High-Ratio	0.29	0.52	0.00
Skewness	0.05	0.13	0.00
Kurtosis	0.03	0.08	0.02
SD Intensity (dB)	0.00	0.10	0.00
SD TVL (sones)	0.05	0.13	0.00
Intensity Decay	0.00	0.00	0.00
TVL Decay	0.04	0.01	0.00

Table 4.28: Eta-squared effect sizes for the effect of manipulation. 'nat' is unaltered speech. 'up5' and 'up10' refer to positive 5 dB and 10 dB manipulations. 'down5' and 'down10' refer to negative 5 dB and 10 dB manipulations.

Table 4.29 presents the post-hoc pairwise Tukey's Honestly Significant Difference (HSD) tests of the ANOVAs comparing spectral manipulation conditions for each measure. Only the comparisons between the unaltered ('Nat') speech and each manipulation are presented as these differences were of primary interest, but all conditions were included in the post-hoc tests and correction for multiple comparisons. Full post-hoc results for each measure are presented in Appendix J.

For the positive 10 dB manipulation, the post-hoc comparison (HSD test) indicated that perceived loudness in the 'Up 10dB' condition (M = 72.36, SD = 17.51) was significantly higher than in the unaltered, 'Nat' condition (M = 61.96, SD = 16.91; HSD = 10.40, p < .001). With regard to the negative 10dB manipulation, the post-hoc comparison indicated that perceived loudness in the 'Down 10dB' condition (M = 52.37, SD = 16.18) was significantly lower than in the unaltered, 'Nat' condition (M = 61.96, SD = 16.91; HSD = 10.40, p < .001).

As expected given the equalization to mean intensity during the process of the manipulation, mean intensity was equal across conditions. Small differences were observed in the means of median intensity, maximum intensity, LKFS and active speech level, as the frequency adjustment causes minor alterations to the intensity contour. However, these differences were very small and not significant.

Conversely, TVL and TVL mean showed significant differences between unaltered speech and both positive and negative 10 dB manipulations. With the positive 10 dB manipulation, results of the post-hoc comparison (HSD test) indicated that TVL in the 'Up10 dB' condition (M = 21.49, SD = 6.23) was significantly higher than in the unaltered, ‘Nat’ condition (M = 17.53, SD = 5.30; HSD = 3.96, p < 3.963). Similarly, TVL mean in the ‘Up10 dB’ condition (M = 15.80, SD = 4.58) was significantly higher than in the unaltered, ‘Nat’ condition (M = 12.78, SD = 3.87; HSD = 3.02, p < 3.021). In the negative 10 dB manipulation, results of the post-hoc comparison indicated that TVL in the ‘Down10 dB’ condition (M = 15.07, SD = 4.57) was significantly higher than in the unaltered, ‘Nat’ condition (M = 17.53, SD = 5.30; HSD = , p <). TVL mean in the ‘Down10 dB’ condition (M = 10.88, SD = 3.33) was also significantly higher than in the unaltered, ‘Nat’ condition (M = 12.78, SD = 3.87; HSD = 1.90, p < 1.905). This pattern reflects the ability of TVL to measure a difference in loudness that is perceived by listeners even in the presence of equal mean intensity. Figure 4.17 presents bar plots of perceived loudness and speech level measures in each manipulation condition to further demonstrate these trends. It is visually apparent that the trends across manipulation conditions were similar for loudness and TVL measures, whereas other speech level measures were essentially stable across conditions.

	Nat - Down 10 dB		Nat - Down 5 dB		Nat - Up 5 dB		Nat - Up 10 dB	
Measure	*Mean Diff.*	*p*	*Mean Diff.*	*p*	*Mean Diff.*	*p*	*Mean Diff.*	*p*
VAS Loudness (%)	9.59	< 0.001	5.35	0.078	4.58	0.183	10.40	< 0.001
Mean Intensity (dB SPL)	0.00	1.000	0.00	1.000	0.00	1.000	0.00	1.000
Median Intensity (dB SPL)	0.10	1.000	0.08	1.000	0.17	1.000	0.44	0.977
Max Intensity (dB SPL)	0.00	1.000	0.00	1.000	-0.01	1.000	0.14	0.999
TVL (sones)	2.46	0.006	1.42	0.278	1.83	0.081	3.96	< 0.001
TVL Mean (sones)	0.53	0.028	0.31	0.437	0.39	0.187	0.85	< 0.001
LKFS	0.08	1.000	0.06	1.000	0.16	0.999	0.51	0.894
Active Speech Level (dB FS)	-0.02	1.000	-0.02	1.000	-0.02	1.000	-0.03	1.000
Tilt (dB)	4.73	< 0.001	2.82	< 0.001	3.64	< 0.001	7.88	< 0.001
Voiced Tilt (dB)	3.49	< 0.001	2.14	0.021	3.15	< 0.001	6.83	< 0.001
Tilt Ratio	0.04	< 0.001	0.03	0.083	0.03	0.047	0.08	< 0.001
Mid-Ratio	0.14	< 0.001	0.07	< 0.001	0.07	< 0.001	0.14	< 0.001
High-Ratio	0.14	< 0.001	0.07	< 0.001	0.07	< 0.001	0.14	< 0.001
Skewness	0.13	1.000	-1.23	0.667	-2.89	0.015	-5.76	< 0.001
Kurtosis	-198.16	0.047	-136.64	0.317	-123.78	0.420	-193.17	0.057
SD Intensity (dB)	-0.85	0.033	-0.47	0.498	-0.55	0.333	-1.14	0.001
SD TVL (sones)	0.53	0.028	0.31	0.437	0.39	0.187	0.85	< 0.001
*Intensity Decay **	3.06	0.989	1.69	0.999	1.95	0.998	4.14	0.967
*TVL Decay **	-0.18	0.972	-0.10	0.997	-0.11	0.995	-0.23	0.933

Table 4.29: Results of Tukey's HSD post-hoc tests on the ANOVAs investigating the effect of spectral manipulation. 'nat' is unaltered speech. 'up5' and 'up10' refer to positive 5 dB and 10 dB manipulations. 'down5' and 'down10' refer to negative 5 dB and 10 dB manipulations. * Mean difference values for intensity decay and TVL decay are expressed in scientific notation (x 10^3).

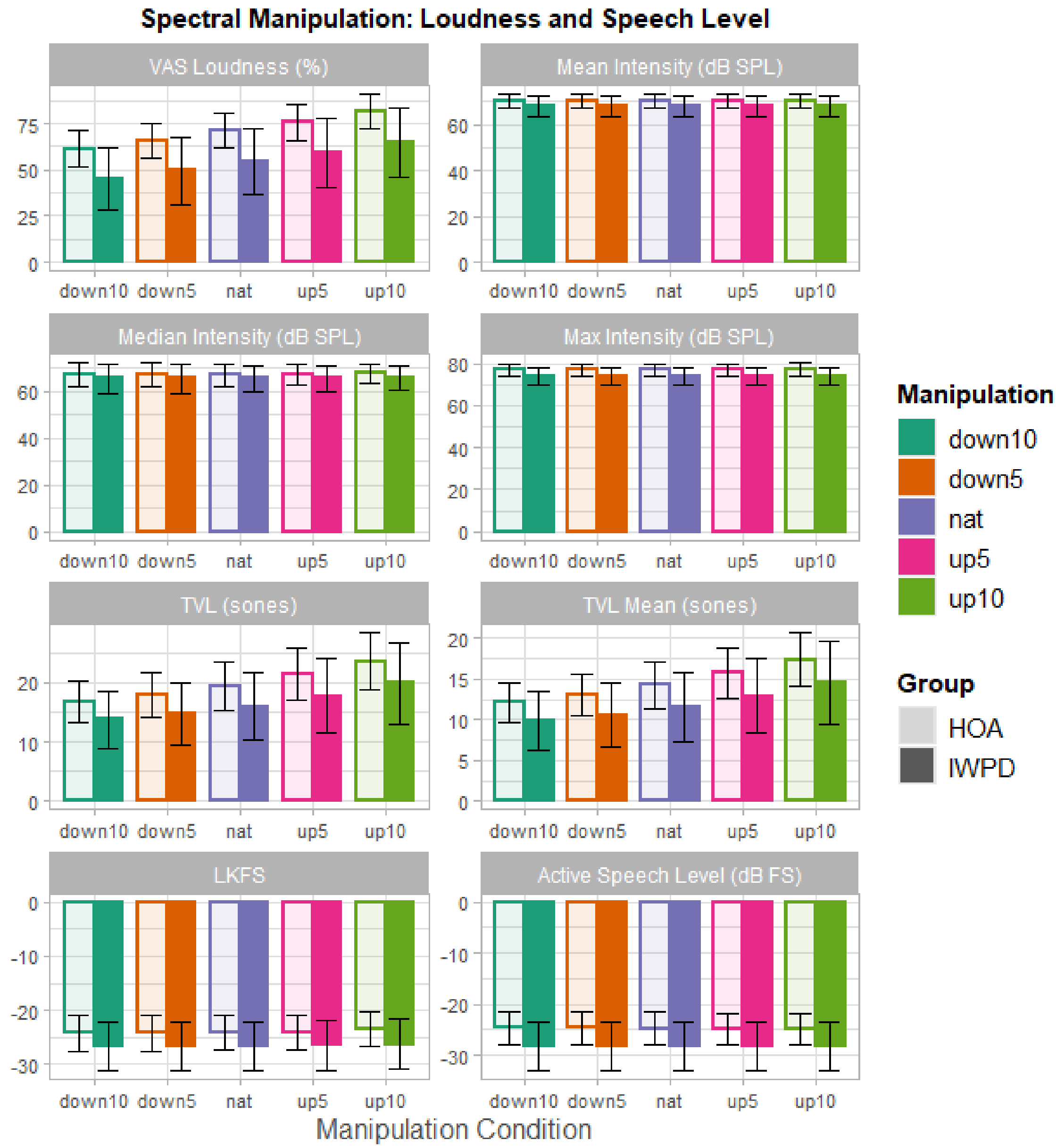


Figure 4.17: Bar plots visualizing the means of perceived loudness and speech level measures in the spectral manipulation experiment for IWPDs and HOAs. Error bars represent ± 1 SD. 'nat' is unaltered speech. 'up5' and 'up10' refer to positive 5 dB and 10 dB manipulations. 'down5' and 'down10' refer to negative 5 dB and 10 dB manipulations.

4.3.2 Loudness Amplification

The post-hoc comparison evaluating the change in perceived loudness in unaltered speech and in the positive 10 dB manipulation condition was of particular interest to this investigation. The extent to which loudness increases following a positive spectral manipulation has important implications for effective amplification of speech, and this relationship was explored in greater detail. Loudness amplification was defined as the difference score of perceived loudness between these conditions ('Up10'-'Nat'). In general, the observed loudness amplification provides evidence that spectral manipulation consistently increases speech loudness. Amplification was normally distributed and was very similar between IWPDs and HOAs, as visible in Figure 4.18.

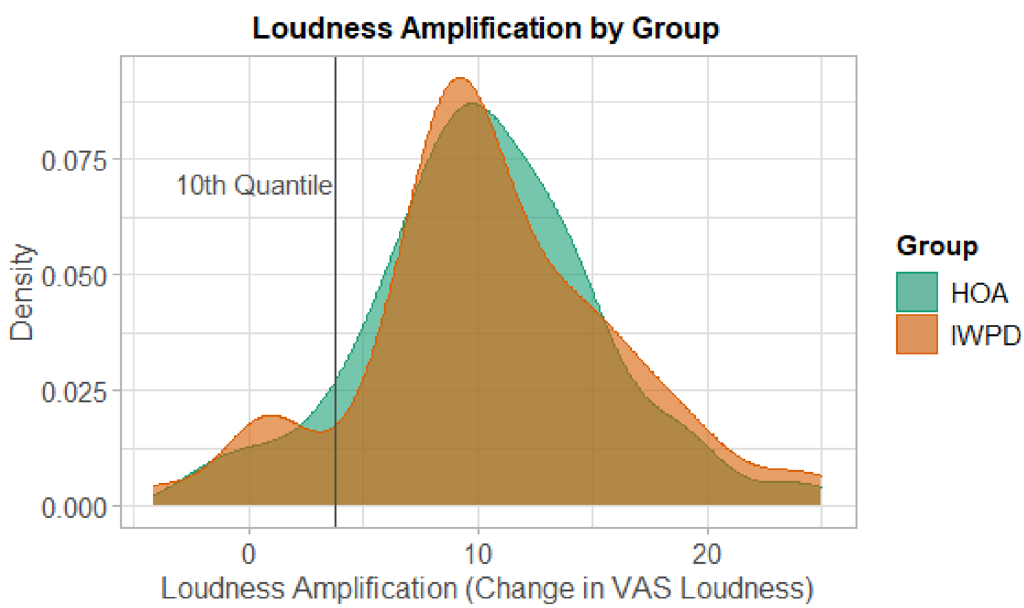


Figure 4.18: Density plots of the amplification in loudness that occurs between unaltered speech and the positive 10 dB spectral manipulation in IWPDs and HOAs. A vertical line identifies the 10th quantile, individuals below which are identified as Low Amplification.

Further investigation of the associations between loudness amplification and acoustic characteristics of speech was of interest to explain why some individuals demonstrated greater loudness amplification than others. Pearson correlations of loudness amplification (perceived loudness difference score; 'Up10' -'Nat') with the acoustic measures in unaltered speech were obtained. Because the spectral manipulation fundamentally alters the spectrum and thus the acoustic measures themselves, acoustic difference scores were not of interest to this investigation. Baseline characteristics of each individual's speech were expected to be associated with observed loudness amplification. For example, it was hypothesized that features like steep tilt might be associated with poorer loudness amplification. However, correlations were weak, and many were not significant. Figure 4.19 provides scatter plots and correlations of loudness amplification and acoustic measures, demonstrating the disparate relationships. Pearson correlations and *p*-values are presented in Table 4.30.

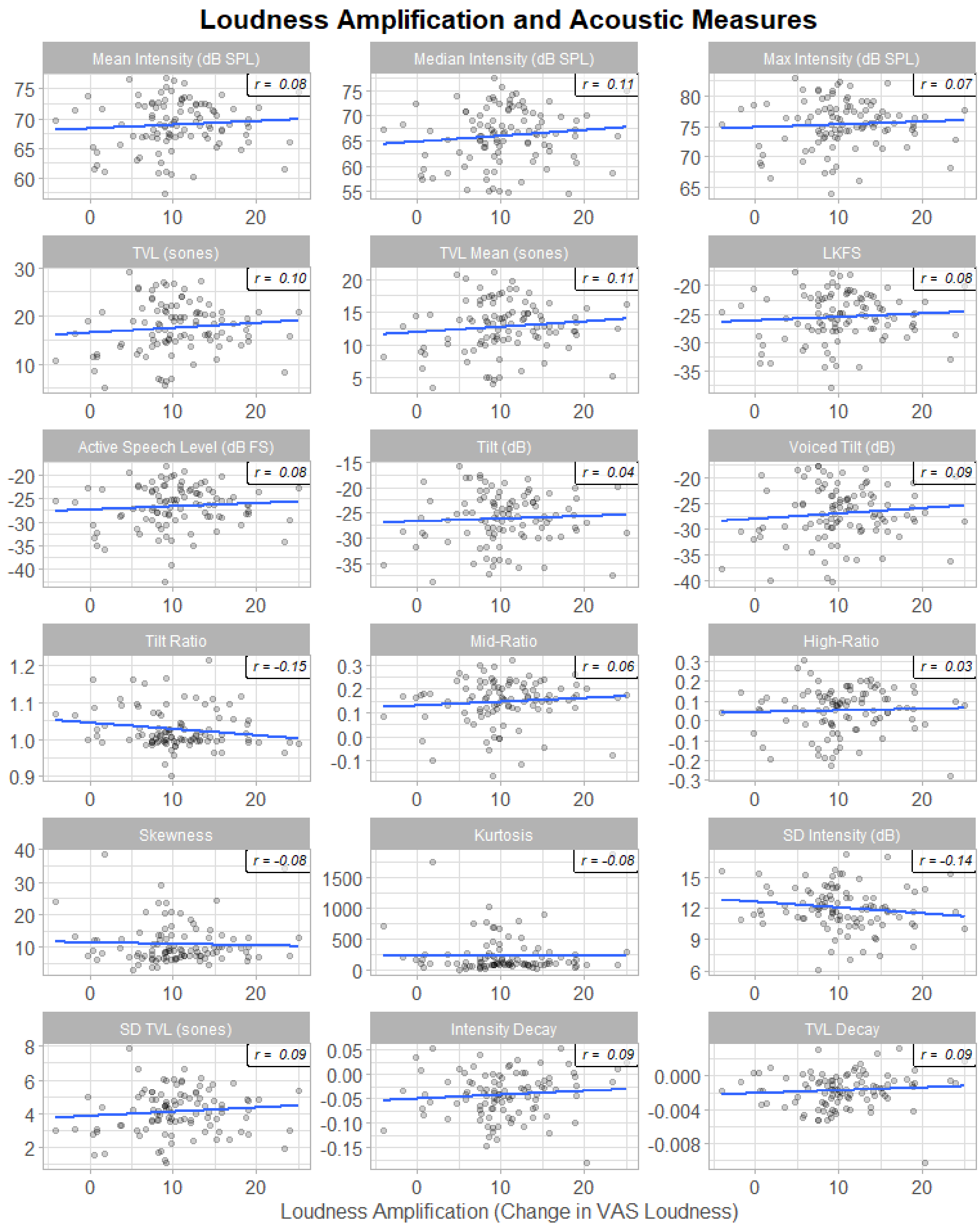


Figure 4.19: Scatter plots presenting the relationships between acoustic measures and the loudness amplification that occurs between unaltered speech and the positive 10 dB spectral manipulation. Trendline presented is the linear regression of each measure predicting loudness amplification. Pearson correlations and p-values are presented within each figure.

Measure	*r*	*N*	*p*
Mean Intensity (dB SPL)	0.08	510	0.068
Median Intensity (dB SPL)	0.11	510	0.010
Max Intensity (dB SPL)	0.07	510	0.127
TVL (sones)	0.10	510	0.031
TVL Mean (sones)	0.11	510	0.017
LKFS	0.08	510	0.072
Active Speech Level (dB FS)	0.08	510	0.058
Tilt (dB)	0.04	510	0.348
Voiced Tilt (dB)	0.09	510	0.050
Tilt Ratio	-0.15	510	< 0.001
Mid-Ratio	0.06	510	0.163
High-Ratio	0.03	510	0.485
Skewness	-0.08	510	0.059
Kurtosis	-0.08	510	0.058
SD Intensity (dB)	-0.14	510	0.002
SD TVL (sones)	0.09	510	0.033
Intensity Decay	0.09	510	0.040
TVL Decay	0.09	510	0.054

Table 4.30: Pearson correlations evaluating the association between acoustic measures and loudness amplification, the difference score between perceived loudness in the positive 10 dB spectral manipulation condition and perceived loudness in unaltered speech.

The weak correlations between amplification and acoustic measures may be a positive indicator for the use of spectral manipulation in achieving effective amplification for speech, as most individuals appear to benefit to some degree from a high-frequency boost. Strong associations between loudness amplification and either speech level or spectral balance measures could mean that spectral manipulation would only be useful for particular speakers. These findings suggest that spectral manipulation is likely to confer benefit for most speakers. This implication will be further explored in Chapter 5.

4.3.2.1 Low Amplification Individuals

Despite weak associations between acoustic measures and loudness amplification overall, it was of interest whether qualitative differences existed among individuals demonstrating poor loudness amplification. A small peak was observed at extremely low values of loudness amplification, particularly among IWPDs. To investigate this group, samples below the 10th quantile were separated into a subgroup called 'Low Amplification.' Statistical examination of this subgroup was not possible because of the small group size, with only 5 HOAs and 6 IWPDs. However, trends were observed descriptively among these individuals. Table 4.32 presents means and standard deviations for each of these groups. The magnitude of the loudness amplification differed considerably in these groups, presented in Table 4.31.

Group	*Amplification Subgroup*	*Loudness Amplification Mean (SD)*
HOA	*Low Amplification*	1.33 (2.16)
HOA	*Normal Amplification*	11.35 (4.06)
IWPD	*Low Amplification*	0.01 (1.93)
IWPD	*Normal Amplification*	11.77 (4.49)

Table 4.31: Means and standard deviation deviations of loudness amplification within the Low Amplification and Normal Amplification groups. Loudness amplification is the difference score between perceived loudness in the positive 10 dB spectral manipulation condition and perceived loudness in unaltered speech.

	Low Amplification		Normal Amplification	
Measure	*HOA*	*IWPD*	*HOA*	*IWPD*
VAS Loudness (%)	69.11 (11.83)	44.26 (14.36)	71.54 (9.54)	55.52 (17.89)
Mean Intensity (dB SPL)	70.30 (3.12)	63.97 (3.20)	70.27 (3.24)	68.37 (4.38)
TVL (sones)	16.75 (3.02)	9.85 (2.61)	19.76 (4.09)	16.70 (5.53)
TVL Mean (sones)	12.60 (2.06)	7.53 (2.23)	14.44 (2.85)	12.07 (4.15)
Tilt (dB)	-28.09 (3.67)	-29.53 (6.97)	-23.74 (4.43)	-27.59 (4.67)
Mid-Ratio	0.16 (0.02)	0.07 (0.11)	0.19 (0.06)	0.12 (0.10)
High-Ratio	0.06 (0.09)	-0.01 (0.12)	0.13 (0.07)	0.00 (0.11)
Skewness	10.07 (2.63)	16.97 (12.16)	8.68 (3.56)	12.00 (6.59)
Kurtosis	171.25 (63.54)	505.34 (660.03)	135.46 (147.74)	294.59 (336.01)

Table 4.32: Means and standard deviation deviations in the unaltered condition of the manipulation experiment for perceived loudness and a selected number of acoustic measures.

In general, low amplification IWPDs were quieter based on TVL, perceived loudness and intensity. Tilt was steeper and kurtosis and skewness were much higher among low amplification IWPDs than among low amplification HOAs, normal amplification IWPDs, and normal amplification HOAs. This may indicate that quiet IWPDs with very steep tilt would have needed a greater gain shift than 10 dB in order to increase their loudness.

The profile of low amplification HOAs was less clear, which may suggest that low amplification HOAs are merely part of the normal distribution's left tail, whereas low amplification IWPDs represent a very small subgroup with more pronounced features. This is consistent with the appearance of a peak among IWPDs, but no peak among HOAs.

Overall, the results of Experiment 2 provide evidence that spectral manipulation effectively increases loudness for both IWPDs and HOAs, with more pronounced changes in loudness with 10 dB shifts relative to 5 dB shifts. The possibility of a low amplification subtype, particularly among IWPDs, warrants further investigation to identify the degree of spectral manipulation required in order to achieve amplification and the speech features that characterize the subtype. Further implications of the spectral manipulation experiment will be discussed in Chapter 5 with regard to several research questions.

5 Discussion

This chapter begins with a review of the investigation's research questions, followed by a detailed integration of findings related to each research question relative to the literature. Clinical implications, limitations, and future directions of this work are presented to the end of this chapter.

The overall aims of this investigation were to obtain deeper insights into the nature of hypophonia with regard to perceived loudness, intelligibility, and acoustic characteristics, evaluate the utility of acoustic models of loudness in the context of hypophonia research, and provide preliminary evidence for the use of spectral manipulation in amplification of hypophonic speech. This investigation examined perceived loudness and intelligibility of connected speech in relation to utterance-level measures of speech level, spectral balance, and speech level variability in what might be considered an optimal listening environment. It is hoped that these insights will guide new directions investigating how the observed relationships change in adverse communication contexts that present greater barriers to IWPDs. Additionally, identified relationships between spectral characteristics and perceived loudness may be an important incorporation into future studies evaluating the nature of hypophonia and outcomes of hypophonia treatment, as well as a potential avenue into management of hypophonia via enhanced amplification devices.

5.1 Overview

As presented in Chapter 2, the primary research questions were the following:

RQ1: Do group differences exist between IWPDs with hypophonia and HOAs in perceived loudness, mean intensity, acoustic loudness, intelligibility, spectral balance, or speech level variability?

RQ2: Are acoustic models of loudness more predictive of perceived loudness than mean intensity?

RQ3: Can perceived loudness be predicted by speech level, spectral balance, or speech level variability?

RQ4: Can differences in perceived loudness between IWPDs and HOAs be explained by acoustic characteristics of their speech?

RQ5: Are loudness ratings collecting using visual analogue scales and direct magnitude estimation consistent and reliable?

RQ6: Do acoustic measures that predict loudness also predict intelligibility?

RQ7: Do manipulations of spectral composition predict perceived loudness ratings?

Each research question, its hypotheses, and its findings will be interpreted relative to the literature in the sections below.

5.2 IWPD-HOA Differences

RQ1: Do group differences exist between IWPDs with hypophonia and HOAs in perceived loudness, mean intensity, acoustic loudness, intelligibility, spectral balance, or speech level variability?

It was hypothesized that several group differences would exist between IWPDs and HOAs. IWPDs were expected to be quieter on the basis of mean intensity, acoustic loudness, and perceived loudness, would show differences in spectral composition, and would be less intelligible. Variability characteristics like standard deviation of intensity and intensity decay were also expected to potentially differ between IWPDs and HOAs.

5.2.1 Perceived Differences

IWPDs were perceived as quieter and less intelligible than HOAs. Reduced speech loudness is the primary characteristic of hypophonia and a marked symptom of hypokinetic dysarthria (Darley et al., 1969, 1969a). As all IWPDs studied in this investigation were noted to present with hypophonia according to their neurologist, it was expected that perceived loudness would be lower among IWPDs.

The present investigation is of benefit to the hypophonia literature due to the inclusion of task-specific ratings of perceived loudness of both IWPDs and HOAs. The majority of studies of hypophonia that have included perceptual ratings did not include HOAs, and studies have differed in the rating methods used. Categorical loudness ratings by speakers, communication partners, experts and experimenters have been reported in several studies of hypophonia, but no control groups were included in these studies (Berke, Gerratt, Kreiman, & Jackson, 1999; Cardoso et al., 2017; De Cock et al., 2007; Constantinescu & Hons, 2010; Evans, Canavan, Foy, Langford, & Proctor, 2012). Similarly, task-specific (Wilson et al., 2020) and generalized (Halpern et al., 2012; Ramig et al., 1995; Sharkawi et al., 2002; Wight & Miller, 2015). VAS loudness ratings have been previously employed in IWPDs, but without HOAs to provide a comparison. Some investigations have required listeners to categorically rate the severity of reduced loudness among IWPDs, with no HOA control groups (Biary, Pimental, & Langenberg, 1988; Cruz et al., 2016; Darley et al., 1969a, 1969b; Sadagopan & Huber, 2007). The present investigation is most similar to the investigation of perceptual and acoustic characteristics of IWPDs and HOAs by Ludlow and Bassich (1984). The authors included a large array of perceptual dimensions and acoustic measures. Of particular interest to this discussion, 3 speech pathology graduate student listeners rated overall loudness level using categorical loudness scaling, with 13 categories ranging from 'too soft' to 'too loud'. Similar to the present investigation, listeners were blind to PD status and samples were presented in a random order. Results of Ludlow and Bassich (1984) identified a large difference (d = 1.20) in overall perceived loudness of IWPDs relative to HOAs. Results of the present investigation are very consistent, also identifying large effect sizes with both VAS (d = 1.15) and DME (d = 1.20) ratings of loudness. The present investigation provides an extension and update of Ludlow and Bassich (1984) with a considerably larger sample size.

Lower intelligibility among IWPDs is also consistent with previous investigations (Chiu et al., 2020; Miller et al., 2007; Tjaden et al., 2014; Weismer et al., 2001). Findings of this investigation add to a body of recent work extending seminal characterizations of Darley et al. (1969a) and supporting that listeners perceive the speech of IWPDs as significantly different from HOAs on a variety of speech dimensions (Anand & Stepp, 2015; Chiu et al., 2020; Cushnie-Sparrow et al., 2018; McKenna & Stepp, 2018). Intelligibility among HOAs was very high, which was expected due to the absence of background noise to reduce speech-to-noise ratio (SNR).

The observed differences in perceived loudness and intelligibility between IWPDs and HOAs in this investigation may provide a conservative estimate, as hypokinetic deficits may be further exacerbated by the addition of background noise. While the majority of studies suggest that IWPDs demonstrate a similar Lombard effect to HOAs (Adams et al., 2006; Adams et al., 2005; Adams & Lang, 1992; Adams et al., 2006a; Stathopoulos et al., 2014), IWPDs generally maintain a lower intensity across conditions relative to HOAs. As a result, IWPDs may be even less intelligible in the presence of background noise due to the widening gap in SNR (Adams et al., 2008; Dykstra et al., 2013). Extending the findings of this investigation to adverse communication contexts, such as the presence of background noise, will further clarify the group differences that exist in perceived loudness and intelligibility.

5.2.2 Speech Level Differences

Consistently, acoustic measures of speech level indicated that IWPDs were quieter than HOAs. As identified in a scoping review recently conducted by this author to characterize the methodological variability of hypophonia studies (Cushnie-Sparrow, Adams, Page, and Parsa, [in prep]), mean intensity has been the most frequently employed measure of hypophonia in the hypophonia literature. Speech intensity of IWPDs is estimated to be, on average, 3-5 dB SPL quieter than HOAs (Adams, Dykstra, et al., 2006; Adams et al., 2005; Adams et al., 2010; Adams, Moon, et al., 2006; Fox & Ramig, 1997; Matheron et al., 2017). The mean difference of mean intensity in the present investigation was 2.9 dB, conservatively consistent with this range. The broad range of severity of hypokinetic dysarthria among the

IWPDs in this investigation may contribute to this conservative estimate. Among low loudness IWPDs, those with perceived loudness more than two standard deviations below the HOA mean, the mean difference in mean intensity relative to HOAs was 6.7 dB.

Group differences of median intensity were generally consistent with mean intensity. A different group interaction on loudness was observed for median intensity compared to other speech level measures, to be discussed in Section 5.5.1. Maximum intensity, which is in this context the maximum intensity observed at the utterance-level rather than an estimate of maximal capacity, did show stronger effects than mean intensity in classification of IWPDs. This may be the result of the incorporation of intensity modulation cues to this measure that are obscured by the mean intensity, consistent with smaller intensity variability in IWPDs. However, intensity variability results in this investigation were not consistent or compelling, and are described further in Section 5.2.4.

Differences between IWPDs and HOAs on TVL, active speech level, and LKFS have not been previously investigated. Results of this investigation indicate that the direction and magnitude of group differences in these measures are consistent with mean intensity and may indeed be more sensitive to IWPD-HOA differences.

Among speech level measures, larger effect sizes of IWPD-HOA differences were obtained via TVL, maximum intensity, and active speech level. Each of these measures incorporates additional cues about speech function above speech level. TVL incorporates spectral information via advanced and detailed filtering designed to approximate hearing processes from the outer ear through the cochlea. Additionally, the use of smoothing similar to automatic-gain control (AGC) incorporates some effects of loudness variability in the overall estimate of loudness. Active speech level, by using a threshold to remove low intensity segments, may incorporate intensity modulation and speech pausing in its estimates of speech level. Maximum intensity also provides clues into intensity modulation. Disrupted spectral balance (Corcoran et al., 2019; Cushnie-Sparrow et al., 2016; Dromey, 2003; Smith & Goberman, 2014; Tjaden et al., 2010), abnormal intensity modulation (Darley et al., 1969a; Ho et al., 2001; Rosen et al., 2005), and abnormal pause behaviour (Alvar, Lee, & Hubera, 2019; Bandini et al., 2015; Hammen & Yorkston, 1996; Huber, Darling, Francis, & Zhang, 2012; Martínez-Sánchez et al., 2016) may all be characteristics of IWPD speech. As a result, it is likely that these measures that incorporate characteristics of spectral balance and speech level variability are more effective discriminators of PD speech by capturing these other hypokinetic dysarthria deficits. Further investigation of these measures is needed, particularly relative to detailed prosodic examination of intensity variation and pausing behaviour. Discussion of IWPD-HOA differences in intensity variability is continued in Section 5.2.4.

5.2.3 Spectral Balance Differences

Overall, IWPDs demonstrated a relatively greater concentration of energy in lower frequencies of the spectrum as identified by steeper (more negative) tilt, lower proportions of mid-frequency (2-5 kHz) and high-frequency (5-8 kHz) energy, higher kurtosis, and higher skewness. This pattern is consistent with previous findings of disrupted spectral balance in both vowels and connected speech of IWPDs (Corcoran et al., 2019; Cushnie-Sparrow et al., 2016; Dromey, 2003; Smith & Goberman, 2014; Tjaden et al., 2010).

Differences in spectral balance between IWPDs and HOAs were particularly pronounced when IWPDs were divided based on low perceived loudness. IWPDs with perceived loudness within 2 SDs of the HOA mean were identified as similar to HOAs in most spectral balance measures, with a notable exception of high-ratio. High-ratio (proportion of 5-8 kHz energy) significantly differed between all subgroups, including the separation of HOAs from IWPDs with HOA-like loudness. The particular importance of this finding is expanded below.

The particular focus on smaller frequency-bands in this investigation provides additional perspectives to previous literature. As demonstrated by this investigation, different patterns can be observed between mid- (2-5 kHz) and high-frequency (5-8 kHz) energy. Depending on the cut-off, these findings can be lumped together into a measure of tilt. In the present study, the tilt cut-off was 1 kHz, such that both mid- and high-frequency were included in the denominator of tilt. The importance of this cut-off when evaluating spectral balance of IWPDs is discussed by Alharbi et al. (2019) and Cannito et al. (2006). Alharbi et al. (2019) investigated sustained vowels of 9 IWPDs pre-post LSVT via spectral and cepstral analyses. Originally, they employed low-high spectral ratio with a cut-off of 4 kHz, as is the default used by the Analysis of Dysphonia in Speech and Voice software (ADSV;

Pentax Medical). Low-high spectral ratio is equivalent to tilt, and the cut-off can similarly vary across investigations. While the authors found significant pre-post differences using cepstral measures, significant differences were not observed for this 4 kHz low-high spectral ratio. Using an adjusted low-high spectral ratio with a cut-off of 2 kHz, group differences emerged. In interpretation of this discrepancy, the authors reference the discussion of Cannito et al. (2006) in their case study of pre-post LSVT vowel harmonic differences. Cannito et al. (2006) found a redistribution of harmonic energy into higher frequencies following LSVT, especially above the second harmonic and below 4 kHz. By including these important frequency differences in the low-frequency portion of the ratio, treatment differences were obscured. Conversely, Watts and Awan (2011) identified significant differences between normal speakers and hypofunctional speakers (including IWPDs) on low-high spectral ratio with a 4 kHz cut-off. Results of the present investigation may contribute to this discussion. Large effect sizes of group were observed for both mid-ratio and high-ratio, consistent with considerable differences in spectral properties throughout mid- and high-frequency ranges. When IWPDs were divided based on their loudness, both mid-ratio and high-ratio significantly differed between the two groups of IWPDs. However, only high-ratio significantly distinguished IWPDs with relatively normal loudness from HOAs. Similarly, high-ratio's large effect sizes and high classification variable importance indicate that it is an effective discriminator of IWPD speech, even when perceived loudness is largely unaffected. This is consistent with findings of Watts and Awan (2011), indicating that relative weakness of energy above 4 kHz is a marked characteristic of hypofunctional speech.

As discussed in Chapter 2, spectral balance has been associated with effortful speech and hyperfunctional speech. Physiological changes associated with effort, including short glottal closing phase, can manifest in acoustic changes by shifting intensity over the spectrum such that additional intensity gained by the increased effort is added to higher frequencies, rather than a flat increase across frequencies (Gauffin & Sundberg, 1989; Glave & Rietveld, 1975; Sluijter & Heuven, 1996). This connection between physiology and acoustic manifestations may also be at the root of other findings of increased high-frequency energy with effortful speech (Eriksson & Traunmuller, 2002; Gauffin & Sundberg, 1989; Liénard & Benedetto, 1999; McKenna & Stepp, 2018; Neel, 2009). Findings of this investigation with regard to spectral balance may provide support to a hypothesis of spectral balance disruptions being the result of laryngeal hypofunction in IWPDs. Further investigation of this hypothesis might include extended examination of the relationships between laryngeal aerodynamics, laryngeal electromyography, effort, and perceived loudness. Additionally, findings of Adams et al. (2012) regarding the consistency between speech intensity estimates of IWPDs and HOAs at the throat and 8 cm from the mouth identified an interaction between PD status and vocal tract intensity transmission, such that only IWPDs were quieter 8 cm from the mouth relative to their throat microphone levels. Abnormalities in vocal tract resonance, including limited oral aperture of IWPDs related to hypokinesia, could further weaken harmonic structure and exacerbate spectral balance abnormalities of IWPDs.

5.2.4 Variability Differences

Measures of the intensity variability and TVL variability did not consistently differ between IWPDs and HOAs. Across statistical methods, speech level variability measures stood out from speech level and spectral balance measures as demonstrated by weak correlations, smaller group differences, and limited predictive performance. Within the subgroup of low loudness IWPDs, SD intensity significantly varied from HOAs, but intensity decay did not.

This was unexpected, as monoloudness is a hallmark perceptual feature of hypokinetic dysarthria (Darley et al., 1969, 1969a). However, this investigation is not alone in failing to find convincing differences between IWPDs and HOAs on intensity decay or intensity variability (Ma et al., 2015; Reyno-Briscoe, 1997; Rosen et al., 2005). Ho et al. (2001) found increased intensity declination among IWPDs in both prolonged vowels and sentence reading, but Rosen et al. (2005) only identified increased intensity declination in diadochokinetic rates. Rosen et al. (2005) emphasized the heterogeneity of IWPDs, indicating that some IWPDs demonstrated high declination despite it not being a consistent group effect. Importantly, Ho et al. (2001) only analyzed sentence samples from individuals capable of producing the sentence on a single breath. It may be informative that some individuals required a breath within the sentence, as IWPDs may take more breaths at minor syntactic boundaries or unrelated to syntax (Huber et al., 2012). Ma et al. (2015) also did not find significant differences in intensity variability between IWPDs and HOAs. Reyno-Briscoe (1997) found that while IWPDs differed from HOAs on perceived monoloudness, they did not vary in acoustic measures of intensity variability. It is possible that acoustic identification of prosodic deficits of IWPDs is challenging and requires different methodological approaches. Robustly capturing intensity variation may also require a more fine-tuned analysis of other prosodic characteristics, such as the incorporation of pausing and breath patterns. Additionally, in the screening of source data to create the pooled dataset used for this investigation, dysfluent speakers were removed. This choice was intended to reduce heterogeneity among samples and allow a clearer focus on loudness and intelligibility in the context of fluent speech. However, this also limits the prosodic variability available in the current data. Only 3 candidate speaker participants were removed for this reason, and it is expected that this does not significantly limit the findings. However, future investigations focused on relationships between hypokinetic dysarthria, perceived variability, and speech level variability should include analyses of more significantly dysfluent speakers.

With TVL's limited application to speech, and no previous application to speech of IWPDs, it was unknown the degree to which TVL variability might capture monoloudness. TVL variability results demonstrated some small differences in the low loudness subgroup analysis, but this may be complicated by the calculation of TVL, to be further expanded in Section 5.3.1. Results of this investigation do not provide robust support for the use of TVL variability as an index of IWPD prosodic deficits, though more detailed examination of this dimension is necessary.

The main findings of RQ1 can be summarized as follows: IWPDs were perceived as quieter and less intelligible than HOAs, and IWPDs were also quieter as measured by all speech level measures. On average, IWPDs had a relatively greater concentration of energy in lower frequencies of the spectrum. Measures of intensity variability and TVL variability did not consistently differ between IWPDs and HOAs. Across measures, the IWPD group was considerably more heterogeneous than HOAs, reflecting variations in severity and presentation of hypokinetic dysarthria. Dividing IWPDs into subgroups based on low perceived loudness can clarify interpretations by reducing this heterogeneity.

5.3 Acoustic Models of Loudness

RQ2: Are acoustic models of loudness more predictive of perceived loudness than mean intensity?

It was hypothesized that acoustic models of loudness would be more predictive of perceived loudness than mean intensity, as they have been designed to take listener factors into account.

5.3.1 TVL

Overall, findings of this investigation indicate that TVL is more predictive of perceived loudness than mean intensity. As TVL was the only robust model of loudness examined in this investigation, findings are consistent with expectations that a model incorporating listener factors improves the prediction of perceived loudness. TVL's long-term loudness maximum (referred to as TVL throughout the results) and long-term loudness mean (referred to as TVL mean) both demonstrated consistently strong associations and predictions of perceived loudness, as well as strong classification variable importance in distinguishing IWPDs from HOAs. The trends between long-term loudness maximum and long-term loudness mean were very similar across the investigation, and they are generally described jointly in the discussion as the TVL measures.

The default overall loudness estimate returned by the model is the long-term loudness maximum, identified in previous studies as a more accurate estimate of loudness than the long-term loudness mean (Marshall & David, 2007; Moore et al., 2016; Zorilă et al., 2016). In contrast to literature expectations, long-term loudness mean (TVL mean) was found to outperform long-term loudness maximum, as demonstrated by slightly stronger correlations with perceived loudness and intelligibility, better predictive performance in the maximal LMER model-building process, and higher classification variable importance. While Zorilă et al. (2016) examined the performance of TVL with respect to speech, loudness matching was employed as the perceptual rating method. It is possible that different perceptual processes are employed in loudness matching and loudness scaling, and that the long-term loudness mean more closely approximates loudness scaling.

As an acoustic model of loudness, TVL is very robust. Careful design and modification of its algorithm over time was intended to hone its performance as a measure of perceived loudness. In addition to the overall results of this investigation, the results of the spectral manipulation experiment are particularly supportive of TVL as an acoustic measure of loudness. Unlike other speech level measures examined, TVL showed a significant effect of spectral manipulation analogous to the observed effect on perceived loudness. This result indicates that TVL is successfully capturing frequency-related spectral contributions to perceived loudness.

More research is needed on the use of TVL variability as an index of prosodic variation and prosodic deficits in IWPDs. The use of standard deviation of TVL's long-term loudness and TVL's short-term loudness decay as indices of monoloudness and loudness decay is novel to this investigation, and it is possible that these measures are not adequately associated with these perceptual dimensions. Additionally, SD TVL and TVL decay correlated more strongly with TVL's long-term maximum and mean than SD intensity and intensity decay correlated with mean and maximum intensity. It is possible that the smoothing incorporated into TVL's algorithm makes these variability estimates inflated by the magnitude of TVL itself, particularly in the case of decay. As discussed in Section 5.2.4, acoustic identification of prosodic deficits of IWPDs has been challenging in previous studies. Clear, consistent effects of speech level variability were not observed in the present study, and it is challenging to interpret this insignificant result in the face of the perceptual prominence of monoloudness as a feature of hypokinetic dysarthria. Given the strong performance of TVL as a measure of loudness, future investigations incorporating perceptual measures of monoloudness and loudness decay may consider examining the relationships between TVL variability and these perceptual dimensions.

The MATLAB code for the current version of TVL (Moore et al., 2018) is freely available, removing a barrier to its use. However, many clinicians and some researchers do not have access to or literacy in MATLAB, which presents a feasibility barrier to its widespread use. Additionally, the computational load of this model is extremely high as a result of its robustness. From the very beginning of its algorithm, TVL obtains six fast Fourier transforms for every millisecond of the sample. For short-duration sounds like a brief pure tone, this is not burdensome, but in the context of clinical speech research, this is extensive in a way that is prohibitive. Samples in the present investigation ranged from 2-8 seconds in length, and calculating TVL required several minutes for each sample, even in the context of higher-than-average computational capacity. Using this measure broadly in clinical speech research and especially in a clinical context will require modifications of this measure to reduce computational load, and subsequent validation of those modifications.

To summarize, the results of this investigation are supportive of the use of TVL as a measure of perceived loudness in clinical contexts and clinical speech research where feasible. TVL is deemed to provide a robust estimate of loudness that captures speech level and spectral balance components.

5.3.2 LKFS and Active Speech Level

LKFS and active speech level provided slightly poorer performance than mean intensity in the prediction of perceived loudness. In the maximal LMER model-building progress predicting loudness, LKFS and active speech level were not selected as speech level predictors because of this poorer performance. Additionally, correlations with perceived loudness were weaker for LKFS and active speech level than for TVL or mean intensity. In the spectral manipulation experiment, LKFS and active speech level were not able to identify the change in loudness despite stable mean intensity, indicating that their algorithms do not adequately capture the spectral characteristics that affect loudness. This finding is less surprising for active speech level than it is for LKFS. The calculation of LKFS includes a sloping high-pass frequency filter emphasizing upper frequencies, which they state is designed to approximate equal-loudness contours. However, the patterns observed in this investigation suggest that this simple filter is not robust enough to capture spectral contributions to perceived loudness. The frequencies emphasized by the sloping high-pass filter are likely to be higher than the optimal speech loudness region, as the upper cut-off of the filter is 14 kHz. LKFS is designed to apply more generally to programme loudness of broadcast material, and this simple filtering likely provides better performance to a broader range of audio materials. Overall, the findings of this investigation do not provide support for the use of LKFS in the context of clinical speech research.

Active speech level's frequency filtering is not intended to specifically measure loudness. It is more generally intended to reduce noise and narrow the frequency-range to key frequencies of speech with a broad focus on the 100 Hz to 8 kHz range. Consequently, as expected, active speech level does not capture the frequency-related spectral contributions to perceived loudness that are emphasized by the spectral manipulation experiment. This algorithm focuses more on speech level variability, obtaining speech level only from active portions of speech, determined based on an intensity threshold. Active speech level may be particularly useful in the context of less predictable speech. The samples included in this investigation have already been carefully selected as mostly fluent speech samples lacking major pauses or mazes. While it may not be a particularly strong predictor of perceived loudness, the design of active speech level could make it a useful alternative or adjunct measure to mean intensity in an investigation focused on broader conversational speech. As discussed in Section 5.3.2, active speech level may be particularly useful for long-term, remote collection of speech (Schalling et al., 2013; Szabo & Hammarberg, 2013; Titze et al., 2007).

Despite modest performance in prediction of loudness, active speech level emerged as an effective discriminator of IWPDs from HOAs in the classification and logistic regression models. As discussed in Section 5.2.2, the use of a threshold to determine active speech may incorporate speech level variability and/or intensity modulation to its estimates, which may be of particular benefit in the context of prosodic deficits in IWPDs. However, as discussed in Section 5.2.4, acoustic identification of prosodic deficits of IWPDs may be challenging and a finer analysis of prosody may be needed to gain insights into the particular merits of active speech level as a measure of speech in IWPDs with hypophonia. Additionally, the VOICEBOX: Speech Processing Toolbox (Brookes, 2020) used to calculate active speech level in MATLAB is freely available, reducing a barrier to its use. However, as discussed with regard to TVL, the need for MATLAB presents a barrier for clinicians and some clinical speech researchers. Results of this investigation indicate that active speech level may not be a particularly useful measure for the prediction of perceived loudness, but may have other benefits as a measure of speech level in the context of clinical speech research, and further investigation is needed to explore these benefits.

The main findings of RQ2 can be summarized as follows: Overall, findings of this investigation indicate that mean intensity does not fully capture the acoustics of perceived loudness and that more robust measures of loudness may be indicated, particularly in clinical speech research. Correlations between perceived loudness and TVL were only slightly stronger than correlations between perceived loudness and mean intensity, but overall performance of TVL was more robust in terms of group effect sizes, predictive performance, and detection of loudness differences following spectral manipulation. TVL measures and active speech level provided better classification performance when separating IWPDs from HOAs. LKFS offered similar predictive value to mean intensity and may not be a useful additional speech level measure in the context of clinical speech research.

5.4 Acoustic Characteristics and Loudness

RQ3: Can perceived loudness be predicted by speech level, spectral balance, or speech level variability?

It was hypothesized that acoustic characteristics like speech level (e.g. mean intensity), spectral balance (e.g. tilt), and variability (e.g. SD intensity) would predict perceived loudness. Specifically, it was hypothesized that low speech level, a relatively greater concentration of low-frequency energy, low speech level standard deviation, and high speech level decay would be associated with lower perceived loudness.

5.4.1 Speech Level

As expected, all speech level measures were positively associated with and predictive of perceived loudness. As loudness is generally described as the psychophysical correlate of sound intensity, this direction of effect was expected. Of interest to this investigation was the relative performance of each speech level measure examined. Strongest associations and best predictive performance in the model-building process were obtained via TVL's long-term loudness mean. As TVL incorporates components of speech level, spectral balance, and variability (via smoothing), TVL's algorithm captures loudness more robustly. As discussed in Section 5.3.1, this robustness comes at the cost of jeopardized practicality of TVL in the context of clinical speech research. However, where feasible, TVL may provide a strong acoustic estimate of perceived loudness.

Among intensity measures, maximum intensity demonstrated slightly stronger correlations with perceived loudness than mean intensity, and median intensity's correlations were weaker than mean intensity. Marginally stronger correlations of maximum intensity might reflect a component of the effect of intensity modulation on the overall perceived loudness. Results of this investigation support the use of mean intensity and maximum intensity as speech level measures associated with perceived loudness, with the caveat that the missing contribution of spectral balance means that these measures are incomplete estimates of loudness.

Relationships between perceived loudness, LKFS, and active speech level were generally similar to mean intensity, and as discussed in Section 5.3.2, the additional components in the algorithms of these measures may not provide a particular benefit for the prediction of perceived loudness of speech.

5.4.2 Spectral Balance

A clear and consistent pattern emphasizing the importance of mid- and high-frequency to perceived loudness was observed. This was expected, as flatter spectral tilt has been associated with greater perceived loudness (Duvvuru & Erickson, 2013; Titze, 2020). The present investigation provides greater detail about the frequencies influencing this relationship.

Greater perceived loudness was associated with flatter (less negative) voiced and overall tilt, higher proportions of mid-frequency (2-5 kHz) and high-frequency (5-8 kHz) energy, lower kurtosis, and lower skewness. These findings all indicate that a relatively greater concentration of energy in the lower frequencies of the spectrum is associated with lower perceived loudness. Kurtosis and skewness describe overall distribution of energy as descriptions of the spectrum, with kurtosis demonstrating the concentration, and skewness describing the relative emphasis of low-frequency energy. Both measures indicated that a broader distribution of energy across the frequency range was associated with greater perceived loudness. Other spectral balance measures required a cut-off dividing energy into frequency ranges. Tilt and voiced tilt expressed the 0-1 kHz energy relative to the 1-10 kHz energy, with voiced tilt investigating the tilt within concatenated voiced-only segments of speech. Findings of tilt and voiced tilt were quite consistent, further reflected by the insignificant correlations between tilt ratio (voiced tilt/overall tilt) and perceived loudness. A greater discrepancy between tilt and voiced tilt could suggest a particular importance of turbulent, high-frequency energy, such as the turbulent energy associated with stop bursts and fricatives. Such a discrepancy was not observed, consistent with the effect of tilt on perceived loudness being driven mostly by a stronger presence of harmonic energy in the high frequencies.

While tilt and voiced tilt were positively associated with perceived loudness, a cut-off of 1 kHz can obscure the more detailed effects of particular frequencies. As discussed in Section 5.2.3, the choice of cut-off between lower and higher frequencies may change the observed effects. Mid-frequency (2-5 kHz) energy consistently demonstrated stronger associations with perceived loudness than high-frequency (5-8 kHz) energy. Similarly, mid-ratio provided better predictive performance in the maximal LMER model-building process than other spectral balance measures. Additionally, the interaction between TVL's long-term mean and mid-ratio in the prediction of loudness indicated that the relationship between TVL and loudness was stronger when mid-frequency energy was weak. Given the strong relationship that has already been observed between TVL and perceived loudness and TVL's incorporation of spectral information, this interaction suggests a particular sensitivity to mid-frequency energy in judgments of perceived loudness. Equal-loudness contours may provide a simple explanation for this result. As displayed in equal-loudness contours, there is a clear increase in sensitivity in the 2-5 kHz frequency range, such that a 3 kHz tone is perceived as louder than a 1 kHz tone at the same dB SPL, and perceived as much louder than a 100 Hz tone at the same dB SPL. The particular importance of mid-frequency energy observed in this investigation may be the result of a fundamental perceptual feature of the human ear in the context of speech. As introduced in Section 2.6.1, an upward shift of energy across the frequency range is associated with the use of intentionally louder, more effortful speech (Alharbi et al., 2019; Cannito et al., 2006; Eriksson & Traunmuller, 2002; Gauffin & Sundberg, 1989; Glave & Rietveld, 1975; McKenna & Stepp, 2018; Neel, 2009; Sluijter & Heuven, 1996). It is possible that these patterns of an increased sensitivity to mid- and high-frequency energy and a greater proportion of these frequencies in effortful speech are not coincidental. Our sensitivity to this frequency range may be directly related to its importance in the loudness, and intelligibility, of the human voice.

5.4.3 Variability

Speech level variability was associated with perceived loudness to a lesser extent, demonstrating weak to moderate correlations. In the maximal LMER model-building process, standard deviation of intensity improved predictive performance. Intensity variability may provide a fine-grained adjustment to perceptions of perceived loudness after the larger contributions of speech level and spectral balance are incorporated. Specifically, standard deviation of intensity positively predicted loudness, indicating that greater variability was associated with greater perceived loudness. This was the expected direction, as it was expected that monoloudness would be associated with lower perceived loudness. Greater intensity modulation, particularly larger peaks in the intensity contour, might increase overall estimates of perceived loudness. To clarify this relationship, future directions might incorporate perceived monoloudness and loudness decay to evaluate the relationships between these perceptual dimensions and the overall perceived loudness. A deeper understanding of the perceptual relationships would facilitate greater acoustic investigation. As discussed in Section 5.3.1, it is unknown the degree to which SD TVL and TVL decay are associated with monoloudness and loudness decay, and further evaluation of these metrics in relation to perceptual measurement is needed.

The main findings of RQ3 can be summarized as follows: Results of this investigation indicate that both speech level and spectral balance are consistently associated with and predictive of perceived loudness. Higher speech level and a relatively greater proportion of mid- (2-5 kHz) and high-frequency (5-8 kHz) energy were associated with and predictive of higher perceived loudness. Variability of intensity was associated with perceived loudness to a lesser extent, demonstrating weak-moderate correlations. However, in a maximal model, standard deviation of intensity improved predictive performance, suggesting that the effect of intensity variability may 'fine-tune' perceived loudness such that reduced intensity modulation decreases the overall perceived loudness. Based on marginal improvement in performance observed for each predictor during the maximal model-building progress, TVL's long-term mean provided the best predictive performance of the speech level measures examined.

5.5 IWPD-HOA Differences: Acoustic Characteristics and Loudness

RQ4: Can differences in perceived loudness between IWPDs and HOAs be explained by acoustic characteristics of their speech?

It was hypothesized that speech level deficits would be associated with greater perceived loudness deficits in IWPDs than HOAs. It was also hypothesized that if present, spectral balance deficits and speech level variability deficits in the speech of IWPDs would significantly contribute to the perceived loudness deficits of hypophonia.

5.5.1 Speech Level

Speech level measures positively predicted loudness in both IWPDs and HOAs, but the relationships between most speech level measures and loudness were stronger in IWPDs. This was expected, as it was hypothesized that concurrent deficits in spectral balance, prosody, articulation, and voice quality in the speech of IWPDs could further influence the overall perceived loudness. As a result of these influences, it was expected that the same dB SPL produced by an IWPD might be perceived as quieter than if an HOA had produced it, widening the gap between speech level and loudness. Significant interactions between group and speech level on perceived loudness were observed for all speech level measures except active speech level. The interaction between group and median intensity in the prediction of loudness showed the opposite pattern of other speech level measures, such that median intensity was more predictive of loudness in HOAs. A greater disparity between mean intensity and median intensity might be reflective of the effects of intensity modulation, as low intensity segments of speech will reduce the median. Further investigation of the associations between mean intensity, median intensity, and perceived loudness should be incorporated alongside more detailed exploration of speech level variability and prosodic deficits.

The strongest interaction was observed between TVL's long-term maximum and group, with IWPDs showing a considerably stronger relationship between TVL and perceived loudness. It is possible that the incorporation of spectral balance characteristics and particularly intensity modulation are factors in the larger interaction of TVL relative to TVL mean. TVL's long-term maximum may better capture the effects of larger intensity peaks in the utterance, and IWPDs may demonstrate a flatter intensity contour consistent with monoloudness, and subsequently may present with lower TVL long-term loudness maximum. The restricted range of speech level among HOAs is an important consideration in the interpretation of these interactions. A smaller range of speech level among HOAs may simply be the result of the absence of hypophonic deficits, in which case, these results are representative of the population of older adults with and without PD. However, it is also possible that an optimal conversational setting of a quiet room with no background noise and a comfortable interlocutor distance is not challenging enough to draw out an appropriate range of loudness from HOAs, which would allow for closer examination of these interactions. As discussed above, IWPDs may struggle to maintain adequate SNR in the presence of background noise or at larger interlocutor distances due to overall lower speech intensity (Adams et al., 2008; Dykstra et al., 2013). Future investigations might consider the incorporation of noise and distance to better evaluate these relationships, particularly between IWPDs and HOAs.

5.5.2 Spectral Balance

Throughout this investigation, IWPDs presented with weaker mid- and high-frequency energy. Weaker mid-frequency energy, in particular, is strongly associated with lower perceived loudness. It is believed that these spectral balance deficits are contributing to the perceived loudness deficits of hypophonia, as expected. However, most spectral balance measures did not demonstrate significant group interactions in their prediction of loudness. In particular, mid-ratio demonstrated very similar relationships with loudness in IWPDs and HOAs, despite the large group differences in mid-ratio. The spectral balance measure demonstrating a significant interaction with group in the prediction of loudness was high-ratio, particularly at low high-ratios. As discussed in Section 5.2.3, high-ratio's group differences were particularly large, and this may be an indicator of hypofunctional speech. High heterogeneity among IWPDs may further complicate the group interactions. This explanation would help clarify why mid-ratio did not show the interaction observed with high-ratio, as high-ratio was observed to significant differ even between HOAs and IWPDs with normal loudness. IWPDs in the low loudness group generally presented with an even clearer picture of weak mid- and high-frequency energy, as represented by much larger subgroup differences and effect sizes. These results indicate that IWPDs with more pronounced hypophonia also present with greater spectral balance deficits. This is consistent with findings of Tjaden et al. (2010), who reported that kurtosis and skewness were positively associated with perceived severity of the speech of IWPDs.

5.5.3 Variability

As discussed above in Sections 5.2.4 and 5.4.3, IWPD-HOA differences on speech level variability and the relationships between speech level variability and perceived loudness are not clear. As a result, strong support for the hypothesis that low speech level variability among IWPDs is associated with greater perceived loudness deficits is not provided by the results of this investigation. Significant interactions with group in the prediction of loudness were observed for standard deviation of TVL's long-term loudness and of the decay TVL's short-term loudness. However, as discussed in Section 5.4.3, it is unclear the degree to which the variability estimates of TVL are driven by TVL itself. As a result, these effects are not deemed to provide strong support for differential effects of variability on loudness among IWPDs. The direction of the interaction of standard deviation of TVL and group was consistent with the expected effect, such that standard deviation of TVL was more positively associated with loudness in IWPDs, particularly at lower values. Conversely, a stronger relationship between TVL decay and loudness was observed in IWPDs, such that flatter (less negative) decay among IWPDs was associated with lower perceived loudness, whereas no effect of TVL decay was observed in HOAs. This is an unexpected effect, given that loudness decay would be expected to decrease overall perceived loudness. However, as discussed in Section 5.4.3, the TVL variability measures examined are novel to this examination and more investigation of their associations with perceptual judgments of monoloudness and loudness decay are needed to clarify these relationships. In summary, this investigation does not provide strong support for the contribution of variability deficits to the overall loudness deficits of IWPDs.

The main findings of RQ4 can be summarized as follows: The observed group differences of IWPDs on speech level and spectral balance measure were clearly and consistently in directions likely to attenuate perceived loudness. Speech level measures positively predicted loudness for both groups, but the relationships between speech level and loudness were stronger in IWPDs, particularly at low speech levels. A restricted range of speech level among HOAs may be a factor to consider in the interpretation of that result. Weak mid- and high-frequency energy in the spectra of IWPDs is a possible contributor to their reduced perceived loudness. However, most spectral balance measures did not demonstrate significant group interactions in their prediction of loudness. High heterogeneity among IWPDs may complicate the group interactions. When IWPDs were divided into subgroups based on their perceived loudness, low loudness IWPDs presented with a clear picture of weak mid- and high-frequency energy. These results are indicative that IWPDs with more pronounced hypophonia also present with greater spectral balance deficits. The relationships between speech level variability and perceived loudness were not clear, and group differences were weak and inconsistent. Further investigation of speech level variability is needed to provide insights into this dimension.

5.6 Perceptual Ratings of Loudness

RQ5: Are loudness ratings collecting using visual analogue scales and direct magnitude estimation consistent and reliable?

It was hypothesized that loudness ratings collected using visual analogue scales and direct magnitude estimation would be consistent with one another and offer similar reliability.

The present investigation provides support for the use of either VAS or DME as loudness scaling methods for clinical speech research. Ratings were consistent between methods, as very high correlations and similar distributions were observed for VAS and DME ratings of perceived loudness.

Both VAS and DME offered excellent reliability, both within and across raters. It was observed that inter-rater reliability was slightly higher than intra-rater reliability with VAS ratings, whereas the reverse was true for DME ratings. Lower inter-rater reliability of direct magnitude estimation may reflect that the method is inherently idiosyncratic, as each listener picks their own scale and increment. Lower intra-rater reliability might be explained by effects of perceptual drift, as each sample may be affected by the neighbouring samples. As a result, listeners' ratings of the duplicated samples may differ based on their context. Randomization of duplication and presentation order is designed to mitigate this effect across listeners, but it may not completely remove the effect. The use of a standard modulus every five samples in the DME collection may have reduced the effect of drift, contributing to higher intra-rater reliability. Despite these small differences, overall observed reliability of both methods was excellent.

Within DME, percent-averaging and geometric means were observed to present consistent results. This support for the use of percent-averaging simplifies the use of DME for ratings of loudness and calculation of reliability from the percent scores.

While both VAS and DME are supported for use in future studies of perceived loudness in hypophonia research based on the results of this investigation, VAS may be more practical for use by clinicians. DME is an effective method of loudness scaling in a research setting, but requires an experimental setup and multiple listeners for informative results. Additionally, DME can be affected by the raters' experience with other scales, such as category loudness scaling (Jesteadt & Joshi, 2013). VAS is more reliable than category scaling, provides better resolution (Karnell et al., 2007; Kreiman et al., 1993), and can be quickly used by clinicians, patients, and communication partners to provide ratings of loudness and speech characteristics.

The main findings of RQ5 can be summarized as follows: Results of this investigation support the use of either visual analogue scales or direct magnitude estimation when obtaining ratings of perceived loudness, as ratings between methods were consistent with one another and both offered high reliability as per ICC. Within direct magnitude estimation, percent-averaging and geometric mean yielded the same results, supporting the use of either method, though percent averaging is simpler and facilitates calculation of reliability scores.

5.7 Loudness and Intelligibility

RQ6: Do acoustic measures that predict loudness also predict intelligibility?

It was hypothesized that measures predicting loudness may also contribute to intelligibility, but that perceived loudness and intelligibility would differ enough that intelligibility ratings could not be considered to encompass loudness.

In this investigation, perceived loudness and intelligibility presented different relationships with acoustic characteristics and different patterns between IWPDs and HOAs. Overall, findings indicate that loudness and intelligibility are distinct outcomes, consistent with expectations. Only moderately strong correlations were observed between loudness and intelligibility. In the maximal LMER model-building process, many of the measures that provided strong prediction of perceived loudness offered poor prediction of intelligibility. This was especially true in HOAs, likely due to a ceiling effect of high intelligibility. A secondary model was pursued among only IWPDs to clarify results. The resulting model predicting intelligibility in IWPDs lacked a significant predictor of speech level. This suggests that with adequate SNR, there was not a significant component of overall audibility on intelligibility among IWPDs. The pattern of weaker contributions of speech level to intelligibility is likely to be particularly prominent in the context of adequate SNR.

As discussed above, IWPDs may struggle to maintain adequate SNR in the presence of background noise due to overall lower speech intensity (Adams et al., 2008; Dykstra et al., 2013). Intelligibility of IWPDs may then be further exacerbated by spectral balance deficits and also by prosodic deficits that may not have been captured by the metrics used in this investigation.

While speech level did not significantly predict intelligibility, the predictors that did contribute to prediction of intelligibility in IWPDs were mid-ratio and high-ratio, reflecting the importance of spectral balance in both perceived loudness and intelligibility. Based on this investigation, it is not known whether spectral balance is only of importance to intelligibility in the context of abnormal spectral balance, or if the weak associations among HOAs resulted from a restricted range of intelligibility.

Correlations between intelligibility and speech level variability were weak and frequently insignificant, and variability did not improve model performance in the prediction of intelligibility. As previously discussed, capturing the prosodic deficits of hypokinetic dysarthria with acoustic measures may be challenging, and the present investigation does not provide evidence of a relationship between speech level variability and perceived intelligibility.

To obtain greater clarity of intelligibility's relationships with speech level, spectral balance, and speech level variability, further investigation of intelligibility and perceived loudness is needed in the context of background noise and in other adverse communication contexts (e.g. interlocutor distance).

The main findings of RQ6 can be summarized as follows: Loudness and intelligibility are distinct outcomes and should be treated as such. Measures that effectively predicted loudness provided poor prediction of intelligibility, suggesting that we cannot generalize between loudness and intelligibility. Both of these outcomes are very relevant to people with hypophonia. Future studies of hypophonia, especially when selecting treatment outcomes, should include both loudness and intelligibility as perceptual indicators of overall effects of hypophonia.

5.8 Spectral Manipulation and Loudness

RQ7: Do manipulations of spectral composition predict perceived loudness ratings?

It was hypothesized that increases and decreases in the gain of mid- and high-frequency energy would increase and decrease loudness, respectively. Flatter tilt (greater proportion of energy in the higher frequencies) was expected to be associated with greater loudness.

Results of this investigation indicate that, as expected, perceived loudness is affected by manipulations of mid- and high-frequency energy, even in the context of equal mean intensity. The direction of this effect was as expected, such that an increase in the relative proportion of mid- and high-frequency energy was associated with an increase in perceived loudness. Differences in loudness between the unaltered speech and the 5 dB positive and negative manipulation conditions were not significant. Significant effects of both 10 dB manipulations were observed, indicating that the magnitude of manipulation changes the magnitude of effect. These findings have important implications, as they support the use of spectral manipulation to achieve improved amplification of speech loudness, and emphasize that the magnitude of manipulation can be increased to increase this effect.

The spectral manipulation conditions affected IWPDs and HOAs to a similar degree. This suggests that the resulting changes in perceived loudness result from the contribution of mid- and high-frequency energy in general, rather than from the mitigation of the speech deficits of IWPDs. However, some individuals demonstrated poor loudness amplification in the positive 10 dB manipulation condition. IWPD members of this group presented with greater speech level deficits and a more pronounced concentration of energy in low frequencies, suggesting that some individuals may require a larger gain shift to benefit from spectral manipulation. Investigation of the spectrographic characteristics of these low amplification IWPDs suggests the role of a poor harmonic structure to the weak amplification. Following the results of this investigation, experimental double-amplification (+ 20 dB gain) was applied to the speech of one of these individuals, and speech loudness was noted to increase following this larger manipulation. Future investigations of the use of these manipulations in the context of disordered speech might incorporate a greater range of gain conditions. This would facilitate the identification of optimal spectral gain for different individuals, and a greater examination of the speech characteristics that predict this optimal spectral gain. These expanded investigations might also identify a possible upper limit to this effect, such that extremely high proportions of mid- and high-frequency energy could be associated with low speech naturalness, which could compromise intelligibility.

The use of mid- and high-frequency manipulations to improve perceptual characteristics of voice is a common practice in audio engineering. Corbett (2015), in a handbook on microphones and mix techniques for audio engineers specializing in music, notes that amplifying the 1-2.5 kHz range can increase clarity. Additionally, the 5 kHz range is associated with 'presence,' which can "give the singer an edge and allow them to cut through the mix" (Corbett, 2015, pg. 191). In the context of speech, this might correspond to enhanced clarity and intelligibility, even in the presence of decreased SNR. Similarly, Ronen (2015) found improvements in the perceived intelligibility of vocals when 6 dB boosts were centered at 2, 5, and 8 kHz, with largest effects when boosts were centered at 2 kHz and 5 kHz. Within the hearing science field, Moore, Füllgrabe, & Stone (2010) found that speech energy at frequencies above 5 kHz significantly improved intelligibility in the presence of spatially separated background noise for normal-hearing and hearing-impaired listeners. Specifically, the investigators were interested in the cut-off frequencies of hearing aids, and found that an increase in cut-off frequency from 5 kHz to 7.5 kHz was associated with an increase in intelligibility, whereas an increase from 7.5 kHz to 10 kHz was not. This corresponds with the higher correlations of high-ratio and intelligibility observed in the present investigation, and emphasizes the need to investigate the effects of this spectral manipulation on intelligibility, as well as on loudness.

Spectral manipulation in the context of clinical speech amplification has been attempted before. The Speech Enhancer, originally produced by Electronic Speech Inc., was a device marketed to improve intelligibility of speech by amplifying the voice, reducing background noise, using spectral alteration to improve segmental perception accuracy and provide auditory feedback via headphones. Evidence for the Speech Enhancer's efficacy is limited, and the device is no longer available. However, preliminary evidence showed that the Speech Enhancer improved intelligibility for some listeners relative to normal presentation and to the Voicette, a speech amplifier (Bain, Ferguson, & Mathisen, 2005), and improved intelligibility in the context of background noise (Weiss, 2002). Early efficacy of the Speech Enhancer provides further support for the pursuit of more research into spectral-boosted amplification in development of new, clinical speech amplification devices. Recent evidence suggests that amplification devices are an efficacious treatment of hypophonia on the basis of perceived intelligibility and SNR, with and without the presence of background noise (Andreetta, Adams, Dykstra, & Jog, 2016; Knowles, Adams, Page, Cushnie-Sparrow, & Jog, 2020). The majority of IWPD and communication partner dyads who participated in the investigation of Knowles et al. (2020) continued using a device following the study, indicating that the benefits of the device were considerable enough for dyads to use them in their communication activities. Neel (2009) found that while amplified speech and loud, effortful speech produced by IWPDs both resulted in improved intelligibility, amplified speech provided less benefit. Incorporation of targeted spectral manipulation could further enhance the clinical benefits of amplification devices as hypophonia treatments.

The main findings of RQ7 can be summarized as follows: Perceived loudness is affected by manipulations of mid- and high-frequency energy. Insignificant changes were observed with 5 dB gain shifts, but significant effects of 10 dB shifts were observed, indicating that the magnitude of manipulation changes the magnitude of effect. Spectral manipulation affected IWPDs and HOAs to a similar degree, suggesting that the resulting changes in perceived loudness are due to the overall contribution of mid- and high-frequency energy to perceived loudness rather than mitigation of a hypokinetic dysarthric deficit. However, some individuals demonstrated poor loudness amplification in the positive 10 dB manipulation condition, and the IWPDs in this group presented with greater speech level deficits and a more pronounced concentration of energy in low-frequencies, suggesting that some individuals may require a larger gain shift to benefit from spectral manipulation. The results of this investigation provide strong preliminary support for the hypothesis that increasing the proportion of mid- and high-frequency (2-10 kHz) energy increases the perceived loudness of speech produced by IWPDs and HOAs. Future investigations of these spectral manipulations incorporating perceptual ratings of perceived intelligibility and adverse communication environments (e.g background noise, interlocutor distance) will provide deeper insights and additional support.

5.9 Limitations

While the present investigation provides strong evidence in support of its hypotheses and has important implications, there are limitations that should be considered in the interpretation of the results. Most notably, these limitations are the restricted communication context with an absence of background noise, limited control and examination of articulation and prosody, and limited clinical assessment of IWPDs as a result of the use of archived audio data. Each of these limitations is further discussed below.

5.9.1 Communication Context

The present investigation focused on the relationships between acoustic characteristics, perceived loudness, and hypokinetic dysarthria in an optimal communication environment. Speech samples were recorded in a quiet- or sound-treated room with a headset microphone placed 6 cm from the mouth. Listeners completed the rating task in a sound-treated room. Background noise was very low in both the recording and listening environments, and speech stimuli were presented using a high-quality audio monitor, with a distance of 1.5 m between the loudspeaker and the listener. This listening environment is analogous to a comfortable interlocutor distance in a quiet room. It was of interest to investigate these research questions in an environment uncomplicated by effects of adverse communication contexts to develop a base from which to expand into future investigations. By reducing effects of SNR on the observed results, a more focused, 'best-case' interpretation of the relationships is observed. This is both a strength and a weakness, and future investigations expanding these research questions to wider communication contexts will mitigate that weakness.

5.9.2 Effects of Articulation

Archived audio data was used for this investigation. In-person participant recruitment was not available at the time of this investigation due to the global pandemic (Cucinotta & Vanelli, 2020). However, the use of archived audio provided a larger sample of participants than would have been available through the originally planned participant recruitment, which mitigates the limitations that it causes. Using archived audio meant that the speech tasks selected for analysis had to be available in all of the source data used to create the pooled data. The consistently available speech tasks were randomized sentence lists via the Sentence Intelligibility Test (SIT; Yorkston, Beukelman, & Tice, 1996) and samples of conversational speech. The absence of a standard sentence or standard reading passage in the available tasks means that this investigation did not have a sentence that was uttered by all participants. Articulatory variability is thus considerable across participants, which makes it challenging to interpret the possible effects of articulation deficits on the overall observed perceived loudness and intelligibility. This limitation could have been mitigated by including a more segmental analysis, such as identifying sibilants, stop consonants, and vowels to better understand spectral properties, but this avenue of analysis was outside of the scope of the current investigation. There are practical advantages to using utterance-level measures, but segmental analyses can provide clearer interpretation of the results observed with utterance-level measures. Future investigations could incorporate a standard sentence or reading passage for greater comparison across individuals, or include a detailed segmental analyses to compare with utterance-level measures.

5.9.3 Effects of Prosody

As discussed above, the unclear and inconsistent effects of speech level variability on perceived loudness in this investigation are challenging to interpret in the absence of thorough prosodic investigation and perceptual ratings of prosodic dimensions. In particular, speech pause analysis would augment and clarify findings of this investigation. Speech pause analysis would provide clearer insights about active speech level as a measure of IWPD speech. Additionally, finer analyses of intensity modulation with a greater focus on the peaks of the intensity contour or descriptions of intensity contours shape (e.g. kurtosis, skewness) might provide new avenues for investigation. Critically, more research is needed to identify acoustic measures that capture the dimensions of monoloudness and loudness decay, as this is a precursor to understanding the effects of speech level variability on overall perceived loudness. This detailed examination was outside of the scope of the present investigation but reflects an important extension of this work that is needed to better characterize hypokinetic dysarthria with acoustic measures.

5.9.4 Clinical Assessment of IWPDs

The clinical assessment of IWPDs was less robust in this investigation as a result of the use of archived data. In the originally planned investigation, IWPDs would have completed the UPDRS at the time of assessment, rather than at the most recent neurological examination. Additionally, MoCA scores would be collected for all participants, whereas these estimates are only available for a subset of IWPDs. A clinical hypophonia severity scale was also planned. This scale included a battery of simple speech tasks probing hypophonia by asking IWPDs to speak at their habitual loudness, higher loudness levels, and greater interlocutor distances, and testing loudness decay by requesting that individuals maintain their loudness over a long counting task. The clinician or administrator of the scale would then provide a 0-3 rating of the individual's performance, yielding estimates of hypophonia severity. Additionally, the IWPD would provide self-ratings of their speech loudness, overall and in specific communication contexts. The inclusion of a clinically applicable assessment battery to a detailed examination of hypophonia characteristics would provide valuable information to clinicians who assess and manage hypophonia in IWPDs. Measures of communication participation (Communication Participation Item Bank short-form; Baylor et al., 2013) and communicative effectiveness (Communicative Effectiveness Survey; Donovan, Velozo, & Rosenbek, 2007) were also included in the planned protocol. It is an unfortunate limitation of the use of archived data that these measures could not be incorporated into the present investigation. However, the important insights gained from the detailed examination of this large sample of IWPDs can be extended by future investigations pairing the key measures identified here with a more detailed clinical assessment of IWPDs.

5.9.5 Listeners and Listening Tasks

The listeners in the present investigation were of a moderate experience level rather than expert listeners, such as experienced speech-language pathologists or speech researchers. All listeners were clinical graduate students studying speech-language pathology, and all had some experience with auditory-perceptual evaluation of speech and voice. However, the listeners were not experienced raters of dysarthria and did not have significant previous experience with the speech of

IWPDs. While perspectives vary with regard to the role of experience in the perceptual ratings of speech (Bain et al., 2005; Eadie & Kapsner-Smith, 2011; Helou et al., 2010; Kuruvilla-Dugdale, Threlkeld, Salazar, Nolan, & Heidrick, 2019; Schliesser, 1985), it is deemed that the listeners who participated in the present investigation provided consistent and reliable ratings of loudness. Experienced raters may use different strategies and mental heuristics when providing ratings of speech parameters (Kreiman, Gerratt, & Precoda, 1990), as their prior experiences shape their internal scales. The semi-naive nature of the listeners in the present investigation may thus be an advantage, as their clinical education and experience is more similar at this time in their career.

The number of listeners in this investigation might be considered modest, with 8 listeners included in the analysis from 10 listeners recruited. Abur, Enos, & Stepp (2019) investigated the relationships between VAS ratings of intelligibility and orthographic transcription in a total of 80 listeners. Their investigation sought to clarify the number of listeners required to achieve good consistency between transcription and scaling, and to identify if listeners needed to rate all samples or if one listener per sample was acceptable. Their findings indicated that strong relationships were observed between transcription intelligibility and scaled intelligibility with at least 2 listeners. While direct translation of these findings to loudness is challenging, it is believed that the number of listeners included in the present investigation was adequate to provide reliable, valid ratings of perceived loudness and intelligibility.

The listening tasks used in this investigation were very controlled, which may affect the ecological validity of the perceptual ratings provided. The listening environment was controlled, with interlocutor distance stable across samples. Samples were randomized, such that sequential samples could be either SIT sentences or conversational speech and would come from different participants. This improves reliability as a research task, but is not representative of a natural communication environment. Additionally, practice trials were not incorporated. Following instructions and orientation to the rating tool and the speech parameter being examined, listeners immediately began providing their perceptual ratings. As this task is very simple, it was expected that listeners would quickly learn the tool and that these early samples would be representative of later ratings. The order of speech samples were randomized, mitigating the effects of order in the average ratings across listeners. Order effects would also be incorporated into intra-rater reliability estimates, which were good to excellent for all listeners. Overall, characteristics of the listeners and listening tasks are not deemed to be major limitations to the findings of this investigation, but they are methodological decisions to consider in the interpretations.

5.10 Clinical Implications

Results of this investigation highlight the importance of spectral balance to perceived loudness of speech. Clinicians interested in collecting acoustic analyses of the speech of individuals with hypophonia may want to incorporate measures of spectral balance into their evaluations to better capture changes in their speech as a result of intervention or disease progression. Recent guidelines for acoustic evaluation of dysarthria and of dysphonia did not include a dimension capturing spectral balance (Patel et al., 2018; Rusz et al., 2021). The present investigation indicates that within the context of hypokinetic dysarthria, spectral balance is an important characteristic and may also be useful for evaluation of other dysarthrias.

The choice of frequency cut-off used in spectral balance measures may be particularly important in the context of hypophonia, as underscored by the findings of Watts and Awan (2011), Cannito et al. (2006), Alharbi et al. (2019), and the present investigation. In the context of speech loudness, 2-5 kHz energy may be of particular importance, and in identification and evaluation of IWPD speech, higher frequencies above 4 kHz may be emphasized. For clinicians with access to ADSV (Pentax Medical), the default frequency cut-off of low-high spectral ratio provided by the program is 4 kHz, which may obscure important spectral information. Adjusting this frequency cut-off may be critical to obtaining a clear picture of the spectral balance characteristics of each client.

Another implication of this work is the support for perceptual ratings of loudness and of intelligibility. Particularly large effect sizes were observed for perceived loudness, reflecting the listener's ability to identify additional characteristics of hypophonic speech. Listener perceptions should be considered as a valuable tool in clinical decision-making. Objective acoustic measures like sound pressure level should be considered as augments to listener perceptions, not replacements or upgrades of them. Additionally, further support is provided for the use of VAS scaling of loudness and intelligibility as a reliable and practical tool.

Where feasible, TVL's long-term loudness mean provides a robust acoustic measure of loudness to supplement a clinician's auditory-perceptual evaluation. Unfortunately, calculation of TVL currently requires MATLAB, which is likely unavailable to most clinicians, and the calculation is also computationally intensive. Improving the practicality of these measures would be of value to clinicians in their assessment and management of hypophonia. Optimizations to TVL could include broadening the time and frequency windows of the algorithm to reduce the total number of fast Fourier transforms and filters required, or rewriting the algorithm in a more efficient programming language.

This investigation also provides preliminary evidence that increasing mid- and high-frequency components of speech increases the perceived loudness despite equal mean intensity. This result suggests that effectiveness of speech amplification would be increased by the incorporation of spectral manipulation. Future investigations of speech amplification devices as a treatment for hypophonia should consider the incorporation of spectral 'boosts' to increase perceived loudness, improving treatment outcomes.

5.11 Future Directions

As discussed in Section 5.9, the use of archived data created limitations to the present investigation. Future investigations with prospective data should incorporate more thorough clinical assessment of IWPDs, including clinical assessment of hypophonia, current estimates of severity and cognitive function, and self-ratings of loudness, communicative participation, and communicative effectiveness. Prospective collection should also include a standard sentence or passage uttered by all participants to simplify the effects of articulation and prosody in at least one exemplar.

The incorporation of adverse communication contexts, including background noise and interlocutor distance would expand the findings of this investigation. The contribution of SNR to perceived loudness is of interest, given the known effects of SNR on intelligibility and the presence of a wider SNR gap among IWPDs (Adams et al., 2008; Dykstra et al., 2013). Additionally, it is unknown whether the spectral balance deficits of IWPDs would have an even stronger effect on perceived loudness and intelligibility in low SNR environments. A future investigation of this dimension might include presenting the same data used in the present investigation mixed with multi-talker background noise to different SNRs, and obtaining perceived loudness and intelligibility ratings in these additional contexts. For example, background noise might be mixed to -2 dB, 0 dB, 2 dB, 5 dB SNR to investigate the IWPD-HOA differences when SNR is stable. Additionally, adding background noise of 65 and 70 dB SPL would provide insights into that wider gap between IWPDs and HOAs, a more ecologically valid comparison. Using both approaches would provide considerable new insights into the relationships between acoustics and loudness in IWPDs and HOAs. Manipulations of interlocutor distance might include rating conditions at 3 m and 6 m of distance between the loudspeaker and the listener to simulate larger interlocutor distances.

Given the positive results of the spectral manipulation experiment in the present investigation, refinements to the spectral manipulation experiment would be indicated to obtain deeper insights. The manipulation in this investigation was a simple gain adjustment, with a sloping increase in gain from 1-2 kHz and a flat gain shift from 2-10 kHz. More advanced filter-bank techniques might specifically target the 2-5 kHz or 5-8 kHz ranges based on the observed importance of these ranges to perceived loudness and intelligibility. Individual customization of spectral manipulation may provide better outcomes than a broad amplification technique. The simple manipulation in the present study would have amplified high-frequency harmonic energy and noise equally. By specifically targeting harmonic energy based on an individual's estimated fundamental frequency, speech loudness could be effectively amplified while reducing distortion related to high-frequency noise. Additionally, further explorations of spectral manipulation should incorporate perceived measures of intelligibility in addition to loudness, as both outcome measures would be useful in determining the benefit of manipulated amplification as a hypophonia treatment. Finally, incorporating background noise into the spectral manipulation experiments will provide even clearer information about the efficacy of spectral manipulation as an augmentation to speech amplification devices for clinical use.

Speech pause analysis and detailed prosodic analysis may be undertaken in a future prospective analysis or as an extension to the pooled data of the present investigation. Perceived measures of monoloudness and loudness decay should be obtained in order to examine the relationships between these prosodic dimensions and the overall perceived loudness. With this perceptual information, acoustic characteristics can be explored that more clearly capture speech level variability than the measures included in the present investigation. Additionally, this would provide helpful insights into the effectiveness of active speech level as a measure that distinguishes IWPDs from HOAs, as this could be the result of pause behaviour, unstressed syllables, or other prosodic characteristics.

5.12 Conclusion

In summary, the results of this investigation provide support for the role of both speech level (e.g. mean intensity) and spectral balance (e.g. tilt) in listeners' judgments of overall perceived loudness of sentence-level and conversational speech of IWPDs and HOAs. Listeners provided consistent and reliable ratings of perceived loudness via visual analogue scales (VAS) and direct magnitude estimation (DME), supporting the use of either technique and helping to bridge between literature and fields that have used each technique. IWPDs and HOAs were observed to differ in the expected directions on perceived and acoustic measures, including perceived loudness, intelligibility, acoustic speech level and spectral balance. IWPDs were observed to be quieter than HOAs as measured by perceptual and acoustic measures, and demonstrated a relatively greater concentration of energy in the lower frequencies. Considerable variability existed among this large group of IWPDs, with larger group differences observed among low loudness IWPDs, who may present members of the PD population with a greater severity of hypophonia. Further research is needed to understand the relationships between speech level variability (e.g. standard deviation of intensity) and perceived loudness. This investigation also supports the use of time-varying loudness (TVL) as an acoustic model of loudness that provides a robust loudness estimate. Additionally, preliminary evidence is obtained that manipulations of spectral balance alter perceived loudness even in the presence of equal speech level, such that an increased proportion of energy above 2 kHz is associated with greater perceived loudness. This finding provides support for further exploration of spectral manipulations in the context of clinical speech amplification devices.

References

Abur, D., Enos, N. M., & Stepp, C. E. (2019). Visual analog scale ratings and orthographic transcription measures of sentence intelligibility in Parkinson's disease with variable listener exposure. *American Journal of Speech-Language Pathology*, *28*(3), 1222–1232. https://doi.org/10.1044/2019_AJSLP-18-0275

Abur, D., Lester-Smith, R. A., Daliri, A., Lupiani, A. A., Guenther, F. H., & Stepp, C. E. (2018). Sensorimotor adaptation of voice fundamental frequency in Parkinson's disease. *PLoS ONE*, *13*(1), e0191839. https://doi.org/10.1371/journal.pone.0191839

Abur, D., Lupiani, A. A., Hickox, A. E., Shinn-Cunningham, B. G., & Stepp, C. E. (2018a). Loudness perception of pure tones in Parkinson's disease. *Journal of Speech, Language, and Hearing Research*, *61*(6), 1487–1496. https://doi.org/10.1044/2018_JSLHR-H-17-0382

Adams, S., Dykstra, A., Abrams, K., Winnell, J., Jenkins, M., & Jog, M. (2006). Conversational speech intensity under different noise conditions in hypophonia and Parkinson's disease. *Canadian Acoustics*, *34*(3), 96–97.

Adams, S. G., & Dykstra, A. (2009). Hypokinetic dysarthria. In M. McNeil (Ed.), *Clinical Management of Sensorimotor Speech Disorders* (2nd ed., pp. 166–186). https://doi.org/10.1007/978-1-62703-556-9

Adams, S. G., Dykstra, A. D., Jenkins, M., & Jog, M. (2008). Speech-to-noise levels and conversational intelligibility in hypophonia and Parkinson's disease. *Journal of Medical Speech-Language Pathology*, *16*(4), 165–172.

Adams, S. G., Dykstra, A. D., & Jog, M. (2012). A comparison of throat and head microphones in a PDA-based evaluation of hypophonia in Parkinson's disease. *Journal of Medical Speech-Language Pathology*, *20*(4), 1–6.

Adams, S. G., Haralabous, O., Dykstra, A., Abrams, K., & Jog, M. (2005). Effects of multi-talker background noise on the intensity of spoken sentences in Parkinson's disease. *Canadian Acoustics*, *33*(3), 94–95.

Adams, S. G., & Lang, A. E. (1992). Can the Lombard effect be used to improve low voice intensity in Parkinson's disease? *European Journal of Disorders of Communication*, *27*, 121–127.

Adams, S. G., Winnell, J., & Jog, M. (2010). Effects of interlocutor distance, multi-talker background noise, and a concurrent manual task on speech intensity in Parkinson's disease. *Journal of Medical Speech-Language Pathology*, *18*(4), 1–8.

Adams, S., Moon, B.-H., Dykstra, A., Abrams, K., Jenkins, M., & Jog, M. (2006a). Effects of multitalker noise on conversational speech intensity in Parkinson's disease. *Journal of Medical Speech-Language Pathology*, *14*(4), 221–228.

Akinwande, M. O., Dikko, H. G., & Samson, A. (2015). Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis. *Open Journal of Statistics*, *05*(07), 754–767. https://doi.org/10.4236/ojs.2015.57075

Alharbi, G. G., Cannito, M. P., Buder, E. H., & Awan, S. N. (2019). Spectral/cepstral analyses of phonation in Parkinson's disease before and after voice treatment: A preliminary study. *Folia Phoniatrica et Logopaedica*, 1–11. https://doi.org/10.1159/000495837

Alvar, A. M., Lee, J., & Hubera, J. E. (2019). Filled pauses as a special case of automatic speech behaviors and the effect of Parkinson's disease. *American Journal of Speech-Language Pathology*, *28*(2 Special Issue), 835–843. https://doi.org/10.1044/2018_AJSLP-MSC18-18-0111

American National Standards Institute, Acoustical Society of America. (2007). Procedure for the computation of loudness of steady sounds (ANSI/ASA S3.4).

Anand, S., & Stepp, C. E. (2015). Listener perception of monopitch, naturalness, and intelligibility for speakers with Parkinson's disease. *Journal of Speech, Language, and Hearing Research*, *58*, 1134–1144. https://doi.org/10.1044/2015

Andreetta, M. D., Adams, S. G., Dykstra, A. D., & Jog, M. (2016). Evaluation of speech amplification devices in Parkinson's disease. *American Journal of Speech-Language Pathology*, *25*, 29–45. https://doi.org/10.1044/2015_AJSLP-15-0008

Appell, J.-E. (2002). *Loudness models for rehabilitative audiology*. Oldenburg University.

Arnold, C., Gehrig, J., Gispert, S., Seifried, C., & Kell, C. A. (2014). Pathomechanisms and compensatory efforts related to parkinsonian speech. *NeuroImage: Clinical*, *4*, 82–97. https://doi.org/10.1016/j.nicl.2013.10.016

Bain, C., Ferguson, A., & Mathisen, B. (2005). Effectiveness of the Speech Enhancer on intelligibility: A case study. *Journal of Medical Speech-Language Pathology*, *13*(2), 85–95.

Baker, K. K., Ramig, L. O., Luschei, E. S., & Smith, M. E. (1998). Thyroarytenoid muscle activity associated with hypophonia in Parkinson disease and aging. *Neurology*, *51*(6), 1592–1598. https://doi.org/10.1212/WNL.51.6.1592

Bandini, A., Giovannelli, F., Orlandi, S., Barbagallo, S. D., Cincotta, M., Vanni, P., Chiaramonti, R., Borgheresi, M., Zaccara, G., & Manfredi, C. (2015). Automatic identification of dysprosody in idiopathic Parkinson's disease. *Biomedical Signal Processing and Control*, *17*. https://doi.org/10.1016/j.bspc.2014.07.006

Baylor, C., Yorkston, K., Eadie, T., Kim, J., Chung, H., & Amtmann, D. (2013). The Communicative Participation Item Bank (CPIB): Item bank calibration and development of a disorder-generic short form. *Journal of Speech, Language, and Hearing Research*, *56*(August), 1190–1209. https://doi.org/10.1044/1092-4388(2012/12-0140)physical

Berke, G. S., Gerratt, B., Kreiman, J., & Jackson, K. (1999). Treatment of Parkinson hypophonia with percutaneous collagen augmentation. *Laryngoscope*, *109*(8), 1295–1299. https://doi.org/10.1097/00005537-199908000-00020

Biary, N., Pimental, P. A., & Langenberg, P. W. (1988). A double-blind trial of clonazepam in the treatment of Parkinsonian dysarthria. *Neurology*, *38*, 255–258.

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer, Version 6.1.16. Retrieved from http://www.praat.org/.

Brajot, F.-X., Shiller, D. M., & Gracco, V. L. (2016). Autophonic loudness perception in Parkinson's disease. *The Journal of the Acoustical Society of America*, *139*(3), 1364–1371. https://doi.org/10.1121/1.4944569

Brand, T., & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *The Journal of the Acoustical Society of America*, *112*(4), 1597–1604. https://doi.org/10.1121/1.1502902

Brandt, J. F., Ruder, K. F., & Shipp, T. (1969). Vocal loudness and effort in continuous speech. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.1911899

Brookes, M. (2020). VOICEBOX: Speech Processing Toolbox for MATLAB. Retrieved from http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1). https://doi.org/10.5334/joc.10

Bunton, K., & Weismer, G. (2002). Segmental level analysis of laryngeal function in persons with motor speech disorders. *Folia Phoniatrica et Logopaedica*, *54*, 223–239.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods and Research*. https://doi.org/10.1177/0049124104268644

Caligiuri, M. P., & Abbs, J. H. (1987). Response properties of the perioral reflex in Parkinson's disease. *Experimental Neurology*, *98*, 563–572. https://doi.org/10.1016/0014-4886(87)90265-2

Cannito, M. P., Suiter, D. M., Wolf, T., Chorna, L., Beverly, D., & Watkins, J. (2006). Vowel harmonic amplitude differences in a speaker with hypokinetic dysarthria before and after treatment. *Journal of Medical Speech-Language Pathology*, *14*(4), 229–234.

Cao, H., Xu, X., Zhao, Y., Long, D., & Zhang, M. (2011). Altered brain activation and connectivity in early Parkinson disease tactile perception. *American Journal of Neuroradiology*, *32*(10), 1969–1974. https://doi.org/10.3174/ajnr.A2672

Cardoso, R., Guimarães, I., Santos, H., Loureiro, R., Domingos, J., Abreu, D. de, … Ferreira, J. (2017). Frenchay dysarthria assessment (FDA-2) in Parkinson's disease: Cross-cultural adaptation and psychometric properties of the European Portuguese version. *Journal of Neurology*, *264*(1), 21–31. https://doi.org/10.1007/s00415-016-8298-6

Chen, Y.-W., & Watson, P. J. (2017). Speech production and sensory impairment in mild Parkinson's disease. *Journal of the Acoustical Society of America*, *141*(5), 3030–3041. https://doi.org/10.1121/1.4980138

Chiu, Y.-F., Neel, A., & Loux, T. (2020). Acoustic characteristics in relation to intelligibility reduction in noise for speakers with Parkinson's disease. *Clinical Linguistics & Phonetics*, 1–15. https://doi.org/10.1080/02699206.2020.1777585

Chou, K. L., Amick, M. M., Brandt, J., Camicioli, R., Frei, K., Gitelman, D., … Uc, E. Y. (2010). A recommended scale for cognitive screening in clinical trials of Parkinson's disease. *Movement Disorders*, *25*(15), 2501–2507. https://doi.org/10.1002/mds.23362

Clark, J. (2012). *Loudness perception and speech intensity control in Parkinson's disease*. University of Western Ontario.

Clark, J. P., Adams, S. G., Dykstra, A. D., Moodie, S., & Jog, M. (2014). *Loudness perception and speech intensity control in Parkinson's disease*. *Journal of Communication Disorders*, *51*, 1-12. https://doi.org/10.1016/j.jcomdis.2014.08.001

De Cock, V. C., Vidailhet, M., Leu, S., Texeira, A., Apartis, E., Elbaz, A., Roze, E., Willer, J.C., Derenne, J.P., Agid, Y., & Arnulf, I. (2007). Restoration of normal motor control in Parkinson's disease during rem sleep. *Brain*, *130*(2), 450–456.

Cohen, J. (1992). A power primer. *Quantitative Methods in Psychology*, *112*(1), 155–159.

Constantinescu, G., Theodoros, D., Russell, T., Ward, E., Wilson, S., & Wootton, R. (2010). Assessing disordered speech and voice in Parkinson's disease: A telerehabilitation application. *International Journal of Language and Communication Disorders, 45*(6), 630-644. https://doi.org/10.3109/13682820903470569

Constantinescu, G., Theodoros, D., Russell, T., Ward, E., Wilson, S., & Wootton, R. (2011). Treating disordered speech and voice in Parkinson's disease online: A randomized controlled non-inferiority trial. *International Journal of Language and Communication Disorders*, *46*(1), 1–16. https://doi.org/10.3109/13682822.2010.484848

Conte, A., Belvisi, D., Tartaglia, M., Cortese, F. N., Baione, V., Battista, E., Zhu, X. Y., Fabbrini, G., & Berardelli, A. (2017). Abnormal temporal coupling of tactile perception and motor action in Parkinson's disease. *Frontiers in Neurology*, *8*, 249. https://doi.org/10.3389/fneur.2017.00249

Conte, A., Khan, N., Defazio, G., Rothwell, J. C., & Berardelli, A. (2013). Pathophysiology of somatosensory abnormalities in parkinson disease. *Nature Reviews Neurology*, *9*, 687–697.

Corbett, I. (2015). *Mic it! Microphones, microphone techniques, and their impact on the final mix*. Routledge. https://doi.org/10.4324/9780203549780

Corcoran, P., Hensman, A., & Kirkpatrick, B. (2019). Glottal flow analysis in parkinsonian speech. In *BIOSIGNALS 2019 - 12th International Conference on Bio-Inspired Systems and Signal Processing, Proceedings; Part of 12th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2019* (pp. 116–123). https://doi.org/10.5220/0007259701160123

Cruz, A. N. da, Beber, B. C., Olchik, M. R., Chaves, M. L. F., de Mello Rieder, C. R., & Dornelles, S. (2016). Aspects of oral communication in patients with Parkinson's disease submitted to deep brain stimulation. *CoDAS*, *28*(4), 480–485. https://doi.org/10.1590/2317-1782/20162015169

Cucinotta, D., & Vanelli, M. (2020). WHO declares COVID-19 a pandemic. *Acta Biomedica*, *91*(1), 157-160. https://doi.org/10.23750/abm.v91i1.9397

Curran-Everett, D. (2013). Explorations in statistics: The analysis of ratios and normalized data. *Advances in Physiology Education*, *37*, 213–219. https://doi.org/10.1152/advan.00053.2013

Cushnie-Sparrow, D., Adams, S., Abeyesekera, A., Pieterman, M., Gilmore, G., & Jog, M. (2018). Voice quality severity and responsiveness to levodopa in Parkinson's disease. *Journal of Communication Disorders*, *76*, 1–10. https://doi.org/10.1016/j.jcomdis.2018.07.003

Cushnie-Sparrow, D., Adams, S., Knowles, T., Leszcz, T. M., & Jog, M. (2016). Effects of multi-talker noise on the acoustics of voiceless stop consonants in Parkinson ' s disease, *3*(1).

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research*, *12*, 462–496. https://doi.org/10.1044/jshr.1203.462

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969a). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*. https://doi.org/10.1044/jshr.1202.246

Desmurget, M., Grafton, S. T., Vindras, P., Grea, H., & Turner, R. S. (2004). Basal ganglia network mediates the control of movement amplitude. *Experimental Brain Research*, *153*(2), 197–209. https://doi.org/10.1007/s00221-003-1593-3

Donovan, N. J., Velozo, C. A., & Rosenbek, J. C. (2007). The Communicative Effectiveness Survey: Investigating its item-level psychometric properties. *Journal of Medical Speech-Language Pathology*, *15*(4), 443–447.

Dromey, C. (2003). Spectral measures and perceptual ratings of hypokinetic dysarthria. *Journal of Medical Speech-Language Pathology*, *11*(2), 85–94.

Dromey, C., & Adams, S. G. (2000). Loudness perception and hypophonia in Parkinson disease. *Journal of Medical Speech-Language Pathology*, *8*(4), 255–259.

Dromey, C., Ramig, L. O., & Johnson, A. B. (1995). Phonatory and articulatory changes associated with increased vocal intensity in Parkinson disease: A case study. *Journal of Speech, Language, and Hearing Research*, *38*(4), 751-764. https://doi.org/10.1044/jshr.3804.751

Dromey, C., & Ramig, L. O. (1998). Intentional changes in sound pressure level and rate: Their impact on measures of respiration, phonation, and articulation. *Journal of Speech, Language, and Hearing Research*, *41*(5), 1003-1018. https://doi.org/10.1044/jslhr.4105.1003

Duvvuru, S., & Erickson, M. (2013). The effect of change in spectral slope and formant frequencies on the perception of loudness. *Journal of Voice*, *27*(6), 691–697. https://doi.org/10.1016/j.jvoice.2013.05.004

Dykstra, A. D., Adams, S. G., & Jog, M. (2013). Examining the conversational speech intelligibility of individuals with hypophonia associated with Parkinson's disease. *Journal of Medical Speech-Language Pathology*, *20*(4), 53–57.

Eadie, T. L., & Kapsner-Smith, M. (2011). The effect of listener experience and anchors on judgments of dysphonia. *Journal of Speech, Language, and Hearing Research*, *54*(2), 430–447. https://doi.org/10.1044/1092-4388(2010/09-0205)

Epstein, M., & Florentine, M. (2006). Loudness of brief tones measured by magnitude estimation and loudness matching. *The Journal of the Acoustical Society of America*, *119*(4), 1943–1945. https://doi.org/10.1121/1.2177592

Eriksson, A., & Traunmuller, H. (2002). Perception of vocal effort and distance from the speaker on the basis of vowel utterances. *Perception & Psychophysics*, *64*(1), 131–139.

Evans, C., Canavan, M., Foy, C., Langford, R., & Proctor, R. (2012). Can group singing provide effective speech therapy for people with Parkinson's disease? *Arts and Health*, *4*(1), 83–95. https://doi.org/10.1080/17533015.2011.584883

Fahn, S., & The Parkinson Study Group. (2004). Levodopa and the progression of Parkinson's disease. *The New England Journal of Medicine*, *351*(24), 2498–2508. https://doi.org/10.1056/NEJMoa033447

Fahn, S., & Poewe, W. (2015). Levodopa: 50 years of a revolutionary drug for Parkinson disease. *Movement Disorders*, *30*(1), 1–3. https://doi.org/10.1002/mds.26122

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behaviour Research Methods*, *39*(2), 175–191.

Fletcher, H., & Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *The Journal of the Acoustical Society of America*, *5*(2), 82–108. https://doi.org/10.1121/1.1915637

Florentine, M. (2011). Loudness. In M. Florentine, A. N. Popper, & R. R. Fay (Eds.), *Loudness* (pp. 1–16). Springer.

Folmer, R. L., Vachhani, J. J., Theodoroff, S. M., Ellinger, R., & Riggins, A. (2017). Auditory processing abilities of Parkinson's disease patients. *BioMed Research International*, *2017*, 2618587. https://doi.org/10.1155/2017/2618587

Ford, M. P., Malone, L. A., Nyikos, I., Yelisetty, R., & Bickel, C. S. (2010). Gait training with progressive external auditory cueing in persons with Parkinson's disease. *Archives of Physical Medicine and Rehabilitation*, *91*(8), 1255–1261. https://doi.org/10.1016/j.apmr.2010.04.012

Fox, C. M., Morrison, C. E., Ramig, L. O., & Sapir, S. (2002). Current perspectives on the Lee Silverman Voice Treatment (LSVT) for individuals with idiopathic Parkinson disease.

American Journal of Speech-Language Pathology, *11*(2), 111–123. https://doi.org/10.1044/1058-0360(2002/012)

Fox, C., & Ramig, L. (1997). Vocal sound pressure level and self-perception of speech and voice in men and women with idiopathic Parkinson disease. *American Journal of Speech-Language Pathology*, *6*, 85–94. Retrieved from http://ajslp.pubs.asha.org/article.aspx?articleid=1774583

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's disease. *Journal of Speech, Language, and Hearing Research*, *44*, 1284–1299.

Gauffin, J., & Sundberg, J. (1989). Spectral correlates of glottal voice source waveform characteristics. *Journal of Speech and Hearing Research*, *32*(3), 556–565. https://doi.org/10.1044/jshr.3203.556

Gescheider, G. A. (1997). *Psychophysics: The fundamentals* (3rd ed.). Lawrence Elbaum.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47,* 103–138.

Glave, R. D., & Rietveld, A. C. M. (1975). Is the effort dependence of speech loudness explicable on the basis of acoustical cues? *The Journal of the Acoustical Society of America*, *58*(4), 875. https://doi.org/10.1121/1.380737

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P.A., Nyenhuis, D., Olanow, W., Rascol, O., Schrag, A., Teresi, J. A., van Hilten, J. J., LaPelle, N. for the Movement Disorder Society UPDRS Revision Task Force. (2008). Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders*, *23*(15), 2129–2170. https://doi.org/10.1002/mds.22340

Goldman, J. G., Williams-Gray, C., Barker, R. A., Duda, J. E., & Galvin, J. E. (2014). The spectrum of cognitive impairment in Lewy body diseases. *Movement Disorders*, *29*(5), 608–621. https://doi.org/10.1002/mds.25866

Gordon, K. R. (2019). How mixed-effects modeling can advance our understanding of learning and memory and improve clinical and educational practice. *Journal of Speech, Language, and Hearing Research*, *62*(3), 507–524. https://doi.org/10.1044/2018_JSLHR-L-ASTM-18-0240

Government of Canada and Neurological Health Charities Canada. *Mapping Connections: An understanding of neurological conditions in Canada*. Ottawa (ON): Public Health Agency of Canada; 2014, Report No.: 140100. Available from: http://www.phac-aspc.gc.ca/publicat/cd-mc/mc-ec/assets/pdf/mc-ec-eng.pdf.

Gulberti, A., Hamel, W., Buhmann, C., Boelmans, K., Zittel, S., Gerloff, C., Westphal, M., Engel, A. K., Schneider, T. R., & Moll, C. K. E. (2015). Subthalamic deep brain stimulation improves auditory sensory gating deficit in Parkinson's disease. *Clinical Neurophysiology*, *126*(3), 565–574. https://doi.org/10.1016/j.clinph.2014.06.046

Halpern, A. E., Ramig, L. O., Matos, C. E. C., Petska-Cabl, J. A., Spielman, J. L., Pogoda, J. M., Gilley, P. M., Sapir, S., Bennett, J. K., & McFarland, D. H. (2012). Innovative technology for the assisted delivery of intensive voice treatment (LSVT® LOUD) for Parkinson disease. *American Journal of Speech-Language Pathology*, *21*(4), 354–367. https://doi.org/10.1044/1058-0360(2012/11-0125)

Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg', J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Oto-Laryngologica, 90,* 441–451.

Hammen, V. L., & Yorkston, K. M. (1996). Speech and pause characteristics following speech rate reduction in hypokinetic dysarthria. *Journal of Communication Disorders*, *29*(6), 429–444.

Hammer, M. J., & Barlow, S. M. (2010). Laryngeal somatosensory deficits in Parkinson's disease: Implications for speech respiratory and phonatory control. *Experimental Brain Research*, *201*(3), 401–409. https://doi.org/10.1007/s00221-009-2048-2

Hammer, M. J., Barlow, S. M., Lyons, K. E., & Pahwa, R. (2011). Subthalamic nucleus deep brain stimulation changes speech respiratory and laryngeal control in Parkinson's disease. *Journal of Neurology*, *257*(10), 1692–1702. https://doi.org/10.1007/s00415-010-5605-5

Hanson, D. G., Gerratt, B. R., & Ward, P. H. (1984). Cinegraphic observations of laryngeal function in Parkinson's disease. *Laryngoscope*, *94*, 348–353. https://doi.org/10.1288/00005537-198403000-00011

Hellman, R. P. (1999). Cross-modality matching: A tool for measuring loudness in sensorineural impairment. *Ear & Hearing*, *20*(3), 193–213.

Helou, L. B., Solomon, N. P., Henry, L. R., Coppit, G. L., Howard, R. S., & Stojadinovic, A. (2010). The role of listener experience on consensus auditory-perceptual evaluation of voice (CAPE-V) ratings of postthyroidectomy voice. *American Journal of Speech-Language Pathology*, *19*, 248–258.

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

Ho, A. K., Bradshaw, J. L., & Iansek, R. (2000). Volume perception in parkinsonian speech. *Movement Disorders*, *15*(6), 1125–1131. https://doi.org/10.1002/1531-8257(200011)15:6<1125::AID-MDS1010>3.0.CO;2-R

Ho, A. K., Bradshaw, J. L., & Iansek, R. (2008). For better or worse: The effect of levodopa on speech in Parkinson's disease. *Movement Disorders*, *23*(4), 574–580. https://doi.org/10.1002/mds.21899

Ho, A. K., Bradshaw, J. L., Iansek, R., & Alfredson, R. (1999). Speech volume regulation in Parkinson's disease: Effects of implicit cues and explicit instructions. *Neuropsychologia*, *37*, 1453–1460. https://doi.org/10.1016/S0028-3932(99)00067-6

Ho, A. K., Iansek, R., & Bradshaw, J. L. (1999a). Regulation of parkinsonian speech volume: The effect of interlocuter distance. *Journal of Neurology Neurosurgery and Psychiatry*, *67*, 199–202. https://doi.org/10.1136/jnnp.67.2.199

Ho, A. K., Iansek, R., & Bradshaw, J. L. (2001). Motor instability in parkinsonian speech intensity. *Neuropsychiatry, Neuropsychology and Behavioral Neurology*, *14*(2), 109–116.

Howell, S., Tripoliti, E., & Pring, T. (2009). Delivering the lee silverman voice treatment (lsvt) by web camera: A feasibility study. *International Journal of Language & Communication Disorders*, *44*(3), 287–300. https://doi.org/10.1080/13682820802033968

Huber, J. E., & Darling, M. (2011). Effect of Parkinson's disease on the production of structured and unstructured speaking tasks: Respiratory physiologic and linguistic considerations. *Journal of Speech, Language, and Hearing Research*, *54*, 33–46.

Huber, J. E., Darling, M., Francis, E. J., & Zhang, D. (2012). Impact of typical aging and Parkinson's disease on the relationship among breath pausing, syntax, and punctuation. *American Journal of Speech-Language Pathology*, *21*, 368–379.

Hunter, E. J., Cantor-Cutiva, L. C., Leer, E. van, Mersbergen, M. van, 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. https://doi.org/10.1044/2019_JSLHR-19-00057

International Organization for Standardization. (2017). *Acoustics — Methods for calculating loudness – Part 2: Moore-Glasberg method* (ISO Standard No. 532-2).

International Telecommunication Union. (2015). *Algorithms to measure audio programme loudness and true-peak audio level* (ITU-R BS. 1770-4).

International Telecommunication Union. (2011). *Objective measurement of speech level* (ITU-T P.56 (12/2011)).

Isshiki, N. (1964). Regulatory mechanism of voice intensity variation. *Journal of Speech, Language, and Hearing Research*, *7*, 17–29.

Jacobson, B. H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., Benninger, M. S., & Newman, C. W. (1997). The voice handicap index (vhi): Development and validation. *American Journal of Speech Language Pathology*, *6*(3), 66–70.

Jankovic, J. (2008). Parkinson's disease: Clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, *79*, 368–376. https://doi.org/10.1136/jnnp.2007.131045

Jesteadt, W., & Joshi, S. N. (2013). Reliability of procedures used for scaling loudness. In *Proceedings of Meetings on Acoustics* (Vol. 19). https://doi.org/10.1121/1.4799962

Jesteadt, W., & Leibold, L. J. (2011). Loudness in the laboratory, Part I: Steady-state sounds. In M. Florentine, A. N. Popper, & R. R. Fay (Eds.), *Loudness* (pp. 109–144). Springer.

Johnson, A. M., & Adams, S. G. (2006). Nonpharmacological management of Parkinson's disease. *Geriatrics and Aging*, *9*(1), 40–43.

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

Kearney, E., Giles, R., Haworth, B., Faloutsos, P., Baljko, M., & Yunusova, Y. (2017). Sentence-level movements in Parkinson's disease: Loud, clear, and slow speech. *Journal of Speech, Language, and Hearing Research*, *60*(December), 1–14. https://doi.org/10.1044/2017_JSLHR-S-17-0075

De Keyser, K., Santens, P., Bockstael, A., Botteldooren, D., Talsma, D., De Vos, S., Van Cauwenberghe, M., Verheugen, F., Corthals, P., & De Letter, M. (2016). The relationship between speech production and speech perception deficits in Parkinson's disease. *Journal of Speech, Language, and Hearing Research*, *59*(5), 915–931. https://doi.org/10.1044/2016_JSLHR-S-15-0197

Knowles, T., Adams, S., Abeyesekera, A., Mancinelli, C., Gilmore, G., & Jog, M. (2018). Deep brain stimulation of the subthalamic nucleus parameter optimization for vowel acoustics and speech intelligibility in Parkinson's disease. *Journal of Speech Language and Hearing Research*, *61*(February), 510–524. https://doi.org/https://doi.org/10.1044/2017_JSLHR-S-17-0157

Knowles, T., Adams, S. G., Page, A., Cushnie-Sparrow, D., & Jog, M. (2020). A comparison of speech amplification and personal communication devices for hypophonia. *Journal of Speech, Language, and Hearing Research*, *63*(8), 2695–2712. https://doi.org/10.23641/asha

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. https://doi.org/10.1016/j.jcm.2016.02.012

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*, 21–40. https://doi.org/doi:10.1044/jshr.3601.21

Kreiman, J., Gerratt, B. R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research*, *33*(1), 103–115. https://doi.org/10.1121/1.2025887

Kuhn, M., & Johnson, K. (2018). *Applied predictive modeling* (1st ed.). Springer.

Kummer, A. W. (2020). *Cleft palate and craniofacial conditions : A comprehensive guide to clinical management.* (Fourth edition.). Jones & Bartlett.

Kuruvilla-Dugdale, M., Threlkeld, K., Salazar, M., Nolan, G., & Heidrick, L. (2019). A comparative study of auditory-perceptual speech measures for the early detection of mild speech impairments. *Seminars in Speech and Language*, *40*(5), 394–405. https://doi.org/10.1055/s-0039-1694997

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Lam, J., & Tjaden, K. (2016). Clear speech variants: An acoustic study in Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 59,* 631-646. https://doi.org/10.1044/2015_JSLHR-S-15-0216

Lane, H. L., Catania, A. C., & Stevens, S. S. (1961). Voice level: Autophonic scale, perceived loudness, and effects of sidetone. *The Journal of the Acoustical Society of America*, *33*(2), 160–167. https://doi.org/10.1121/1.1908608

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research*, *14*, 677–709.

Lansford, K. L., & Liss, J. M. (2014). Vowel acoustics in dysarthria: Speech disorder diagnosis and classification. *Journal of Speech Language and Hearing Research*, *57*(1), 57. https://doi.org/10.1044/1092-4388(2013/12-0262)

Lei, Y., Ozdemir, R. A., & Perez, M. A. (2018). Gating of sensory input at subcortical and cortical levels during grasping in humans. *The Journal of Neuroscience*, *38*(33), 7237–7247. https://doi.org/10.1523/jneurosci.0545-18.2018

Liénard, J.-S., & Benedetto, M.-G. D. (1999). Effect of vocal effort on spectral properties of vowels. *The Journal of the Acoustical Society of America*, *106*(1), 411–422. https://doi.org/10.1121/1.428140

Liu, H., Wang, E. Q., Metman, L. V., & Larson, C. R. (2012). Vocal responses to perturbations in voice auditory feedback in individuals with Parkinson’s disease. *PLoS ONE*, *7*(3), e33629. https://doi.org/10.1371/journal.pone.0033629

Logemann, J. A., Fisher, H. B., Boshes, B., & Blonsky, E. R. (1978). Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *Journal of Speech and Hearing Disorders*, *43*, 47–57. https://doi.org/10.1044/jshd.4301.47

Lombard, E. (1911). Le signe de l’elevation de la voix. *Annales Des Maladies de L’Oreille et Du Larynx*, *37*(2), 101–109.

Ludlow, C. L., & Bassich, C. J. (194AD). Relationships between perceptual ratings and acoustic measures of hypokinetic speech. In M. R. McNeil, J. C. Rosenbek, & A. E. Aronson (Eds.), *The Dysarthrias: Physiology, Acoustics, Perception, Management* (pp. 163–192). College Hill.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, *23*, 151–169. https://doi.org/10.1146/annurev.publheath.23.100901.140546

Ma, J. K.-Y., Schneider, C. B., Hoffmann, R., & Storch, A. (2015). Speech prosody across stimulus types for individuals with Parkinson’s disease. *Journal of Parkinson’s Disease*, *5*(2), 291–299. https://doi.org/10.3233/JPD-140451

Ma, J. K.-Y., Whitehill, T., & Cheung, K. S.-K. (2010). Dysprosody and stimulus effects in cantonese speakers with Parkinson’s disease. *International Journal of Language and Communication Disorders*, *45*(6), 645–655.

Marks, L. E. (1968). *Stimulus-range, number of categories, and form of the category-scale. Source: The American Journal of Psychology* (Vol. 81, pp. 467–479). Retrieved from https://about.jstor.org/terms

Marks, L. E., & Florentine, M. (2011). Measurement of loudness, part i: Methods, problems, and pitfalls. In M. Florentine, A. N. Popper, & R. R. Fay (Eds.), *Loudness* (pp. 17–56). Springer.

Marshall, A., & David, P. (2007). A semantic differential study of low amplitude supersonic aircraft noise and other transient sounds. *International Congress on Acoustics*, Madrid.

Martínez-Sánchez, F., Meilán, J. J. G., Carro, J., Íñiguez, C. G., Millian-Morell, L., Pujante Valverde, I. M., López-Alburquerque, T., & López, D. E. (2016). Speech rate in Parkinson's disease: A controlled study. *Neurologia, 31*(7), 466-472.

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. https://doi.org/10.1016/j.jvoice.2014.06.015

Matheron, D., Stathopoulos, E. T., Huber, J. E., & Sussman, J. E. (2017). Laryngeal aerodynamics in healthy older adults and adults with Parkinson's disease. *Journal of Speech, Language, and Hearing Research*, *60*(3), 1–15. https://doi.org/DOI: 10.1044/2016_JSLHR-S-14-0314

MATLAB. (2020). Version 9.8.0.1380330 (R2020a). Natick, Massachusetts: The MathWorks Inc.

McCaig, C. M., Adams, S. G., Dykstra, A. D., & Jog, M. (2016). Effect of concurrent walking and interlocutor distance on conversational speech intensity and rate in Parkinson's disease. *Gait & Posture*, *43*, 132–136. https://doi.org/10.1016/j.gaitpost.2015.09.011

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). https://doi.org/10.1044/2018_JSLHR-S-18-0205

McKenna, V. S., Murray, E. S., Lien, Y. A. 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. https://doi.org/10.1044/2016_JSLHR-S-15-0406

McKenna, V. S., & Stepp, C. E. (2018). The relationship between acoustical and perceptual measures of vocal effort. *The Journal of the Acoustical Society of America*, *144*(3), 1643–1658. https://doi.org/10.1121/1.5055234

McRobert, H., Bryan, M. E., & Tempest, W. (1965). Magnitude estimation of loudness. *Journal of Sound and Vibration*, *2*(4), 391–401. https://doi.org/10.1016/0022-460X(65)90117-3

Miller, N., Allcock, L., Jones, D., Noble, E., Hildreth, A. J., & Burn, D. J. (2007). Prevalence and pattern of perceived intelligibility changes in Parkinson's disease. *Journal of Neurology, Neurosurgery and Psychiatry*, *78*(11), 1188–1190. https://doi.org/10.1136/jnnp.2006.110171

Mollaei, F., Shiller, D. M., & Gracco, V. L. (2013). Sensorimotor adaptation of speech in Parkinson's disease. *Movement Disorders*, *28*(12). https://doi.org/10.1002/mds.25588

Moore, B. C. J., Füllgrabe, C., & Stone, M. A. (2010). Effect of spatial separation, extended bandwidth, and compression speed on intelligibility in a competing-speech task. *The Journal of the Acoustical Society of America*, *128*(1), 360–371. https://doi.org/10.1121/1.3436533

Moore, B. C. J., Glasberg, B. R., & Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, *45*(4), 224–240. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=10272

Moore, B. C. J., Glasberg, B. R., Varathanathan, A., & Schlittenlacher, J. (2016). A loudness model for time-varying sounds incorporating binaural inhibition. *Trends in Hearing*, *20*, 1–16.

Moore, B. C. J., Jervis, M., Harries, L., & Schlittenlacher, J. (2018). Testing and refining a loudness model for time-varying sounds incorporating binaural inhibition. *The Journal of the Acoustical Society of America*, *143*(3), 1504–1513. https://doi.org/10.1121/1.5027246

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Neel, A. T. (2009). Effects of loud and amplified speech on sentence and word intelligibility in Parkinson disease. *Journal of Speech, Language, and Hearing Research*, *52*(4), 1021–1033. https://doi.org/10.1044/1092-4388(2008/08-0119)

Obeso, J. A., Rodriguez-Oroz, M. C., Goetz, C. G., Marin, C., Kordower, J. H., Rodriguez, M., Hirsch, E. C., Farrer, M., Schapira, A. H. V., & Halliday, G. (2010). Missing pieces in the Parkinson's disease puzzle. *Nature Medicine*, *16*(6), 653–661. https://doi.org/10.1038/nm.2165

Okun, M. S. (2012). Deep-brain stimulation for Parkinson's disease. *New England Journal of Medicine*, *367*(16), 1529–1538. https://doi.org/10.1056/NEJMct1208070

Oliveira, R. M., Gurd, J. M., Nixon, P., & Marshall, J. C. (1997). Micrographia in Parkinson's disease: The effect of providing external cues. *Journal of Neurology, Neurosurgery & Psychiatry*, *63*, 429–433.

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*(August), 887–905.

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., & Milroy, R. (1982). The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *The Journal of the Acoustical Society of America*, *72*(6), 1788–1803. https://doi.org/10.1121/1.388652

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ramig, L., Halpern, A., Spielman, J., Fox, C., & Freeman, K. (2018). Speech treatment in Parkinson's disease: Randomized controlled trial (RCT). *Movement Disorders*, *33*(11), 1777–1791. https://doi.org/10.1002/mds.27460

Ramig, L. O., Countryman, S., O'Brien, C., Hoehn, M., & Thompson, L. (1996). Intensive speech treatment for patients with Parkinson's disease: Short- and long-term comparison of two techniques. *Neurology*, *47*(6), 1496–1504. https://doi.org/10.1212/WNL.47.6.1496

Ramig, L. O., Countryman, S., Thompson, L., & Horii, Y. (1995). Comparison of two forms of intensive speech treatment for Parkinson disease. *Journal of Speech and Hearing Research*, *38*, 1232–1251.

Ramig, L. O., Pawlas, A. A., & Countryman, S. (1995a). *The Lee Silverman Voice Treatment: A practical guide to treating the voice and speech disorders in Parkinson disease*. National Center for Voice; Speech.

Ramig, L. O., Sapir, S., Countryman, S., Pawlas, A. A., O'brien, C., Hoehn, M., & Thompson, L. L. (2001). Intensive voice treatment (LSVT®) for patients with Parkinson's disease: A 2 year follow up. *Journal of Neurology, Neurosurgery & Psychiatry*, *71*, 493–498. https://doi.org/10.1136/jnnp.71.4.493

Rascol, O., Sabatini, U., Chollet, F., Celsis, P., Montastruc, J. L., Marc-Vergnes, J. P., & Rascol, A. (1992). Supplementary and primary sensory motor area activity in Parkinson's disease: Regional cerebral blood flow changes during finger movements and effects of apomorphine. *Archives of Neurology*, *49*(2), 144–148. https://doi.org/10.1001/archneur.1992.00530260044017

Rasetshwane, D. M., Trevino, A. C., Gombert, J. N., Liebig-Trehearn, L., Kopun, J. G., Jesteadt, W., Neely, S. T., & Gorga, M. P. (2015). Categorical loudness scaling and equal-loudness contours in listeners with normal hearing and hearing loss. *The Journal of the Acoustical Society of America*, *137*(4), 1899–1913. https://doi.org/10.1121/1.4916605

Rennies, J., Holube, I., & Verhey, J. L. (2013). Loudness of speech and speech-like signals. *Acta Acustica United with Acustica*. https://doi.org/10.3813/AAA.918609

Reyno-Briscoe, K. A. (1997). *Acoustic correlates of monopitch and monoloudness in Parkinson's disease*. University of Western Ontario.

Richardson, K. C., & Sussman, J. E. (2019). Intensity resolution in individuals with Parkinson's disease: Sensory and auditory memory limitations. *Journal of Speech, Language, and Hearing Research*, *62*(9), 3564–3581. https://doi.org/10.1044/2019_JSLHR-H-18-0424

Ronen, Y. (2015). Vocal clarity in the mix: Techniques to improve the intelligibility of vocals. In *Audio Engineering Society 139th Convention*. Retrieved from http://www.aes.org/e-lib.

Rosen, K. M., Kent, R. D., & Duffy, J. R. (2005). Task-based profile of vocal intensity decline in Parkinson's disease. *Folia Phoniatrica et Logopaedica*, *57*(1), 28–37. https://doi.org/10.1159/000081959

Rosenthal, A. L., Lowell, S. Y., & Colton, R. H. (2014). Aerodynamic and acoustic features of vocal effort. *Journal of Voice*, *28*(2), 144–153. https://doi.org/10.1016/j.jvoice.2013.09.007

Rusz, J., Čmejla, R., Růžičková, H., Klempíř, J., Majerová, V., Picmausová, J., … Růžička, E. (2013). Evaluation of speech impairment in early stages of Parkinson's disease: A prospective study with the role of pharmacotherapy. *Journal of Neural Transmission*, *120*(2), 319–329. https://doi.org/10.1007/s00702-012-0853-4

Rusz, J., Tykalova, T., Ramig, L. O., & Tripoliti, E. (2021). Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders. *Movement Disorders, 36*(4), 803-814. https://doi.org/10.1002/mds.28465

Sadagopan, N., & Huber, J. E. (2007). Effects of loudness cues on respiration in individuals with Parkinson's disease. *Movement Disorders*, *22*(5), 651–659. https://doi.org/10.1002/mds.21375

Sapir, S. (2014). Multiple factors are involved in the dysarthria associated with Parkinson's disease: A review with implications for clinical practice and research. *Journal of Speech Language and Hearing Research*, *57*, 1330–1343. https://doi.org/10.1044/2014_JSLHRS130039

Sapir, S., Ramig, L. O., & Fox, C. M. (2011). Intensive voice treatment in Parkinson's disease: Lee Silverman Voice Treatment. *Expert Review of Neurotherapeutics*, *7*(5), 1–15.

Sapir, S., Spielman, J. L., Ramig, L. O., Story, B. H., & Fox, C. (2007). Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: Acoustic and perceptual findings. *Journal of Speech Language and Hearing Research*, *50*(4), 899. https://doi.org/10.1044/1092-4388(2007/064)

Schalling, E., Gustafsson, J., Ternström, S., Wilén, F. B., & Södersten, M. (2013). Effects of tactile biofeedback by a portable voice accumulator on voice sound level in speakers with Parkinson's disease. *Journal of Voice*, *27*(6), 729–737. https://doi.org/10.1016/j.jvoice.2013.04.014

Schliesser, H. F. (1985). Psychological scaling of speech by students in training compared to that by experienced speech-language pathologists. *Perceptual and Motor Skills*, *61*, 1299–1302. https://doi.org/10.2466/pms.1985.61.3f.1299

Schneider, J. S., Diamond, S. G., & Markham, C. H. (1986). Deficits in orofacial sensorimotor function in Parkinson's disease. *Annals of Neurology*, *19*, 275–282. https://doi.org/10.1002/ana.410190309

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means, *34*(4), 216–221.

Senthinathan, A., Adams, S., Page, A. D., & Jog, M. (2021). Speech intensity response to altered intensity feedback in individuals with Parkinson's disease. *Journal of Speech, Language, and Hearing Research*, *April, published online ahead of issue*.

El Sharkawi, A., Ramig, L. O., Pauloski, B. R., Rademaker, A. W., Smith, C. H., Pawlas, A., Baum, S., & Werner, C. (2002). Swallowing and voice effects of Lee Silverman Voice Treatment (LSVT): A pilot study. *Journal of Neurology, Neurosurgery & Psychiatry*, *72*, 31–36.

Sluijter, A. M. C., & Heuven, V. J. van. (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America*, *100*(4), 2471–2485. https://doi.org/10.1121/1.417955

Sluijter, A. M. C., Heuven, V. J. van, & Pacilly, J. J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. *The Journal of the Acoustical Society of America*, *101*(1), 503–513. https://doi.org/10.1121/1.417994

Smith, L. K., & Goberman, A. M. (2014). Long-time average spectrum in individuals with Parkinson disease. *NeuroRehabilitation*. https://doi.org/10.3233/NRE-141102

Solomon, N. P., & Hixon, T. J. (1993). Speech breathing in Parkinson's disease. *Journal of Speech and Hearing Research*, *36*(2), 294–310.

Solomon, N. P., & Robin, D. A. (2005). Perceptions of effort during handgrip and tongue elevation in Parkinson's disease. *Parkinsonism and Related Disorders*, *11*, 353–361. https://doi.org/10.1016/j.parkreldis.2005.06.004

Sonderegger, M., Wagner, M., & Torreira, F. (2018). *Quantitative methods for linguistic data*. Available online at: http://people.linguistics.mcgill.ca/ morgan/book/.

Soulodre, G. A. (2004). *Evaluation of objective loudness meters*. Retrieved from www.aes.org.

Spencer, K. A., Morgan, K. W., & Blond, E. (2009). Dopaminergic medication effects on the speech of individuals with Parkinson's disease. *Journal of Medical Speech-Language Pathology*, *17*(3), 125–144.

Stathopoulos, E. T., Huber, J. E., Richardson, K., Kamphaus, J., DeCicco, D., Darling, M., Fulcher, K., & Sussman, J. E. (2014). Increased vocal intensity due to the Lombard effect in speakers with Parkinson's disease: Simultaneous laryngeal and respiratory strategies. *Journal of Communication Disorders*, *48*, 1–17. https://doi.org/10.1016/j.jcomdis.2013.12.001

Stevens, S. S. (1946). *On the theory of scales of measurement*. *New Series* (Vol. 103, pp. 677–680).

Stevens, S. S. (1955). The measurement of loudness. *Journal of the Acoustical Society of America*, *27*(5), 815–829. https://doi.org/10.1121/1.1908048

Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *Source: The American Journal of Psychology, Autumn-Winter*, *69*(1), 664–689. Retrieved from https://about.jstor.org/terms

Suzuki, Y., & Takeshima, H. (2004). Equal-loudness-level contours for pure tones. *The Journal of the Acoustical Society of America*, *116*(2), 918–933. https://doi.org/10.1121/1.1763601

Szabo, A., & Hammarberg, B. (2013). Speaking fundamental frequency and phonation time during work and leisure time in vocally healthy preschool teachers measured with a voice accumulator. *Folia Phoniatrica et Logopaedica*, *65*, 84–90. https://doi.org/10.1159/000354673

Švec, J. G., & Granqvist, S. (2018). Tutorial and guidelines on measurement of sound pressure level (SPL) in voice and speech. *Journal of Speech Language and Hearing Research*, *61*(March), 441–461.

Tamburin, S., Fiaschi, A., Idone, D., Lochner, P., Manganotti, P., & Zanette, G. (2003). Abnormal sensorimotor integration is related to disease severity in Parkinson's disease: A TMS study. *Movement Disorders*, *18*(11), 1316–1324.

Theodoros, D. G., Constantinescu, G., Russell, T. G., Ward, E. C., Wilson, S. J., & Wootton, R. (2006). Treating the speech disorder in Parkinson's disease online. *Journal of Telemedicine and Telecare*, *12*(S3), 88–91. https://doi.org/10.1258/135763306779380101

Tindall, L. R., Huebner, R. A., Stemple, J. C., & Kleinert, H. L. (2008). Videophone-delivered voice therapy: A comparative analysis of outcomes to traditional delivery for adults with Parkinson's disease. *Telemedicine and E-Health*, *14*(10), 1070–1077. https://doi.org/10.1089/tmj.2008.0040

Titze, I. R. (2020). The effect of single harmonic tuning on vocal loudness. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2020.02.018

Titze, I. R., Hunter, E. J., Švec, J. G., Titze, I. R., & Hunter, E. J. (2007). Voicing and silence periods in daily and weekly vocalizations of teachers voicing and silence periods in daily and weekly vocalizations. *Journal of the Acoustical Society of America, 121*(1), 469-478. https://doi.org/10.1121/1.2390676

Titze, I.R., & Palaparthi, A. (2020). Vocal loudness variation with spectral slope. *Journal of Speech, Language, and Hearing Research, 63*(1), 74-82. https:://doi.org/10.1044/2019_JSLHR-19-00018

Tjaden, K., Lam, J., & Wilding, G. (2013). Vowel acoustics in Parkinson's disease and multiple sclerosis: Comparison of clear, loud, and slow speaking conditions. *Journal of Speech, Language, and Hearing Research*, *56*(5), 1485–1502. https://doi.org/10.1044/1092-4388(2013/12-0259)

Tjaden, K., & Martel-Sauvageau, V. (2017). Consonant acoustics in Parkinson's disease and multiple sclerosis: Comparison of clear and loud speaking conditions. *American Journal of Speech-Language Pathology*, *26*(2Special Issue), 569–582. https://doi.org/10.1044/2017_AJSLP-16-0090

Tjaden, K., Sussman, J. E., Liu, G., & Wilding, G. (2010). Long-term average spectral (LTAS) measures of dysarthria and their relationship to perceived severity. *Journal of Medical Speech-Language Pathology*, *18*(4), 125–132.

Tjaden, K., Sussman, J. E., & Wilding, G. E. (2014). Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in Parkinson's disease and multiple sclerosis. *Journal of Speech, Language, and Hearing Research*, *57*(3), 779–792. https://doi.org/10.1044/2014_JSLHR-S-12-0372

Tjaden, K., & Wilding, G. (2004). Rate and loudness manipulations in dysarthria: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*, *47*, 766–783.

Tjaden, K., & Wilding, G. (2011). Effects of speaking task on intelligibility in Parkinson's disease. *Clinical Linguistics and Phonetics*, *25*(2), 155–168. https://doi.org/10.3109/02699206.2010.520185

Troche, J., Troche, M. S., Berkowitz, R., Grossman, M., & Reilly, J. (2012). Tone discrimination as a window into acoustic perceptual deficits in Parkinson's disease. *American Journal of Speech-Language Pathology*, *21*(3), 258–263. https://doi.org/10.1044/1058-0360(2012/11-0007)

Vitale, C., Marcelli, V., Allocca, R., Santangelo, G., Riccardi, P., Erro, R., Amboni, M., Pellecchia, M. T., Cozzolino, A., Longo, K., Picillo, M., Moccia, M., Agosti, V., Sorrentino, G., Cavaliere, M., Marciano, E., & Barone, P. (2012). Hearing impairment in Parkinson's disease: Expanding the nonmotor phenotype. *Movement Disorders*, *27*(12), 1530–1535. https://doi.org/10.1002/mds.25149

Walsh, B., & Smith, A. (2012). Basic parameters of articulatory movements and acoustics in individuals with Parkinson's disease. *Movement Disorders*, *27*(7), 843–850. https://doi.org/10.1002/mds.24888

Walshe, M., Miller, N., Leahy, M., & Murray, A. (2008). Intelligibility of dysarthric speech: Perceptions of speakers and listeners. *International Journal of Language and Communication Disorders*, *43*(6), 633–648. https://doi.org/10.1080/13682820801887117

Watts, C. R. (2016). A retrospective study of long-term treatment outcomes for reduced vocal intensity in hypokinetic dysarthria throat disorders. *BMC Ear, Nose and Throat Disorders*, *16*(2), 1–7. https://doi.org/10.1186/s12901-016-0022-8

Watts, C. R., & Awan, S. N. (2011). Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts. *Journal of Speech, Language, and Hearing Research*, *54*(6), 1525–1537. https://doi.org/10.1044/1092-4388(2011/10-0209)

Weingartová, L., & Volín, J. (2014). Short-term spectral slope measures and their sensitivity to speaker, vowel identity and prominence. *Akustické Listy*, *20*(1), 5–12.

Weismer, G., Jeng, J.-Y., Laures, J. S., Kent, R. D., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica, 53*, 1-18.

Weismer, G., & Laures, J. S. (2002). Direct magnitude estimates of speech intelligibility in dysarthria. *Journal of Speech, Language, and Hearing Research*, *45*(3), 421–433. https://doi.org/10.1044/1092-4388(2002/033)

Weiss, L. (2002). Speech intelligibility in quiet and noise environments with the Speech Enhancer (TM) amplification and natural speech. *Journal of Medical Speech-Language Pathology*, *10*(4), 327–331.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Wight, S., & Miller, N. (2015). Lee Silverman Voice Treatment for people with Parkinson's: Audit of outcomes in a routine clinic. *International Journal of Language & Communication Disorders*, *50*(2), 215–225. https://doi.org/10.1111/1460-6984.12132

Wilson, C., Page, A. D., & Adams, S. G. (2020). Listener ratings of effort, speech intelligibility, and loudness of individuals with Parkinson's disease and hypophonia. *Canadian Journal of Speech-Language Pathology and Audiology*, *44*(2), 33–48.

Yunusova, Y., Kearney, E., Kulkarni, M., Haworth, B., Baljko, M., & Faloutsos, P. (2017). Game-based augmented visual feedback for enlarging speech movements in Parkinson's disease. *Journal of Speech Language and Hearing Research*, *60*(6S), 1818. https://doi.org/10.1044/2017_JSLHR-S-16-0233

Zitser, J., Casaletto, K. B., Staffaroni, A. M., Sexton, C., Weiner-Light, S., Wolf, A., . . . Kramer, J. H. (2021). Mild motor signs matter in typical brain aging: The value of the UPDRS score within a functionally intact cohort of older adults. *Frontiers in Aging Neuroscience, 13* doi:10.3389/fnagi.2021.594637

Zorilă, T.-C., Stylianou, Y., Flanagan, S., & Moore, B. C. J. (2016). Effectiveness of a loudness model for time-varying sounds in equating the loudness of sentences subjected to different forms of signal processing. *The Journal of the Acoustical Society of America*, *140*(1), 402–408. https://doi.org/10.1121/1.4955005

Zwicker, E., & Scharf, B. (1965). A model of loudness summation. *Psychological Review*, *72*(1), 3–26. https://doi.org/10.1037/h0021703

Appendix

Appendix A: Approval by the Ethical Review Board

Western Research

Date: 8 May 2020

To: Dr. Scott Adams

Project ID: 115159

Study Title: Listener perception of loudness and acoustic correlates of hypophonia in Parkinson's disease

Application Type: HSREB Initial Application

Review Type: Delegated

Meeting Date / Full Board Reporting Date: 19/May/2020

Date Approval Issued: 08/May/2020

REB Approval Expiry Date: 08/May/2021

Dear Dr. Scott Adams

The Western University Health Science Research Ethics Board (HSREB) has reviewed and approved the above mentioned study as described in the WREM application form, as of the HSREB Initial Approval Date noted above. This research study is to be conducted by the investigator noted above. All other required institutional approvals must also be obtained prior to the conduct of the study.

Documents Approved:

Document Name	Document Type	Document Date
Classroom script	Email Script	05/May/2020
Email recruitment	Email Script	08/Apr/2020
Intake Form	Paper Survey	08/Apr/2020
LOI Listeners	Written Consent/Assent	05/May/2020
Perceptual Rating Software	Other Data Collection Instruments	08/Apr/2020
Protocol	Protocol	08/Apr/2020

Documents Acknowledged:

Document Name	Document Type	Document Date
Rationale References	References	08/Apr/2020

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

REB members involved in the research project do not participate in the review, discussion or decision.

The Western University HSREB operates in compliance with, and is constituted in accordance with, the requirements of the TriCouncil Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2); the International Conference on Harmonisation Good Clinical Practice Consolidated Guideline (ICH GCP); Part C, Division 5 of the Food and Drug Regulations; Part 4 of the Natural Health Products Regulations; Part 3 of the Medical Devices Regulations and the provisions of the Ontario Personal Health Information Protection Act (PHIPA 2004) and its applicable regulations. The HSREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000940.

Please do not hesitate to contact us if you have any questions.

Sincerely,

Patricia Sargeant, Ethics Officer [redacted] on behalf of Dr. Philip Jones, HSREB Vice-Chair

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

Appendix B: Software Information and Package Citations

Praat

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer, Version 6.1.16. Retrieved from http://www.praat.org/.

Voiced tilt was obtained via concatenated voiced segments created using the same approach as the Acoustic Voice Quality Index (AVQI). Praat scripts for this technique were published as supplemental material in Maryn and Weenink (2015).

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.

MATLAB

MATLAB. (2018). Version 9.8.0.1380330 (R2018a). Natick, Massachusetts: The MathWorks Inc.

Brookes, M. (2020). VOICEBOX: Speech Processing Toolbox for MATLAB. Retrieved from http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

Moore, B., Jervis, M., Harries, L. & Schlittenlacher, J. (2018). Testing and refining a loudness model for time-varying sounds incorporating binaural inhibition. The Journal of the Acoustical Society of America, 143(3), 1504-1513. Related model code retrieved from: https://www.psychol.cam.ac.uk/hearing.

Technical Standards

TVL, LKFS, and active speech level were determined in accordance with relevant technical standards.

American National Standards Institute, Acoustical Society of America. (2007). Procedure for the computation of loudness of steady sounds (ANSI/ASA S3.4).

International Organization for Standardization. (2017). *Acoustics — Methods for calculating loudness – Part 2: Moore-Glasberg method* (ISO Standard No. 532-2).

International Telecommunication Union. (2011). *Objective measurement of speech level* (ITU-T P.56 (12/2011)).

International Telecommunication Union. (2015). *Algorithms to measure audio programme loudness and true-peak audio level* (ITU-R BS. 1770-4).

G * Power

G*Power (version 3.0) was used for calculation of statistical power.

Faul, F., Erdfelder, E., Lang, A-G., Buchner, A. (2007). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Journal of Behaviour Research, 41*(4), 1149-1160.

Faul, F., Erdfelder, E., Lang, A-G., Buchner, A. (2009). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Journal of Behaviour Research, 39*(2), 175-191.

R

R Core Team (2021). R: A language and environment for statistical computing, Version 4.0.4. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: https://www.R-project.org/.

R Packages

Packages are cited below as per the function `citation`. All packages were up to date as of March 31, 2021.

`arm`: *Data analysis using regression and multilevel/hierarchical models*. Version 1.11-2. Gelman, A., & Su, Y.-S.

`broom`*: Convert statistical objects into tidy tibbles*. Version 0.7.5. Robinson, D., Hayes, A., & Couch, S.

`broom.mixed`: *Tidying methods for mixed models*. Version 0.2.6. Bolker, B., & Robinson, D. http://github.com/bbolker/broom.mixed

`captioner`: *Numbers figures and creates simple captions.* Version 2.2.3. Alathea, L. https://github.com/adletaw/captioner

`car`: *Companion to Applied Regression*. Version 3.0-10. Fox, J., & Weisberg, S. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

`caret`: *Classification and regression training*. Version 6.0-86. Kuhn, M.

`corrplot`: *Visualization of a correlation matrix*. Version 0.84. Wei, T., & Simko, V. https://github.com/taiyun/corrplot

`cowplot`: *Streamlined plot theme and plot annotations for ggplot2*. Version 1.1.1. Wilke, C. O.

`dplyr`: *A grammar of data manipulation.* Version 1.0.5. Wickham, H., Francois, R., Henry, L., Muller, K.

`e1071`: *Misc functions of the department of statistics, probability theory*. Version 1.7-6. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F.

`effectsize`: *Indices of effect size and standardized parameters*. Version 0.4.4. Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815. https://easystats.github.io/effectsize/

`emmeans`: *Estimated marginal means, aka least-squares means*. Version 1.5.4. Lenth, R. V. https://github.com/rvlenth/emmeans

`flextable`: *Functions for tabular reporting*. Version 0.6.4. Gohel, D.

`ggeffects`: *Create tidy data frames of marginal effects for ggplot from model outputs*. Version 1.0.2. Lüdecke, D. *Journal of Open Source Software*, *3*(26), 772. https://doi.org/10.21105/joss.00772. https://strengejacke.github.io/ggeffects/

`ggforce`: *Accelerating ggplot2*. Version 0.3.3. Pedersen, T. L.

`ggplot2`: *Elegant graphics for data analysis*. Version 3.3.3. Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., … Dunnington, D. https://ggplot2.tidyverse.org

`ggpubr`: *ggplot2 based publication ready plots.* Version 0.4.0. Kassambara, A. https://rpkgs.datanovia.com/ggpubr/

`gridExtra`: *Miscellaneous functions for "grid" graphics*. Version 2.3. Auguie, B.

`Hmisc`: *Harrell miscellaneous*. Version 4.5-0. Harrell, F. E., Jr.

`knitr`: *A general-purpose package for dynamic report generation in R.* Version 1.31. Xie, Y.

`lme4`: *Fitting linear mixed-effects models using lme4.* Version 1.1-26. Bates, D., Maechler, M., Bolker, B., Walker, S. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01.

`lmerTest`: *Tests in linear mixed effects models*. Version 3.1-3. Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13. https://github.com/runehaubo/lmerTestR

`officer`: *Manipulation of Microsoft Word and PowerPoint documents.* Version 0.3.17. Gohel, D.

`party`: *A laboratory for recursive partytioning*. Version 1.3-7. Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. http://party.R-forge.R-project.org

`performance`: *Assessment of regression models performance*. Version 0.7.0. Lüdecke, D., Makowski, D., Waggoner, P., Patil, I., & Ben-Shachar, M. S. https://easystats.github.io/performance/

`pls`: *Partial least squares and principal component regression*. Version 2.7-3. Mevik, B.-H., Wehrens, R., & Liland, K. H.

`pROC`: *An open-source package for R and S+ to analyze and compare ROC curves.* Version 1.17.0.1. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Muller, M. *BMC Bioinformatics, 12,* 77. https://doi.org/10.1186/1471-2105-12-77

`psych`: *Procedures for psychological, psychometric, and personality research*. Version 2.0.12. Revelle, W. https://personality-project.org/r/psych/

`RColorBrewer`: *ColorBrewer palettes*. Version 1.1-2. Neuwirth, E.

`readr`: *Read rectangular text data.* Version 1.4.0. Wickham, H., & Hester, J.

`readxl`: *Read excel files*. Version 1.3.1. Wickham, H., & Bryan, J.

`rmarkdown`: *Dynamic documents for R.* Version 2.7. Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., … Iannone, R.

`rms`: *Regression modeling strategies*. Version 6.2-0. Harrell, F. E. Jr.

`rpart`: *Recursive partitioning and regression trees*. Version 4.1-15. Therneau, T., & Atkinson, B.

`rpart.plot`: *Plot rpart models: An enhanced version of plot.rpart*. Version 3.0.9. Milborrow, S. http://www.milbo.org/rpart-plot/index.html

`sjPlot`: *Data visualization for statistics in social science*. Version 2.8.7. Lüdecke, D. https://strengejacke.github.io/sjPlot/

`stringr`: *Simple, consistent wrappers for common string operations*. Version 1.4.0. Wickham, H.

`tidyr`: *Tidy messy data.* Version 1.1.3. Wickham, H. `tidyverse`: *Easily install and load the tidyverse*. Version 1.3.0. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686.

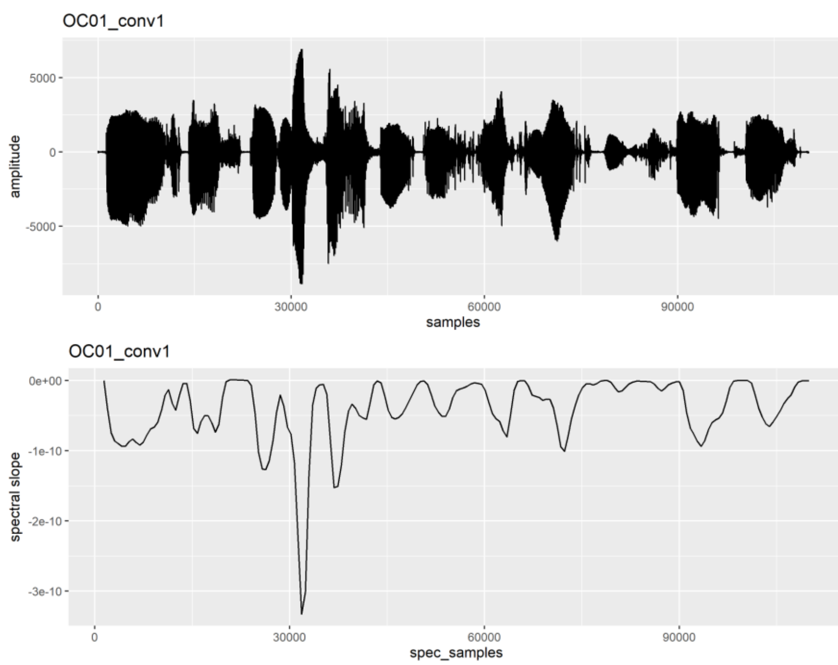
`tuneR`: *Analysis of music and speech*. Version 1.3.3. Ligges, U.

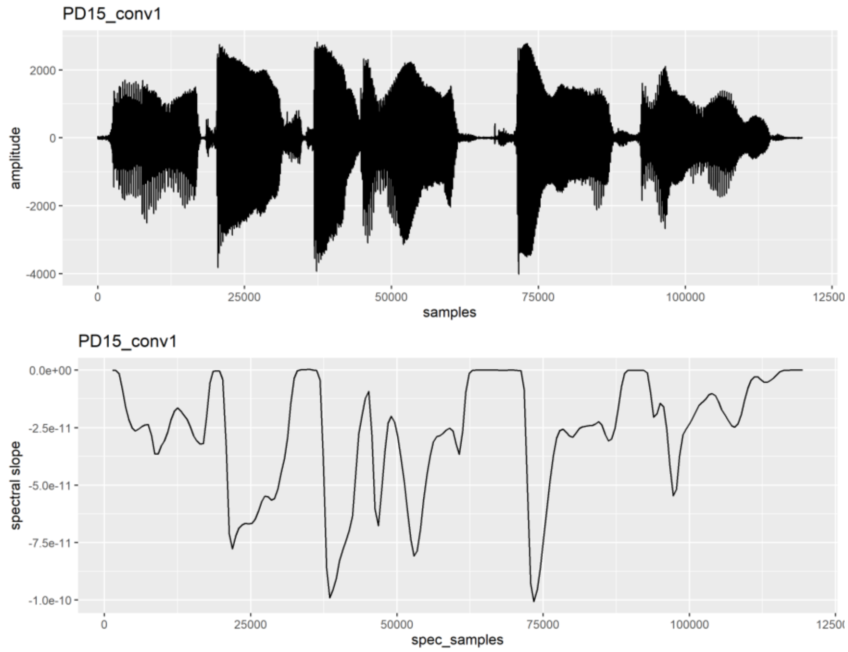
Appendix C: Dynamic Measures

Dynamic measures investigating spectral balance over the course of the utterance were employed in preliminary analyses. These measures were obtained through a combination of MATLAB, Praat, and R. It was of interest to explore the variability of spectral balance at the utterance level and investigate the relationships with loudness. However, dynamic measures were not included in the final version of this study as presented in the body of this dissertation, as they were found to consistently under-perform relative to LTAS versions of measures. Due to the structure of the data set used for this investigation, articulatory control was not possible as there was no standard sentence uttered by all participants. As a result, articulatory variability caused by the random collection of uttered sentences creates noise that clouds the interpretation of these measures. It was deemed that dynamic measures were not compatible with the available data. However, interesting trends were observed when visualizing the plots of these dynamic measures between IWPDs and HOAs. The dynamic measures employed in preliminary analyses are being presented in this appendix as it is recommended that future investigations consider the addition of a standard sentence and the inclusion of these measures. Future investigations might also consider the use of a temporal processing network approach such a long short-term memory network (LSTM), rather than summative measures (e.g. mean, interquartile range, kurtosis). These approaches are beyond the scope of this investigation but would provide valuable information about the relationship between variability of spectral balance and overall perception of loudness.

Spectral Slope

Spectral slope was obtained from a spectrogram (obtained in MATLAB) of each sample with 125 ms windows and 25 ms advancement for every 5 Hz increment from 70 Hz to 10 kHz. Each value (in dB/Hz) is a power spectral density estimate for that time window and that frequency. In R, summary measures from each spectrogram were obtained and they were plotted for visualization. Spectral slopes were first smoothed by averaging, yielding a frequency array of spectral densities for each timepoint. A slope of that array was called the spectral slope. For analysis with other parameters, these slopes can be further summarized into the peak (minimum) and inter-quartile range to describe the overall pattern of values across time.

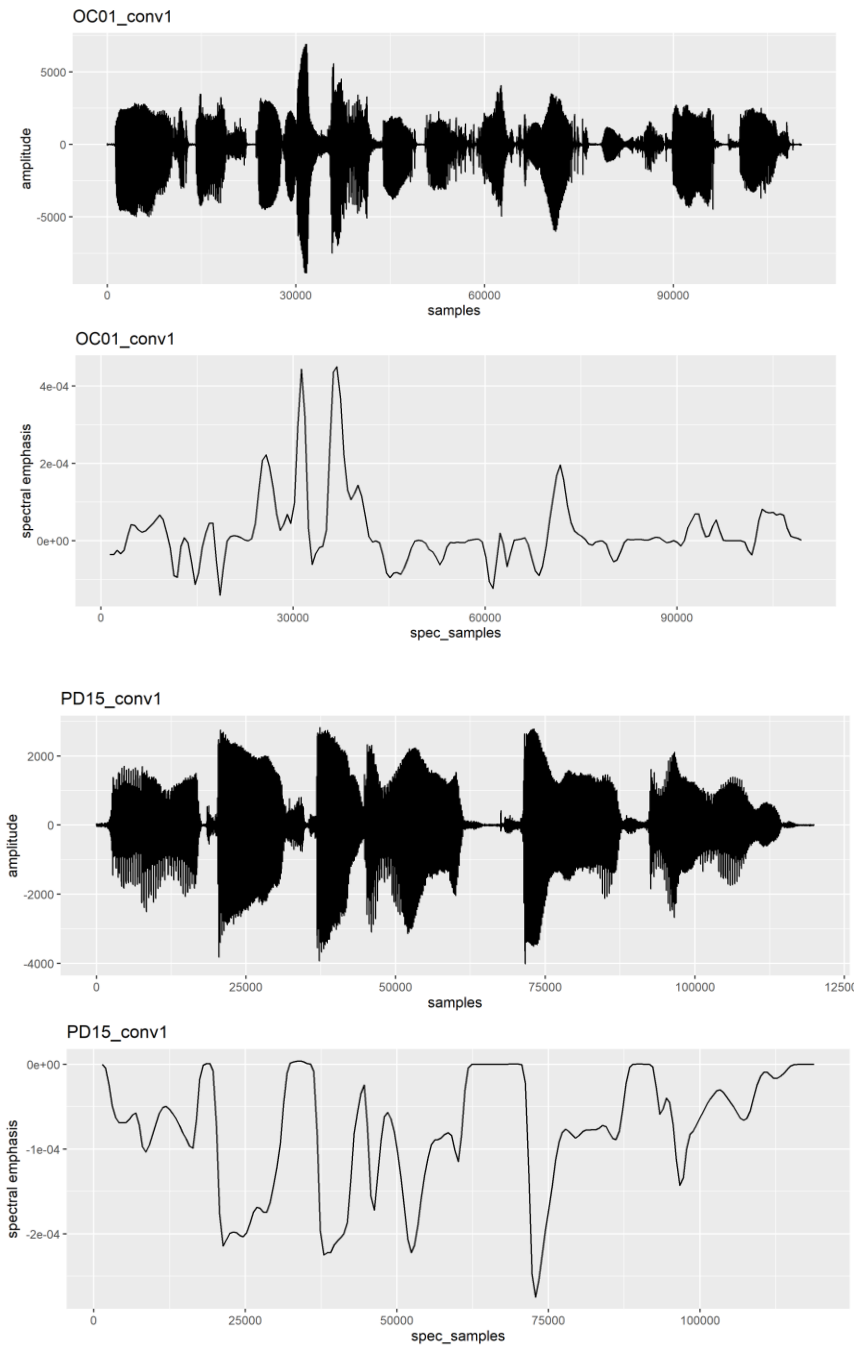




Spectral slope over time with the amplitude waveform for a conversational speech sample from a HOA (OC01) and an IWPD (PD15).

Spectral Emphasis

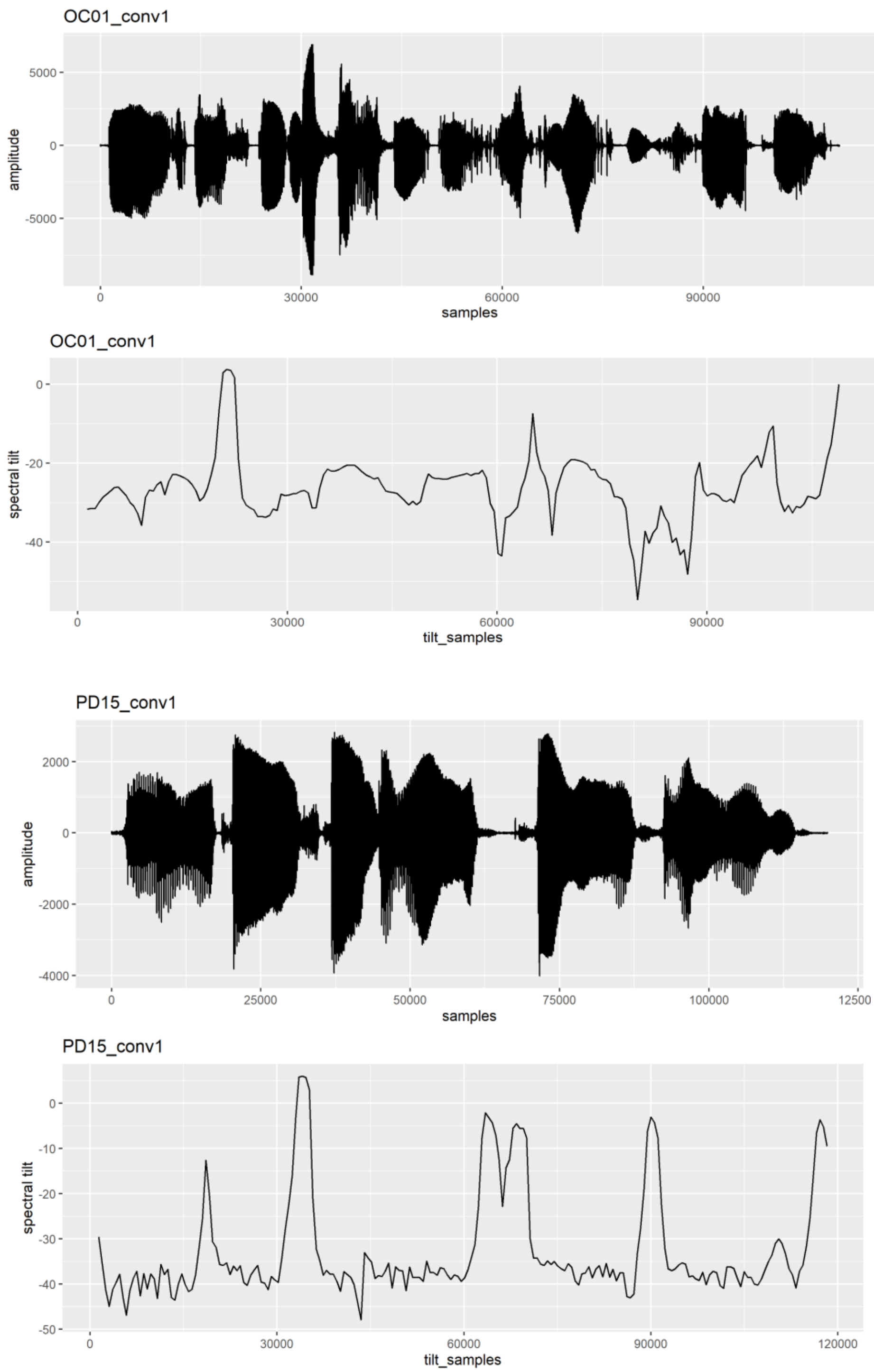
Spectral emphasis is the difference between the total energy across frequencies and the energy in the pitch band (0 Hz to (1.43 * F0) Hz) for that speaker, as per Weingartová and Volín (2014). Fundamental frequency was obtained from the LTAS in Praat. Spectral emphasis was obtained using the same spectrograms outlined for the spectral slope measures and was similarly visualized. Spectral emphasis was calculated for each timepoint (125 ms windows, with 25 ms advancement). For analysis with other parameters, this emphasis contour can be summarized into minimum, maximum, interquartile range (variability), and root-mean square (overall magnitude). The figure below presents visualizations of spectral emphasis for a HOA and an IWPD.



Spectral emphasis over time with the amplitude waveform for a conversational speech sample from a HOA (OC01) and an IWPD (PD15).

Dynamic Tilt

In addition to the LTAS tilt used in the main investigation, tilt measures were obtained over time. Tilt estimates were obtained in Praat from 125 ms windows with 25 ms advancement, calculated in the same way as overall tilt (energy in 0-1kHz vs energy in 1-10kHz). In R, these were visualized over time. Dynamic tilt can be summarized for analysis with other parameters via mean, standard deviation, skewness, and kurtosis. Dynamic tilt in particular may be complicated by articulatory effects (e.g. different tilt for /s/ than for vowels). The figure below presents visualizations of spectral emphasis for a HOA and an IWPD.



Dynamic tilt with the amplitude waveform for a conversational speech sample from a HOA (OC01) and an IWPD (PD15).

Appendix D: Testing of Assumptions

Shapiro-Wilk tests of normality and Levene's tests for equality of variance within Experiment 1 are presented in tables below. Shapiro-Wilk tests and Levene's tests of group (IWPD-HOA) variance for perceptual measures in Experiment 1 are based on values averaged within each participant across the 4 speech tasks presented to listeners. Shapiro-Wilk tests and Levene's tests of group (IWPD-HOA) variance for acoustic measures in Experiment 1 are based on values averaged within each participant across all 14 speech tasks (N = 102). For all measures in Experiment 1, Levene's tests were also run on the subgrouping analysis based on all valu es without participant-averaging (perceptual N = 408, acoustic N = 1424), as the subgrouping analysis was ultimately conducted with LMER and did not use participant-averaged values. Homoscedasticity of the underlying data is not an assumption of LMER; however, the observed heteroscedasticity, considerable inequality in group sizes, and repeated-measures nature of the data were driving factors in the decision to use LMER to evaluate these groups as opposed to ANOVA.

	Shapiro-Wilk		**Levene: Group**		**Levene: Subgroup**	
Measure	*W*	*p*	*F*	*p*	*F*	*p*
VAS Loudness	0.928	< 0.001	18.80	< 0.001	8.16	< 0.001
Intelligibility	0.848	< 0.001	24.29	< 0.001	79.83	< 0.001
DME Percent	0.959	0.003	14.02	< 0.001	1.04	0.355
DME Geometric	0.959	0.003	13.02	< 0.001	1.55	0.214

Tests of normality and of homogeneity of variance for perceptual measures.

	Shapiro-Wilk		Levene: Group		Levene: Subgroup	
Measure	*W*	*p*	*F*	*p*	*F*	*p*
Mean Intensity	0.971	0.023	7.02	0.009	5.33	0.005
Median Intensity	0.911	< 0.001	11.93	< 0.001	50.40	< 0.001
Max Intensity	0.978	0.089	3.95	0.050	3.20	0.041
TVL	0.974	0.044	9.11	0.003	0.13	0.882
TVL Mean	0.965	0.008	9.50	0.003	3.14	0.044
LKFS	0.969	0.018	6.96	0.010	3.89	0.021
Active Speech Level	0.940	< 0.001	12.16	< 0.001	8.95	< 0.001
Tilt	0.955	0.001	5.71	0.019	44.20	< 0.001
Voiced Tilt	0.953	0.001	6.85	0.010	26.40	< 0.001
Tilt Ratio	0.856	< 0.001	0.53	0.467	23.05	< 0.001
Mid-Ratio	0.871	< 0.001	9.89	0.002	156.70	< 0.001
High-Ratio	0.945	< 0.001	7.02	0.009	40.28	< 0.001
Skewness	0.703	< 0.001	11.87	< 0.001	260.72	< 0.001
Kurtosis	0.554	< 0.001	9.26	0.003	214.34	< 0.001
SD Intensity	0.954	0.001	7.37	0.008	38.24	< 0.001
SD TVL	0.977	0.076	8.18	0.005	2.11	0.122
Intensity Decay	0.954	0.001	0.18	0.668	0.19	0.827
TVL Decay	0.925	< 0.001	0.27	0.606	26.08	< 0.001

Tests of normality and of homogeneity of variance for acoustic measures.

Shapiro-Wilk tests of normality and Levene's tests for equality of variance within Experiment 2 are presented below in Table D.3. Values were not averaged as only one speech task was presented per listener per manipulation condition. Levene's tests were performed evaluating homoscedasticity between groups (IWPD-HOA) and across manipulation conditions. While heteroscedasticity is still observed on the basis of group and in manipulation condition for some measures, ANOVA was maintained as the method of evaluating these conditions as Kruskal-Wallis tests are still affected by heteroscedasticity and result in a loss of power.

	Shapiro-Wilk		**Levene: Group**		**Levene: Manipulation**	
Measure	*W*	*P*	*F*	*p*	*F*	*p*
VAS Loudness	0.968	< 0.001	38.00	< 0.001	0.14	0.969
Mean Intensity	0.977	< 0.001	19.94	< 0.001	0.00	1.000
Median Intensity	0.982	< 0.001	10.29	0.001	0.39	0.818
Max Intensity	0.978	< 0.001	24.95	< 0.001	0.03	0.998
TVL	0.995	0.114	14.68	< 0.001	2.94	0.020
TVL Mean	0.995	0.083	19.13	< 0.001	2.49	0.043
LKFS	0.976	< 0.001	17.74	< 0.001	0.00	1.000
Active Speech Level	0.964	< 0.001	27.20	< 0.001	0.00	1.000
Tilt	0.994	0.062	0.67	0.413	1.08	0.365
Voiced Tilt	0.994	0.038	0.04	0.848	0.04	0.997
Tilt Ratio	0.774	< 0.001	2.61	0.107	22.36	< 0.001
Mid-Ratio	0.981	< 0.001	11.01	< 0.001	0.60	0.664
High-Ratio	0.989	< 0.001	9.54	0.002	0.29	0.883
Skewness	0.737	< 0.001	16.64	< 0.001	6.59	< 0.001
Kurtosis	0.357	< 0.001	13.07	< 0.001	6.15	< 0.001
SD Intensity	0.995	0.097	34.04	< 0.001	0.04	0.997
SD TVL	0.995	0.101	18.86	< 0.001	2.51	0.041
Intensity Decay	0.988	< 0.001	8.40	0.004	1.63	0.166
TVL Decay	0.969	< 0.001	0.07	0.790	1.42	0.224

Tests of normality and of homogeneity of variance within the spectral manipulation experiment.

Appendix E: Correlation Tables

Full details of Pearson correlation analyses are presented in the tables below. Correlations were run on values averaged within participants. For correlations involving perceptual measures, values were averaged across the 4 speech tasks presented to listeners. For correlations between acoustic measures, values were averaged across all 14 speech tasks.

		r	*N*	*p*
DME (Percent)	*VAS Loudness*	0.987	102	< 0.001
DME (Geometric)	*VAS Loudness*	0.986	102	< 0.001
DME (Percent)	*DME (Geometric)*	0.999	102	< 0.001
Intelligibility	*VAS Loudness*	0.749	102	< 0.001
Intelligibility	*DME (Percent)*	0.767	102	< 0.001
Intelligibility	*DME (Geometric)*	0.774	102	< 0.001

Pearson correlations between perceptual measures, averaged within each participant across the 4 tasks provided to listeners.

		r	N	p
Maximum Intensity	*Mean Intensity*	0.976	102	< 0.001
Median Intensity	*Mean Intensity*	0.911	102	< 0.001
Mean Intensity	*TVL*	0.919	102	< 0.001
TVL (Long-Term Mean)	*Mean Intensity*	0.929	102	< 0.001
Mean Intensity	*LKFS*	0.996	102	< 0.001
Mean Intensity	*Active Speech Level*	0.955	102	< 0.001
Tilt	*Mean Intensity*	0.381	102	< 0.001
Voiced Tilt	*Mean Intensity*	0.412	102	< 0.001
TiltRatio	*Mean Intensity*	-0.138	102	0.1679
Mid-Ratio	*Mean Intensity*	0.784	102	< 0.001
High-Ratio	*Mean Intensity*	0.628	102	< 0.001
Skewness	*Mean Intensity*	-0.519	102	< 0.001
Kurtosis	*Mean Intensity*	-0.522	102	< 0.001
SD Intensity	*Mean Intensity*	-0.183	102	0.0651
Intensity Decay	*Mean Intensity*	-0.123	102	0.2196
Short-Term TVL Decay	*Mean Intensity*	-0.447	102	< 0.001
SD TVL	*Mean Intensity*	0.894	102	< 0.001
SD TVL	*TVL*	0.971	102	< 0.001
Short-Term TVL Decay	*TVL*	-0.561	102	< 0.001
Short-Term TVL Decay	*TVL (Long-Term Mean)*	-0.559	102	< 0.001
SD TVL	*TVL (Long-Term Mean)*	0.947	102	< 0.001
Voiced Tilt	*Tilt*	0.97	102	< 0.001
Tilt	*Skewness*	-0.885	102	< 0.001
Tilt	*Kurtosis*	-0.813	102	< 0.001
Mid-Ratio	*Tilt*	0.833	102	< 0.001
Mid-Ratio	*Skewness*	-0.863	102	< 0.001
Mid-Ratio	*Kurtosis*	-0.825	102	< 0.001
High-Ratio	*Tilt*	0.759	102	< 0.001
High-Ratio	*Skewness*	-0.695	102	< 0.001
High-Ratio	*Kurtosis*	-0.7	102	< 0.001
SD TVL	*SD Intensity*	-0.159	102	0.1107

Pearson correlations between acoustic measures, averaged within each participant across the 14 tasks provided to listeners.

		r	*N*	*p*
Mean Intensity	*VAS Loudness*	0.908	102	< 0.001
Median Intensity	*VAS Loudness*	0.836	102	< 0.001
Maximum Intensity	*VAS Loudness*	0.911	102	< 0.001
TVL Maximum	*VAS Loudness*	0.956	102	< 0.001
TVL Mean	*VAS Loudness*	0.972	102	< 0.001
LKFS	*VAS Loudness*	0.897	102	< 0.001
Active Speech Level	*VAS Loudness*	0.889	102	< 0.001
Tilt	*VAS Loudness*	0.673	102	< 0.001
Voiced Tilt	*VAS Loudness*	0.698	102	< 0.001
TiltRatio	*VAS Loudness*	-0.028	102	0.782
Mid-Ratio	*VAS Loudness*	0.915	102	< 0.001
High-Ratio	*VAS Loudness*	0.773	102	< 0.001
Skewness	*VAS Loudness*	-0.731	102	< 0.001
Kurtosis	*VAS Loudness*	-0.685	102	< 0.001
SD Intensity	*VAS Loudness*	-0.266	102	0.007
SD TVL	*VAS Loudness*	0.871	102	< 0.001
Intensity Decay	*VAS Loudness*	-0.038	102	0.701
Short-Term TVL Decay	*VAS Loudness*	-0.406	102	< 0.001
Mean Intensity	*DME (Percent)*	0.893	102	< 0.001
Median Intensity	*DME (Percent)*	0.827	102	< 0.001
Maximum Intensity	*DME (Percent)*	0.899	102	< 0.001
TVL Maximum	*DME (Percent)*	0.953	102	< 0.001
TVL Mean	*DME (Percent)*	0.972	102	< 0.001
LKFS	*DME (Percent)*	0.880	102	< 0.001
Active Speech Level	*DME (Percent)*	0.876	102	< 0.001
Tilt	*DME (Percent)*	0.670	102	< 0.001
Voiced Tilt	*DME (Percent)*	0.690	102	< 0.001
TiltRatio	*DME (Percent)*	-0.010	102	0.920
Mid-Ratio	*DME (Percent)*	0.899	102	< 0.001
High-Ratio	*DME (Percent)*	0.779	102	< 0.001

		r	*N*	*p*
Skewness	*DME (Percent)*	-0.712	102	< 0.001
Kurtosis	*DME (Percent)*	-0.666	102	< 0.001
SD Intensity	*DME (Percent)*	-0.246	102	0.013
SD TVL	*DME (Percent)*	0.855	102	< 0.001
Intensity Decay	*DME (Percent)*	-0.056	102	0.578
Short-Term TVL Decay	*DME (Percent)*	-0.430	102	< 0.001
Mean Intensity	*Intelligibility*	0.533	102	< 0.001
Median Intensity	*Intelligibility*	0.611	102	< 0.001
Maximum Intensity	*Intelligibility*	0.555	102	< 0.001
TVL Maximum	*Intelligibility*	0.655	102	< 0.001
TVL Mean	*Intelligibility*	0.689	102	< 0.001
LKFS	*Intelligibility*	0.504	102	< 0.001
Active Speech Level	*Intelligibility*	0.558	102	< 0.001
Tilt	*Intelligibility*	0.742	102	< 0.001
Voiced Tilt	*Intelligibility*	0.699	102	< 0.001
TiltRatio	*Intelligibility*	0.203	102	0.040
Mid-Ratio	*Intelligibility*	0.772	102	< 0.001
High-Ratio	*Intelligibility*	0.749	102	< 0.001
Skewness	*Intelligibility*	-0.767	102	< 0.001
Kurtosis	*Intelligibility*	-0.730	102	< 0.001
SD Intensity	*Intelligibility*	-0.406	102	< 0.001
SD TVL	*Intelligibility*	0.519	102	< 0.001
Intensity Decay	*Intelligibility*	-0.146	102	0.142
Short-Term TVL Decay	*Intelligibility*	-0.453	102	< 0.001

Pearson correlations between perceptual and acoustic measures, averaged within each participant across the 4 tasks provided to listeners.

IWPD

		r	*N*	*p*
DME (Percent)	*VAS Loudness*	0.985	56	< 0.001
DME (Geometric)	*VAS Loudness*	0.983	56	< 0.001
DME (Percent)	*DME (Geometric)*	0.999	56	< 0.001
Intelligibility	*VAS Loudness*	0.682	56	< 0.001
Intelligibility	*DME (Percent)*	0.703	56	< 0.001
Intelligibility	*DME (Geometric)*	0.713	56	< 0.001

Pearson correlations between perceptual measures, averaged within each IWPD across the 4 tasks provided to listeners.

		r	*N*	*p*
Maximum Intensity	*Mean Intensity*	0.976	56	< 0.001
Median Intensity	*Mean Intensity*	0.891	56	< 0.001
Mean Intensity	*TVL*	0.926	56	< 0.001
TVL (Long-Term Mean)	*Mean Intensity*	0.934	56	< 0.001
Mean Intensity	*LKFS*	0.994	56	< 0.001
Mean Intensity	*Active Speech Level*	0.948	56	< 0.001
Tilt	*Mean Intensity*	0.451	56	< 0.001
Voiced Tilt	*Mean Intensity*	0.493	56	< 0.001
TiltRatio	*Mean Intensity*	-0.294	56	0.028
Mid-Ratio	*Mean Intensity*	0.797	56	< 0.001
High-Ratio	*Mean Intensity*	0.65	56	< 0.001
Skewness	*Mean Intensity*	-0.574	56	< 0.001
Kurtosis	*Mean Intensity*	-0.558	56	< 0.001
SD Intensity	*Mean Intensity*	-0.239	56	0.0764
Intensity Decay	*Mean Intensity*	-0.09	56	0.5117
Short-Term TVL Decay	*Mean Intensity*	-0.48	56	< 0.001

		r	N	p
SD TVL	*Mean Intensity*	0.912	56	< 0.001
SD TVL	*SD Intensity*	-0.2	56	0.1394
Short-Term TVL Decay	*Intensity Decay*	0.726	56	< 0.001
Voiced Tilt	*Tilt*	0.973	56	< 0.001
Tilt	*Skewness*	-0.895	56	< 0.001
Tilt	*Kurtosis*	-0.833	56	< 0.001
Mid-Ratio	*Tilt*	0.873	56	< 0.001
Mid-Ratio	*Skewness*	-0.882	56	< 0.001
Mid-Ratio	*Kurtosis*	-0.841	56	< 0.001
High-Ratio	*Tilt*	0.725	56	< 0.001
High-Ratio	*Skewness*	-0.676	56	< 0.001
High-Ratio	*Kurtosis*	-0.708	56	< 0.001

Pearson correlations between acoustic measures, averaged within each IWPD across the 14 tasks provided to listeners.

		r	N	p
Mean Intensity	*VAS Loudness*	0.925	56	< 0.001
Maximum Intensity	*VAS Loudness*	0.907	56	< 0.001
Median Intensity	*VAS Loudness*	0.82	56	< 0.001
TVL	*VAS Loudness*	0.963	56	< 0.001
TVL (Long-Term Mean)	*VAS Loudness*	0.972	56	< 0.001
LKFS	*VAS Loudness*	0.909	56	< 0.001
Active Speech Level	*VAS Loudness*	0.866	56	< 0.001
Tilt	*VAS Loudness*	0.653	56	< 0.001
Voiced Tilt	*VAS Loudness*	0.7	56	< 0.001
TiltRatio	*VAS Loudness*	-0.237	56	0.0791
Mid-Ratio	*VAS Loudness*	0.906	56	< 0.001
High-Ratio	*VAS Loudness*	0.738	56	< 0.001

		r	*N*	*p*
Skewness	*VAS Loudness*	-0.728	56	< 0.001
Kurtosis	*VAS Loudness*	-0.678	56	< 0.001
SD Intensity	*VAS Loudness*	-0.203	56	0.1325
Intensity Decay	*VAS Loudness*	0.044	56	0.745
Short-Term TVL Decay	*VAS Loudness*	-0.285	56	0.0334
SD TVL	*VAS Loudness*	0.912	56	< 0.001
Mean Intensity	*DME (Percent)*	0.908	56	< 0.001
Maximum Intensity	*DME (Percent)*	0.886	56	< 0.001
Median Intensity	*DME (Percent)*	0.814	56	< 0.001
TVL	*DME (Percent)*	0.958	56	< 0.001
TVL (Long-Term Mean)	*DME (Percent)*	0.973	56	< 0.001
LKFS	*DME (Percent)*	0.889	56	< 0.001
Active Speech Level	*DME (Percent)*	0.848	56	< 0.001
Tilt	*DME (Percent)*	0.643	56	< 0.001
Voiced Tilt	*DME (Percent)*	0.689	56	< 0.001
TiltRatio	*DME (Percent)*	-0.238	56	0.0775
Mid-Ratio	*DME (Percent)*	0.883	56	< 0.001
High-Ratio	*DME (Percent)*	0.725	56	< 0.001
Skewness	*DME (Percent)*	-0.708	56	< 0.001
Kurtosis	*DME (Percent)*	-0.659	56	< 0.001
SD Intensity	*DME (Percent)*	-0.173	56	0.2017
Intensity Decay	*DME (Percent)*	0.026	56	0.8483
Short-Term TVL Decay	*DME (Percent)*	-0.315	56	0.018
SD TVL	*DME (Percent)*	0.896	56	< 0.001
Mean Intensity	*Intelligibility*	0.503	56	< 0.001
Maximum Intensity	*Intelligibility*	0.489	56	< 0.001
Median Intensity	*Intelligibility*	0.588	56	< 0.001

		r	N	p
TVL	*Intelligibility*	0.61	56	< 0.001
TVL (Long-Term Mean)	*Intelligibility*	0.637	56	< 0.001
LKFS	*Intelligibility*	0.465	56	< 0.001
Active Speech Level	*Intelligibility*	0.464	56	< 0.001
Tilt	*Intelligibility*	0.701	56	< 0.001
Voiced Tilt	*Intelligibility*	0.674	56	< 0.001
TiltRatio	*Intelligibility*	0.017	56	0.9002
Mid-Ratio	*Intelligibility*	0.746	56	< 0.001
High-Ratio	*Intelligibility*	0.643	56	< 0.001
Skewness	*Intelligibility*	-0.734	56	< 0.001
Kurtosis	*Intelligibility*	-0.704	56	< 0.001
SD Intensity	*Intelligibility*	-0.314	56	0.0184
Intensity Decay	*Intelligibility*	-0.041	56	0.765
Short-Term TVL Decay	*Intelligibility*	-0.324	56	0.0148
SD TVL	*Intelligibility*	0.521	56	< 0.001

Pearson correlations between perceptual and acoustic measures, averaged within each IWPD across the 4 tasks provided to listeners.

HOA

		r	N	p
DME (Percent)	*VAS Loudness*	0.973	46	< 0.001
DME (Geometric)	*VAS Loudness*	0.972	46	< 0.001
DME (Percent)	*DME (Geometric)*	1	46	< 0.001
Intelligibility	*VAS Loudness*	0.21	46	0.1613
Intelligibility	*DME (Percent)*	0.296	46	0.0461
Intelligibility	*DME (Geometric)*	0.304	46	0.0402

Pearson correlations between perceptual measures, averaged within each HOA across the 4 tasks provided to listeners.

		r	*N*	*p*
Maximum Intensity	*Mean Intensity*	0.97	46	< 0.001
Median Intensity	*Mean Intensity*	0.954	46	< 0.001
Mean Intensity	*TVL*	0.842	46	< 0.001
TVL (Long-Term Mean)	*Mean Intensity*	0.865	46	< 0.001
Mean Intensity	*LKFS*	0.998	46	< 0.001
Mean Intensity	*Active Speech Level*	0.959	46	< 0.001
Tilt	*Mean Intensity*	-0.269	46	0.0712
Voiced Tilt	*Mean Intensity*	-0.199	46	0.1855
TiltRatio	*Mean Intensity*	-0.218	46	0.1455
Mid-Ratio	*Mean Intensity*	0.555	46	< 0.001
High-Ratio	*Mean Intensity*	0.225	46	0.1334
Skewness	*Mean Intensity*	0.279	46	0.0602
Kurtosis	*Mean Intensity*	0.275	46	0.0645
SD Intensity	*Mean Intensity*	0.311	46	0.0353
Intensity Decay	*Mean Intensity*	-0.015	46	0.9189
Short-Term TVL Decay	*Mean Intensity*	-0.136	46	0.3677
SD TVL	*Mean Intensity*	0.79	46	< 0.001
Voiced Tilt	*Tilt*	0.95	46	< 0.001
Tilt	*Skewness*	-0.918	46	< 0.001
Tilt	*Kurtosis*	-0.905	46	< 0.001
Mid-Ratio	*Tilt*	0.526	46	< 0.001
Mid-Ratio	*Skewness*	-0.532	46	< 0.001
Mid-Ratio	*Kurtosis*	-0.532	46	< 0.001
High-Ratio	*Tilt*	0.612	46	< 0.001
High-Ratio	*Skewness*	-0.505	46	< 0.001
High-Ratio	*Kurtosis*	-0.558	46	< 0.001

Pearson correlations between acoustic measures, averaged within each HOA across the 14 tasks provided to listeners.

		r	N	p
Mean Intensity	*VAS Loudness*	0.824	46	< 0.001
Maximum Intensity	*VAS Loudness*	0.868	46	< 0.001
Median Intensity	*VAS Loudness*	0.758	46	< 0.001
TVL	*VAS Loudness*	0.93	46	< 0.001
TVL (Long-Term Mean)	*VAS Loudness*	0.95	46	< 0.001
LKFS	*VAS Loudness*	0.845	46	< 0.001
Active Speech Level	*VAS Loudness*	0.849	46	< 0.001
Tilt	*VAS Loudness*	0.317	46	0.0321
Voiced Tilt	*VAS Loudness*	0.388	46	0.0078
TiltRatio	*VAS Loudness*	-0.129	46	0.393
Mid-Ratio	*VAS Loudness*	0.797	46	< 0.001
High-Ratio	*VAS Loudness*	0.469	46	0.001
Skewness	*VAS Loudness*	-0.253	46	0.09
Kurtosis	*VAS Loudness*	-0.223	46	0.136
SD Intensity	*VAS Loudness*	0.17	46	0.2577
Intensity Decay	*VAS Loudness*	0.1	46	0.5088
Short-Term TVL Decay	*VAS Loudness*	-0.226	46	0.1317
SD TVL	*VAS Loudness*	0.826	46	< 0.001
Mean Intensity	*DME (Percent)*	0.808	46	< 0.001
Maximum Intensity	*DME (Percent)*	0.859	46	< 0.001
Median Intensity	*DME (Percent)*	0.74	46	< 0.001
TVL	*DME (Percent)*	0.928	46	< 0.001
TVL (Long-Term Mean)	*DME (Percent)*	0.948	46	< 0.001
LKFS	*DME (Percent)*	0.83	46	< 0.001
Active Speech Level	*DME (Percent)*	0.835	46	< 0.001
Tilt	*DME (Percent)*	0.329	46	0.0257
Voiced Tilt	*DME (Percent)*	0.384	46	0.0084
TiltRatio	*DME (Percent)*	-0.084	46	0.5768

		r	*N*	*p*
Mid-Ratio	*DME (Percent)*	0.785	46	< 0.001
High-Ratio	*DME (Percent)*	0.524	46	< 0.001
Skewness	*DME (Percent)*	-0.26	46	0.0805
Kurtosis	*DME (Percent)*	-0.25	46	0.0941
SD Intensity	*DME (Percent)*	0.184	46	0.2219
Intensity Decay	*DME (Percent)*	0.081	46	0.5916
Short-Term TVL Decay	*DME (Percent)*	-0.234	46	0.1173
SD TVL	*DME (Percent)*	0.812	46	< 0.001
Mean Intensity	*Intelligibility*	-0.158	46	0.2932
Maximum Intensity	*Intelligibility*	-0.081	46	0.5937
Median Intensity	*Intelligibility*	-0.128	46	0.3968
TVL	*Intelligibility*	0.232	46	0.1214
TVL (Long-Term Mean)	*Intelligibility*	0.213	46	0.1554
LKFS	*Intelligibility*	-0.115	46	0.4474
Active Speech Level	*Intelligibility*	0	46	0.9992
Tilt	*Intelligibility*	0.683	46	< 0.001
Voiced Tilt	*Intelligibility*	0.595	46	< 0.001
TiltRatio	*Intelligibility*	0.336	46	0.0225
Mid-Ratio	*Intelligibility*	0.276	46	0.0633
High-Ratio	*Intelligibility*	0.605	46	< 0.001
Skewness	*Intelligibility*	-0.631	46	< 0.001
Kurtosis	*Intelligibility*	-0.655	46	< 0.001
SD Intensity	*Intelligibility*	-0.19	46	0.2055
Intensity Decay	*Intelligibility*	-0.189	46	0.2076
Short-Term TVL Decay	*Intelligibility*	-0.354	46	0.0158
SD TVL	*Intelligibility*	0.168	46	0.2648

Pearson correlations between perceptual and acoustic measures, averaged within each HOA across the 4 tasks provided to listeners.

Appendix F: LMER Group Interaction Model Tables

LMER model tables are provided for each model, contributing additional detail to the models described and presented visually in Chapter 4. Results in the table below are scaled, whereas the figures previously presented were not scaled to allow for clearer visualization of trends relative to the original scale of each measure. For interpretation of model coefficients, means and standard deviations for each measure are presented below.

Measure	*Mean (SD)*
Mean Intensity	67.86 (4.30)
Median Intensity	64.74 (6.46)
Max Intensity	75.07 (4.29)
TVL	17.05 (5.29)
TVL Mean	12.08 (3.91)
LKFS	-26.63 (4.23)
Active Speech Level	-27.64 (4.76)
Tilt	-26.97 (4.72)
Voiced Tilt	-27.84 (4.82)
Tilt Ratio	1.03 (0.06)
Mid-Ratio	0.12 (0.10)
High-Ratio	0.04 (0.12)
Skewness	11.89 (7.31)
Kurtosis	281.16 (430.01)
SD Intensity	12.63 (2.43)
SD TVL	3.37 (1.14)
Intensity Decay *	-20.46 (26.47)
TVL Decay *	-1.11 (1.32)

Means and standard deviations based on values from all 14 speech tasks.
* Values for intensity decay and TVL decay are expressed in scientific notation (x 10^3).

Mean Intensity

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	64.59	61.79 – 67.40	< 0.001
meanInt	26.15	22.98 – 29.31	< 0.001
groupIWPD	-9.49	-12.21 – -6.78	< 0.001
*meanInt * groupIWPD*	5.19	1.50 – 8.88	0.0062
SD (Intercept): participant	6.20		
SD (Intercept): task	1.98		
SD Observation: Residual	5.00		
N participant	102		
N task	4		
N observations	408		
AIC	2,681.24		
Conditional R2	0.930		
Marginal R2	0.812		

Median Intensity

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	66.10	62.44 – 69.77	< 0.001
medianInt	20.67	15.73 – 25.60	< 0.001
groupIWPD	-14.16	-18.84 – -9.47	< 0.001
*medianInt * groupIWPD*	-8.08	-13.33 – -2.84	0.0027
SD (Intercept): participant	11.34		
SD (Intercept): task	1.10		
SD Observation: Residual	6.45		
N participant	102		
N task	4		

Median Intensity

Predictors	*Estimates*	*95% CI*	*p*
N observations	408		
AIC	2,945.53		
Conditional R2	0.866		
Marginal R2	0.447		

Maximum Intensity

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	64.80	62.33 – 67.28	< 0.001
maxInt	21.20	17.92 – 24.48	< 0.001
groupIWPD	-9.07	-12.02 – -6.12	< 0.001
*maxInt * groupIWPD*	8.78	4.46 – 13.10	< 0.001
SD (Intercept): participant	6.57		
SD (Intercept): task	1.15		
SD Observation: Residual	5.80		
N participant	102		
N task	4		
N observations	408		
AIC	2,783.27		
Conditional R2	0.897		
Marginal R2	0.761		

TVL

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	64.76	63.02 – 66.49	< 0.001
TVL	21.96	19.40 – 24.51	< 0.001

TVL

Predictors	*Estimates*	*95% CI*	*p*
groupIWPD	-8.22	-10.21 – -6.23	< 0.001
*TVL * groupIWPD*	13.26	9.91 – 16.61	< 0.001
SD (Intercept): participant	3.94		
SD (Intercept): task	0.91		
SD Observation: Residual	5.43		
N participant	102		
N task	4		
N observations	408		
AIC	2,661.41		
Conditional R2	0.915		
Marginal R2	0.868		

TVL Mean

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	62.99	60.09 – 65.89	< 0.001
TVLlong_mean	28.56	25.86 – 31.25	< 0.001
groupIWPD	-6.01	-7.79 – -4.23	< 0.001
*TVLlong_mean * groupIWPD*	7.27	4.13 – 10.41	< 0.001
SD (Intercept): participant	3.22		
SD (Intercept): task	2.62		
SD Observation: Residual	5.18		
N participant	102		
N task	4		
N observations	408		
AIC	2,608.64		

TVL Mean

Predictors	*Estimates*	*95% CI*	*p*
Conditional R2	0.928		
Marginal R2	0.882		

LKFS

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	65.02	62.32 – 67.72	< 0.001
lkfs	25.06	21.89 – 28.23	< 0.001
groupIWPD	-10.22	-13.04 – -7.41	< 0.001
*lkfs * groupIWPD*	5.78	1.98 – 9.58	0.003
SD (Intercept): participant	6.48		
SD (Intercept): task	1.74		
SD Observation: Residual	5.15		
N participant	102		
N task	4		
N observations	408		
AIC	2,707.68		
Conditional R2	0.925		
Marginal R2	0.799		

Active Speech Level

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	63.90	60.96 – 66.85	< 0.001
activeSL	26.14	22.84 – 29.44	< 0.001
groupIWPD	-8.50	-11.77 – -5.24	< 0.001
*activeSL * groupIWPD*	3.04	-0.82 – 6.91	0.1238

Active Speech Level

Predictors	*Estimates*	*95% CI*	*p*
SD (Intercept): participant	7.64		
SD (Intercept): task	1.70		
SD Observation: Residual	4.98		
N participant	102		
N task	4		
N observations	408		
AIC	2,715.66		
Conditional R2	0.929		
Marginal R2	0.754		

Tilt

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	66.94	62.30 – 71.58	< 0.001
tilt	11.18	6.24 – 16.13	< 0.001
groupIWPD	-13.95	-19.21 – -8.69	< 0.001
*tilt * groupIWPD*	3.30	-3.27 – 9.86	0.3253
SD (Intercept): participant	12.48		
SD (Intercept): task	2.54		
SD Observation: Residual	7.07		
N participant	102		
N task	4		
N observations	408		
AIC	3,023.85		
Conditional R2	0.854		

Tilt

Predictors	*Estimates*	*95% CI*	*p*
Marginal R2	0.380		

Voiced Tilt

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	66.98	62.69 – 71.27	< 0.001
voicedTilt	12.85	8.67 – 17.03	< 0.001
groupIWPD	-13.74	-18.58 – -8.90	< 0.001
*voicedTilt * groupIWPD*	5.57	-0.15 – 11.28	0.0569
SD (Intercept): participant	11.63		
SD (Intercept): task	2.39		
SD Observation: Residual	6.71		
N participant	102		
N task	4		
N observations	408		
AIC	2,978.31		
Conditional R2	0.868		
Marginal R2	0.453		

Tilt Ratio

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	70.02	64.65 – 75.39	< 0.001
tiltRatio	-4.53	-7.24 – -1.82	0.0012
groupIWPD	-20.41	-26.34 – -14.49	< 0.001
*tiltRatio * groupIWPD*	-2.76	-6.89 – 1.37	0.1912
SD (Intercept): participant	14.76		

Tilt Ratio

Predictors	*Estimates*	*95% CI*	*p*
SD (Intercept): task	3.15		
SD Observation: Residual	6.94		
N participant	102		
N task	4		
N observations	408		
AIC	3,047.02		
Conditional R2	0.872		
Marginal R2	0.269		

Skewness

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	67.33	62.44 – 72.23	< 0.001
skew	-10.76	-19.51 – -2.00	0.0165
groupIWPD	-14.96	-20.20 – -9.72	< 0.001
*skew * groupIWPD*	-1.95	-11.41 – 7.52	0.687
SD (Intercept): participant	11.95		
SD (Intercept): task	2.82		
SD Observation: Residual	7.30		
N participant	102		
N task	4		
N observations	408		
AIC	3,035.52		
Conditional R2	0.839		
Marginal R2	0.382		

Kurtosis

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	66.43	60.44 – 72.42	< 0.001
kurt	-17.79	-35.95 – 0.37	0.0555
groupIWPD	-15.42	-21.80 – -9.05	< 0.001
*kurt * groupIWPD*	12.35	-6.07 – 30.76	0.1896
SD (Intercept): participant	13.63		
SD (Intercept): task	3.09		
SD Observation: Residual	7.31		
N participant	102		
N task	4		
N observations	408		
AIC	3,060.38		
Conditional R2	0.847		
Marginal R2	0.289		

Mid-Ratio

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	63.04	60.37 – 65.71	< 0.001
mid_ratio	26.42	21.97 – 30.87	< 0.001
groupIWPD	-6.97	-10.14 – -3.80	< 0.001
*mid_ratio * groupIWPD*	2.47	-2.65 – 7.59	0.3456
SD (Intercept): participant	6.75		
SD (Intercept): task	1.17		
SD Observation: Residual	6.14		

Mid-Ratio

Predictors	*Estimates*	*95% CI*	*p*
N participant	102		
N task	4		
N observations	408		
AIC	2,823.30		
Conditional R2	0.889		
Marginal R2	0.750		

High-Ratio

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	67.47	63.19 – 71.76	< 0.001
high_ratio	7.03	2.45 – 11.61	0.0028
groupIWPD	-13.36	-18.49 – -8.22	< 0.001
*high_ratio * groupIWPD*	9.06	3.06 – 15.07	0.0033
SD (Intercept): participant	11.86		
SD (Intercept): task	1.96		
SD Observation: Residual	7.21		
N participant	102		
N task	4		
N observations	408		
AIC	3,025.45		
Conditional R2	0.839		
Marginal R2	0.391		

SD Intensity

Predictors	Estimates	95% CI	p
(Intercept)	69.56	64.25 – 74.88	< 0.001
sdInt	0.14	-3.27 – 3.55	0.9366
groupIWPD	-19.10	-25.12 – -13.09	< 0.001
sdInt * groupIWPD	-2.44	-6.57 – 1.69	0.2482
SD (Intercept): participant	14.94		
SD (Intercept): task	2.95		
SD Observation: Residual	7.24		
N participant	102		
N task	4		
N observations	408		
AIC	3,074.61		
Conditional R2	0.861		
Marginal R2	0.247		

SD TVL

Predictors	Estimates	95% CI	p
(Intercept)	67.71	64.29 – 71.14	< 0.001
TVLlong_sd	13.90	10.94 – 16.86	< 0.001
groupIWPD	-15.17	-18.78 – -11.55	< 0.001
TVLlong_sd * groupIWPD	7.57	3.57 – 11.56	< 0.001
SD (Intercept): participant	8.56		
SD (Intercept): task	2.18		
SD Observation: Residual	6.51		
N participant	102		
N task	4		

SD TVL

Predictors	Estimates	95% CI	p
N observations	408		
AIC	2,905.22		
Conditional R2	0.858		
Marginal R2	0.595		

Intensity Decay

Predictors	Estimates	95% CI	p
(Intercept)	69.56	64.14 – 74.97	< 0.001
decay	0.30	-2.19 – 2.80	0.811
groupIWPD	-19.28	-25.32 – -13.23	< 0.001
decay * groupIWPD	-1.87	-5.09 – 1.34	0.255
SD (Intercept): participant	15.06		
SD (Intercept): task	3.11		
SD Observation: Residual	7.23		
N participant	102		
N task	4		
N observations	408		
AIC	3,076.69		
Conditional R2	0.863		
Marginal R2	0.243		

TVL Decay

Predictors	Estimates	95% CI	p
(Intercept)	69.31	64.34 – 74.29	< 0.001
TVLshort_decay	-2.00	-4.23 – 0.23	0.0796

TVL Decay

Predictors	*Estimates*	*95% CI*	*p*
groupIWPD	-18.47	-24.41 – -12.53	< 0.001
*TVLshort_decay * groupIWPD*	-4.56	-7.84 – -1.28	0.0068
SD (Intercept): participant	14.79		
SD (Intercept): task	2.37		
SD Observation: Residual	7.05		
N participant	102		
N task	4		
N observations	408		
AIC	3,056.37		
Conditional R2	0.866		
Marginal R2	0.262		

Measure	δ
Mean Intensity (dB SPL)	0.63
Median Intensity (dB SPL)	0.62
Max Intensity (dB SPL)	0.99
TVL (sones)	1.96
TVL Mean (sones)	1.10
LKFS	0.68
Active Speech Level (dB FS)	0.33
Tilt (dB)	0.23
Voiced Tilt (dB)	0.41
Tilt Ratio	0.17
Mid-Ratio	0.27
High-Ratio	0.65
Skewness	0.14
Kurtosis	0.78
SD Intensity (dB)	0.14
SD TVL (sones)	0.69
Intensity Decay	0.11
TVL Decay	0.28

Table F.20: Effect sizes of group interaction terms based on estimated marginal means (EMM) are calculated as the mean difference divided by the square root of the sum of all variances, as an extension to Cohen's d as per Westfall et al. (2014).

Appendix G: Model Building

The process of building each maximal model (LMER and logistic regression) is summarized in Chapter 4, along with presentation of the final model. The progression from baseline to maximal models for each maximal model presented in Chapter 4 is presented below.

For each comparison step in which candidate models are entertained, an AIC table is presented showing the new candidates and the previous models to which these are being compared. In these tables, bolded models represent a significant improvement in performance based on LRT (*p* < .05) between a candidate model and its previous nested iteration. Italicized models represent those with singular fits or non-convergence. VIF tables are presented with interim models and where VIF was a primary decision-making factor.

Loudness

Baseline Model

VAS Loudness: Baseline

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	69.54	64.08 – 75.01	< 0.001
groupIWPD	-19.32	-25.35 – -13.28	< 0.001
SD (Intercept): participant	15.05		
SD (Intercept): task	3.20		
SD Observation: Residual	7.23		
N participant	102		
N task	4		

VAS Loudness: Baseline			
Predictors	*Estimates*	*95% CI*	*p*
N observations	408		
AIC	3,079.07		
Conditional R2	0.863		
Marginal R2	0.243		

Speech Level

Model	*AIC*
Baseline	3,079.070
Mean Intensity	2,689.560
Median Intensity	2,956.225
Max Intensity	2,799.323
TVL	2,715.180
TVL Mean	2,628.667
LKFS	2,717.409
Active Speech Level	2,719.213

VAS Loudness: Speech Level			
Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	61.99	59.04 – 64.93	< 0.001
TVLlong_mean	32.92	31.17 – 34.67	< 0.001
groupIWPD	-5.56	-7.47 – -3.64	< 0.001
SD (Intercept): participant	3.71		
SD (Intercept): task	2.66		
SD Observation: Residual	5.17		
N participant	102		

VAS Loudness: Speech Level			
Predictors	*Estimates*	*95% CI*	*p*
N task	4		
N observations	408		
AIC	2,628.67		
Conditional R2	0.927		
Marginal R2	0.869		

Speech Level Combinations

Model	*AIC*
TVL Mean	2,628.667
Mean Intensity	2,689.560
LKFS	2,717.409
Active Speech Level	2,719.213
TVL Mean + Mean Intensity	2,589.014
TVL Mean + Active Speech Level	2,591.979
TVL Mean + LKFS	2,593.182

Predictor	*VIF*
TVLlong_mean	5.15
meanInt	4.90
Group	1.17

Predictor	*VIF*
TVLlong_mean	3.87
activeSL	3.81

Predictor	*VIF*
Group	1.16

Predictor	*VIF*
TVLlong_mean	4.58
Lkfs	4.32
Group	1.17

VIF was acceptable at this stage, but with addition of spectral balance predictors VIF exceeded threshold so a combination was not maintained.

VIF with TVL Mean and Active Speech Level with Mid-Ratio:

Predictor	*VIF*
TVLlong_mean	6.08
activeSL	4.00
mid_ratio	3.09
Group	1.22

VIF with only TVL Mean and Mid-Ratio:

Predictor	*VIF*
TVLlong_mean	3.15
mid_ratio	3.25
Group	1.23

Spectral Balance

Model	*AIC*
TVL Mean	2,628.667
TVL Mean + Tilt	2,625.577
TVL Mean + Voiced Tilt	2,624.539
TVL Mean + Tilt Ratio	2,629.492
TVL Mean + Mid-Ratio	2,585.668
TVL Mean + High-Ratio	2,617.098
TVL Mean + Skewness	2,612.546
TVL Mean + Kurtosis	2,617.636

VAS Loudness: Spectral Balance

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	61.25	58.31 – 64.18	< 0.001
TVLlong_mean	26.21	23.57 – 28.85	< 0.001
mid_ratio	9.27	6.57 – 11.97	< 0.001
groupIWPD	-4.20	-5.96 – -2.45	< 0.001
SD (Intercept): participant	3.21		
SD (Intercept): task	2.71		
SD Observation: Residual	5.00		
N participant	102		
N task	4		
N observations	408		
AIC	2,585.67		
Conditional R2	0.934		
Marginal R2	0.887		

Predictor	*VIF*
TVLlong_mean	3.15
mid_ratio	3.25
Group	1.23

Spectral Balance Combinations

Model	*AIC*
TVL Mean + Mid-Ratio	2,585.668
TVL Mean + Mid-Ratio + Tilt	2,582.233
TVL Mean + Mid-Ratio + Voiced Tilt	2,585.742
TVL Mean + Mid-Ratio + Skewness	2,585.222
TVL Mean + Mid-Ratio + Kurtosis	2,585.901
TVL Mean + Mid-Ratio + High-Ratio	2,585.062

No combinations were maintained.

Variability

Model	*AIC*
TVL Mean + Mid-Ratio	2,585.668
TVL Mean + Mid-Ratio + SD Intensity	2,565.796
TVL Mean + Mid-Ratio + SD TVL	2,579.262
TVL Mean + Mid-Ratio + Intensity Decay	2,582.067
TVL Mean + Mid-Ratio + TVL Decay	2,584.641

VAS Loudness: Variability

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	61.41	58.77 – 64.05	< 0.001

VAS Loudness: Variability

Predictors	*Estimates*	*95% CI*	*p*
TVLlong_mean	26.62	24.04 – 29.19	< 0.001
mid_ratio	9.71	7.08 – 12.34	< 0.001
sdInt	2.96	1.71 – 4.21	< 0.001
groupIWPD	-4.50	-6.21 – -2.80	< 0.001
SD (Intercept): participant	3.08		
SD (Intercept): task	2.40		
SD Observation: Residual	4.90		
N participant	102		
N task	4		
N observations	408		
AIC	2,565.80		
Conditional R2	0.936		
Marginal R2	0.896		

Predictor	*VIF*
TVLlong_mean	3.18
mid_ratio	3.28
sdInt	1.08
Group	1.24

Interactions

Model	*AIC*
TVL Mean + Mid-Ratio + SD Intensity	2,565.796
TVL Mean * Mid-Ratio + SD Intensity	2,504.751
TVL Mean + Mid-Ratio * SD Intensity	2,557.965

Model	*AIC*
TVL Mean * SD Intensity + Mid-Ratio	2,557.603
TVL Mean * Group + Mid-Ratio + SD Intensity	2,545.649
TVL Mean + Mid-Ratio * Group + SD Intensity	2,550.849
TVL Mean + Mid-Ratio + SD Intensity * Group	2,565.389
TVL Mean * Mid-Ratio + TVL Mean * Group + SD Intensity	2,502.496
TVL Mean * Mid-Ratio + Mid-Ratio * Group + SD Intensity	2,502.733

The combination of interactions between TVL Mean/Mid-Ratio and TVL Mean/Group significantly improved performance over TVL Mean/Mid-Ratio alone, but VIF exceeded threshold and the combination was not maintained.

Predictor	*VIF*
TVLlong_mean	6.77
mid_ratio	4.70
Group	1.34
sdInt	1.11
TVLlong_mean:mid_ratio	1.80
TVLlong_mean:group	4.00

VAS Loudness: Interactions

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	63.22	60.87 – 65.57	< 0.001
TVLlong_mean	30.37	27.90 – 32.84	< 0.001
mid_ratio	4.14	1.42 – 6.87	0.0031
sdInt	3.28	2.13 – 4.44	< 0.001
groupIWPD	-4.32	-5.78 – -2.86	< 0.001
*TVLlong_mean * mid_ratio*	-8.79	-10.87 – -6.70	< 0.001

VAS Loudness: Interactions

Predictors	*Estimates*	*95% CI*	*p*
SD (Intercept): participant	2.38		
SD (Intercept): task	2.10		
SD Observation: Residual	4.68		
N participant	102		
N task	4		
N observations	408		
AIC	2,504.75		
Conditional R2	0.943		
Marginal R2	0.916		

Predictor	*VIF*
TVLlong_mean	3.81
mid_ratio	4.65
sdInt	1.09
Group	1.25
TVLlong_mean:mid_ratio	1.42

Random Slopes

Model	*AIC*
TVL Mean * Mid-Ratio + SD Intensity	2,504.751
TVL Slope	2,495.569
Mid-Ratio Slope	2,504.538
SD Intensity Slope	2,499.988

Final Model

Predictor	*VIF*
TVL Mean	2.57
Mid-Ratio	2.81
SD Intensity	1.05
Group	1.17
TVL Mean * Mid-Ratio	1.14

VAS Loudness

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	63.03	60.71 – 65.35	< 0.001
TVL Mean	30.19	27.64 – 32.73	< 0.001
Mid-Ratio	4.55	1.92 – 7.18	< 0.001
SD Intensity	3.02	1.89 – 4.14	< 0.001
Group (IWPD)	-3.65	-4.98 – -2.31	< 0.001
*TVL Mean * Mid-Ratio*	-8.39	-10.97 – -5.82	< 0.001
SD: Participant Intercept	1.88		
SD: TVL Slope	4.45		
SD: Task Intercept	2.12		
SD: Residual	4.47		
N participant	102		
N task	4		
N observations	408		
AIC	2,495.57		
Conditional R2	0.947		
Marginal R2	0.913		

Measure	*Mean (SD)*
TVL Mean	12.08 (3.91)
Mid-Ratio	0.12 (0.10)
SD Intensity	12.63 (2.43)

Intelligibility

Baseline Model

Intelligibility: Baseline

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	86.44	80.39 – 92.48	< 0.001
groupIWPD	-26.80	-33.37 – -20.22	< 0.001
SD (Intercept): participant	16.37		
SD (Intercept): task	3.66		
SD Observation: Residual	8.00		
N participant	102		
N task	4		
N observations	408		
AIC	3,158.43		
Conditional R2	0.878		
Marginal R2	0.340		

Speech Level

Model	*AIC*
Baseline	3,158.433
Mean Intensity	3,125.205

Model	*AIC*
Median Intensity	3,144.637
Max Intensity	3,134.911
TVL	3,139.141
TVL Mean	3,139.141
LKFS	3,127.683
Active Speeech Level	3,125.911

Intelligibility: Speech Level

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	84.23	79.41 – 89.06	< 0.001
meanInt	11.63	7.98 – 15.29	< 0.001
groupIWPD	-22.78	-28.98 – -16.59	< 0.001
SD (Intercept): participant	15.04		
SD (Intercept): task	1.66		
SD Observation: Residual	7.86		
N participant	102		
N task	4		
N observations	408		
AIC	3,125.21		
Conditional R2	0.876		
Marginal R2	0.417		

Speech Level Combinations

Model	*AIC*
Mean Intensity	3,125.205
Active Speech Level	3,125.911

Model	*AIC*
TVL Mean	3,121.885
Mean Intensity + Active Speech Level	3,121.110
Mean Intensity + TVL Mean	3,118.001
Active Speech Level + TVL Mean	3,117.575

Predictor	*VIF*
meanInt	8.75
activeSL	8.85
Group	1.06

Predictor	*VIF*
meanInt	4.39
TVLlong_mean	4.47
Group	1.06

Predictor	*VIF*
activeSL	3.42
TVLlong_mean	3.44
Group	1.07

VIF was acceptable at this stage for two of the combination models, but with addition of spectral balance predictors VIF exceeded threshold so a combination was not maintained.

Spectral Balance

Model	*AIC*
Mean Intensity	3,125.205
Mean Intensity + Tilt	3,094.701
Mean Intensity + Voiced Tilt	3,115.520
Mean Intensity + Tilt Ratio	3,111.155
Mean Intensity + Mid-Ratio	3,085.227
Mean Intensity + High-Ratio	3,094.410
Mean Intensity + Skewness	3,094.049
Mean Intensity + Kurtosis	3,097.439

Intelligibility: Spectral Balance

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	81.68	77.74 – 85.62	< 0.001
meanInt	3.46	-0.87 – 7.79	0.1187
mid_ratio	16.65	11.87 – 21.44	< 0.001
groupIWPD	-18.14	-23.47 – -12.80	< 0.001
SD (Intercept): participant	12.44		
SD (Intercept): task	0.75		
SD Observation: Residual	7.87		
N participant	102		
N task	4		
N observations	408		
AIC	3,085.23		
Conditional R2	0.868		
Marginal R2	0.539		

Predictor	*VIF*
meanInt	1.86
mid_ratio	1.95
Group	1.10

Spectral Balance Combinations

Model	*AIC*
Mean Intensity + Mid-Ratio	3,085.227
Mean Intensity + High-Ratio	3,094.410
Mean Intensity + Mid-Ratio + High-Ratio	3,071.321
Mean Intensity + Mid-Ratio + Skewness	3,075.758
Mean Intensity + Mid-Ratio + Tilt	3,077.905
Mean Intensity + High-Ratio + Skewness	3,075.613
Mean Intensity + High-Ratio + Tilt	3,079.987
Mean Intensity + Mid-Ratio + High-Ratio + Skewness	3,064.145
Mean Intensity + Mid-Ratio + High-Ratio + Tilt	3,067.296

Intelligibility: Spectral Balance Combinations

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	80.09	76.38 – 83.79	< 0.001
meanInt	4.89	0.45 – 9.33	0.032
mid_ratio	9.53	3.54 – 15.52	0.002
high_ratio	6.53	2.60 – 10.46	0.0012
skew	-5.62	-10.06 – -1.19	0.0134
groupIWPD	-15.23	-20.34 – -10.12	< 0.001
SD (Intercept): participant	11.49		
SD (Intercept): task	0.55		

Intelligibility: Spectral Balance Combinations

Predictors	*Estimates*	*95% CI*	*p*
SD Observation: Residual	7.85		
N participant	102		
N task	4		
N observations	408		
AIC	3,064.15		
Conditional R2	0.870		
Marginal R2	0.592		

Predictor	*VIF*
meanInt	2.17
mid_ratio	3.37
high_ratio	1.67
Skew	1.71
Group	1.17

Variability

Model	*AIC*
Mean Intensity + Mid-Ratio + High-Ratio + Skewness	3,064.145
Mean Intensity + Mid-Ratio + High-Ratio + Skewness + SD Intensity	3,063.996
Mean Intensity + Mid-Ratio + High-Ratio + Skewness + SD TVL	3,061.231
Mean Intensity + Mid-Ratio + High-Ratio + Skewness + Intensity Decay	3,064.117
Mean Intensity + Mid-Ratio + High-Ratio + Skewness + TVL Decay	3,064.062

Intelligibility: Variability

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	79.98	76.26 – 83.71	< 0.001
meanInt	6.60	1.37 – 11.83	0.0139
mid_ratio	9.69	3.69 – 15.69	0.0017
high_ratio	6.26	2.32 – 10.21	0.002
skew	-6.36	-10.89 – -1.83	0.0062
TVLlong_sd	-2.54	-6.12 – 1.04	0.1656
groupIWPD	-15.04	-20.15 – -9.93	< 0.001
SD (Intercept): participant	11.40		
SD (Intercept): task	0.74		
SD Observation: Residual	7.85		
N participant	102		
N task	4		
N observations	408		
AIC	3,061.23		
Conditional R2	0.870		
Marginal R2	0.594		

Predictor	*VIF*
meanInt	2.88
mid_ratio	3.29
high_ratio	1.67
Skew	1.78
TVLlong_sd	1.98
Group	1.18

Interactions

Model	*AIC*
Mean Intensity + Mid-Ratio + High-Ratio + Skewness + SD TVL	3,061.231
Mean Intensity * Group + Mid-Ratio + High-Ratio + Skewness + SD TVL	3,054.047
Mean Intensity + Mid-Ratio * Group + High-Ratio + Skewness + SD TVL	3,048.083
Mean Intensity + Mid-Ratio + High-Ratio * Group + Skewness + SD TVL	3,053.907
Mean Intensity + Mid-Ratio + High-Ratio + Skewness * Group + SD TVL	3,056.697
Mean Intensity + Mid-Ratio + High-Ratio + Skewnessness + SD TVL * Group	3,054.021
Mean Intensity * Mid-Ratio + High-Ratio + Skewness + SD TVL	3,056.994
Mean Intensity + Mid-Ratio * High-Ratio + Skewness + SD TVL	3,054.118
Mean Intensity + Mid-Ratio + High-Ratio * Skewness + SD TVL	3,058.335
Mean Intensity * High-Ratio + Mid-Ratio + Skewness + SD TVL	3,058.406
Mean Intensity * Skewness + Mid-Ratio + High-Ratio + SD TVL	3,055.836
Mean Intensity + Mid-Ratio * Group + Mid-Ratio * High-Ratio + Skewness + SD TVL	3,045.466

All interactions improved model performance, and combinations demonstrated appropriate VIF. However, in the combination model, the interaction between mid-ratio and high-ratio was no longer significant so it was not maintained.

Predictor	*VIF*
meanInt	2.90
mid_ratio	5.38

Predictor	*VIF*
high_ratio	1.65
Group	1.24
Skew	2.05
TVLlong_sd	2.08
mid_ratio:high_ratio	1.67
mid_ratio:group	4.07

Intelligibility: Interactions

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	82.27	78.36 – 86.17	< 0.001
meanInt	5.32	0.09 – 10.55	0.047
mid_ratio	2.39	-4.95 – 9.72	0.5244
groupIWPD	-16.74	-21.81 – -11.67	< 0.001
high_ratio	6.00	2.09 – 9.91	0.0028
skew	-4.85	-9.44 – -0.26	0.039
TVLlong_sd	-1.43	-5.06 – 2.20	0.4399
*mid_ratio * groupIWPD*	12.16	4.97 – 19.36	0.001
SD (Intercept): participant	11.02		
SD (Intercept): task	0.93		
SD Observation: Residual	7.80		
N participant	102		
N task	4		
N observations	408		
AIC	3,048.08		
Conditional R2	0.873		

Intelligibility: Interactions

Predictors	*Estimates*	*95% CI*	*p*
Marginal R2	0.616		

Predictor	*VIF*
meanInt	2.85
mid_ratio	4.96
Group	1.24
high_ratio	1.66
Skew	1.90
TVLlong_sd	2.04
mid_ratio:group	3.16

Random Slopes

Model	*AIC*
Mean Intensity + Mid-Ratio * Group + High-Ratio + Skewness + SD TVL	3,048.083
Mean Intensity Slope	3,042.219
Mid-Ratio Slope	3,031.576
High-Ratio Slope	3,039.629
Skewness Slope	3,038.340
SD TVL Slope	3,043.125
Mean Intensity and Mid-Ratio Slopes	3,032.332

Final Model

Predictor	*VIF*
Mean Intensity	2.41
Mid-Ratio	3.67

Predictor	*VIF*
Group	1.48
High-Ratio	1.42
Skewness	1.63
SD TVL	1.91
Mid-Ratio * Group	2.57

Intelligibility

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	83.25	79.37 – 87.14	< 0.001
Mean Intensity	4.27	-0.89 – 9.42	0.1054
Mid-Ratio	0.92	-6.72 – 8.57	0.813
Group (IWPD)	-16.94	-21.70 – -12.17	< 0.001
High-Ratio	6.50	2.66 – 10.35	0.001
Skewness	-2.39	-7.28 – 2.51	0.3406
SD TVL	-0.83	-4.44 – 2.78	0.654
*Mid-Ratio * Group*	13.71	5.66 – 21.75	0.0011
SD: Participant Intercept	9.93		
SD: Mid-Ratio	9.67		
SD: Task Intercept	1.38		
SD: Residual	7.54		
N participant	102		
N task	4		
N observations	408		
AIC	3,031.58		
Conditional R2	0.871		
Marginal R2	0.592		

Measure	*Mean (SD)*
Mean Intensity	67.86 (4.30)
Mid-Ratio	0.12 (0.10)
High-Ratio	0.04 (0.12)
Skewness	11.89 (7.31)
SD TVL	3.37 (1.14)

Intelligibility: IWPD

For interpretation of model coefficients, means and standard deviations of each measure within IWPDs are presented in the table below.

Measure	*Mean (SD)*
Mean Intensity	66.56 (4.66)
Median Intensity	63.12 (7.63)
Max Intensity	73.56 (4.43)
TVL	15.27 (5.55)
TVL Mean	10.71 (4.18)
LKFS	-27.88 (4.57)
Active Speech Level	-29.32 (5.14)
Tilt	-28.61 (4.95)
Voiced Tilt	-29.27 (5.16)
Tilt Ratio	1.02 (0.06)
Mid-Ratio	0.08 (0.11)
High-Ratio	-0.02 (0.12)
Skewness	14.11 (8.86)
Kurtosis	397.51 (544.93)
SD Intensity	12.96 (2.70)
SD TVL	3.12 (1.21)
*Intensity Decay **	-18.92 (27.09)
*TVL Decay **	-0.90 (1.23)

Means and standard deviations among IWPDs based on values from all 14 speech tasks. * Values for intensity decay and TVL decay are expressed in scientific notation (x 10^3).

Baseline Model

IWPD Intelligibility: Baseline

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	59.64	53.85 – 65.43	< 0.001
SD (Intercept): participant	21.46		
SD Observation: Residual	10.51		
N participant	56.00		
N observations	224		
AIC	1,851		
Conditional R2	0.807		
Marginal R2	0.000		

Speech Level

Model	*AIC*
Baseline	1,851.208
Mean Intensity	1,802.089
Median Intensity	1,829.627
Max Intensity	1,815.382
TVL	1,813.344
TVL Mean	1,813.344
LKFS	1,804.516
Active Speech Level	1,805.781

IWPD Intelligibility: Speech Level

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	62.48	57.38 – 67.58	< 0.001

IWPD Intelligibility: Speech Level

Predictors	*Estimates*	*95% CI*	*p*
meanInt	18.27	13.34 – 23.20	< 0.001
SD (Intercept): participant	18.66		
SD Observation: Residual	9.57		
N participant	56		
N observations	224		
AIC	1,802.09		
Conditional R2	0.829		
Marginal R2	0.178		

Speech Level Combinations

Model	*AIC*
Mean Intensity	1,802.089
Active Speech Level	1,805.781
LKFS	1,804.516
Mean Intensity + Active Speech Level	1,797.986
Mean Intensity + LKFS	1,797.189
Active Speech Level + LKFS	1,799.781

Predictor	*VIF*
meanInt	8.13
activeSL	8.13

Predictor	*VIF*
meanInt	22.93
Lkfs	22.93

Predictor	*VIF*
activeSL	12.90
Lkfs	12.90

Spectral Balance

Model	*AIC*
Mean Intensity	1,802.089
Mean Intensity + Tilt	1,772.929
Mean Intensity + Voiced Tilt	1,786.941
Mean Intensity + Tilt Ratio	1,793.781
Mean Intensity + Mid-Ratio	1,770.445
Mean Intensity + High-Ratio	1,780.047
Mean Intensity + Skewness	1,785.246
Mean Intensity + Kurtosis	1,785.757

IWPD Intelligibility: Spectral Balance

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	64.68	60.46 – 68.91	< 0.001
meanInt	4.88	-1.63 – 11.39	0.1433
mid_ratio	21.20	14.15 – 28.25	< 0.001
SD (Intercept): participant	14.94		
SD Observation: Residual	9.41		
N participant	56		

IWPD Intelligibility: Spectral Balance

Predictors	*Estimates*	*95% CI*	*p*
N observations	224		
AIC	1,770.44		
Conditional R2	0.820		
Marginal R2	0.367		

Predictor	*VIF*
meanInt	1.99
mid_ratio	1.99

Spectral Balance Combinations

Model	*AIC*
Mean Intensity + Mid-Ratio	1,770.445
Mean Intensity + High-Ratio	1,780.047
Mean Intensity + Mid-Ratio + High-Ratio	1,762.796
Mean Intensity + Mid-Ratio + Tilt	1,762.924
Mean Intensity + High-Ratio + Tilt	1,764.329

IWPD Intelligibility: Spectral Balance Combinations

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	65.58	61.32 – 69.85	< 0.001
meanInt	4.27	-2.20 – 10.74	0.197
mid_ratio	16.45	8.41 – 24.49	< 0.001
high_ratio	8.08	1.24 – 14.93	0.0216
SD (Intercept): participant	14.86		

IWPD Intelligibility: Spectral Balance Combinations

Predictors	*Estimates*	*95% CI*	*p*
SD Observation: Residual	9.31		
N participant	56		
N observations	224		
AIC	1,762.80		
Conditional R2	0.826		
Marginal R2	0.382		

Predictor	*VIF*
meanInt	2.00
mid_ratio	2.64
high_ratio	1.79

Variability

Model	*AIC*
Mean Intensity + Mid-Ratio + High-Ratio	1,762.796
Mean Intensity + Mid-Ratio + High-Ratio + SD Intensity	1,761.391
Mean Intensity + Mid-Ratio + High-Ratio + SD TVL	1,760.647
Mean Intensity + Mid-Ratio + High-Ratio + Intensity Decay	1,762.148
Mean Intensity + Mid-Ratio + High-Ratio + TVL Decay	1,761.659

Interactions

Model	*AIC*
Mean Intensity + Mid-Ratio + High-Ratio	1,762.796
Mean Intensity * Mid-Ratio + High-Ratio	1,759.475
Mean Intensity * High-Ratio + Mid-Ratio	1,758.708

Model	*AIC*
Mean Intensity + Mid-Ratio * High-Ratio	1,760.202

Random Slopes

Model	*AIC*
Mean Intensity + Mid-Ratio + High-Ratio	1,762.796
Mean Intensity Slope	1,764.739
Mid-Ratio Slope	1,761.551
High-Ratio Slope	1,766.184
Mean Intensity Removed	1,766.622
Mean Intensity Removed, By-Task Intercept Reincorporated	1,768.596

Final Model

Predictor	*VIF*
Mid-Ratio	1.76
High-Ratio	1.76

IWPD Intelligibility

Predictors	*Estimates*	*95% CI*	*p*
(Intercept)	65.60	61.37 – 69.83	< 0.001
Mid-Ratio	19.41	12.76 – 26.06	< 0.001
High-Ratio	8.42	1.57 – 15.27	0.0169
SD: Participant Intercept	14.53		
SD: Task Intercept	0.51		
SD: Residual	9.37		
N participant	56		
N task	4		

IWPD Intelligibility

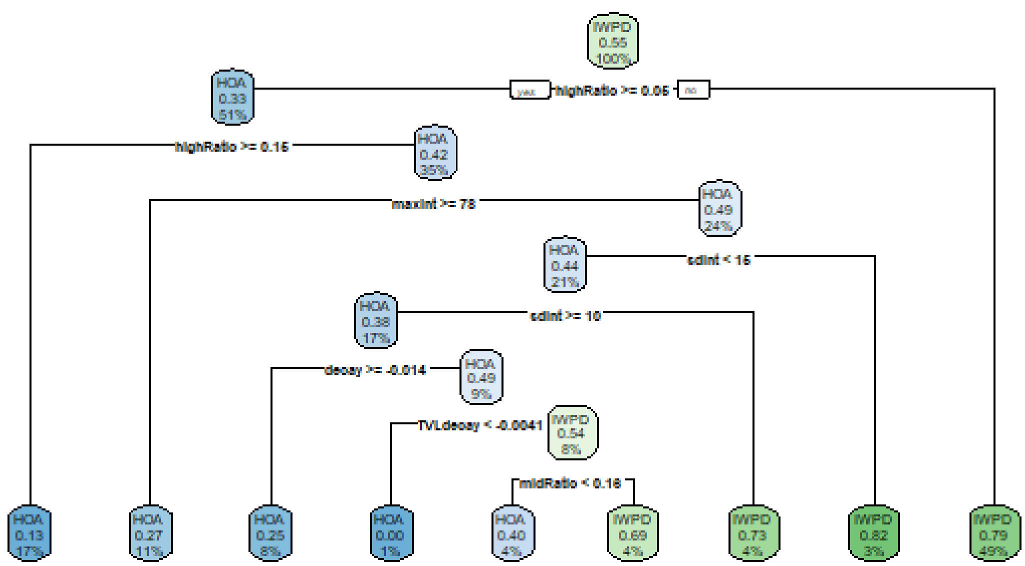
Predictors	*Estimates*	*95% CI*	*p*
N observations	224		
AIC	1,768.60		
Conditional R2	0.820		
Marginal R2	0.385		

Measure	*Mean (SD)*
Mid-Ratio	0.08 (0.11)
High-Ratio	-0.02 (0.12)

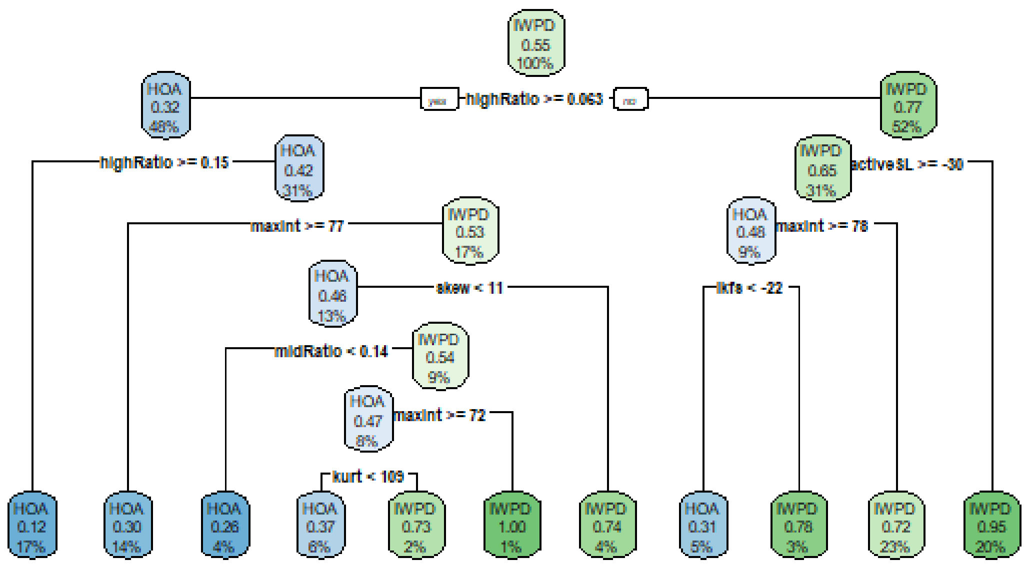
Appendix H: Classification Trees

Each of the 10 classification decision trees described in Chapter 4 are presented below. All trees are equally valid, so each should be considered when interpreting trends. Decision trees are unstable and prone to overfit, and the trends are more important that the individual splits or cut-offs of one particular tree.

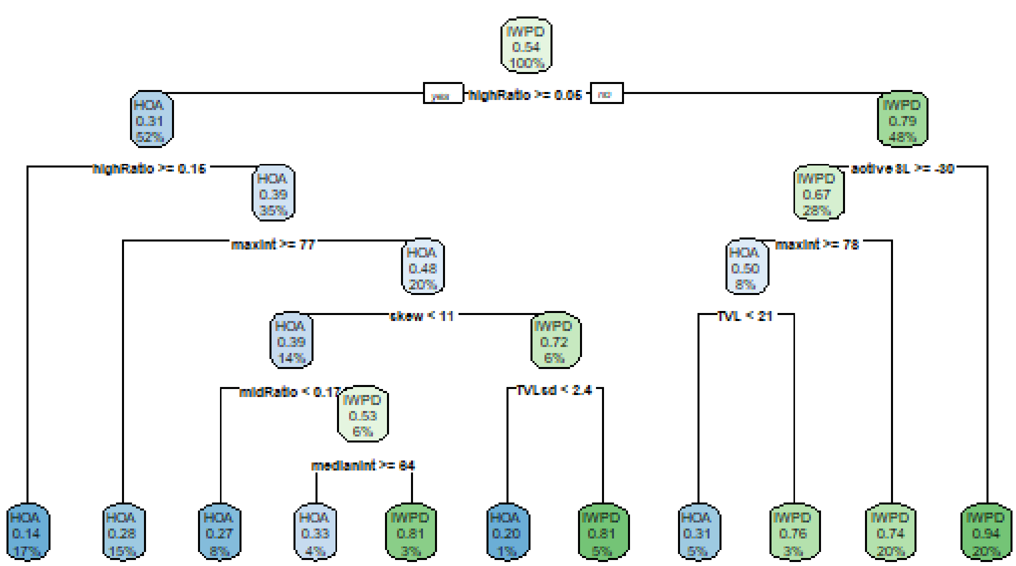
Tree 1



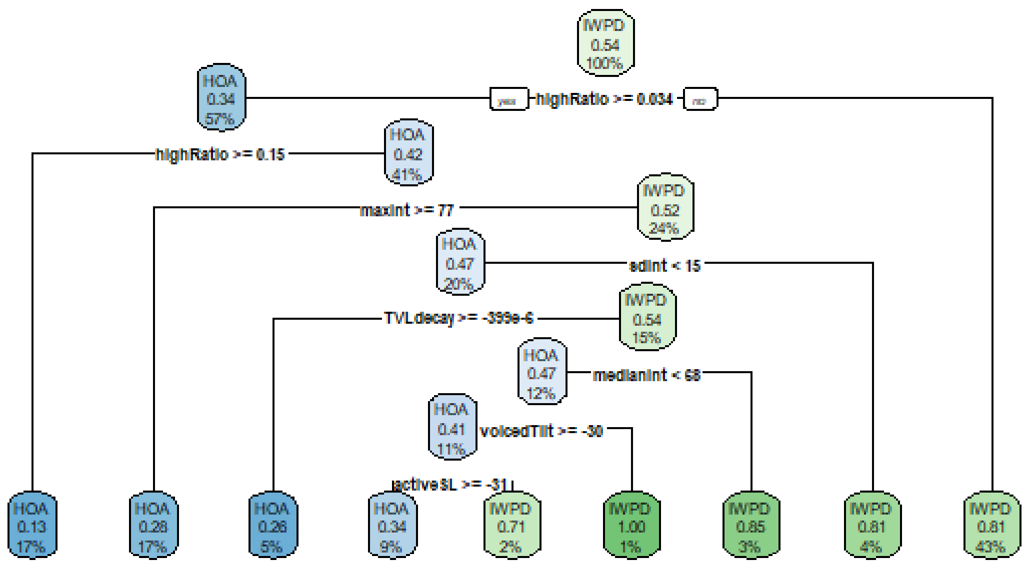
Tree 2


IWPD 0.55 100%
yes
highRatio >= 0.063
no
HOA 0.32 48%
IWPD 0.77 52%
highRatio >= 0.15
HOA 0.42 31%
IWPD 0.65 31%
activeSL >= -30
maxint >= 77
IWPD 0.53 17%
HOA 0.48 9%
maxint >= 78
HOA 0.46 13%
skew < 11
lkfs < -22
midRatio < 0.14
IWPD 0.54 9%
HOA 0.47 8%
maxint >= 72
kurt < 109
HOA 0.12 17%
HOA 0.30 14%
HOA 0.26 4%
HOA 0.37 6%
IWPD 0.73 2%
IWPD 1.00 1%
IWPD 0.74 4%
HOA 0.31 5%
IWPD 0.78 3%
IWPD 0.72 23%
IWPD 0.95 20%

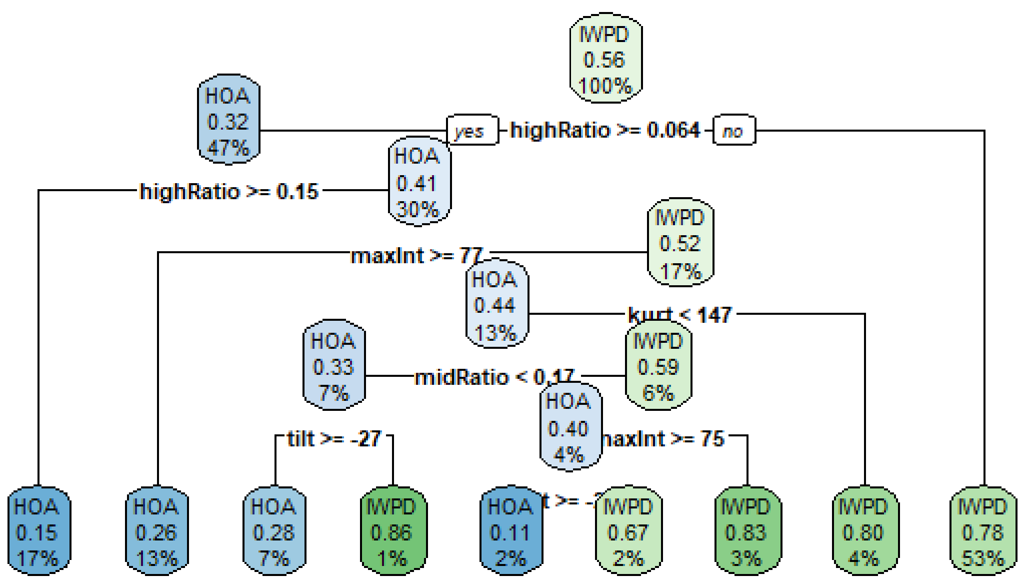
Tree 3


IWPD 0.54 100%
yes
highRatio >= 0.05
no
HOA 0.31 52%
IWPD 0.79 48%
highRatio >= 0.16
HOA 0.39 35%
activeSL >= -30
IWPD 0.67 28%
maxint >= 77
HOA 0.48 20%
HOA 0.50 8%
maxint >= 78
HOA 0.39 14%
skew < 11
IWPD 0.72 6%
TVL < 21
midRatio < 0.17
IWPD 0.53 6%
TVLsd < 2.4
medianint >= 84
HOA 0.14 17%
HOA 0.28 15%
HOA 0.27 8%
HOA 0.33 4%
IWPD 0.81 3%
HOA 0.20 1%
IWPD 0.81 5%
HOA 0.31 5%
IWPD 0.76 3%
IWPD 0.74 20%
IWPD 0.94 20%

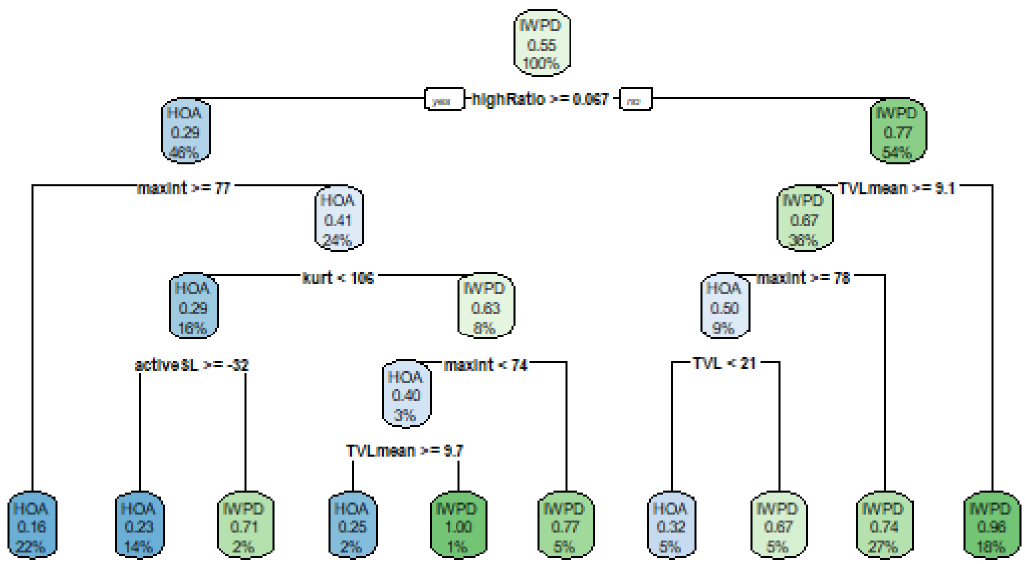
Tree 4


IWPD 0.54 100%
yes
highRatio >= 0.034
no
HOA 0.34 57%
highRatio >= 0.15
HOA 0.42 41%
maxInt >= 77
IWPD 0.52 24%
HOA 0.47 20%
sdInt < 15
TVLdecay >= -399e-6
IWPD 0.54 15%
HOA 0.47 12%
medianInt < 68
HOA 0.41 11%
voicedTilt >= -30
activeSL >= -31
HOA 0.13 17%
HOA 0.28 17%
HOA 0.26 5%
HOA 0.34 9%
IWPD 0.71 2%
IWPD 1.00 1%
IWPD 0.85 3%
IWPD 0.81 4%
IWPD 0.81 43%

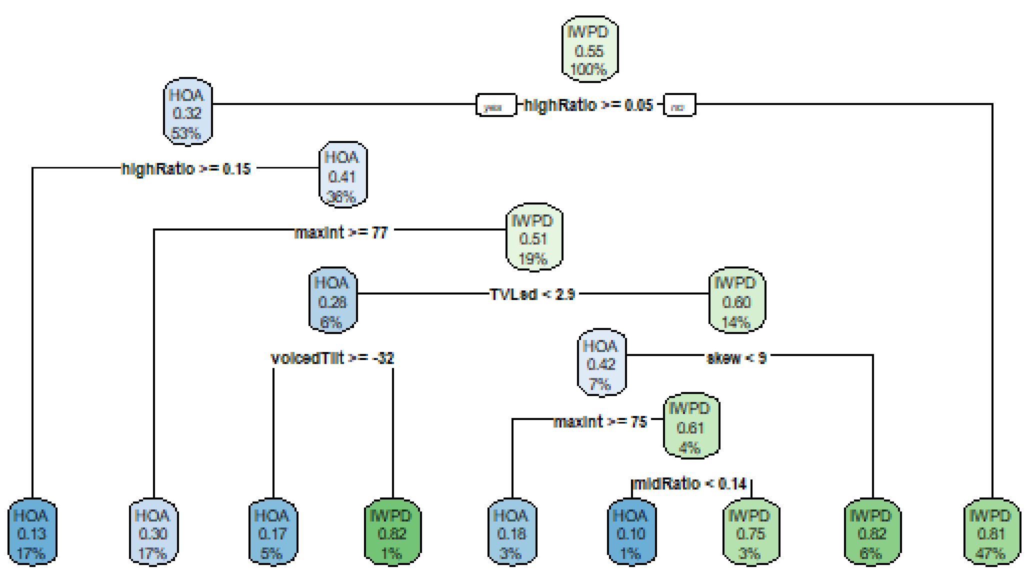
Tree 5


IWPD 0.56 100%
yes
highRatio >= 0.064
no
HOA 0.32 47%
highRatio >= 0.15
HOA 0.41 30%
maxInt >= 77
IWPD 0.52 17%
HOA 0.44 13%
kurt < 147
HOA 0.33 7%
midRatio < 0.17
IWPD 0.59 6%
HOA 0.40 4%
tilt >= -27
HOA 0.15 17%
HOA 0.26 13%
HOA 0.28 7%
IWPD 0.86 1%
HOA 0.11 2%
IWPD 0.67 2%
IWPD 0.83 3%
IWPD 0.80 4%
IWPD 0.78 53%

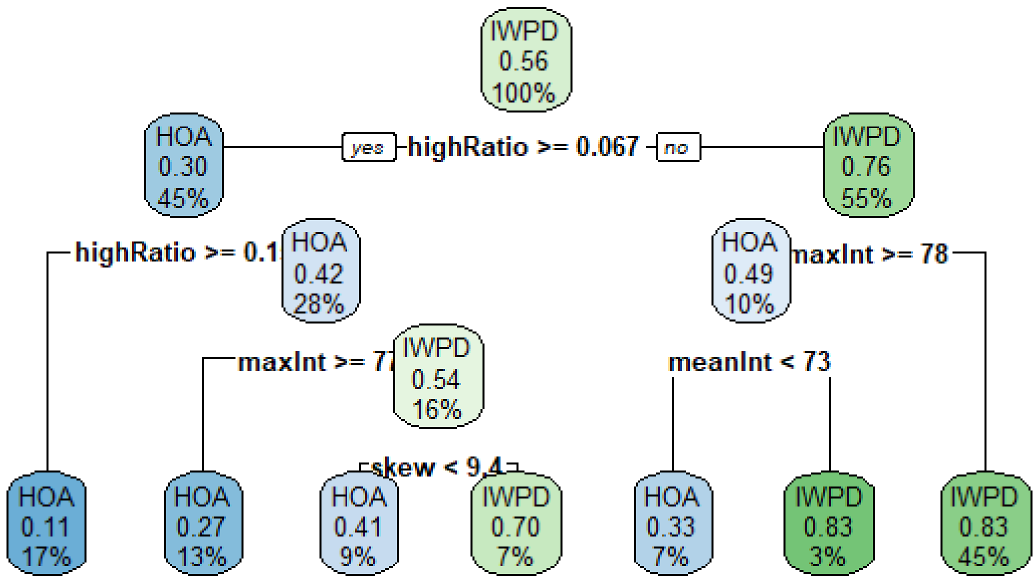
Tree 6


IWPD 0.55 100%
yes
highRatio >= 0.067
no
HOA 0.29 46%
IWPD 0.77 54%
maxInt >= 77
HOA 0.41 24%
TVLmean >= 9.1
IWPD 0.67 36%
kurt < 106
HOA 0.29 16%
IWPD 0.63 8%
maxInt >= 78
HOA 0.50 9%
activeSL >= -32
maxInt < 74
HOA 0.40 3%
TVL < 21
TVLmean >= 9.7
HOA 0.16 22%
HOA 0.23 14%
IWPD 0.71 2%
HOA 0.25 2%
IWPD 1.00 1%
IWPD 0.77 5%
HOA 0.32 5%
IWPD 0.67 5%
IWPD 0.74 27%
IWPD 0.96 18%

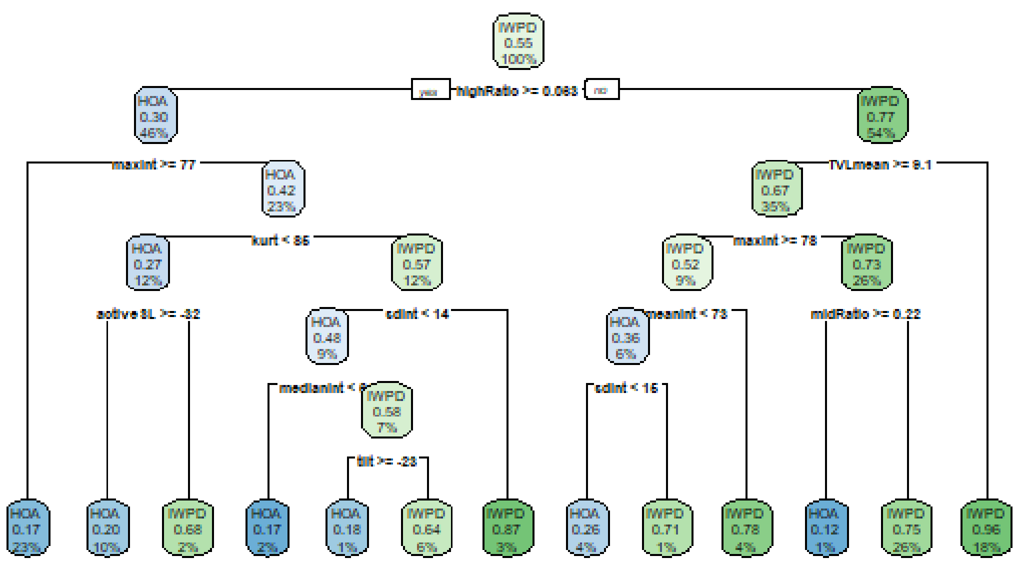
Tree 7


IWPD 0.55 100%
HOA 0.32 53%
yes
highRatio >= 0.05
no
highRatio >= 0.15
HOA 0.41 36%
maxInt >= 77
IWPD 0.51 19%
HOA 0.28 6%
TVLsd < 2.9
IWPD 0.60 14%
HOA 0.42 7%
skew < 9
voicedTilt >= -32
maxInt >= 75
IWPD 0.61 4%
midRatio < 0.14
HOA 0.13 17%
HOA 0.30 17%
HOA 0.17 5%
IWPD 0.82 1%
HOA 0.18 3%
HOA 0.10 1%
IWPD 0.75 3%
IWPD 0.82 6%
IWPD 0.81 47%

Tree 8


IWPD
0.56
100%
yes
highRatio >= 0.067
no
HOA
0.30
45%
IWPD
0.76
55%
highRatio >= 0.1
HOA
0.42
28%
HOA
0.49
10%
maxInt >= 78
maxInt >= 77
IWPD
0.54
16%
meanInt < 73
skew < 9.4
HOA
0.11
17%
HOA
0.27
13%
HOA
0.41
9%
IWPD
0.70
7%
HOA
0.33
7%
IWPD
0.83
3%
IWPD
0.83
45%

Tree 9


IWPD
0.55
100%
yes
highRatio >= 0.068
no
HOA
0.30
46%
IWPD
0.77
54%
maxInt >= 77
HOA
0.42
23%
IWPD
0.67
35%
TVLmean >= 9.1
HOA
0.27
12%
kurt < 86
IWPD
0.57
12%
IWPD
0.52
9%
maxInt >= 78
IWPD
0.73
26%
active SL >= -32
HOA
0.48
9%
sdInt < 14
HOA
0.36
6%
meanInt < 73
midRatio >= 0.22
medianInt < 6
IWPD
0.58
7%
sdInt < 16
tilt >= -23
HOA
0.17
23%
HOA
0.20
10%
IWPD
0.68
2%
HOA
0.17
2%
HOA
0.18
1%
IWPD
0.64
6%
IWPD
0.87
3%
HOA
0.26
4%
IWPD
0.71
1%
IWPD
0.78
4%
HOA
0.12
1%
IWPD
0.75
26%
IWPD
0.96
18%

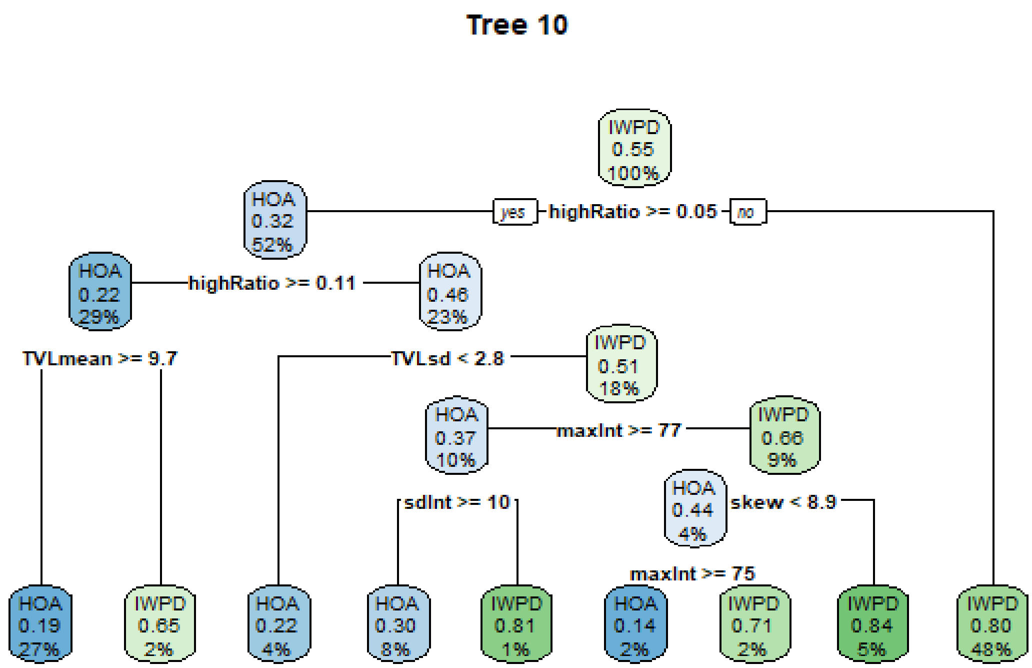


Figure H.1: Classification trees predicting PD status (IWPD vs. HOA) based on acoustic measures (N = 1424).

Appendix J: Spectral Manipulation

Measure	*Group*	*Down 10 dB*	*Down 5 dB*	*Unaltered Speech*	*Up 5 dB*	*Up 10 dB*
VAS Loudness	HOA	61.08 (9.76)	65.28 (9.43)	71.27 (9.69)	75.58 (9.49)	81.54 (9.72)
VAS Loudness	IWPD	45.22 (16.95)	49.49 (18.29)	54.31 (17.78)	59.11 (18.37)	64.83 (18.91)
Mean Intensity	HOA	70.27 (3.19)	70.27 (3.19)	70.27 (3.19)	70.27 (3.19)	70.27 (3.19)
Mean Intensity	IWPD	67.90 (4.46)	67.90 (4.46)	67.90 (4.46)	67.90 (4.46)	67.90 (4.46)
Median Intensity	HOA	66.81 (5.20)	66.85 (5.13)	66.95 (4.90)	67.18 (4.57)	67.62 (4.24)
Median Intensity	IWPD	65.26 (6.00)	65.27 (5.95)	65.33 (5.79)	65.44 (5.58)	65.57 (5.30)
Max Intensity	HOA	76.97 (2.80)	76.97 (2.79)	76.96 (2.79)	76.94 (2.84)	77.15 (3.00)
Max Intensity	IWPD	73.92 (4.08)	73.92 (4.08)	73.93 (4.09)	73.93 (4.12)	74.02 (4.16)
TVL	HOA	16.73 (3.52)	17.89 (3.76)	19.43 (4.07)	21.38 (4.45)	23.53 (4.83)
TVL	IWPD	13.69 (4.89)	14.65 (5.24)	15.96 (5.69)	17.71 (6.22)	19.82 (6.77)
TVL Mean	HOA	12.16 (2.43)	13.06 (2.60)	14.24 (2.81)	15.73 (3.06)	17.37 (3.32)
TVL Mean	IWPD	9.82 (3.61)	10.56 (3.88)	11.58 (4.22)	12.92 (4.63)	14.52 (5.07)
LKFS	HOA	-24.25 (3.26)	-24.22 (3.25)	-24.15 (3.24)	-23.96 (3.20)	-23.52 (3.13)
LKFS	IWPD	-26.68 (4.50)	-26.66 (4.50)	-26.62 (4.50)	-26.49 (4.53)	-26.21 (4.58)
Active Speech Level	HOA	-24.68 (3.09)	-24.68 (3.09)	-24.71 (3.10)	-24.76 (3.11)	-24.86 (3.18)
Active Speech Level	IWPD	-28.23 (4.79)	-28.23 (4.79)	-28.24 (4.79)	-28.24 (4.81)	-28.18 (4.80)

Measure	*Group*	*Down 10 dB*	*Down 5 dB*	*Unaltered Speech*	*Up 5 dB*	*Up 10 dB*
Tilt	HOA	-28.70 (5.25)	-26.92 (4.86)	-24.21 (4.53)	-20.63 (4.35)	-16.42 (4.28)
Tilt	IWPD	-32.73 (5.89)	-30.71 (5.37)	-27.80 (4.92)	-24.10 (4.69)	-19.85 (4.61)
Voiced Tilt	HOA	-28.60 (5.18)	-27.28 (4.98)	-25.15 (4.79)	-22.08 (4.69)	-18.36 (4.52)
Voiced Tilt	IWPD	-31.78 (5.08)	-30.40 (5.09)	-28.25 (5.19)	-25.04 (5.13)	-21.40 (5.35)
Tilt Ratio	HOA	1.00 (0.02)	1.01 (0.03)	1.04 (0.05)	1.07 (0.08)	1.13 (0.14)
Tilt Ratio	IWPD	0.97 (0.04)	0.99 (0.04)	1.02 (0.05)	1.04 (0.07)	1.08 (0.10)
Mid-Ratio	HOA	0.04 (0.06)	0.12 (0.06)	0.19 (0.06)	0.25 (0.06)	0.32 (0.05)
Mid-Ratio	IWPD	-0.03 (0.11)	0.04 (0.11)	0.12 (0.10)	0.19 (0.10)	0.26 (0.09)
High-Ratio	HOA	-0.02 (0.07)	0.05 (0.07)	0.12 (0.07)	0.19 (0.07)	0.25 (0.07)
High-Ratio	IWPD	-0.15 (0.12)	-0.08 (0.11)	-0.01 (0.11)	0.07 (0.11)	0.13 (0.10)
Skewness	HOA	9.21 (4.81)	10.07 (3.93)	8.83 (3.47)	6.42 (3.22)	4.10 (2.44)
Skewness	IWPD	11.98 (12.01)	13.75 (10.19)	12.53 (7.39)	9.25 (4.98)	5.92 (3.12)
Kurtosis	HOA	231.50 (179.47)	208.47 (161.68)	139.35 (141.02)	70.70 (100.83)	28.34 (48.75)
Kurtosis	IWPD	602.42 (1252.21)	509.27 (757.50)	317.17 (380.15)	148.11 (162.13)	56.50 (59.55)
SD Intensity	HOA	12.83 (1.64)	12.42 (1.60)	11.95 (1.60)	11.44 (1.62)	10.94 (1.64)
SD Intensity	IWPD	12.93 (2.38)	12.59 (2.39)	12.12 (2.42)	11.53 (2.44)	10.86 (2.43)
SD TVL	HOA	3.86 (0.89)	4.11 (0.93)	4.44 (0.99)	4.85 (1.07)	5.31 (1.14)

Measure	*Group*	*Down 10 dB*	*Down 5 dB*	*Unaltered Speech*	*Up 5 dB*	*Up 10 dB*
SD TVL	IWPD	3.32 (1.25)	3.53 (1.34)	3.82 (1.45)	4.19 (1.58)	4.65 (1.73)
*Intensity Decay **	HOA	-48.77 (43.43)	-46.84 (40.66)	-44.68 (37.90)	-42.42 (35.37)	-40.09 (33.46)
*Intensity Decay **	IWPD	-41.52 (56.11)	-40.61 (54.11)	-39.30 (51.32)	-37.60 (47.84)	-35.53 (43.87)
*TVL Decay **	HOA	-1.96 (1.65)	-2.08 (1.71)	-2.21 (1.82)	-2.37 (2.00)	-2.53 (2.23)
*TVL Decay **	IWPD	-1.22 (2.00)	-1.28 (2.07)	-1.34 (2.17)	-1.42 (2.31)	-1.50 (2.49)

Means and standard deviations for perceived loudness and acoustic measures within the spectral manipulation experiment, separated by group (IWPD-HOA). 'Up5' and 'Up10' refer to positive 5 dB and 10 dB manipulations. 'Down5' and 'Down10' refer to negative 5 dB and 10 dB manipulations. * Mean and standard deviation values for intensity decay and TVL decay are expressed in scientific notation (x 10^3).

Full results of each ANOVA investigating the effects of group and manipulation in Experiment 2 are presented below. Each ANOVA table is followed by the Tukey's Honestly Significant Difference (HSD) post-hoc comparisons for that measure.

VAS Loudness

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	25,465.29	6,366.32	28.77	< 0.001
Group	1	33,799.71	33,799.71	152.75	< 0.001
*Manip * Group*	4	27.02	6.75	0.03	0.998
Residuals	500	110,636.79	221.27		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Loudness	*down5-down10*	4.24	-1.47 – 9.94	0.251
Loudness	*nat-down10*	9.59	3.88 – 15.29	< 0.001
Loudness	*up5-down10*	14.16	8.46 – 19.87	< 0.001
Loudness	*up10-down10*	19.99	14.29 – 25.69	< 0.001
Loudness	*nat-down5*	5.35	-0.35 – 11.05	0.078
Loudness	*up5-down5*	9.93	4.22 – 15.63	< 0.001
Loudness	*up10-down5*	15.75	10.05 – 21.45	< 0.001
Loudness	*up5-nat*	4.58	-1.13 – 10.28	0.183
Loudness	*up10-nat*	10.40	4.70 – 16.10	< 0.001
Loudness	*up10-up5*	5.83	0.12 – 11.53	0.043

Mean Intensity

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	0.00	0.00	0.00	1.000
Group	1	710.75	710.75	45.76	< 0.001
*Manip * Group*	4	0.00	0.00	0.00	1.000

Mean Intensity

	df	*SS*	*MS*	*F*	*p*
Residuals	500	7,766.49	15.53		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Mean Intensity	*down5-down10*	0.00	-1.51 – 1.51	1.000
Mean Intensity	*nat-down10*	0.00	-1.51 – 1.51	1.000
Mean Intensity	*up5-down10*	0.00	-1.51 – 1.51	1.000
Mean Intensity	*up10-down10*	-0.00	-1.51 – 1.51	1.000
Mean Intensity	*nat-down5*	0.00	-1.51 – 1.51	1.000
Mean Intensity	*up5-down5*	0.00	-1.51 – 1.51	1.000
Mean Intensity	*up10-down5*	-0.00	-1.51 – 1.51	1.000
Mean Intensity	*up5-nat*	0.00	-1.51 – 1.51	1.000
Mean Intensity	*up10-nat*	-0.00	-1.51 – 1.51	1.000
Mean Intensity	*up10-up5*	-0.00	-1.51 – 1.51	1.000

Median Intensity

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	20.52	5.13	0.18	0.949
Group	1	367.89	367.89	12.90	< 0.001
*Manip * Group*	4	4.18	1.04	0.04	0.997
Residuals	500	14,255.00	28.51		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Median Intensity	*down5-down10*	0.02	-2.02 – 2.07	1.000
Median Intensity	*nat-down10*	0.10	-1.95 – 2.15	1.000
Median Intensity	*up5-down10*	0.27	-1.78 – 2.32	0.996
Median Intensity	*up10-down10*	0.54	-1.51 – 2.59	0.951

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Median Intensity	*nat-down5*	0.08	-1.97 – 2.12	1.000
Median Intensity	*up5-down5*	0.24	-1.80 – 2.29	0.998
Median Intensity	*up10-down5*	0.52	-1.53 – 2.56	0.958
Median Intensity	*up5-nat*	0.17	-1.88 – 2.22	0.999
Median Intensity	*up10-nat*	0.44	-1.61 – 2.49	0.977
Median Intensity	*up10-up5*	0.27	-1.77 – 2.32	0.996

Max Intensity

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	1.57	0.39	0.03	0.998
Group	1	1,176.17	1,176.17	91.10	< 0.001
*Manip * Group*	4	0.21	0.05	0.00	1.000
Residuals	500	6,455.17	12.91		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Max Intensity	*down5-down10*	-0.00	-1.38 – 1.38	1.000
Max Intensity	*nat-down10*	-0.00	-1.38 – 1.38	1.000
Max Intensity	*up5-down10*	-0.01	-1.38 – 1.37	1.000
Max Intensity	*up10-down10*	0.14	-1.24 – 1.51	0.999
Max Intensity	*nat-down5*	0.00	-1.38 – 1.38	1.000
Max Intensity	*up5-down5*	-0.00	-1.38 – 1.37	1.000
Max Intensity	*up10-down5*	0.14	-1.24 – 1.52	0.999
Max Intensity	*up5-nat*	-0.01	-1.38 – 1.37	1.000
Max Intensity	*up10-nat*	0.14	-1.24 – 1.51	0.999
Max Intensity	*up10-up5*	0.14	-1.23 – 1.52	0.999

TVL

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	2,695.25	673.81	25.63	< 0.001
Group	1	1,481.15	1,481.15	56.35	< 0.001
*Manip * Group*	4	8.17	2.04	0.08	0.989
Residuals	500	13,142.90	26.29		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
TVL	*down5-down10*	1.04	-0.92 – 3.01	0.594
TVL	*nat-down10*	2.46	0.50 – 4.43	0.006
TVL	*up5-down10*	4.30	2.33 – 6.26	< 0.001
TVL	*up10-down10*	6.43	4.46 – 8.39	< 0.001
TVL	*nat-down5*	1.42	-0.55 – 3.39	0.278
TVL	*up5-down5*	3.25	1.29 – 5.22	< 0.001
TVL	*up10-down5*	5.38	3.42 – 7.35	< 0.001
TVL	*up5-nat*	1.83	-0.13 – 3.80	0.081
TVL	*up10-nat*	3.96	2.00 – 5.93	< 0.001
TVL	*up10-up5*	2.13	0.16 – 4.09	0.026

TVL Mean

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	1,582.60	395.65	28.43	< 0.001
Group	1	874.12	874.12	62.82	< 0.001
*Manip * Group*	4	4.59	1.15	0.08	0.988
Residuals	500	6,957.74	13.92		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
TVL Mean	*down5-down10*	0.22	-0.27 – 0.72	0.725
TVL Mean	*nat-down10*	0.53	0.04 – 1.03	0.028
TVL Mean	*up5-down10*	0.93	0.43 – 1.42	< 0.001
TVL Mean	*up10-down10*	1.39	0.89 – 1.88	< 0.001
TVL Mean	*nat-down5*	0.31	-0.19 – 0.80	0.437
TVL Mean	*up5-down5*	0.70	0.21 – 1.20	0.001
TVL Mean	*up10-down5*	1.16	0.67 – 1.66	< 0.001
TVL Mean	*up5-nat*	0.39	-0.10 – 0.89	0.187
TVL Mean	*up10-nat*	0.85	0.36 – 1.35	< 0.001
TVL Mean	*up10-up5*	0.46	-0.03 – 0.95	0.083

Active Speech Level

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	0.23	0.06	0.00	1.000
Group	1	1,533.60	1,533.60	90.09	< 0.001
*Manip * Group*	4	0.97	0.24	0.01	1.000
Residuals	500	8,511.70	17.02		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Active Speech Level	*down5-down10*	-0.00	-1.59 – 1.58	1.000
Active Speech Level	*nat-down10*	-0.02	-1.60 – 1.56	1.000
Active Speech Level	*up5-down10*	-0.04	-1.62 – 1.54	1.000
Active Speech Level	*up10-down10*	-0.05	-1.64 – 1.53	1.000
Active Speech Level	*nat-down5*	-0.02	-1.60 – 1.57	1.000
Active Speech Level	*up5-down5*	-0.04	-1.62 – 1.54	1.000
Active Speech Level	*up10-down5*	-0.05	-1.63 – 1.53	1.000

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Active Speech Level	*up5-nat*	-0.02	-1.60 – 1.56	1.000
Active Speech Level	*up10-nat*	-0.03	-1.62 – 1.55	1.000
Active Speech Level	*up10-up5*	-0.01	-1.59 – 1.57	1.000

LKFS

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	23.86	5.96	0.38	0.826
Group	1	799.78	799.78	50.30	< 0.001
*Manip * Group*	4	1.12	0.28	0.02	1.000
Residuals	500	7,950.17	15.90		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
LKFS	*down5-down10*	0.02	-1.51 – 1.55	1.000
LKFS	*nat-down10*	0.08	-1.45 – 1.61	1.000
LKFS	*up5-down10*	0.23	-1.29 – 1.76	0.994
LKFS	*up10-down10*	0.58	-0.94 – 2.11	0.834
LKFS	*nat-down5*	0.06	-1.47 – 1.58	1.000
LKFS	*up5-down5*	0.21	-1.32 – 1.74	0.996
LKFS	*up10-down5*	0.56	-0.97 – 2.09	0.853
LKFS	*up5-nat*	0.16	-1.37 – 1.69	0.999
LKFS	*up10-nat*	0.51	-1.02 – 2.04	0.894
LKFS	*up10-up5*	0.35	-1.18 – 1.88	0.971

Tilt

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	10,460.44	2,615.11	107.96	< 0.001
Group	1	1,691.37	1,691.37	69.82	< 0.001

Tilt

	df	*SS*	*MS*	*F*	*p*
*Manip * Group*	4	6.50	1.63	0.07	0.992
Residuals	500	12,111.62	24.22		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Tilt	*down5-down10*	1.91	0.02 – 3.80	0.045
Tilt	*nat-down10*	4.73	2.85 – 6.62	< 0.001
Tilt	*up5-down10*	8.38	6.49 – 10.26	< 0.001
Tilt	*up10-down10*	12.61	10.72 – 14.50	< 0.001
Tilt	*nat-down5*	2.82	0.94 – 4.71	< 0.001
Tilt	*up5-down5*	6.47	4.58 – 8.35	< 0.001
Tilt	*up10-down5*	10.70	8.81 – 12.59	< 0.001
Tilt	*up5-nat*	3.64	1.76 – 5.53	< 0.001
Tilt	*up10-nat*	7.88	5.99 – 9.76	< 0.001
Tilt	*up10-up5*	4.23	2.35 – 6.12	< 0.001

Voiced Tilt

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	7,088.74	1,772.19	70.25	< 0.001
Group	1	1,199.67	1,199.67	47.55	< 0.001
*Manip * Group*	4	0.75	0.19	0.01	1.000
Residuals	500	12,613.81	25.23		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Voiced Tilt	*down5-down10*	1.35	-0.57 – 3.28	0.306
Voiced Tilt	*nat-down10*	3.49	1.57 – 5.42	< 0.001
Voiced Tilt	*up5-down10*	6.64	4.71 – 8.57	< 0.001

Measure	Contrast	HSD	95% CI	p
Voiced Tilt	up10-down10	10.32	8.39 – 12.24	< 0.001
Voiced Tilt	nat-down5	2.14	0.21 – 4.07	0.021
Voiced Tilt	up5-down5	5.29	3.36 – 7.21	< 0.001
Voiced Tilt	up10-down5	8.97	7.04 – 10.89	< 0.001
Voiced Tilt	up5-nat	3.15	1.22 – 5.07	< 0.001
Voiced Tilt	up10-nat	6.83	4.90 – 8.75	< 0.001
Voiced Tilt	up10-up5	3.68	1.75 – 5.60	< 0.001

Tilt-Ratio

	df	SS	MS	F	p
Manipulation	4	0.88	0.22	43.44	< 0.001
Group	1	0.11	0.11	22.76	< 0.001
Manip * Group	4	0.02	0.00	0.83	0.505
Residuals	500	2.53	0.01		

Measure	Contrast	HSD	95% CI	p
Tilt Ratio	down5-down10	0.02	-0.01 – 0.04	0.448
Tilt Ratio	nat-down10	0.04	0.01 – 0.07	< 0.001
Tilt Ratio	up5-down10	0.07	0.04 – 0.10	< 0.001
Tilt Ratio	up10-down10	0.12	0.09 – 0.14	< 0.001
Tilt Ratio	nat-down5	0.03	0.00 – 0.05	0.083
Tilt Ratio	up5-down5	0.05	0.03 – 0.08	< 0.001
Tilt Ratio	up10-down5	0.10	0.07 – 0.13	< 0.001
Tilt Ratio	up5-nat	0.03	0.00 – 0.05	0.047
Tilt Ratio	up10-nat	0.08	0.05 – 0.10	< 0.001
Tilt Ratio	up10-up5	0.05	0.02 – 0.08	< 0.001

Mid-Ratio

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	5.03	1.26	173.57	< 0.001
Group	1	0.60	0.60	82.39	< 0.001
*Manip * Group*	4	0.00	0.00	0.15	0.962
Residuals	500	3.62	0.01		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Mid-Ratio	*down5-down10*	0.07	0.04 – 0.10	< 0.001
Mid-Ratio	*nat-down10*	0.14	0.11 – 0.18	< 0.001
Mid-Ratio	*up5-down10*	0.21	0.18 – 0.25	< 0.001
Mid-Ratio	*up10-down10*	0.28	0.25 – 0.31	< 0.001
Mid-Ratio	*nat-down5*	0.07	0.04 – 0.10	< 0.001
Mid-Ratio	*up5-down5*	0.14	0.11 – 0.17	< 0.001
Mid-Ratio	*up10-down5*	0.21	0.18 – 0.24	< 0.001
Mid-Ratio	*up5-nat*	0.07	0.04 – 0.10	< 0.001
Mid-Ratio	*up10-nat*	0.14	0.10 – 0.17	< 0.001
Mid-Ratio	*up10-up5*	0.07	0.03 – 0.10	< 0.001

High-Ratio

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	5.02	1.26	137.39	< 0.001
Group	1	1.90	1.90	207.98	< 0.001
*Manip * Group*	4	0.00	0.00	0.12	0.976
Residuals	500	4.57	0.01		

Measure	Contrast	HSD	95% CI	p
High-Ratio	down5-down10	0.07	0.04 – 0.11	< 0.001
High-Ratio	nat-down10	0.14	0.11 – 0.18	< 0.001
High-Ratio	up5-down10	0.21	0.18 – 0.25	< 0.001
High-Ratio	up10-down10	0.28	0.24 – 0.32	< 0.001
High-Ratio	nat-down5	0.07	0.04 – 0.11	< 0.001
High-Ratio	up5-down5	0.14	0.11 – 0.18	< 0.001
High-Ratio	up10-down5	0.21	0.17 – 0.24	< 0.001
High-Ratio	up5-nat	0.07	0.03 – 0.11	< 0.001
High-Ratio	up10-nat	0.14	0.10 – 0.17	< 0.001
High-Ratio	up10-up5	0.07	0.03 – 0.10	< 0.001

Skewness

	df	SS	MS	F	p
Manipulation	4	3,230.86	807.72	18.74	< 0.001
Group	1	1,106.97	1,106.97	25.69	< 0.001
Manip * Group	4	60.95	15.24	0.35	0.842
Residuals	500	21,544.94	43.09		

Measure	Contrast	HSD	95% CI	p
Skewness	down5-down10	1.36	-1.15 – 3.88	0.574
Skewness	nat-down10	0.13	-2.38 – 2.65	1.000
Skewness	up5-down10	-2.76	-5.27 – -0.24	0.024
Skewness	up10-down10	-5.63	-8.15 – -3.11	< 0.001
Skewness	nat-down5	-1.23	-3.75 – 1.29	0.667
Skewness	up5-down5	-4.12	-6.64 – -1.60	< 0.001
Skewness	up10-down5	-6.99	-9.51 – -4.48	< 0.001
Skewness	up5-nat	-2.89	-5.41 – -0.37	0.015

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Skewness	*up10-nat*	-5.76	-8.28 – -3.25	< 0.001
Skewness	*up10-up5*	-2.87	-5.39 – -0.36	0.016

Kurtosis

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	11,272,265.76	2,818,066.44	10.72	< 0.001
Group	1	4,607,707.07	4,607,707.07	17.52	< 0.001
*Manip * Group*	4	2,121,995.24	530,498.81	2.02	0.091
Residuals	500	131,475,857.46	262,951.71		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Kurtosis	*down5-down10*	-61.53	-258.11 – 135.06	0.912
Kurtosis	*nat-down10*	-198.16	-394.75 – -1.58	0.047
Kurtosis	*up5-down10*	-321.94	-518.53 – -125.36	< 0.001
Kurtosis	*up10-down10*	-391.34	-587.92 – -194.75	< 0.001
Kurtosis	*nat-down5*	-136.64	-333.22 – 59.95	0.317
Kurtosis	*up5-down5*	-260.42	-457.00 – -63.83	0.003
Kurtosis	*up10-down5*	-329.81	-526.40 – -133.23	< 0.001
Kurtosis	*up5-nat*	-123.78	-320.36 – 72.81	0.420
Kurtosis	*up10-nat*	-193.17	-389.76 – 3.41	0.057
Kurtosis	*up10-up5*	-69.39	-265.98 – 127.19	0.870

SD Intensity

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	256.40	64.10	14.63	< 0.001
Group	1	0.96	0.96	0.22	0.640
*Manip * Group*	4	1.08	0.27	0.06	0.993

SD Intensity

	df	*SS*	*MS*	*F*	*p*
Residuals	500	2,190.11	4.38		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
SD Intensity	*down5-down10*	-0.38	-1.18 – 0.43	0.702
SD Intensity	*nat-down10*	-0.85	-1.65 – -0.04	0.033
SD Intensity	*up5-down10*	-1.39	-2.20 – -0.59	< 0.001
SD Intensity	*up10-down10*	-1.99	-2.79 – -1.19	< 0.001
SD Intensity	*nat-down5*	-0.47	-1.27 – 0.33	0.498
SD Intensity	*up5-down5*	-1.02	-1.82 – -0.22	0.005
SD Intensity	*up10-down5*	-1.61	-2.41 – -0.81	< 0.001
SD Intensity	*up5-nat*	-0.55	-1.35 – 0.25	0.333
SD Intensity	*up10-nat*	-1.14	-1.95 – -0.34	0.001
SD Intensity	*up10-up5*	-0.59	-1.40 – 0.21	0.255

SD TVL

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	125.22	31.31	18.82	< 0.001
Group	1	46.91	46.91	28.19	< 0.001
*Manip * Group*	4	0.28	0.07	0.04	0.997
Residuals	500	831.89	1.66		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
SD TVL	*down5-down10*	0.22	-0.27 – 0.72	0.725
SD TVL	*nat-down10*	0.53	0.04 – 1.03	0.028
SD TVL	*up5-down10*	0.93	0.43 – 1.42	< 0.001
SD TVL	*up10-down10*	1.39	0.89 – 1.88	< 0.001

Measure	*Contrast*	*HSD*	*95% CI*	*p*
SD TVL	*nat-down5*	0.31	-0.19 – 0.80	0.437
SD TVL	*up5-down5*	0.70	0.21 – 1.20	0.001
SD TVL	*up10-down5*	1.16	0.67 – 1.66	< 0.001
SD TVL	*up5-nat*	0.39	-0.10 – 0.89	0.187
SD TVL	*up10-nat*	0.85	0.36 – 1.35	< 0.001
SD TVL	*up10-up5*	0.46	-0.03 – 0.95	0.083

Intensity Decay

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	0.00	0.00	0.40	0.807
Group	1	0.00	0.00	1.93	0.165
*Manip * Group*	4	0.00	0.00	0.01	1.000
Residuals	500	1.04	0.00		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
Intensity Decay	*down5-down10*	0.00	-0.02 – 0.02	1.000
Intensity Decay	*nat-down10*	0.00	-0.01 – 0.02	0.989
Intensity Decay	*up5-down10*	0.01	-0.01 – 0.02	0.935
Intensity Decay	*up10-down10*	0.01	-0.01 – 0.02	0.792
Intensity Decay	*nat-down5*	0.00	-0.02 – 0.02	0.999
Intensity Decay	*up5-down5*	0.00	-0.01 – 0.02	0.979
Intensity Decay	*up10-down5*	0.01	-0.01 – 0.02	0.892
Intensity Decay	*up5-nat*	0.00	-0.02 – 0.02	0.998
Intensity Decay	*up10-nat*	0.00	-0.01 – 0.02	0.967
Intensity Decay	*up10-up5*	0.00	-0.02 – 0.02	0.997

TVL Decay

	df	*SS*	*MS*	*F*	*p*
Manipulation	4	0.00	0.00	0.64	0.636
Group	1	0.00	0.00	22.64	< 0.001
*Manip * Group*	4	0.00	0.00	0.08	0.989
Residuals	500	0.00	0.00		

Measure	*Contrast*	*HSD*	*95% CI*	*p*
TVL Decay	*down5-down10*	-0.00	0.00 – 0.00	0.999
TVL Decay	*nat-down10*	-0.00	0.00 – 0.00	0.972
TVL Decay	*up5-down10*	-0.00	0.00 – 0.00	0.848
TVL Decay	*up10-down10*	-0.00	0.00 – 0.00	0.619
TVL Decay	*nat-down5*	-0.00	0.00 – 0.00	0.997
TVL Decay	*up5-down5*	-0.00	0.00 – 0.00	0.948
TVL Decay	*up10-down5*	-0.00	0.00 – 0.00	0.788
TVL Decay	*up5-nat*	-0.00	0.00 – 0.00	0.995
TVL Decay	*up10-nat*	-0.00	0.00 – 0.00	0.933
TVL Decay	*up10-up5*	-0.00	0.00 – 0.00	0.995

Vita

Name:	Daryn Cushnie-Sparrow
Education:	Western University London, Ontario, Canada 2013-2016 BA Honours 2016-2021 MClSc/PhD
Selected Honours and Awards:	Ontario Graduate Scholarship 2016-2017, 2019-2020 Parkinson Society of Southwestern Ontario 2018-2019
Related Work Experience:	Teaching Assistant 2018-2021 Lecturer, 2020 Research Assistant, 2016-2021
Selected Publications:	Knowles, T., Adams, S.G., Page, A., Cushnie-Sparrow, D., Jog, M. (2020). A comparison of speech amplification and personal communication devices for hypophonia. Journal of Speech, Language, and Hearing Research, 63(8), 2695-2712. Cushnie-Sparrow, D., Adams, S.G., Abeyesekera, A., Pieterman, M., Gilmore, G., Jog, M. (2018). Voice quality severity and responsiveness to levodopa in Parkinson's disease. Journal of Communication Disorders, 76, 1-10. Cushnie-Sparrow, D., Adams, S. G., Knowles, T., Leszcz, T., Jog, M. (2016). Effects of multi-talker noise on the acoustics of voiceless consonants in Parkinson's disease. Western Papers in Linguistics 3(1).