Health monitoring with voice analysis: Acoustic correlates of heart failure, irregular pitch periods, and dysphonia

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Division of Medical Sciences

Speech and Hearing Bioscience and Technology

have examined a dissertation entitled

Health monitoring with voice analysis: acoustic correlates of heart failure, irregular pitch periods, and dysphonia

presented by Olivia Mae Murton

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature: _____

Typed Name: Dr. Satrajit Ghosh

Signature: Jordan Green
Jordan Green (Nov 2, 2020 12:16 EST)

Typed Name: Dr. Jordan Green

Signature: _____

Typed Name: Dr. Carol Espy-Wilson

Signature: _____

Typed Name: Dr. Larry Allen

Date: October 13, 2020

Signature:

Email: larry.allen@cuanshutz.edu

Health monitoring with voice analysis: Acoustic correlates of heart failure, irregular pitch periods, and dysphonia

A dissertation presented

by

Olivia Mae Murton

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Speech and Hearing Bioscience and Technology

Harvard University

Cambridge, Massachusetts

October 2020

Dissertation Advisor: Daryush D. Mehta Olivia Mae Murton

Health monitoring with voice analysis: Acoustic correlates of heart failure, irregular pitch periods, and dysphonia

Abstract

Voice and speech production relies on complex interactions of linguistic and cognitive systems, neuromotor pathways, respiration, and airflow through the vocal tract. Voice can reveal disruptions to any of those systems, so it can be used to non-invasively detect and monitor illness. This thesis examines three interrelated applications of voice analysis for health monitoring. The first application investigates acoustic voice features as a biomarker for acute decompensated heart failure (ADHF), a serious escalation of heart failure symptoms including breathlessness and fatigue. ADHF-related systemic fluid accumulation in the lungs and laryngeal tissues is hypothesized to affect voice acoustics. A set of daily voice samples from 52 patients undergoing inpatient ADHF treatment is analyzed to identify vocal biomarkers for ADHF and examine the trajectory of voice change during treatment. Data from an audio microphone and from a neck-surface vibration sensor are also compared. Results indicate that speakers have more stable phonation, more creaky voice, faster speech rates, and longer phrases after ADHF treatment compared to pre-treatment. These findings motivate work on two additional acoustic features: irregular pitch periods (IPPs), which contribute to the perception of creaky voice, and cepstral peak prominence (CPP), a measure of dysphonia and phonatory stability. To that end, the second application uses voice recordings from healthy speakers and compares the output of an existing algorithm for creaky voice detection to hand labels of IPPs. A perceptually relevant creak probability threshold is determined. These results are useful for voice monitoring of

ADHF, since speakers produced more IPPs after ADHF treatment. In the third application, CPP thresholds that distinguish speakers with and without voice disorders are determined separately for continuous speech and sustained vowels using two widely-used voice analysis programs. These normative CPP values provide an objective dysphonia indicator to aid evaluation of voice and other disorders. For example, CPP tended to improve with ADHF treatment for patients whose pre-treatment CPP was relatively low. Together, these projects present a novel method of monitoring ADHF using vocal biomarkers and develop a more-detailed understanding of relevant voice features. Proposed future work includes prospective at-home monitoring of patients at risk for ADHF.

Table of Contents

Acknowledgments

This work was supported by grants from the NIH National Institute on Deafness and Other Communication Disorders (T32 DC000038), the NIH National Heart, Lung, and Blood Institute (F31 HL143824), the Voice Health Institute, the Center for Assessment Technology and Continuous Health, and the Dennis Klatt Memorial Fund at the Speech Communication Group in the MIT Research Laboratory of Electronics.

First, to my advisor, Daryush Mehta, who welcomed and mentored me into the world of voice science with unwavering encouragement and boundless enthusiasm: thank you for believing in this work and in me.

Thank you to the members of my advising committee—Satra Ghosh, Bill Dec, and Stefanie Shattuck-Hufnagel—who answered many questions, connected me to resources, and provided ongoing support during and beyond this project. I also want to thank Bob Hillman for all his help on scientific, academic, and clinical questions; John Guttag for advice on the machine learning and other data analysis; and everyone in the MGH Cardiology department who made the heart failure study possible. I am only one of the many beneficiaries of your hard work.

Thank you to everyone at the Voice Center for welcoming me into your professional and personal lives. Special thanks to my companions in the scholararium—AJ Ortiz, Katie Marks, and Laura Toles—for your countless acts of friendship and food.

Thank you to my SHBT cohort— Hannah Goldberg, Justin Fleming, Kameron Clayton, Malinda MacPherson, and Meenakshi Asokan—and my honorary cohort-member Jeanne Gallée. I am so lucky to have learned all that I've learned with you.

Thank you to the MIT Linguistics department and my undergraduate advisors, Adam Albright and Stefanie Shattuck-Hufnagel, who taught me to see beauty and complexity in the structure of language. Thank you also to all of my teachers, formal and otherwise, who have inspired me, challenged me, and shown me new ways of looking at the world.

Thank you to my friends and family, near and far—I could not have asked for better companions on this journey. Thank you especially to my parents and my sister for your unlimited love and encouragement. You are the very best.

And of course, thank you to Stephen Eltinge, who traveled 40,000 miles in support of our life together.

List of tables and figures

Tables

Figures

Chapter 1. Introduction

Producing voice relies on a complex interaction of cognitive and physical processes, including linguistic and cognitive systems, neuromotor pathways, and the physics of airflow through the airway and vocal tract. Voice can potentially reveal disruptions to any of those systems, so it is a good candidate for non-invasive detection and monitoring of many health concerns. In particular, voice has been used to identify vocal and neurological diseases including vocal hyperfunction (Mehta et al., 2015), depression (Williamson et al., 2013), Parkinson's disease (Holmes et al, 2000), and amyotrophic lateral sclerosis (Horwitz-Martin et al., 2016). Voice is also being used as a biomarker for systemic diseases such as coronary artery disease (Maor et al., 2016), pulmonary hypertension (Sara et al., 2020), and heart failure (HF) (Maor et al. 2020). This thesis examines three interrelated applications of voice monitoring for health.

1.1 Heart failure

Heart failure is the primary diagnosis for over 1 million hospitalizations each year in the US (Gheorghiade & Pang, 2009). It is the most common cause of hospitalization in Americans over 65 years of age (Desai & Stevenson, 2012). Although HF can have many causes, it broadly involves a reduction in the heart's ability to move blood through the body. In response, compensatory mechanisms, including increased blood pressure and fluid retention, maintain homeostasis and adequate blood supply (Kemp & Conte, 2012). However, those compensatory mechanisms can also fail, leading to acute decompensated heart failure (ADHF). ADHF is an acute crisis requiring immediate intervention. Its primary symptom is edema, or fluid accumulation in the lungs or other body tissues; other symptoms include dyspnea (shortness of breath) and fatigue (Boorsma et al., 2020). Most hospitalizations for ADHF involve patients who

already have HF diagnoses (Joseph et al., 2009). A major goal of HF management, therefore, is to predict and prevent episodes of ADHF.

As compensatory mechanisms begin to fail and excess fluid begins to accumulate, a patient's weight increases. Although a weight increase due to edema can signal an impending decompensation episode, the increase may not be detected in time for intervention to succeed. There is currently an unmet need for a reliable signal that can be identified earlier than weight increase, but can be measured just as easily at home. This thesis investigates voice as a noninvasive biomarker of HF status, hypothesizing that the vocal folds' thin tissue layers are particularly sensitive to HF-related edema.

Chapter 2 presents a pilot study that analyzes voice samples from ten patients undergoing inpatient treatment for ADHF. Each patient's pre-treatment and post-treatment voices are compared, and acoustic voice features that show consistent change across speakers in response to treatment are identified. That study is expanded in Chapter 5, which applies additional statistical and machine learning analyses to a larger data set of 52 patients with ADHF. Pre- and post-treatment voices are distinguished using logistic classifiers, and the acoustic voice features with the greatest predictive power in those classifiers are identified. The day-to-day trajectories of voice change in response to ADHF treatment are also examined.

The work in Chapter 2 prompted additional inquiry into two specific vocal features: irregular pitch periods (IPPs) and cepstral peak prominence (CPP). Unexpectedly, the proportion of IPPs in patients' speech increased with HF treatment. Although most other vocal measures tended toward improved voice quality after treatment, IPPs can be associated with pathological voice quality (Holmberg et al., 2001). Chapter 3 investigates the algorithm that was used to detect IPPs and presents a new threshold for improved IPP detection. Separately, CPP tended to

improve with HF treatment only for patients whose pre-treatment CPP was low. Chapter 4 analyzes voice samples from speakers with and without voice disorders to identify a clinically-relevant normative CPP value.

1.2 Irregular pitch periods

IPPs are a type of non-modal phonation that are frequently perceived as creaky voice or vocal fry. In typical voices, they tend to occur at word and phrase boundaries and in some specific prosodic contours (Pierrehumbert & Talkin, 1992). In addition to this syntactic and prosodic information, IPPs can also indicate a speaker's emotions (Gobl, 2003), dialect (Henton, 1986), and identity (Bőhm & Shattuck-Hufnagel, 2009). More broadly, IPPs are clinically-significant indicators of voice disorders (Holmberg et al., 2001) and were found to be produced more frequently after treatment for ADHF (Murton et al., 2017).

Although they are useful for many kinds of voice and speech analysis, identifying IPPs by hand is time-consuming and requires training. Reliable automatic detection of IPP is therefore a compelling goal (Redi & Shattuck-Hufnagel, 2001). In Chapter 3, we investigate the performance of an automatic creaky voice detection algorithm (Drugman et al., 2014). The data set used here consists of conversations recorded by eight American English speakers with healthy vocal status. The algorithm's output is compared to IPP labels created by experienced human labelers (*American English Map Task*, 1999). We identify an appropriate creak probability threshold for the algorithm output that best aligns with the human-rated IPP labels. The resulting threshold is also applied in Chapter 5 to detect IPPs in speakers with ADHF. This work can be used to further investigate the varying acoustic realizations of IPPs, to relate IPP usage to specific prosodic contexts, and to understand relationships between health and IPP production.

1.3 Dysphonia

CPP has been increasingly recommended for clinical use as an objective measure of vocal breathiness and overall dysphonia, or altered vocal quality (Patel et al., 2018). CPP has been shown to distinguish between dysphonic and non-dysphonic voices for a variety of disorders and speaking tasks (Fraile & Godino-Llorente, 2014). In particular, lower CPP tends to indicate greater dysphonia severity. Despite its increasing clinical use, there is limited published work that specifies which CPP values indicate a voice disorder, and how those thresholds vary with speech task or analysis method. The absence of a clearly-defined clinical CPP threshold makes it challenging for clinicians and patients to interpret CPP values, especially ones generated under different conditions.

In Chapter 4, we use two widely-used voice analysis programs to determine which CPP cutoff values best identify probable voice disorders. This section consists of two related experiments. Experiment 1 uses a database of 345 speakers with and without voice disorders (Massachusetts Eye and Ear Infirmary, 1994). For each voice analysis program, CPP thresholds that best indicate a voice disorder are identified separately for sustained vowels and continuous speech samples. Experiment 2 evaluates CPP's ability to predict dysphonia severity based on auditory-perceptual judgments by trained listeners (Kempster et al., 2009). Linear relationships between CPP and dysphonia severity rating are identified for sustained vowels and continuous speech. These results are applicable for clinicians who use CPP to clinically assess and monitor patients with voice disorders. Additionally, the CPP values determined for speakers with ADHF in Chapter 5 are evaluated in light of these thresholds.

Taken together, these projects present a novel method of monitoring HF status using vocal biomarkers as well as a more-detailed understanding of two particularly relevant voice

features. In addition to monitoring of HF and related cardiovascular disease, this work can be used to improve detection and treatment of voice and speech disorders. Further, recent early work by several groups attempts to use voice to diagnose and track the spread of coronavirus disease 2019, or COVID-19 (Adans-Dester et al., 2020; Quatieri et al., 2020). Some findings from this thesis, particularly those reflecting a speaker's degree of breathlessness, may be useful for identifying and monitoring disease progression in COVID-19 and other respiratory illnesses.

1.4 Thesis outline

Chapter 2 presents a statistical analysis of a pilot data set of daily voice samples from 10 patients being treated for ADHF. Each patient's pre- and post-treatment voices are compared. That work identified several voice features that responded to HF treatment, which led to additional investigation that is described in Chapter 5. Chapter 2 has been peer reviewed and was previously published as Murton et al. (2017).

Chapter 3 identifies a creak probability threshold for an algorithm that detects IPPs. IPPs are difficult to detect by hand, but are useful for voice analysis because they can carry information about grammatical, pragmatic, and clinical aspects of speech. Drugman et al. (2014) developed an artificial neural network that uses acoustic features to generate frame-by-frame probabilities of creaky voice. Here, we determine a perceptually relevant creak probability threshold by comparing that algorithm's outputs to hand labels of IPPs. Chapter 3 has been peer reviewed and was previously published as Murton et al. (2019).

Chapter 4 presents an application of cepstral peak prominence (CPP) to diagnose voice disorders. We determine CPP values that best distinguish speakers with and without voice disorders in continuous speech and sustained vowel tasks. We also use a large set of auditory-perceptual ratings from trained listeners to estimate linear relationships between CPP and

perceived overall dysphonia. This set of normative CPP values aids clinical voice evaluation by providing an objective indicator of dysphonia. Chapter 4 has been peer reviewed and was previously published as Murton et al. (2020).

Chapter 5 extends the pilot study from Chapter 2 to include voice samples from 52 patients undergoing inpatient treatment for ADHF. Statistical and machine learning techniques are used to identify vocal biomarkers that correlate with HF status. In addition to comparing patients' pre- and post-treatment voices, this work also includes analysis of multiple timepoints throughout the treatment to understand the trajectory of voice change. Finally, traditional audio microphone recordings are compared to recordings from a neck-skin acceleration sensor. This acceleration sensor preserves speaker privacy and is more robust to background noise, which is useful for monitoring voice in daily life outside the lab.

Chapter 2. Acoustic speech analysis of patients with decompensated heart failure: a pilot study

Reproduced, with the permission of the Acoustical Society of America, from

Murton, O. M., Hillman, R. E., Mehta, D. D., Semigran, M., Daher, M., Cunningham, T., Verkouw, K., Tabtabai, S., Steiner, J., Dec, G. W., & Ausiello, D. (2017). Acoustic speech analysis of patients with decompensated heart failure: A pilot study. The Journal of the Acoustical Society of America, 142(4), EL401–EL407. https://doi.org/10.1121/1.5007092

Author Contributions

R. Hillman and D. Ausiello developed the initial study and hypotheses. M. Semigran, M. Daher, T. Cunningham, K. Verkouw, S. Tabtabai, and J. Steiner collected the data. O. Murton carried out the data analysis and interpretation with the guidance of D. Mehta, R. Hillman, M. Semigran, M. Daher, G. Dec, and D. Ausiello. O. Murton wrote the manuscript with input from the other authors. All authors approved the final published version.

Abstract

This pilot study used acoustic speech analysis to monitor patients with heart failure (HF), which is characterized by increased intracardiac filling pressures and peripheral edema. HF-related edema in the vocal folds and lungs is hypothesized to affect phonation and speech respiration. Acoustic measures of vocal perturbation and speech breathing characteristics were computed from sustained vowels and speech passages recorded daily from ten patients with HF undergoing inpatient diuretic treatment. After treatment, patients displayed a higher proportion of automatically identified creaky voice, increased fundamental frequency, and decreased cepstral peak prominence variation, suggesting that speech biomarkers can be early indicators of HF.

2.1 Introduction

Heart failure (HF) is a large and growing concern, consuming significant clinician time and major healthcare expenditures. Preventing the escalation of existing disease is critical to avoiding frequent hospitalizations, lowering healthcare costs, and reducing patient mortality. The 30-day readmission rate after a HF–related hospitalization is 24%, increasing to over 50% within six months (Desai & Stevenson, 2012). There are over 5.3 million HF patients in the U.S., with 550,000 new cases diagnosed annually and almost 300,000 deaths each year (O'Connor et al., 2005).

Patients with HF can be stable for long periods of time. However, progression of the underlying disease or other changes in health can cause those compensatory strategies to fail. This loss of stability is known as *decompensated* HF and is an acute crisis requiring immediate intervention (Gheorghiade & Pang, 2009). The course of a patient's disease may include multiple episodes of decompensation separated by stable periods of varying duration (Desai & Stevenson, 2012). Thus, a major goal of HF management is to maintain stability by predicting and preventing episodes of decompensation.

A hallmark of HF is the presence of *edema*, which is characterized by swelling from fluid retention in body tissues. Edema is typically found in the legs and feet, but can also occur in the lungs and throughout the body. Leading HF specialists have emphasized the importance of managing edema in patients and the need for technologies that identify edema early (Gheorghiade & Pang, 2009). As decompensation approaches, HF-related edema causes enough of a weight increase—often termed *volume overload*—to be detectable with regular weighing, which patients can do easily and non-invasively at home (Joseph et al., 2009). Weight is used for first-line monitoring of volume status and as an early warning for decompensation. However,

edema-related weight increases occur relatively late in the disease progression and may not be detected in time to prevent an episode of decompensation (Joseph et al., 2009). There are also surgically implantable devices which may predict impending decompensation by monitoring pressures within the heart and lungs (Abraham et al., 2011). However, these monitors are expensive and invasive. There is currently an unmet need for a reliable signal that can be identified earlier than weight increase, but can be measured just as easily at home by patients or caregivers.

Voice has the potential to provide an easily-obtained, non-invasive way to monitor physiological changes throughout the body, as long as those changes also affect the larynx (Van Stan et al., 2017). In particular, the vocal folds consist of thin tissue layers that might be particularly sensitive to HF-related edema. For example, Verdolini et al. (2002) found that a dose of the diuretic Lasix (furosemide), which is widely used to treat decompensated HF, induced a 23% increase in the phonation threshold pressure in healthy adults. Since the amount of edema required to measurably change the voice is small (whereas the amount needed to increase body weight is large), voice monitoring may allow us to detect and track HF-related edema at an earlier stage than weight does.

In this pilot study, we analyzed the voices of HF patients as they underwent treatment for decompensated HF and returned to a stable clinical state. Decompensation is treated with diuretics, which cause patients to rapidly lose fluid and decrease their edema until they reach their *target weight* (typical body weight without extra fluid). This decrease in volume overload during treatment mirrors the increasing edema that occurs during the asymptomatic pre-decompensation phase. Similarly, voice changes during treatment may also mirror corresponding changes before decompensation. By looking for changes in voice that correlate with in-hospital

improvements, we can use this new knowledge to direct future studies that may predict decompensation in stable HF patients at home. Our approach is based on hypotheses about how voice and speech production may be impacted by HF pathophysiology. This approach is fundamentally different from studies with similar goals that have used traditional machine learning methods to analyze voice acoustic signals, because such investigations have often lacked hypotheses and theories related to the underlying physical systems (Maor et al., 2018).

The goal was to determine whether measurable changes in voice occur during treatment for acute decompensated HF. Specifically, we hypothesized that the fluid retention, dyspnea, and fatigue associated with HF will cause increased edema and surface dehydration of the vocal folds' phonatory mucosa and reduced respiratory support for speech production, resulting in increased phonatory irregularity, decreased fundamental frequency, decreased durations of speech breath groups, and increased pausing.

2.2 Methods

Ten patients (8 male, 2 female) with acute decompensated HF were enrolled in the study with a mean ± standard deviation (SD) age of 70 ± 13 years. Inclusion criteria called for patients to be at least 4.5 kg above their target weight on admission and expected to need diuretics for over 48 hours. The average length of stay was 7 ± 3 days during which weight was measured each day and the average total weight loss was 8.5 ± 5 kg. Blood levels of N-terminal pro-brain b-type natriuretic peptide (NT-proBNP) were tested at the beginning and end of each patient's hospitalization since high levels of NT-proBNP have been associated with HF (Dao et al., 2001). The average change (last minus first measurement) in NT-proBNP level was −1379 ± 3100 pg/mL. Patients also used visual analog scales to evaluate their dyspnea and global symptoms

from 0 (worst) to 100 (best). Average changes in dyspnea and global symptom rating were 8.4 ± 21 and 22 ± 30, respectively.

Each day, patients recorded a standard speech protocol consisting of four different utterance types: sustained vowels, CAPE-V sentences (Kempster et al., 2009), the Rainbow Passage (Fairbanks, 1960), and spontaneous speech. Recording sustained vowels allows us to evaluate a patient's ability to produce stable, regular phonation. The other tasks, which elicit connected speech, are designed to evaluate voice characteristics under more natural speaking conditions and provide information about prosody and speech breathing that cannot be recovered from short non-linguistic vowel productions. The CAPE-V sentences and the Rainbow Passage are widely used in voice research to elicit various speech contexts, including soft/hard glottal attacks, vowel coarticulation, and nasality. Both an acoustic microphone (H1 Handy Recorder, Zoom Corporation, Tokyo, Japan) and neck-mounted accelerometer (BU-27135; Knowles Corp., Itasca, IL) were used to record the speech tasks (Mehta et al., 2016). This paper presents analysis of the microphone signal, which was sampled at 44.1 kHz and 16-bit quantization.

All acoustic recordings were high-pass filtered at 70 Hz to remove low-frequency noise artifacts. We then computed the change in voice and speech measures from the first recording session to the last.

Each sustained vowel recording was divided into overlapping 1-second segments starting 100 ms apart. Praat (Boersma & Weenink, 2018) was used to extract the 1-second segment with the lowest percent jitter (degree of cycle-to-cycle variation in glottal pulse spacing) that was used for all further vowel analyses. The segment with the lowest jitter was selected to compare each patient's best voice sample and to minimize artifacts from environmental noise or irregular voicing at onsets and offsets.

The fundamental frequency (F0) was extracted using Praat's cross-correlation pitch-tracking method (40-ms Hanning window every 3.3 ms). For each vowel utterance, the F0 track was used to generate a Praat voice report including mean F0, median F0, F0 standard deviation (SD), and jitter percent. Following Awan & Roy (2006), a low-high spectral energy ratio was calculated from the spectrum as the band energy difference, in dB, between 0–4 kHz and 4–10 kHz to reflect perceptual ratings of dysphonia severity. The Praat pitch track was also used to calculate mean F0, median F0, and F0 SD for each Rainbow Passage recording. To minimize potential distortion from pitch tracking artifacts, F0 measures were only calculated over frames with F0 between the 5^{th} and 95^{th} percentile for each recording.

Cepstral peak prominence (CPP) and CPP SD were calculated for both vowel and Rainbow Passage utterances following Awan et al. (2010). CPP was computed as the difference, in dB, between the magnitude of the highest peak and the noise floor in the power cepstrum (window length 40.96 ms, computed every 10.24 ms). The location of the CPP was limited to quefrencies between 3.3 and 16.7 ms (equivalent to F0 between 60 and 300 Hz). For Rainbow Passage utterances, CPP was only calculated on voiced frames (those with a non-zero F0 in the Praat pitch track). These measures were used to test our hypothesis that, pre-treatment, vocal fold edema would cause increased acoustic perturbation and decreased F0.

The probability of creak was computed as a measure of glottal cycle irregularity related to vocal fry for sentence and Rainbow Passage recordings. This method uses short-term power contours, intra-frame periodicity, and inter-pulse similarity, as well as measures indicating secondary or widely-spaced glottal pulses as inputs to an artificial neural network that generates frame-by-frame creak probabilities (Ishi et al., 2008; Kane et al., 2013). Creak percent for an utterance was given by (# creaky frames) / (# voiced frames) × 100, where # voiced frames was

the sum of the number of frames considered voiced by Praat and the number of frames with a very high creak probability (above 0.8). Adding high-probability creak frames accounted for frames with highly irregular phonation—often on the margins of voiced regions—where Praat's pitch tracking did not identify an underlying F0. Creaky frames were defined as voiced frames with creak probability above 0.3. The creak threshold of 0.3 was based on an iterative process involving perceptual evaluation of the acoustic signal and balanced the requirement of including frames with clear creaky voice quality and minimizing artifacts from regions with low creak probability (pseudo-periodic signal).

Rainbow Passage readings were transcribed and aligned to sound files using the Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008). Intra-passage pause durations longer than 500 ms were assumed to exhibit an inhalation; thus, a speech breath group was defined between successive inhalations, excluding the pause duration. Respiratory-related measures within each Rainbow Passage reading included total breath group duration (sum of the durations of each breath group); breath group duration SD (SD of all the breath group durations); and mean, max, and SD of the number of phonemes per breath group (based on the phoneme counts given by the forced aligner). Unlike the measures described earlier, which evaluated phonatory irregularity, these measures were used to test the effects of edema on speech breathing support.

2.3 Results

Since the goal was to determine voice measures that correlated with cardiac status (for which weight is a proxy), Table 2.1 lists acoustic speech measures that trended in the same direction for a majority of patients. Creak percent increased after treatment in the sentence and Rainbow Passage conditions for eight and nine patients, respectively. Total breath group duration within the Rainbow Passage decreased, and measures related to the number of phonemes per

breath group increased, for seven of ten patients. Measures of acoustic irregularity (sustained-vowel F0 SD, jitter percent, and CPP SD) decreased for most patients. Median F0 increased for a majority of patients in the Rainbow Passage condition (but not the sustained-vowel task). Most patients required less speaking time to read the Rainbow Passage after treatment and produced more phonemes per phrase. These results indicate that patients spoke faster or breathed less frequently after HF treatment (potential effects of increased familiarity with the reading passage require further investigation). As hypothesized, these preliminary results indicate that most patients showed increased irregularity and decreased F0 at admission compared to discharge.

Table 2.1. Summary of measures for which seven or more patients changed in the same direction during hospitalization. Includes first-day mean values across all ten patients, means and ranges of changes from first to last day, and *p*-values from paired, two-tailed t-tests comparing first and last day values. *p < 0.05.

Measure name	Utterance type	Majority change	Day 1 mean	Mean change	Range of change	*p*
Creak (%)	passage	9 +	3.0	6.9	[−0.69, 33]	0.065
F0 SD (Hz)	vowel	8 –	3.6	−0.95	[−6.5, 1.2]	0.20
Jitter (%)	vowel	8 –	0.88	−0.25	[−1.5, 0.25]	0.18
CPP SD (dB)	vowel	8 –	2.5	−0.41	[−1.1, 0.15]	0.022*
Creak (%)	sentence	8 +	3.6	5.5	[−0.75, 19]	0.031*
F0 SD (Hz)	passage	8 +	16	2.4	[−24, 16]	0.50
Low-high spectral ratio	vowel	7 +	25	1.8	[−8.7, 12]	0.39
Median F0 (Hz)	passage	7 +	141	3.7	[−24, 20]	0.38
Total breath group duration (s)	passage	7 –	35	−1.9	[−9.8, 1.5]	0.11
Mean phonemes per phrase	passage	7 +	20	1.1	[−7.9, 12]	0.49
Max phonemes per phrase	passage	7 +	41	1.7	[−33, 16]	0.73
SD phonemes per phrase	passage	7 +	9.9	0.89	[−7.7, 6.4]	0.48

Figure 2.1 displays each patient's creak percent on their first and last day of treatment. Creak percent was higher on the last day than the first for nine patients in the Rainbow Passage condition (mean change 6.9 pp, range −0.69–33 pp, $p = 0.065$) and eight patients on the CAPE-V sentence condition (mean change 5.5 pp, range −0.75–19 pp, $p = 0.031$).

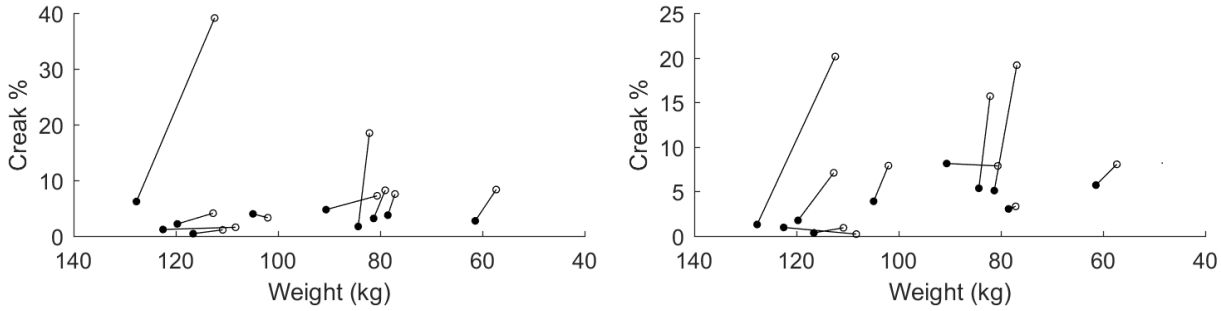

Figure 2.1. Creak percent versus weight for each patient's Rainbow Passage (left) and CAPE-V sentences (right), with first-day (filled) and last-day (open) values. Note the horizontal axis shows weight decreasing toward the right, indicating direction of HF status improvement.

Figure 2.2, which is based on data from Rainbow Passage recordings, shows median F0 increasing in seven patients and F0 SD increasing in eight. The mean change in median F0 was 3.7 Hz (0.54 semitones), whereas the mean change in F0 SD was 2.4 Hz (0.23 semitones). Figure 2.3 is based on data from sustained vowel recordings. The left panel shows that CPP increased in six of ten patients. However, the CPP increased for all four patients who began with a comparatively low CPP (< 20 dB). The mean change among these patients was 3.9 dB, compared to −0.7 dB among those who started with CPP > 20 dB (indicating good voice quality to begin with). The right panel shows decreases in sustained-vowel F0 SD for eight of ten patients after treatment. The mean change in F0 SD was −1 Hz, with one patient having an F0 SD 6.5 Hz lower on their last day.

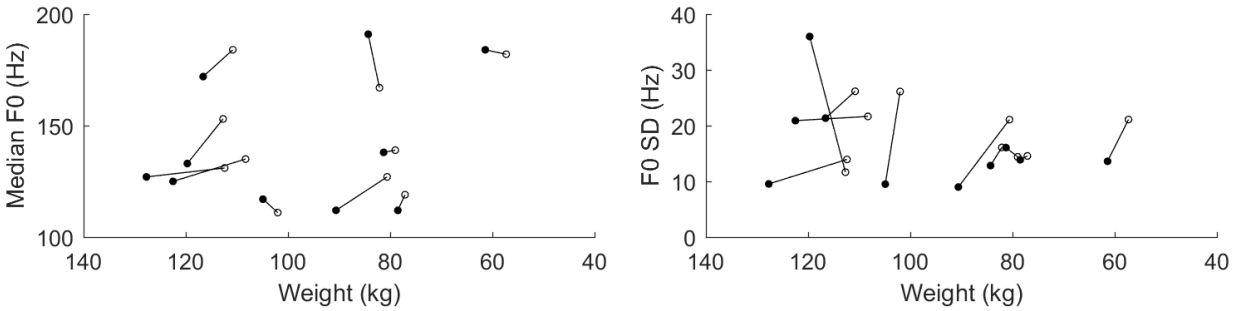

Figure 2.2. Median (left) and SD (right) F0 across voiced frames for each patient's Rainbow Passage.

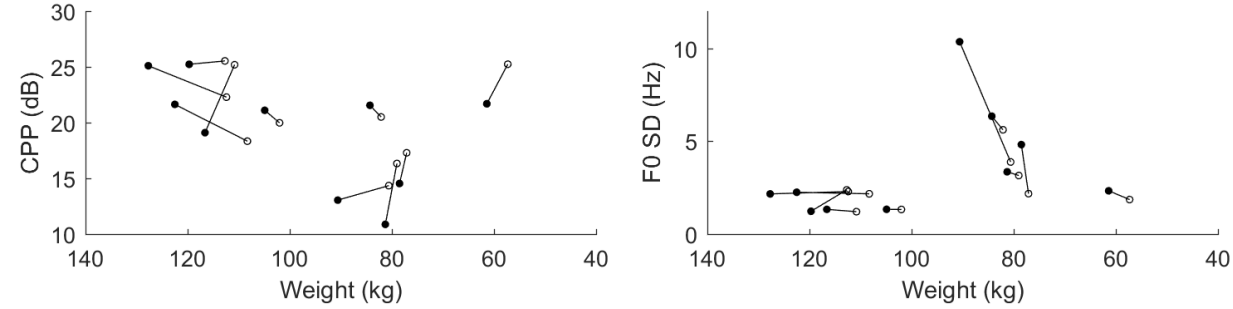

Figure 2.3. Mean CPP (left) and F0 SD (right) vs weight across all sustained-vowel utterances in each patient's first- and last-day recordings.

2.4 Discussion

This pilot analysis studied patients who were being treated for acute decompensated HF and monitored their voices as their volume status improved. The ultimate goal, though, is the reverse: to monitor stable HF patients and identify impending decompensation before it becomes a crisis. Our main purpose in this pilot project was to demonstrate that voice features have the potential to reflect HF status and may have use in at-home or continuous ambulatory monitoring contexts.

Voice measures examined can be divided into two categories based on their response to treatment. The first and most promising category for future analysis includes measures for which most patients exhibited large changes in the same direction. Creak percent increased for most

people in both the Rainbow Passage and sentence conditions, with large effect sizes (mean change 6.9 pp and 5.5 pp, respectively, relative to starting values of 3.0% and 3.6%). Also, median F0 in the Rainbow Passage task increased after treatment for seven of the ten patients. In the sustained-vowel condition, CPP SD decreased for eight of ten patients with another fairly large effect size (mean change −0.4 dB relative to a mean starting value of 2.5 dB).

The second category of measures includes those for which most patients saw large changes due to treatment, but not all in the same direction. For example, median F0, maximum and mean phonemes per breath group (in the Rainbow Passage) and low-high spectral ratio (in the vowel condition) had small average changes relative to baseline because large increases in some patients were cancelled out by large decreases in others. These measures may be useful for HF monitoring in future if we can understand what causes a given patient's voice to change in a given direction.

Creaky voice quality can be associated with a pathological voice condition (Holmberg et al., 2001). However, unlike most other voice quality measures, creak is also partially governed by the linguistic system, and some amount of creak can be normal in any voice (Gordon & Ladefoged, 2001). It is possible that HF-related edema in the vocal folds affects the phonatory mechanism by reducing the capability of the vocal folds to become slack or flaccid (necessary for creaky voice production), which decreases the likelihood of creaky voice. Further examination of the distributions of creak in reading passages would allow us to see where creaky regions are relative to linguistically-expected ones.

The direction of CPP change during treatment was not uniform; six of ten patients had higher CPP at discharge than at admission. However, four patients began their hospitalization with CPPs below 20 dB, and all improved by discharge. Our eventual goal is to predict

decompensation in stable patients, rather than observing the return to stability after decompensation. Our CPP results indicate that, if most stable patients have high CPP, some patients will have lower CPPs during decompensation and some will not. Reduced CPP might be a sign of increasing fluid volume in some HF patients who exhibit a large change.

A potential confound for acoustic voice analysis is that the voice production mechanism can compensate for physiological vocal fold changes, such as increased edema. Therefore, changes in voice output may not be directly related to the amount of edema in the vocal folds since compensatory mechanisms allow talkers to produce voices of similar quality even under different physiological conditions (Hillman et al., 1989). In this population, direct visualization of edema in the larynx via laryngoscopy is not feasible due to their comprised medical condition and increased fragility.

Additionally, seven of the ten patients in this study were 65 years of age or older. Older patients may have aging-related deterioration of vocal function that is unrelated to HF. For example, a common cause of degraded voice quality in older individuals is presbyphonia, which is described as vocal fold bowing (incomplete closure) secondary to loss of tissue and/or diminished laryngeal neuromotor control (Mueller, 1997). HF-related edema might actually have the potential to improve voice quality by bulking up the vocal folds and facilitating better contact and closure (reduced bowing) during phonation.

2.5 Conclusion

Reliable methods of monitoring fluid overload in stable HF patients are critically important for preventing decompensation. Voice monitoring has the potential to provide a non-invasive, easily-obtained way for patients to track their own status. Several measures of voice

quality, including creak percent, F0, and CPP SD, may correlate well with improvements in HF symptoms during decompensation treatment.

Chapter 3. Identifying a creak probability threshold for an irregular pitch period detection algorithm

Reproduced, with the permission of the Acoustical Society of America, from

Murton, O., Shattuck-Hufnagel, S., Choi, J.-Y., & Mehta, D. D. (2019). Identifying a creak probability threshold for an irregular pitch period detection algorithm. *The Journal of the Acoustical Society of America*, *145*(5), EL379–EL385. https://doi.org/10.1121/1.5100911

Author Contributions

All authors contributed to developing the study and hypotheses. S. Shattuck-Hufnagel provided the previously-collected data. O. Murton carried out the data analysis and interpretation with the guidance of S. Shattuck-Hufnagel and D. Mehta. O. Murton wrote the manuscript with input from the other authors. All authors approved the final published version.

Abstract

Irregular pitch periods (IPPs) are associated with grammatically, pragmatically, and clinically significant types of nonmodal phonation, but are challenging to identify. Automatic detection of IPPs is desirable because accurately hand-identifying IPPs is time-consuming and requires training. We evaluated an algorithm developed for creaky voice analysis to automatically identify IPPs in recordings of American English conversational speech. To determine a perceptually relevant threshold probability, frame-by-frame creak probabilities were compared to hand labels, yielding a threshold of approximately 0.02. These results indicate generally good agreement between hand-labeled IPPs and automatic detection, calling for future work investigating effects of linguistic and prosodic context.

3.1 Introduction

In typical healthy speakers, non-modal phonation patterns occur in a wide variety of speech contexts, including at word boundaries, at phrase boundaries, and in certain prosodic contours. The type of non-modal phonation that is characterized by irregular pitch periods (IPPs) can also provide information about emotional content, dialect, speaker identity, and health status. IPPs exhibit varied acoustic realizations, including irregular spacing of pitch periods, single glottal pulses, and local decreases in F0 or amplitude (Dilley et al., 1996). Discussions of IPPs, creaky voice, and similar phonation types have often encountered terminological challenges. Researchers have used differing terms to describe these phenomena and have proposed a variety of theories about the mechanisms by which they occur and their functions in speech (Gerratt & Kreiman, 2001; Keating et al., 2015). Here, we compare a strictly acoustic measure (the output of an algorithm) to labels created by human raters using a combination of visual and auditory-perceptual criteria.

In some languages, including American English, IPPs carry linguistic significance because of their association with lexical information, word or phrase boundaries, and prosodic prominences. Often corresponding to word boundaries, common locations for IPPs in American English include (1) the first few pitch periods of vowels or sonorant consonants at the beginning of an intonational phrase; (2) the last portion of an intonational phrase, particularly one with a low boundary tone; (3) the first few pitch periods at the onset of a high- or low-pitch accented syllable; and (4) the nucleus of a low-pitch-accented syllable where F0 is particularly low (Pierrehumbert & Talkin, 1992). Additionally, in this dialect of English, IPPs can also convey lexical information by signaling a /t/ that occurs word-finally (e.g. *cat*) or between a stressed vowel and sonorant consonant (e.g. *butler*) (Redi & Shattuck-Hufnagel, 2001).

IPPs are pragmatically significant when they are used to indicate speaker identity (Bőhm & Shattuck-Hufnagel, 2009), dialect (Henton, 1986), and emotional state (Gobl, 2003). Speakers can vary widely from each other in their usage of IPPs, including both their baseline IPP usage across speech contexts and their grammatically-driven IPP usage in specific prosodic locations. Some speakers use IPPs more rarely than others, and speakers vary in their usage of IPPs to mark prosodic boundaries (Dilley et al., 1996). Although these differences present challenges for inter-speaker analyses of IPP usage, they also mean that IPPs may be useful for detecting unique aspects of specific speakers, like individual identity and dialect.

Finally, IPPs are clinically significant indicators of voice disorders and systemic diseases including acute decompensated heart failure (Holmberg et al., 2001; Murton et al., 2017). Holmberg et al. (2001) used the perceptual term "scrape" to group together vocal fry, roughness, and irregular phonation. They found that patients with voice disorders presented with significantly less scrape after voice therapy compared to their pre-therapy baseline voices. Additionally, Murton et al. (2017) used the detection algorithm discussed here to analyze the voices of patients undergoing treatment for acute heart failure. They found that these patients displayed a higher proportion of creaky voice after completing acute heart failure treatment, compared to their pre-treatment voices. IPPs are also relevant to automatic speech recognition and other systems that are aimed at extracting this information automatically.

In sum, episodes of IPP carry many different kinds of information, but despite their utility, identifying IPPs by hand is time-consuming and requires training. This difficulty makes reliable automatic detection of this phenomenon a compelling goal (Redi & Shattuck-Hufnagel, 2001). Previous work has used frame-based features, including short-term power, intra-frame periodicity, inter-pulse similarity, subharmonic energy, and glottal pulse peakiness, as inputs to

an artificial neural network (ANN) (Kane et al., 2013; Ishi et al., 2008; Drugman et al., 2014). This ANN then assigns creak probabilities from 0 to 1 at regularly spaced time intervals across an acoustic recording. In contrast to the continuously varying creak probability, hand-labeling of IPPs or creaky voice is a binary decision—a frame is either in a creaky/IPP region or not. To be compared to hand labels, the algorithm's continuous output needs to be converted into a binary classification decision by identifying an appropriate creak probability threshold.

Drugman et al. (2014) performed this threshold identification process with a multi-lingual data set containing creaky voice hand labels. Their results indicated that creak probability thresholds that maximized the F1 score for each speaker were in the range of 0.3 for all speakers. In this project, we perform a similar analysis with a data set that consists of conversations recorded by eight American English speakers and was hand-labeled for IPP regions by experienced labelers. Our goal is to determine whether the threshold identified in previous work is perceptually meaningful in our data set before we continue on to additional analyses based on IPP detection.

3.2 Methods

The American English Map Task corpus consists of acoustic recordings (16 kHz sampling rate) of 16 dyadic conversations by eight American English speakers who each wore a close-talk lapel microphone (*American English Map Task*, 1999). The speakers were all female, aged 18–22 years, and were familiar with each other. The recordings consisted of casual conversational speech during which one speaker (instruction giver) provided instruction to the other (instruction follower) to navigate through a map. Each speaker was an instruction giver in two conversations and instruction follower in two additional conversations. Only the data from

instruction givers was included in the analysis. The data include four speakers recorded in two conversations, and four in a single conversation, for a total of 12 recordings.

For those 12 conversations, IPP regions were hand-labeled in Praat by trained raters following a standard labeling procedure similar to the one described in Dilley et al. (1996). Raters identified speech regions with both (1) an auditory perception of non-modal voice quality associated with IPPs and (2) a visible irregularity in temporal spacing of periods in the acoustic waveform. Transcripts were also created by hand in Praat, indicating the time regions corresponding to each spoken word.

IPP regions were automatically identified using an algorithm developed for creaky voice analysis (Kane et al., 2013; Ishi et al., 2008; Drugman et al., 2014). Window size settings were 25 ms for linear predictive coding analysis, 4 ms for short-term power, and 32 ms for intra-frame periodicity. Every 1 ms, the ANN generated a creak probability from 0 to 1, inclusive. Figure 3.1 illustrates the waveform, transcript, IPP hand labels, and creak probability contour for a short sample of the recorded speech. Only frames that occurred within a word were included; frames that occurred during silence, non-speech noise, or the instruction-follower's speech were discarded.

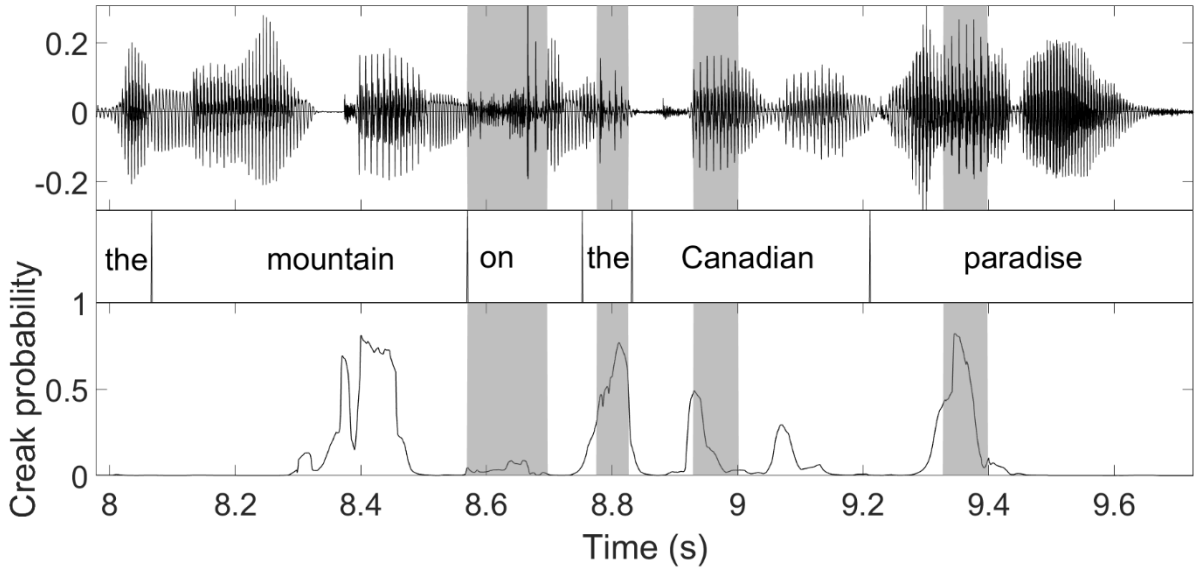

Figure 3.1. Acoustic waveform (above) and automatically detected creak probability contour (below) from a section of Conversation 1 (Speaker 1). Shaded areas indicate hand-labeled IPP regions.

Because the outputs from the creaky voice algorithm varied continuously while the hand labels were binary, our goal was to choose a binary classification rule with the creak probability threshold that most closely aligned with the binary hand labels. Therefore, we used a variety of performance metrics to identify several different candidate thresholds and compared the classifier's performance at each threshold. We applied these performance metrics to a data set derived from concatenating all 12 conversations together. We swept thresholds from 0 to 1 and calculated the following performance metrics at each threshold: true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), positive predictive value (PPV), accuracy, and F1 score. The F1 score is based on the numbers of true positives, false negatives, and false positives. It is especially appropriate for evaluating performance when the positive class is comparatively rare in the data set, as in the case of laughter (Scherer et al., 2009) or creaky voice/IPP usage. We also plotted the FPR against TPR at each threshold to generate a receiver operating characteristic (ROC) curve, and calculated the area under the ROC curve (AUC) for the combination of all 12

conversations as well as each speaker's conversations individually. AUC is a measure of overall performance, where an AUC closer to 1 indicates better classification.

We compared creak probability thresholds across all speakers using four threshold criteria: (1) threshold at which TNR and TPR were equal (the equal error rate, or EER); (2) threshold yielding maximum F1 score; (3) threshold yielding maximum accuracy score; and (4) 0.3 creak probability, for comparison to results from Drugman et al. (2014). Finally, to examine differences in algorithmic performance across speakers, we visualized the creak probability distributions of frames that were or were not hand-labeled as containing IPPs for each speaker.

3.3 Results

The instruction-givers' speech was recorded and labeled in 12 conversations, which totaled 98 minutes of recording. Approximately half of this recording time (48.5 minutes) consisted of the instruction-giver speaking. Approximately 11% (5.2 minutes) of this speaking time was hand-labeled as containing IPP segments, with a range of 6% to 18%. The variation in recording, speaking, and IPP region duration is reported in Table 3.1.

Table 3.1. Duration (in seconds) of total recording, speech segments (i.e. total time of all words as labeled in the Praat transcript), and IPP regions for each speaker who acted as an instruction giver. The percentage of speaking time that occurred in IPP regions is also reported.

Speaker	Recording time (s)	Speaking time (s)	IPP time (s)	IPP % of speaking time
1	726	301	32	11%
2	532	319	28	8.8%
3	524	296	26	8.8%
4	733	434	62	14%
5	953	496	29	5.9%
6	446	245	44	18%
7	1053	413	62	15%
8	894	405	30	7.5%

Figure 3.2 illustrates how the four creak probability threshold candidates were derived using the contours of the calculated performance metrics at each threshold. The intersection point of TNR and TPR gave the TPR = TNR threshold, and the location of the accuracy and F1 maxima gave the Max Accuracy and Max F1 thresholds. These contours were also used to obtain the values of multiple performance metrics at each candidate threshold. The threshold probabilities that this process identified, and the values of various performance metrics at each threshold, are presented in Table 3.2. The maximum F1 score of 0.549 was located at a threshold of 0.0202, which was used for additional analysis of individual speakers.

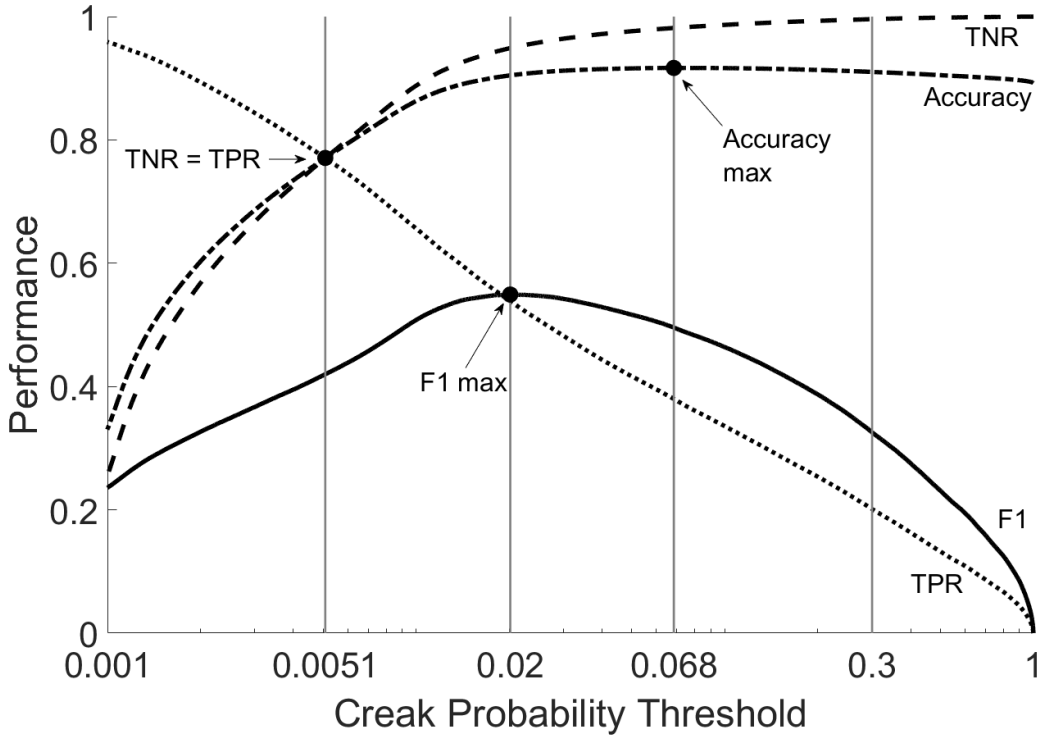

Figure 3.2. Classifier performance in terms of various metrics (TNR, TPR, accuracy, and F1) for all conversations concatenated. The four creak probability threshold candidates (at TNR = TPR, max accuracy, max F1, and 0.3) are indicated with vertical lines. Note that the horizontal axis is logarithmic for visualization purposes.

Table 3.2. Creak probability thresholds and performance metrics corresponding to four threshold criteria. Data is based on the group combining all speaker conversations.

Criterion	Creak threshold	F1	Accuracy	PPV	TPR	TNR
Max F1	0.0202	0.549	0.904	0.560	0.539	0.949
Max Accuracy	0.0684	0.496	0.917	0.713	0.340	0.981
TPR = TNR	0.0051	0.420	0.771	0.289	0.771	0.771
0.3	0.3	0.326	0.910	0.854	0.201	0.996

Using the TPR and FPR (given by 1 – TNR) in Figure 3.2 yielded a ROC curve with an AUC of 0.853. The AUCs for equivalent ROC curves based on each speaker's conversations individually ranged from 0.797 to 0.926, with a mean of 0.864.

Overall, there were large differences in the distribution of creak probabilities for frames that were labeled as being in IPP regions compared to those that were not. However, the shapes of the probability distributions in IPP regions varied considerably by speaker, as shown in Figure 3.3. Additionally, we calculated the F1 score for each speaker at the 0.0202 threshold, yielding a mean F1 of 0.539 (standard deviation of 0.084).

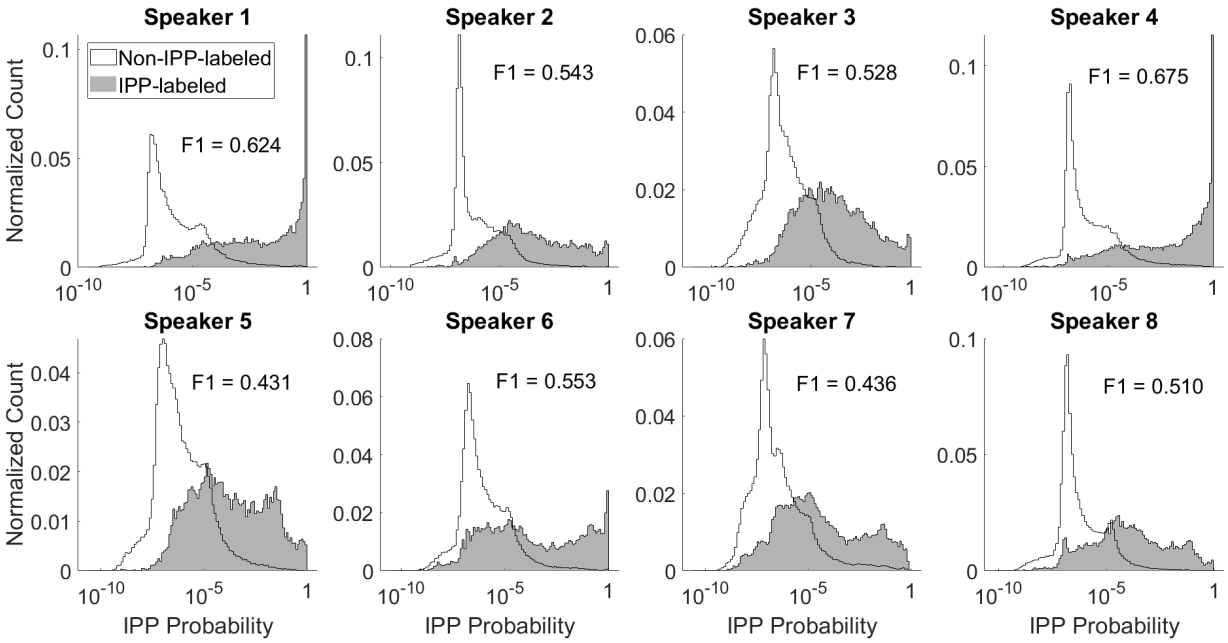

Figure 3.3. Creak probability distributions of frames that were and were not hand-labeled as being in IPP regions for each speaker. Total bin counts are normalized to sum to 1 within each IPP label (IPP or non-IPP) for each speaker. Text indicates each speaker's F1 score at the 0.0202 threshold, which gave the maximum F1 score for all conversations combined.

3.4 Discussion

Overall, we found that a lower threshold for creak probability than stated in the literature yielded satisfactory alignment with hand labels of IPP segments. While no single threshold maximized all performance metrics, our goal was to identify a threshold that best balanced the necessary tradeoffs among different metrics. Because most speech frames were not creaky, it was important to choose a detection threshold carefully. Detection of rare events is more challenging than events that occur with more even probability, since the number of false positives can easily overwhelm the number of true positives if the threshold is set too low. However, setting the threshold too high risks missing many of the true positive tokens.

We chose the F1 score as the performance metric that best addressed this problem. The two higher thresholds (maximum accuracy and 0.3) under-identified creaky frames, leading to high PPV but low actual detection rates (TPR). In contrast, the lower threshold that equalized

TPR and TNR had good detection rates, but the correspondingly high number of false positives made the PPV unacceptably low. A threshold located at the maximum F1 score gave high accuracy and identified over half of the hand-labeled creaky frames while still keeping the PPV above 0.5. Like Drugman et al. (2014) we selected the maximum F1 score to find a threshold for detection of creaky frames in running speech. However, in our data set, the maximum F1 scores were typically achieved with a creak probability threshold of ~0.02, whereas Drugman et al. (2014) found their F1 maxima using a creak probability threshold of ~0.3. Further, the detection output is intended to indicate the probability of a given frame lying in a creaky/IPP region. Therefore, a naively chosen threshold probability would be 0.5; i.e., a frame would be considered to fall in an IPP region if the probability assigned by the detector were above chance. Here, instead, we find that frames assigned a probability of only 0.02 aligned well with hand-labeled IPP regions; in other words, that a frame with just a 3% "probability" of falling in an IPP region can actually be expected to do so. A threshold of 0.02 is so low that it is effectively meaningless as a "probability" of creak/IPP. This unexpected finding suggests that the algorithm's output may be reflecting some underlying phenomenon that is different from creaky or IPP phonation, and calls for further investigation.

Although the detection algorithm performed well at the lower threshold, the large difference between our results and those from Drugman et al. (2014) suggests that our input data set is different from theirs in some important way. That difference is likely related to our hand-labeling techniques and/or our speaker populations. The hand-labeling process of identifying creaky regions both auditorily and visually is similar to the process described by Drugman et al. (2014), but many factors could have caused differences in output despite that broad similarity. Our speakers were all American English speakers, while Drugman et al. (2014) analyzed data

sets of speakers speaking Swedish, Finnish, and Japanese in addition to American English. The acoustic realizations of IPPs may differ across languages and cultures, causing the algorithm to interact differently with our speakers than with speakers of other languages.

For example, 11% of speech in our data set was hand-labeled as being in an IPP region, compared to approximately 6% of speech in the Drugman corpora. Speakers of American English in the Drugman corpus tended to have higher proportions of creaky/IPP-labeled speech (7.7%) compared to non-English speakers (5.7%), which may explain why our proportion of IPP-labeled speech was higher. Like previous studies on different corpora, our own results (Figure 3.3) indicate that even speakers within a demographic group (in this case, young female American English speakers) show different distributions of detection probability in speech frames that are labeled in IPP regions. Our results indicate that group-specific patterns of IPP usage and acoustic production may affect the performance of this automatic detection algorithm and should be considered when applying it to novel speaker populations.

The results presented here raise a number of questions for future investigation. First, these results are based on frame-by-frame comparisons of hand and automatic labels. However, IPP regions typically span many frames, so this frame-by-frame comparison may be unnecessarily strict. For example, if the hand labelers and algorithm identified 1-second IPP regions with 80% overlap, it might be sufficient to treat the algorithm as having detected that IPP region, without penalizing it for the false positive and false negative frames on the region's boundaries. This approach would require determining a minimum overlap amount necessary for "detection" of an IPP region, as well as an appropriate probability threshold for the classifier. Taking this more generous event-based approach to IPP detection might give a more accurate picture of the agreement between human and automatic labels.

Second, as discussed previously, IPPs can have varied acoustic realizations across speakers and speech contexts. Drugman et al. (2014) identified three acoustic patterns of creaky voice production and reported different detection rates for each pattern. Examining our data in light of these or other patterns might help explain our classifier's performance in different speech contexts, across speakers, and in comparison to other experimental results. For example, our data set is also labeled for prosodic structure with Tones and Break Indices (ToBI) labels that capture phrasing and prominence (Beckman et al., 2006). Using automatic detection results to distinguish different acoustic realizations of IPPs could expand on previous work regarding the interactions between IPP realization and prosodic location (Dilley et al., 1996; Gerratt & Kreiman, 2001; Keating et al., 2015) and help to better understand the algorithm's performance in a variety of prosodic contexts. Additionally, because we classified frames with an output probability above 0.02 as IPP frames, our data set contains IPP frames with a very wide range of assigned probabilities. It is possible that different IPP probability ranges correspond to distinct acoustic realizations of IPP, or have other meaningful differences between them.

Finally, IPP-like voice characteristics are known to interact with voice disorders (Holmberg et al., 2001) and systemic disease (Murton et al., 2017). Applying this algorithm to clinical populations could provide more information about how these disorders affect IPP production.

3.5 Conclusion

IPPs in speech are informative, but challenging to detect because they require time and training to label by hand. We investigated appropriate detection thresholds for an algorithm that uses multiple acoustic features input to an ANN to automatically yield creak probabilities that were hypothesized to be related to IPP behavior. This work has implications for use in future

work investigating different productions and realizations of IPPs, relating IPP usage to specific prosodic contexts, and understanding the relationships between IPP usage and human health states.

Chapter 4. Cepstral peak prominence values for clinical voice evaluation

Published as

Murton, O., Hillman, R., & Mehta, D. (2020). Cepstral peak prominence values for clinical voice evaluation. *American Journal of Speech-Language Pathology*, *29*(3), 1596–1607. https://doi.org/10.1044/2020_AJSLP-20-00001

Author Contributions

R. Hillman developed the study. O. Murton carried out the data analysis and interpretation with the guidance of R. Hillman and D. Mehta. O. Murton wrote the manuscript with input from the other authors. All authors approved the final published version.

Abstract

Purpose

The goal of this study was to employ frequently used analysis methods and tasks to identify values for cepstral peak prominence (CPP) that can aid clinical voice evaluation. Experiment 1 identified CPP values to distinguish speakers with and without voice disorders. Experiment 2 was an initial attempt to estimate auditory-perceptual ratings of overall dysphonia severity using CPP values.

Method

CPP was computed using the Analysis of Dysphonia in Speech and Voice (ADSV) program and Praat. Experiment 1 included recordings from 295 patients with medically diagnosed voice disorders and 50 vocally healthy control speakers. Speakers produced sustained /a/ vowels and the English-language Rainbow Passage. CPP cutoff values that best distinguished patient and control speakers were identified. Experiment 2 analyzed recordings from 32 English speakers with varying dysphonia severity and provided preliminary validation of the Experiment 1 cutoffs. Speakers sustained the /a/ vowel and read four sentences from the Consensus Auditory-Perceptual Evaluation of Voice protocol. Trained listeners provided auditory-perceptual ratings of overall dysphonia for the recordings, which were estimated using CPP values in a linear regression model whose performance was evaluated using the coefficient of determination (r^2).

Results

Experiment 1 identified CPP cutoff values of 11.46 dB (ADSV) and 14.45 dB (Praat) for the sustained /a/ vowels, and 6.1 dB (ADSV) and 9.3 dB (Praat) for the Rainbow Passage. CPP values below those thresholds indicated the presence of a voice disorder with up to 94.5%

accuracy. In Experiment 2, CPP values estimated ratings of overall dysphonia with r^2 values up to 0.74.

Conclusion

The CPP cutoff values identified in Experiment 1 provide normative reference points for clinical voice evaluation based on sustained /a/ vowels and the Rainbow Passage. Experiment 2 provides an initial predictive framework that can be used to relate CPP values to the auditory perception of overall dysphonia severity based on sustained /a/ vowels and CAPE-V sentences.

4.1 Introduction

Recent work in acoustic voice analysis has increasingly supported the cepstral peak prominence, or CPP, as an objective measure of breathiness and overall dysphonia. In 2018, guidance from the American Speech-Language-Hearing Association (ASHA) recommended CPP as a tool for "measuring the overall level of noise in the vocal signal" and as "a general measure of dysphonia" (Patel et al., 2018). In this recommendation, CPP replaces previous measures of acoustic perturbation, including jitter, shimmer, and harmonics-to-noise ratio. Those traditional measures can only be extracted from sustained vowels and rely on fundamental frequency computation, which may not be reliable for voices with more than moderate dysphonia. In contrast, CPP can be extracted from connected speech as well as from sustained vowels and does not require direct computation of the fundamental frequency.

A growing body of work has demonstrated CPP's ability to differentiate perceptually dysphonic and non-dysphonic voices across languages, disorder types, and speaking tasks. Many of these findings in English speakers are reviewed by Fraile & Godino-Llorente (2014), which also provides an overview of the algorithms underlying CPP computation. Other work has also examined CPP in languages other than English, including Spanish (Núñez-Batalla et al., 2019; Delgado-Hernández et al., 2019), Korean (Lee et al., 2019; Yu et al., 2018), and Turkish (Aydinli et al., 2019). In general, these studies find that lower CPP values are well-correlated with increases in dysphonia severity, based on auditory-perceptual judgments.

The clinical use of CPP is currently limited by a lack of objective guidelines that specify when values are likely to indicate abnormality. Such guidelines would increase the potential for CPP to function as a screening measure (i.e., probability of a voice disorder being present) and make it easier for clinicians to meaningfully interpret CPP, particularly with respect to treatment-

related changes (i.e., helping determine whether post-treatment vocal function and/or voice quality more closely approximate normal). Ideally such guidelines would be based on the analysis methods and tasks that are most frequently used for clinical voice evaluation and include cutoff values/thresholds for detecting the presence or absence of a voice disorder, as well information about how CPP values relate to dysphonia severity.

4.1.1 CPP conceptualization

The cepstrum typically used in voice and speech analysis is given by the inverse Fourier transform of the acoustic spectrum. This process can be intuitively understood as a "spectrum of a spectrum." First, the waveform is Fourier-transformed into the spectral domain. Then, the logarithm of that spectrum is taken and another (inverse) Fourier transform is performed into the "cepstral" domain (Fraile & Godino-Llorente, 2014; Heman-Ackah et al., 2003). The horizontal axis of a spectrum shows a range of frequencies. By analogy, the horizontal axis of the cepstrum is a time-like dimension termed *quefrency*, an anagram of *frequency*, just as *cepstrum* is an anagram of *spectrum* (Oppenheim & Schafer, 2004). The periodic harmonic peaks in the spectrum are represented as a single large peak (and its harmonics) in the cepstrum around a quefrency corresponding to the period of the voice signal. The height (i.e., "prominence") of that peak relative to a regression line through the overall cepstrum is called the cepstral peak prominence, or CPP, and is typically reported in units of decibels (dB). CPP values therefore fall into a continuous range, where lower values are typically correlated with greater levels of dysphonia.

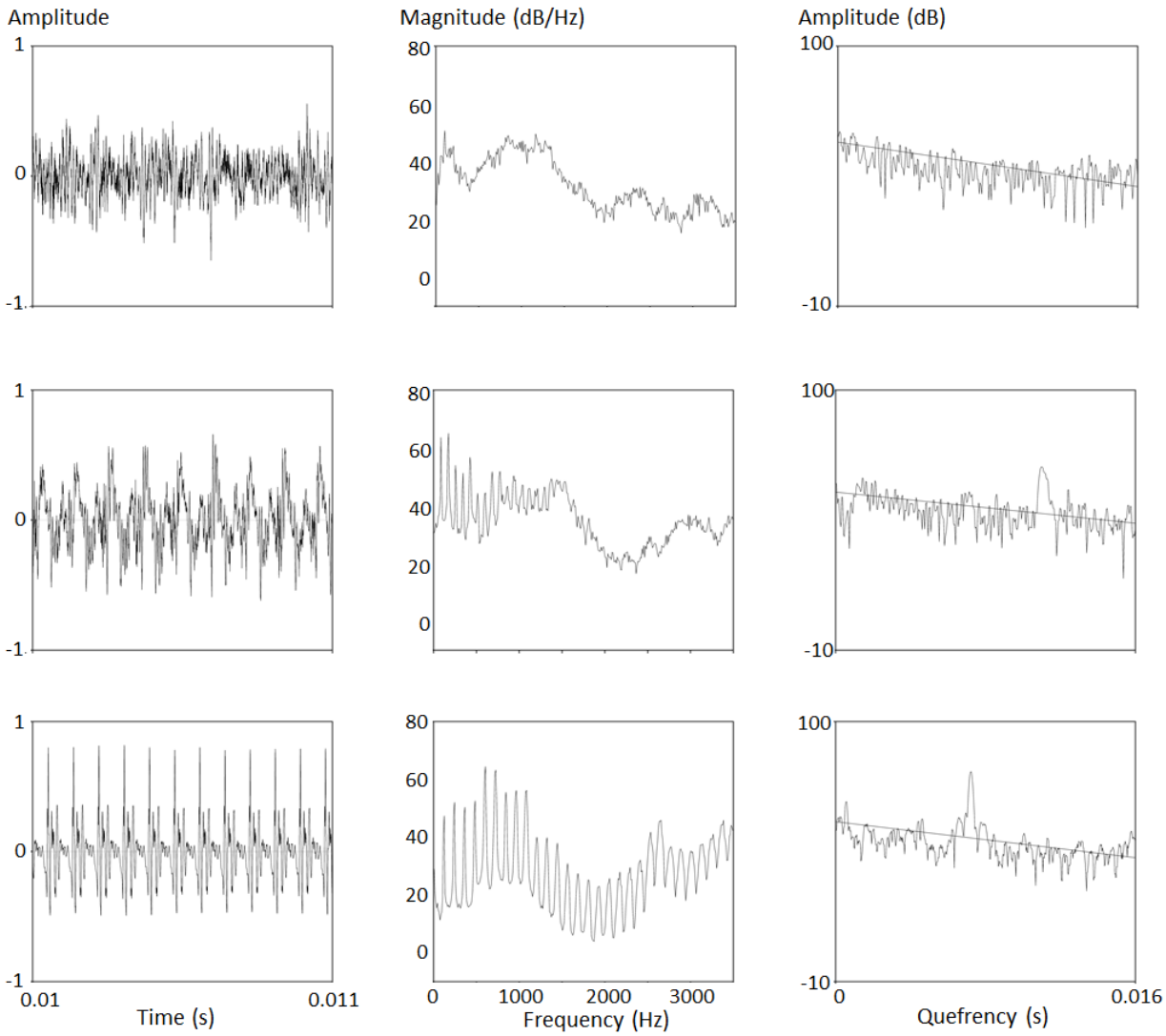

Figure 4.1. Waveform (left), spectrum (center) and cepstrum (right) from speakers with aphonia (top row), non-aphonic but disordered voice quality (center row), and no voice disorder (bottom row).

Figure 4.1 illustrates this CPP calculation process for /a/ vowels from three speakers exhibiting a typical, dysphonic, and aphonic voice, respectively. For each row, the leftmost image shows a section of the /a/ vowel's waveform. The center image shows the vowel's spectrum, and the rightmost image shows the vowel's cepstrum and the regression line used in calculating the cepstrum. The vowel from the top row comes from an aphonic speaker, so no harmonics are visible in the spectrum and no peak is apparent in the cepstrum. The vowel in the

bottom row comes from a speaker with no voice disorder, so harmonics are prominent in the spectrum and the CPP is well above the regression line. The vowel in the middle row was produced by a non-aphonic speaker with disordered voice quality, so the CPP height is lower relative to that of the typical speaker.

To distinguish normal from disordered voices, the continuous range of CPP values needs to be divided into groups at one or more thresholds. Each potential CPP threshold could yield a different sensitivity (true positive rate, TPR) and specificity (true negative rate, TNR). In general, a desirable threshold will have high values for both of these quantities (although it may not always possible to maximize both TPR and TNR at the same time). For any test, a threshold should be chosen such that the rate and type of errors are both acceptable for that test's goal.

4.1.2 Previous work

Table 4.1 summarizes prior studies that examined the ability of CPP to distinguish healthy from pathological voices. In general, they did so by obtaining auditory-perceptual ratings (typically of overall severity and/or breathiness) and correlating those perceptual values with objective CPP values. They often also established CPP cutoff thresholds by dividing the perceptual ratings into two or more categories of dysphonia severity and determining CPP's performance in classifying voices into those groups.

Many of these studies included both sustained vowels and continuous speech tasks. In all of those cases, CPP thresholds were lower for continuous speech tasks compared to sustained vowels. This general finding suggests that it is important to keep speech tasks consistent when comparing CPP values across recordings. Additionally, the choice of algorithm for calculating CPP is critically important, as different algorithms produce values in different ranges. The three major CPP computation algorithms in these studies are Hillenbrand & Houde's (1996) algorithm

for calculating smoothed CPP (CPPS), Praat's CPPS algorithm (Boersma & Weenink, 2018), and the CPP computation method in the Analysis of Dysphonia in Speech and Voice (*ADSV*, 2019).

Two studies by Heman-Ackah et al. (2003, 2014) used Hillenbrand's algorithm (1996) to identify CPPS thresholds that distinguish voices with severe dysphonia from voices with mild or no dysphonia. These studies provide a basis for comparison of threshold CPP values for sustained vowels and running speech in English speakers. However, because they excluded speakers with moderate dysphonia, it is not clear that those thresholds are appropriate for use in all speakers.

Studies by Núñez-Batalla et al. (2019) and Delgado-Hernández et al. (2019) used Praat's CPPS algorithm to investigate CPP in Spanish speakers with and without diagnoses of voice disorders. Núñez-Batalla et al. (2019) identified normative CPPS values by computing the averages and standard deviations of the control groups' CPPS values for each task, rather than identifying threshold values to separate speakers with and without voice disorders. Delgado-Hernández et al. (2019) used two distinct configurations of Praat—the default configuration (Configuration 1) and the configuration used to calculate the Acoustic Voice Quality Index (Configuration 2)—to find CPPS cutoff thresholds based on auditory-perceptual ratings of overall severity. However, that result has not been replicated with English-language speakers or with other voice analysis programs (e.g. ADSV) that clinicians may also use.

Several studies have used the ADSV program to analyze CPP in Korean (Lee et al., 2019; Yu et al., 2018) and Turkish (Aydinli et al., 2019) speakers. Lee et al. (2019) reported CPP values that distinguished speakers with varying levels of dysphonia, while Yu et al. (2018) reported CPP values that separated speakers with and without dysphonia. Aydinli et al. (2019)

found lower CPP values in Turkish-speaking children with nodules compared to age- and sex-matched controls, but did not report specific CPP thresholds to distinguish those populations.

In related work, Awan et al. (2016) attempted to find clinically relevant cutoff values for the Cepstral Spectral Index of Dysphonia (CSID), a computational estimate of dysphonia severity that incorporates CPP and measures of spectral energy. They defined three groups of "voice disordered" patients according to different criteria: (1) "dysphonia-positive" patients according to auditory-perceptual ratings by trained speech-language pathology students; (2) "laryngoscopic-positive" patients based on signs and symptoms visible on laryngeal stroboscopy; and (3) "VHI-positive" patients with a value greater than 12 on the 30-item Voice Handicap Index (Jacobson et al., 1997). Awan et al. (2016) found that CSID best distinguished dysphonia-positive participants from dysphonia-negative ones. CSID was less accurate for the laryngoscopic and VHI classifications. The VHI classification is arguably the least relevant comparison to CPP, since CPP is not designed to reflect a speaker's self-perception of vocal health/function.

As illustrated by the preceding review, CPP values can vary widely with different speaking tasks and computation algorithms. This variation arises in part because tasks differ in their degree of voicing, and computation algorithms differ in how they treat unvoiced segments. Watts et al. (2017) compared CPP values from Praat and ADSV. English and Flemish speakers produced sustained vowels (/a/) and continuous speech. The Flemish vowel and sentence recordings had correlation coefficients of 0.93 in ADSV and 0.88 in Praat, while the English vowels and sentences had correlation coefficients of 0.92 and 0.96 respectively.

A major difference between these algorithms is ADSV's use of a voicing activity detector. In ADSV, frames with negative CPP (i.e., cepstral peak below the regression line) are

not considered for analysis. ADSV's use of the voicing activity detector might explain why the correlation was higher for the English sentence, which was almost fully voiced, than for the Flemish sentence, which contained many unvoiced segments. Unvoiced segments are not typically periodic and are likely to have very low CPP. This result suggests the need to use the same speaking tasks when comparing CPP values from continuous speech, especially when not using a voicing activity detector. Speech samples with different degrees of voicing may yield artificially different CPP values.

Table 4.1. Summary of previous work identifying clinically relevant cepstral peak prominence (CPP) cutoff values.

Author	Year	Language	Study size	CPP method	Group classification	Sustained vowel CPP cutoff	Running speech CPP cutoff
Heman-Ackah et al.	2003	English	281 patients (176F/105M)	CPPS (Hillenbrand)	Perceptually mild vs. severe dysphonia	10 dB	5 dB
Heman-Ackah et al.	2014	English	835 patients; 50 controls	CPPS (Hillenbrand)	Perceptually normal vs. dysphonic	n/a	4.0 dB
Núñez-Batalla et al.	2018	Spanish	72 patients; 52 controls	CPPS (Praat)	Normative values (not cutoff values)	Female: 16.0 dB Male: 16.4 dB	Female: 7.9–11.3 dB Male: 7.8–10.9 dB
Yu et al.	2018	Korean	214 patients (142F/72M); 74 controls (47F/27M)	ADSV CPP	Perceptually normal vs. dysphonic	12 dB	7 dB
Aydinli	2019	Turkish	27 patients; 27 controls (40M/14F, pediatric)	ADSV CPP	Nodules diagnosis vs. normal voices	No thresholds, but found significantly lower CPP in pediatric speakers with nodules vs. age- and sex-matched controls for most, but not all, speaking tasks.	
Delgado-Hernández et al.	2019	Spanish	136 patients; 47 controls	CPPS (Praat) in two configurations	Perceptually normal vs. dysphonic	Configuration 1: 23.62 dB 2: 13.96 dB	Configuration 1: 18.4 dB 2: 8.37 dB
Lee et al.	2019	Korean	1029 patients (512M/517F)	ADSV CPP	Normal vs. mild	10 dB	7.7 dB
					Mild vs. moderate	7.5 dB	5.4 dB
					Moderate vs. severe	4.1 dB	2.9 dB

4.1.3 Current work

In this study, we follow up on the recent ASHA recommendation to use CPP in the clinical assessment of voice (Patel et al., 2018). Unlike previous studies, we use ADSV and Praat to analyze two English-language data sets and identify CPP cutoff values to detect probable voice disorders. To our knowledge, no published work has proposed clinically relevant CPP cutoff values for English speakers based on these widely used voice analysis software products. ADSV is a commercially available and supported product widely used by clinicians. Praat is free software available online that is increasingly being used for clinical assessment of voice and speech due to its ease of use, graphical user interface, and scripting features. Hillenbrand & Houde's (1996) algorithm has been used implemented for research study and not typically for clinical use—potentially due to lack of support and a user-friendly interface—and therefore is not evaluated in the present work.

In Experiment 1, we investigate CPP as a screening tool to predict the presence of a voice disorder using a voice database (Massachusetts Eye and Ear Infirmary, 1994) that has been analyzed in many other studies. In Experiment 2, we evaluate the performance of CPP to predict the auditory perception of dysphonia severity using a smaller data set of acoustic recordings that has been rigorously evaluated by trained listeners using ASHA's recommended protocol for the CAPE-V (Kempster et al., 2009). Our goal is to aid practitioners who wish to use the objective measure of CPP as part of their clinical assessment and monitoring of patients with voice disorders.

Experiment 1: CPP cutoff values for detecting the potential presence of a voice disorder

4.2 Methods

4.2.1 Data

The MEEI Voice Disorders Database consists of recordings from 687 patients diagnosed with voice disorders and 53 vocally healthy control speakers (Massachusetts Eye and Ear Infirmary, 1994). The speakers were recorded between 1992 and 1994 at the Massachusetts Eye and Ear Infirmary's Voice and Speech Lab and Kay Elemetrics (now part of PENTAX Medical). The database consists of sustained /a/ vowel productions from 657 of the patients and Rainbow Passage readings from 662 of the patients. Only one second of each sustained vowel and the first 12 seconds of each Rainbow Passage are available in the database.

In this study, we excluded three control speakers who had a history of smoking. We also excluded a total of 392 patients with the following classifications: 5 were classified as "normal", 51 were classified as post-surgery or post-therapy, 89 were missing a diagnosis, and 247 were classified generically as "pathological voice". Patients with the generic "pathological voice" classification were excluded because including individuals that lack a definitive/standard diagnosis would introduce uncertainty about the composition and integrity of the pathological data set and make the results less clinically interpretable or applicable. After these exclusions, the database consisted of 295 voice patients and 50 controls. The voice patients' primary diagnoses are presented in Table 4.2. Table 4.3 summarizes these speakers' demographics.

Table 4.2. Primary diagnoses of the 295 voice patients in the MEEI Voice Disorders Database whose recordings remained in the analysis after exclusion criteria were applied in Experiment 1.

Primary Diagnosis	Count
Neurological (92)	
Paralysis/paresis	62
Spasmodic dysphonia	19
Other neurological	11
Muscle tension dysphonia	49
Nodules or polyps	41
Lesions (including cyst, mass, dysplasia)	41
Edema (including Reinke's edema)	41
Scar and/or trauma	15
Presbyphonia	8
Partial laryngectomy	5
Other (arthritis, tuberculosis, laryngocele)	3
Total	**295**

Table 4.3. Sex and age distributions of speakers in the MEEI Voice Disorders Database analyzed in Experiment 1.

	Female	Male	Median age (years)	Age range (years)
Controls	30	20	36.5	22–59
Patients	183	112	45	13–93

All of the speakers in our data set produced both the sustained vowel and the Rainbow Passage, except for four voice patients who produced only the sustained vowel. Therefore, our data set consisted of 345 vowel recordings (295 patients/50 controls) and 341 Rainbow Passage recordings (291 patients/50 controls).

4.2.2 Acoustic and statistical analysis

Each recording was analyzed in ADSV (version 3.4.2) using the program's default settings. The "CPP/EXP Mean (dB)" parameter was extracted to yield CPP for each recording.

Each recording was also analyzed in Praat (version 6.0.40) using a PowerCepstrogram (60 Hz pitch floor, 2 ms time step, 5 kHz maximum frequency, and pre-emphasis from 50 Hz). CPPS was calculated from each PowerCepstrogram with the following settings: Subtract tilt before smoothing = "no"; time averaging window = 0.01 s; quefrency averaging window = 0.001 s; peak search pitch range = 60 – 330 Hz; tolerance = 0.05; interpolation = "Parabolic"; tilt line quefrency range = 0.001 – 0 (no upper bound) s; line type = "Straight"; fit method = "Robust". These settings are identical to those used by Watts et al. (2017) and Brockmann-Bauser et al. (2019).

As discussed above, previous studies have found substantially different CPP cutoff thresholds for sustained vowels and continuous speech. Therefore, the sustained vowel and Rainbow Passage recordings were treated separately throughout the analysis. For each task, we identified every CPP value that any participant produced on that task and calculated several performance metrics based on each value. These performance metrics included TPR, TNR, false positive rate (FPR), positive predictive value (PPV), accuracy, and Youden's J index (J), which is given by sensitivity + specificity – 1 (= TPR + TNR – 1). Therefore, Youden's J is 1 only when neither false positives nor false negatives are present. It is also not affected by the relative sizes of the positive and negative groups (Youden, 1950). That property is useful for studies in which most people in a study fall into the same class. For example, studies based on people who present to voice clinics are likely to have many more dysphonic voices than controls.

We plotted the series of TPRs against the FPRs to generate a receiver operating characteristic (ROC) curve and calculated the area under the ROC curve (AUC) to evaluate the overall classification performance. An AUC closer to 1 indicates better classification performance. The CPP threshold yielding the maximum Youden index was also identified for both the sustained vowels and running speech. This analysis was performed for the ADSV CPP and Praat CPPS calculations separately.

4.3 Results

4.3.1 ADSV CPP

Figure 4.2 (top row) shows the distribution of ADSV-based CPP values in participants with and without voice disorders for the sustained vowel and continuous speech conditions. The histograms are normalized such that the heights of each condition's bars sum to 1. Inspecting these histograms shows that CPPs from controls' and patients' voices typically fall into distinct ranges, with patients' voices showing much wider variation than controls'. Figure 4.2 (bottom row) shows the ROC curves for the sustained vowel and continuous speech conditions, with the CPP cutoff value indicating maximum Youden's index labeled on each.

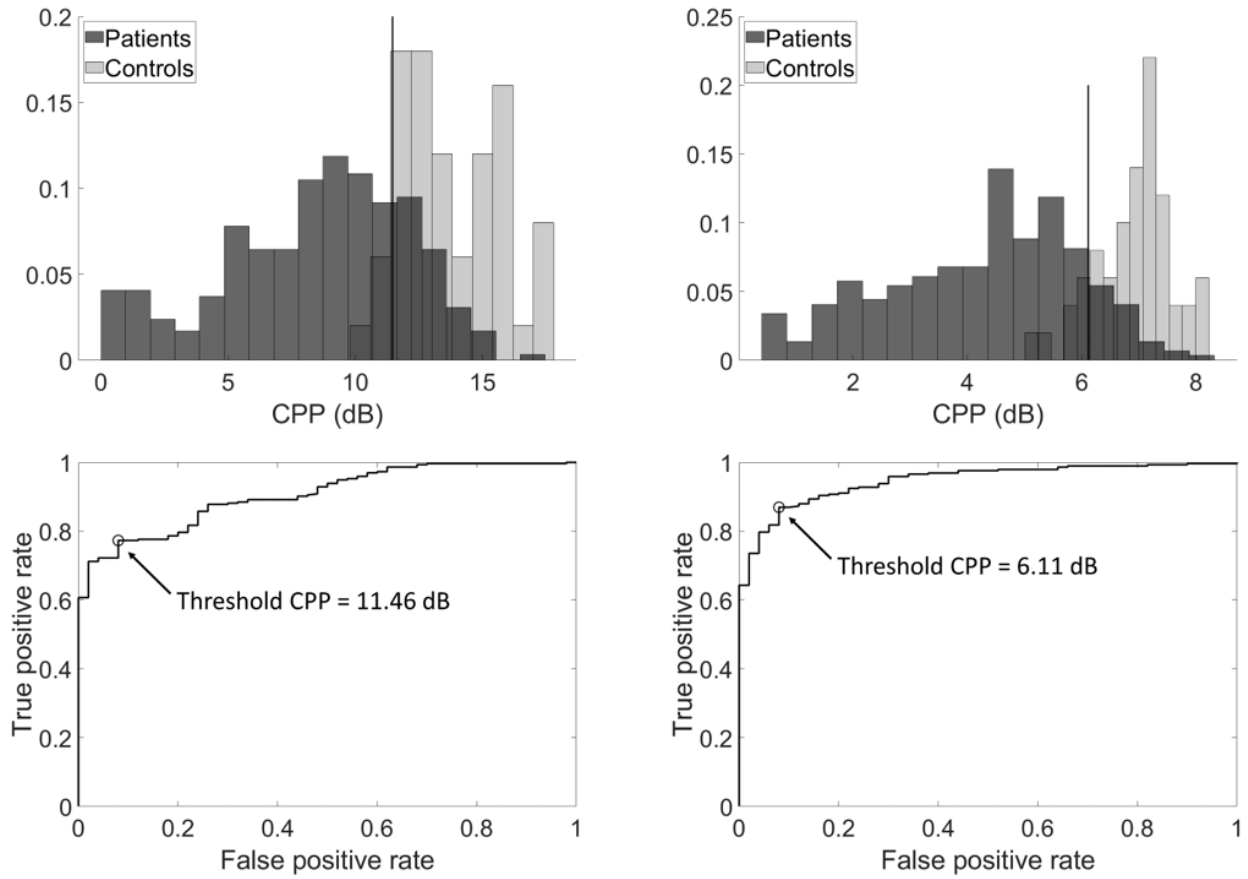

Figure 4.2. Top row: Histogram of ADSV1-based cepstral peak prominence (CPP) values from patients with voice disorders (dark) and vocally healthy individuals (light), for sustained vowels (left) and continuous speech (right). Total bin counts sum to 1 within each group. Vertical lines indicate thresholds derived from the maximum Youden's index. Bottom row: Receiver operating characteristic curves plotting true positive versus false positive rates at various CPP thresholds for sustained vowels (left) and continuous speech (right). The "positive" class is the patient group. Open circles indicate the CPP threshold given by the maximum Youden's index.

4.3.2 Praat CPPS

Figure 4.3 shows the distributions of Praat-based CPPS values (top row) and ROC curves (bottom row) for the sustained vowel and continuous speech conditions as described for Figure 4.2. The histograms are normalized such that the heights of each condition's bars sum to 1. Like the ADSV-based CPP values, Praat CPPS separates control and patient voices well, with a wider range of CPPS values for patients' voices than for controls'.

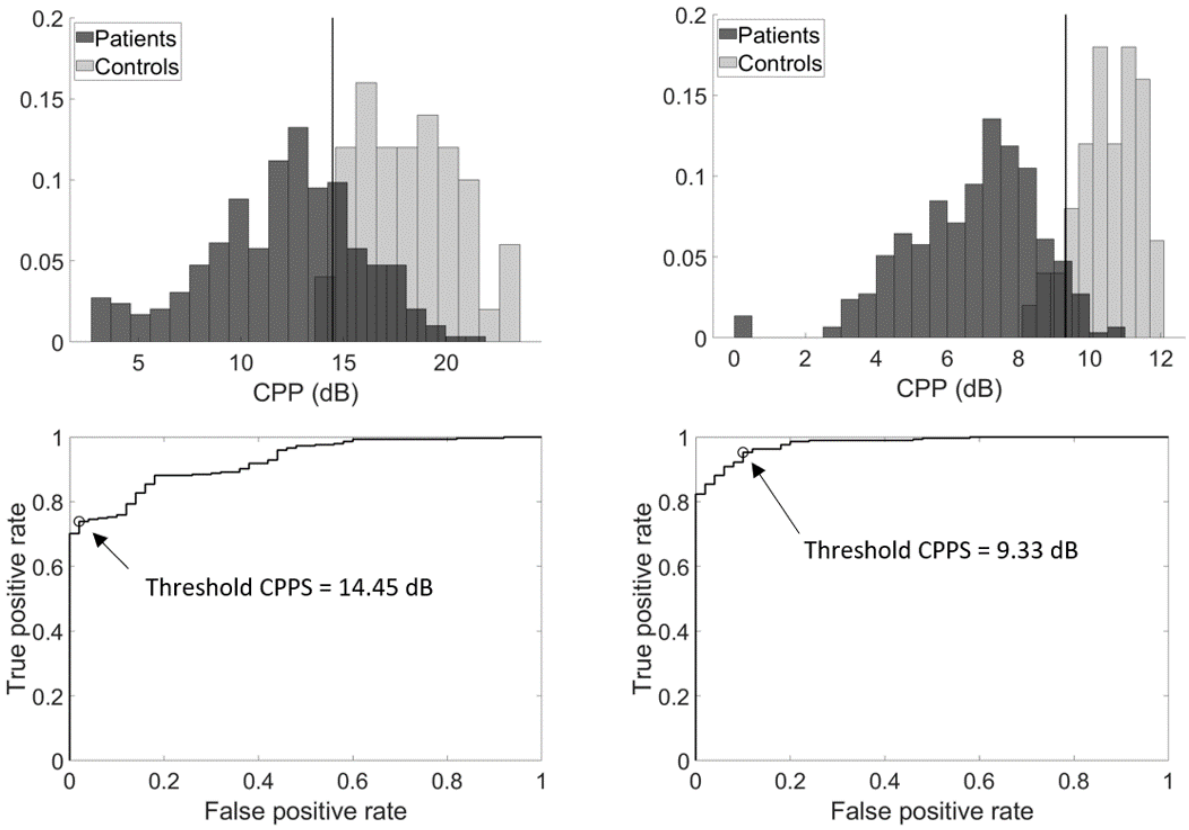

Figure 4.3. Top row: Histogram of Praat-based smoothed cepstral peak prominence (CPPS) values from patients with voice disorders (dark) and vocally healthy individuals (light), for sustained vowels (left) and continuous speech (right). Total bin counts sum to 1 within each group. Vertical lines indicate thresholds derived from the maximum Youden's index. The "positive" class is the patient group. Bottom row: Receiver operating characteristic curves plotting true positive versus false positive rates at various CPPS thresholds for sustained vowels (left) and continuous speech (right). Open circles indicate the threshold given by the maximum Youden's index.

Table 4.4 summarizes the ADSV CPP and Praat CPPS threshold values for sustained vowels and the Rainbow Passage. It also includes performance metrics including ROC AUC, accuracy, TPR, FPR, TNR, PPV, and Youden's J for each threshold value.

Table 4.4. Threshold values and performance measures for the ADSV-based CPP and Praat-based CPPS classifiers.

	ADSV CPP		**Praat CPPS**	
	Sustained vowels	Rainbow Passage	Sustained vowels	Rainbow Passage
Threshold	11.5 dB	6.1 dB	14.5 dB	9.3 dB
ROC AUC	0.91	0.95	0.93	0.98
Accuracy	79.4%	87.7%	77.4%	94.5%
TPR	0.77	0.87	0.74	0.95
FPR	0.08	0.08	0.02	0.10
TNR	0.92	0.92	0.98	0.90
PPV	0.98	0.98	0.99	0.98
Youden's J	0.69	0.79	0.72	0.85

Experiment 2: Estimating dysphonia severity using CPP values

4.4 Methods

4.4.1 Database

In Experiment 2, we analyzed a data set consisting of 32 speakers that was first published in Awan et al. (2010). This data set includes 24 speakers with voice disorders (12F/12M) and 8 speakers with typical voices (4F/4M) based on auditory-perceptual judgment and self-report. The data set also contains auditory-perceptual judgments of voice from 25 trained speech-language pathology graduate student listeners (Kempster et al., 2009). In contrast, the Experiment 1 data set contains only a binary categorization of speakers with and without voice disorder diagnoses. Experiment 2's large set of listener ratings provides a valuable opportunity to quantitatively relate CPP to auditory-perceptual judgments of voice, using a continuous scale rather than the binary decision from Experiment 1. The voice-related diagnoses of the 24 speakers with voice disorders are summarized in Table 4.5. Each speaker produced a sustained /a/ vowel and four of

the six CAPE-V sentences targeting various voicing behaviors: easy onsets (S2: “How hard did he hit him?”), full voicing (S3: “We were away a year ago”), hard glottal attacks (S4: “We eat eggs every Easter”), and voiceless stops (S6: “Peter will keep at the peak”) (Kempster et al., 2009).

Table 4.5. Diagnoses of speakers in Experiment 2, from Awan et al. (2010).

Diagnosis	Count	Male	Female
Paralysis/paresis	8	4	4
Muscle tension dysphonia	4	2	2
Cyst	4	1	3
Nodules/polyp	3	1	2
Papilloma	2	2	0
Amyloidosis	1	1	0
Cancer	1	1	0
Reinke’s edema	1	0	1
Total	**24**	**12**	**12**

Twenty-five trained speech-language pathology graduate students participated in five separate listening sessions to produce five ratings of each utterance for overall severity, roughness, breathiness, and strain according to the CAPE-V evaluation criteria. These listener ratings resulted in 125 ratings (25 listeners x 5 ratings each) for each sustained vowel and CAPE-V sentence. Measures of inter- and intra-rater reliability indicated that the listeners accurately distinguished between dysphonia severity levels and provided highly reliable ratings. The listener rating process is described in more detail in Awan et al. (2010). Final ratings of each dysphonia category were computed as the mean over all 125 ratings for that category. For this study, overall severity was selected as the auditory-perceptual category to be estimated using CPP values.

4.4.2 Acoustic and statistical analysis

Following the same procedure as our analysis of the MEEI corpus, each recording was analyzed in ADSV using the program's default settings. The "CPP/EXP Mean (dB)" parameter was extracted to yield a CPP value for each recording. Additionally, for each recording, all 125 listener ratings of overall severity were averaged to yield a single measure of perceived overall dysphonia severity. Separately, the Praat-based CPPS was calculated using the parameters described in Experiment 1.

Linear regression models were computed using the MATLAB `fitlm` function to assess CPP's ability to estimate the perceptual ratings of overall severity. These models were computed separately for the vowels and each of the four CAPE-V sentences, resulting in five regression models of the form $predicted\ severity\ rating\ = m * CPP + b$. For each model, the coefficient of determination (r^2) was computed, and 95% prediction intervals were calculated as:

$$PI_{95} = y_i \pm t_{crit} * RMSE * \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}}$$

where n represents the number of observations in the model, y_i represents the model's prediction given x_i , t_{crit} represents the critical t value for $n-2$ observations at a 95% significance level, and *RMSE* is the root-mean-squared error of the regression model.

Finally, the sustained vowels' CPP cutoff scores from Experiment 1 (11.46 dB for ADSV and 14.45 dB for Praat) were used to classify the patient and control speakers in Experiment 2 as an initial of validation of the cutoff scores in an independent data set. TPR, TNR, FPR, PPV, and accuracy were calculated for the vowel-based ADSV and Praat cutoffs separately. Note that FPR and TNR always sum to 1, so if one value is high then the other will be low.

4.5 Results

4.5.1 ADSV CPP

Figure 4.4 shows best-fit regression lines and PI_{95} linking ADSV-based CPP to the mean listener rating of overall severity for each task separately. In general, CPP and overall severity were well-correlated, with r^2 ranging from 0.5 to 0.74, depending on the task. As expected, the r^2 value of 0.71 for the /a/ vowels is very close to the r^2 of 0.70 found by Awan et al. (2010) on this data set using a similar version of ADSV. Subfigure titles include the regression line equation relating listener ratings to CPP, and the r^2 value is indicated in each subfigure. The x-axis tick labels show the CPP values that correspond to each y-axis label based on the regression model. For example, a CPP value of 2.7 on an /a/ vowel corresponds to a mean listener rating of 80 on the CAPE-V overall severity scale (Kempster et al., 2009). Dashed lines indicate PI_{95} for each point on the regression line.

The CPP threshold value of 11.46 dB from Experiment 1 yielded an Experiment 2 accuracy of 68.8% (22/32), TPR 87.5% (21/24), TNR 12.5% (1/8), PPV 75% (21/28), and FPR 87.5% (7/8).

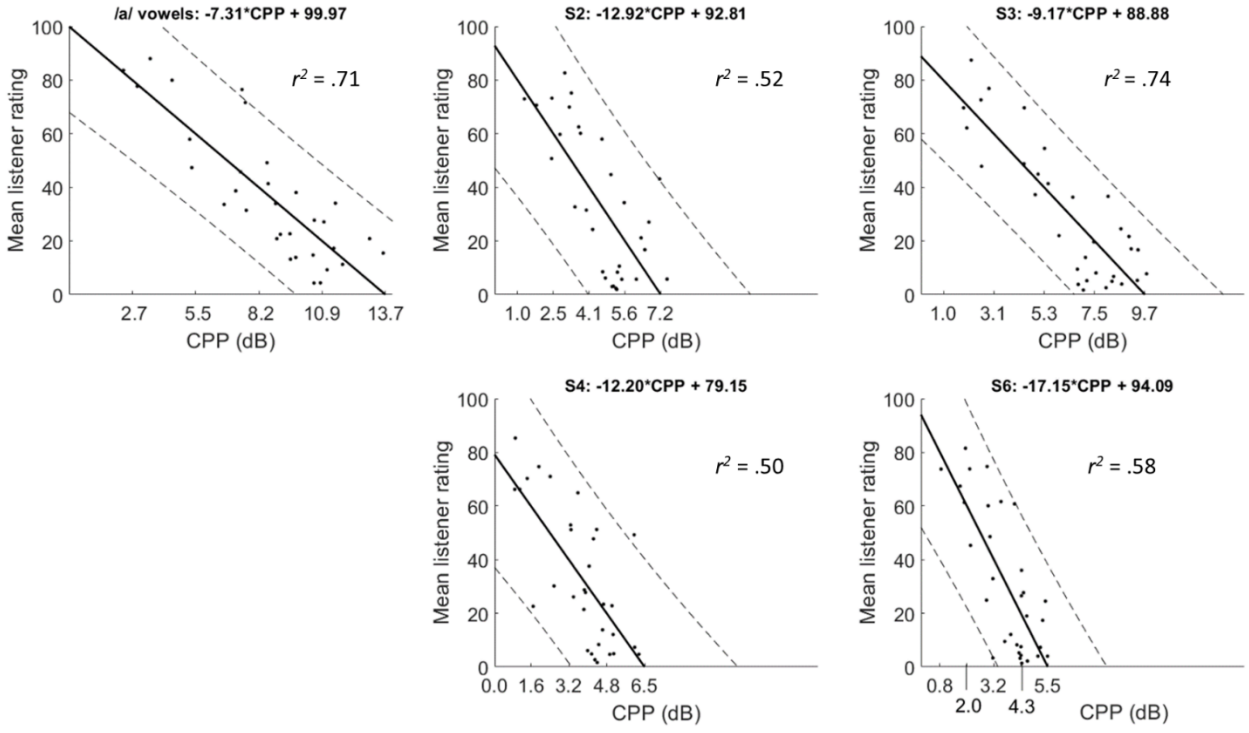

Figure 4.4. Correlations between ADSV-based cepstral peak prominence (CPP) and listener rating of overall severity for each speaking task. Solid lines indicate best-fit regression line, and dashed lines show 95% prediction intervals. X-axis tick labels show the CPP values corresponding to each y-axis tick label based on the regression line.

4.5.2 Praat CPPS

Figure 4.5 shows regression lines and PI_{95} relating Praat CPPS to mean listener ratings of overall severity. The figure was generated following the same procedure used to create Figure 4.4, with the regression line, r^2 value, and PI_{95} indicated on each subfigure. The r^2 values relating overall severity to CPPS ranged from 0.38 to 0.72.

The CPPS threshold value of 14.45 dB from Experiment 1 yielded an Experiment 2 accuracy of 75% (24/32), TPR 79% (19/24), TNR 62.5% (5/8), PPV 86.4% (19/22), and FPR 37.5% (3/8).

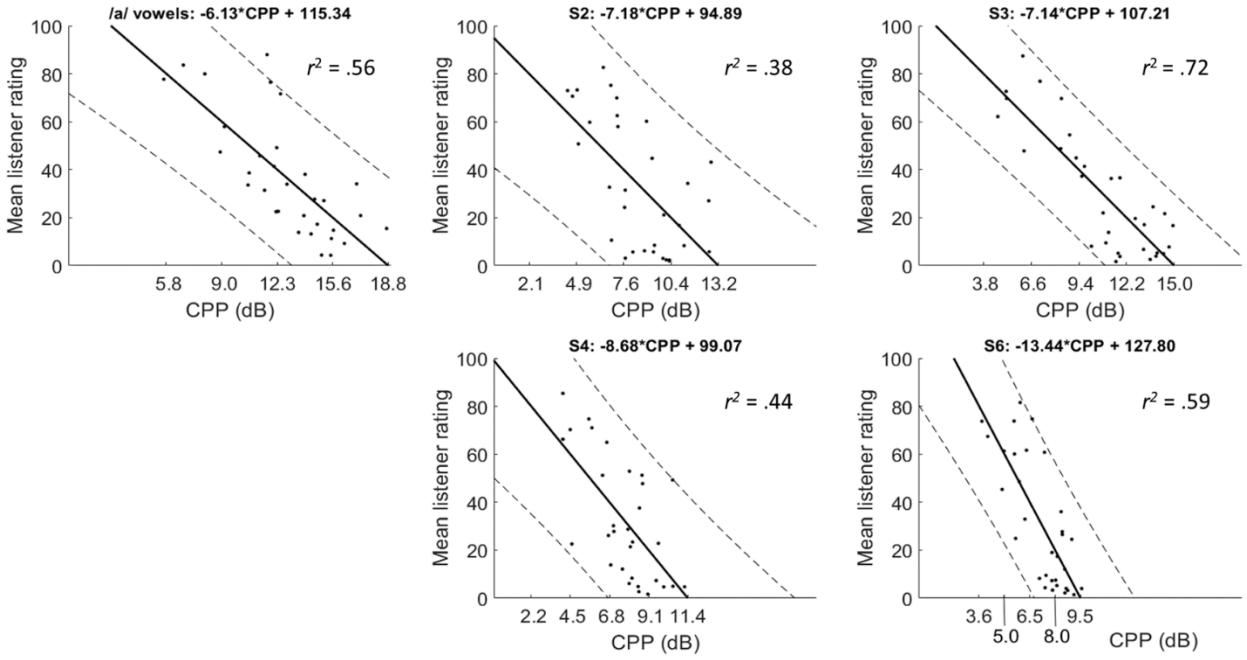

Figure 4.5. Correlations between Praat smoothed cepstral peak prominence (CPPS) and listener rating of overall severity for each speaking task. Solid lines indicate best-fit regression line, and dashed lines show 95% prediction intervals. X-axis tick labels show the CPPS values corresponding to each y-axis tick label based on the regression line.

4.6 Discussion

CPP is widely understood to be an accurate predictor of dysphonia severity. To our knowledge, this is the first study to identify CPP values that are based on using both ADSV and Praat-based analysis methods on the same well-controlled databases of English speakers.

4.6.1 CPP cutoff values and comparisons to previous studies

Our results from Experiment 1 suggest that ADSV-based CPP values below 11.46 dB for sustained vowels and below 6.1 dB for the Rainbow Passage should be considered indicative of a voice disorder. Praat-based CPPS values below 14.45 dB for sustained vowels or below 9.3 dB for continuous speech indicate a high probability of the presence of a voice disorder. The cutoff values indicated here represent only one possible estimate of a CPP clinical cutoff, so values in the near vicinity of the cutoff should be given further consideration when used clinically. There

are several other possible methods of determining an appropriate cutoff threshold that could be applied in future work to balance specificity and sensitivity in different ways (Habibzadeh et al., 2016; Unal, 2017).

Separately, our results from Experiment 2 suggest quantitative relationships between CPP values and perceptual ratings of dysphonia severity levels. The regression lines above each plot in Figure 4.4 and Figure 4.5 can be used to predict dysphonia severity based on ADSV CPP (Figure 4.4) or Praat CPPS (Figure 4.5). For example, if a clinician used Praat to analyze a speaker's /a/ value and found a CPPS value of 10 dB, the predicted CAPE-V overall severity rating would be approximately 54: $-6.13 * 10 + 115.34 = 54.04$. That said, the data set for Experiment 2 is relatively small, and the PI_{95} ranges are fairly large, so these results should be applied with caution and call for additional study with larger databases.

Our results are comparable to those from previous similar studies. For example, Yu et al. (2018) found ADSV-based CPP thresholds of approximately 12 dB for sustained vowels and 7 dB for running speech in Korean speakers. Heman-Ackah et al. (2003) used Hillenbrand's algorithm with English speakers and found that thresholds of 10 dB for sustained vowels and 5 dB for running speech distinguished mild from severe dysphonia. These thresholds are somewhat lower than ours, but our Experiment 1 cutoff values are intended to distinguish patients with voice disorders vs. vocally healthy controls instead of mild vs. severe dysphonia. Additionally, our Praat CPPS thresholds are similar to the ones identified by Delgado-Hernández et al. (2019), whose Praat "Configuration 2" settings yielded CPP thresholds of 13.96 dB for sustained vowels and 8.37 dB for continuous speech.

4.6.2 Validation of Experiment 1 thresholds using Experiment 2 data set

We used the cutoff values for /a/ vowels from Experiment 1 to classify the patient vs. control voices from Experiment 2. We did not perform that validation for the connected speech because the speaking tasks were different (Rainbow Passage vs. CAPE-V sentences) and the resulting CPP values could not be directly compared. The accuracy scores for ADSV-based CPP and Praat-based CPPS were similar, at 68.8% for ADSV and 71.9% for Praat. The TPR was high for both ADSV (21/24) and Praat (18/24), but the TNR was higher for Praat (5/8) than for ADSV (1/8). Although the TNR for ADSV seems low, there were four control speakers whose ADSV-based CPP values were close to the cutoff value (< 1 dB below). The remaining three control speakers were the same ones who were below the Praat CPPS threshold. Those speakers also had the highest overall severity ratings of the control group, as judged by the trained listeners. Overall, these results indicate that the Experiment 1 CPP cutoffs classified most of the Experiment 2 speakers accurately and that clinicians should use particular caution when interpreting CPP values that are close to the clinical cutoff thresholds.

4.6.3 Choices of task and computation algorithm are important

An important finding from Experiment 2 is the wide variation in CPP values between the different CAPE-V sentence tasks. This result suggests that between and within speaker comparisons of CPP values for continuous speech should be based on the same speech material (e.g. same sentences from the CAPE-V or reading passage). Within a single speech task, speakers are likely to be similar to each other in their production of non-voiced elements like consonants and pauses, so changes in speakers' vocal quality can be directly observed (Hillenbrand & Houde, 1996).

One possible explanation for the variation in running speech CPP values is the differing amounts of voicing in each CAPE-V sentence. All four sentences tended to have lower CPP values than the sustained vowels. Sentence 3 was fully voiced ("We were away a year ago") and tended to have the highest CPP values of the four sentences. In contrast, Sentence 6 had many voiceless stops ("Peter will keep at the peak") and tended to have the lowest CPP values. This pattern also occurs for the Praat-based CPPS values, with Sentence 3 tending to have the highest CPPS and Sentence 6 having the lowest.

These results suggest that unvoiced frames are being included in the ADSV-based CPP and Praat-based CPPS calculations. A sentence with many voiceless consonants, especially stop consonants, is likely to contain many unvoiced frames with very low CPP. Including these frames in a calculation of an utterance's average CPP can artificially lower the overall CPP. Praat's CPPS calculation does not include voicing activity detection, but ADSV's CPP algorithm does. Although ADSV does incorporate voicing detection, our results suggest that ADSV's voice activity detector may not filter out all the unvoiced frames in an utterance. As noted in Awan et al. (2010), incomplete or no voicing detection could cause the observed differences between CPP values for sentences with and without unvoiced segments. Improving voice activity detection could reduce the effects of unvoiced frames on a CPP calculation and facilitate comparison between CPPs of different speech tasks.

The use of voicing detection is particularly complicated for voices with aphonia, particularly those with intermittent aphonia. Frames that do not contain voicing due to aphonia, pausing, voiceless consonants, etc., yield low CPP values. If voicing detection is inaccurate or not used, those non-voiced frames will be included in the computation of average CPP, so including aphonic segments will tend to decrease the average CPP. In that case, the low CPP

accurately reflects a noise-like or aphonic voice quality that is clinically meaningful. If very accurate voicing detection is applied, however, only voiced frames will be included in the average CPP calculation, which might be a very small percentage of the speech. The CPP in this case could be high, if the non-aphonic voiced segments are periodic, but would not represent the speech as a whole. Paradoxically, then, accurate voicing detection can actually lead to a higher-than-expected CPP if the voice contains intermittent aphonia.

Ideally, CPP would be calculated only over frames that were intended to be voiced. The CAPE-V fully-voiced sentence ("We were away a year ago") can be used for this purpose. More generally, if frames that were not intended to be voiced (including pauses and voiceless consonants) could be accurately excluded, then utterances with different phonemes could be compared. Aphonic segments, which occur during speech that is intended to have voicing, would be included in the CPP computation and lower the result. Applying this criterion automatically would require very accurate segment-level automatic speech recognition, so it may not currently be realistic. Alternatively, CPP algorithms could require that a certain fraction of the recording be voiced in order to calculate CPP (e.g., if the speech is 95% aphonic, no CPP would be returned). Further research would be needed to identify the appropriate fraction of voicing needed to calculate an accurate CPP.

Notably, the r^2 values for the /a/ vowel regression model were considerably different for the ADSV-based CPP ($r^2 = 0.71$) and Praat-based CPPS ($r^2 = 0.56$). This discrepancy is likely due in part to a single point from one speaker, whose Praat CPPS was relatively high (11.7 dB) but whose ADSV CPP was substantially lower (3.5 dB). This speaker's /a/ vowel received a high overall severity rating of 88 from the trained listeners. The vowel's phonation is characterized by irregular, widely spaced pulses which were perceived, in this case, as significant strain (mean

CAPE-V rating of 90 from trained listeners) and vocal fry. That phonation pattern is likely to be the cause of the discrepancy between the Praat and ADSV CPP values. The algorithms differ in windowing, smoothing, and other parameter settings, and those differences may be particularly sensitive to some property of this specific phonation pattern. Practically, both the Praat CPPS and ADSV CPP values for this speaker were below the clinical cutoff values, so this speaker would have been categorized as having a voice disorder with either program. Still, this finding suggests a need for future work investigating how various CPP computation methods respond to different voice qualities, particularly non-modal ones.

4.6.4 CPP interpretation in context

CPP is just one in a set of objective and subjective measures that have been recommended for use in clinical voice assessment and as such should be considered/interpreted in the context of the other recommended measures which include additional acoustic parameters, and aerodynamic assessment as well as subjective listener ratings, medical exam findings (including laryngeal endoscopic imaging), and patient self-report (Patel et al., 2018). In this study, we identified thresholds below which CPP values were associated with the presence of a voice disorder. However, it is conceivable that some voice disorders may lead to abnormally high CPP (e.g., some manifestations of vocal hyperfunction) which a single threshold would not take into account. Recent work on this topic by Awan & Awan (2020) has indicated that rough voices with a strong subharmonic component may exhibit high CPP values and may benefit from a two-stage analysis method that considers the relative heights of cepstral peaks in high and low quefrency ranges. Future work could additionally determine whether it is necessary or possible to also establish upper boundaries for the clinical application of CPP.

In addition to the computation software and speech task, CPP values may be affected by vowel quality, loudness, or a speaker's sex and age. Awan et al. (2012) found that low vowels (e.g., /a/ and /æ/) tended to have higher CPP values than high vowels (e.g., /i/ and /u/) did. Clinicians should ensure that vowel quality is as similar as possible when comparing CPPs based on sustained vowels. Future work could also identify appropriate CPP cutoff values for sustained vowels other than /a/. Additionally, Awan et al. (2012) found that CPP increases significantly with increases in loudness. These increases are likely due to naturally increased glottal closure and reduced perturbation at higher loudness levels, and do not reflect changes in underlying dysphonia or voice disorder. Additionally, male speakers tended to have higher CPP than female speakers, possibly because of increased loudness in their normal speaking voices. Similarly, Brockmann-Bauser et al. (2019) found that Praat-based CPPS increased significantly with loudness for both patients with and without voice disorders. Clinicians should therefore use caution when comparing CPP values based on speech samples with different loudness levels.

The patients in the MEEI database ranged in age from 13 to 93 years, while the age range of the vocally healthy speakers fell in a smaller bracket (22–59 years). Age is known to often bring voice changes (Mueller, 1997). It may be useful to establish separate normative values for older adults to help distinguish normal aging from disordered voice. This data set did not contain old enough control speakers to establish different norms for older age ranges, but future work could address this question.

4.7 Conclusion

The goal of this study was to employ two frequently used analysis methods and tasks to identify values for cepstral peak prominence (CPP) that can aid clinical voice evaluation, including cutoff thresholds for detecting the presence or absence of a voice disorder and

information about how CPP values relate to auditory-perceptual ratings of overall severity of dysphonia. Results from Experiment 1 suggest that ADSV-based CPP values below 11.46 dB (for sustained /a/ vowels) and below 6.1 dB (for the Rainbow Passage) are strongly indicative of the presence of a voice disorder. Corresponding Praat-based CPPS values were 14.45 dB and 9.3 dB, respectively. Experiment 2 results suggest strong relationships between CPP values and auditory-perceptual ratings of overall severity of dysphonia. Future work could include larger sample sizes to further investigate the relationship between CPP and dysphonia severity, further examination of voicing activity detection in CPP calculation, and investigation into different thresholds for speakers in different age ranges.

Chapter 5. Acoustic speech analysis of patients with decompensated heart failure

5.1 Introduction

5.1.1 Heart failure: introduction and physiology

Heart failure (HF) is the primary diagnosis for over 1 million hospitalizations each year in the US (Gheorghiade & Pang, 2009). Its prevalence in American adults over 65 years of age is as high as 10% (Joseph et al., 2009). In that population, HF is the most common cause of hospitalization (Desai & Stevenson, 2012), causing 20% of hospitalizations (Jessup & Brozena, 2003). Because of its increasing frequency later in life, HF is the most common and most expensive diagnosis group for Medicare patients (Gheorghiade & Pang, 2009).

HF is a complex disease with a multifaceted pathophysiology. Broadly, HF occurs when the heart's ability to pump blood around the body is impaired. The reduction in heart function sets off a chain of compensatory mechanisms to maintain adequate blood supply. The heart becomes larger, with thicker muscle walls. There are also changes in activation of the sympathetic nervous system, which increases heart rate and blood pressure, and of the renin-angiotensin-aldosterone system, which causes retention of sodium and water. These changes initially counteract the decline in heart function by increasing blood supply, and they are adaptive for short-term situations like intense exercise or serious injury. However, in the long term, these alterations to homeostasis can lead to further deterioration of heart function that cannot be compensated for (Kemp & Conte, 2012). Symptoms of HF include dyspnea (shortness of breath), especially with exercise or when lying down; edema or swelling, especially in the lower limbs; and fatigue (Boorsma et al., 2020; Kemp & Conte, 2012). Structural and functional cardiac abnormalities, changes to hormonal regulation and sympathetic nervous system activation, and comorbid conditions can all contribute to the development or exacerbation of HF

(Jessup & Brozena, 2003). Coronary artery disease, hypertension, diabetes, and kidney disease are all common comorbidities (Joseph et al., 2009).

5.1.2 Acute decompensated HF

Patients with HF can remain stable and out of the hospital for long periods of time, but many triggers can also exacerbate HF symptoms. Acute decompensated heart failure (ADHF) refers to "the sudden or gradual onset of the signs or symptoms of heart failure" (Joseph et al., 2009). At least 80% of patients with ADHF have an existing diagnosis of chronic HF (Joseph et al., 2009), and most have at least one "serious" comorbidity (Jessup & Brozena, 2003). The mean age of patients with ADHF is approximately 75 years (Gheorghiade & Pang, 2009).

The primary symptom of ADHF is *congestion*, in which fluid accumulates in blood vessels and, eventually, between tissues of the body (Gheorghiade & Pang, 2009). ADHF involves failures in the compensatory mechanisms described above, which normally maintain homeostasis including appropriate blood supply. These failures lead to fluid retention and increased volume in the circulatory system, which increases pressure on blood vessel walls and filling pressures in the heart. In ADHF, pressure builds high enough that compensatory mechanisms fail, and osmotic pressure pushes fluid out of blood vessels and into the interstitial spaces between tissues (Boorsma et al., 2020). Excess interstitial fluid, or *edema*, can occur in the lungs or systemically throughout the body, especially in the lower limbs (Gheorghiade & Pang, 2009; Boorsma et al., 2020).

Each hospitalization for ADHF increases the risk of subsequent hospitalization and overall mortality (Gheorghiade & Pang, 2009; Joseph et al., 2009). After discharge from a hospitalization for ADHF, 24% of patients are readmitted within 30 days. That rate rises to 50% for readmissions within 6 months. As many as 75% of the 30-day readmissions may be

preventable, and increased follow-up and monitoring in the immediate post-discharge phase appears to be particularly effective in reducing readmissions (Desai & Stevenson, 2012). The course of a patient's recovery during the first post-discharge weeks has been shown to predict readmission and mortality in the 60- to 90-day window after discharge (Gheorghiade & Pang, 2009). Preventing ADHF and associated hospitalization is therefore a major goal of care for patients with chronic HF.

The ejection fraction (EF) describes the fraction of blood pumped out of the left ventricle in one heart contraction (Kemp & Conte, 2012). HF can be divided into two subtypes based on EF, which occur at roughly equal rates (Sharma & Kass, 2014). In heart failure with preserved ejection fraction (HFpEF), the heart retains a normal EF above 50% (Borlaug & Paulus, 2011). In heart failure with reduced ejection fraction (HFrEF), the ejection fraction is below 40% (Miranda et al., 2016). Although the symptoms of HFrEF and HFpEF during decompensation are similar, the two diseases have different underlying physiology. Compared to HFrEF, little is known about the pathophysiology of HFpEF (Fang, 2016). Evidence-based drug regimens for HFrEF, including angiotensin-converting enzyme inhibitors, beta blockers, and aldosterone antagonists, have not been shown to be effective against HFpEF (Miranda et al., 2016). Notably, during diuretic treatment, patients with HFpEF tend to experience large changes in filling pressure in response to small changes in volume, which makes safe and effective diuresis comparatively difficult. In contrast, patients with HFrEF can lose large amounts of fluid rapidly without such an effect on filling pressures (Miranda et al., 2016).

During hospitalizations for ADHF, congestion can be managed with intravenous diuretics. These drugs reduce edema and congestion by promoting fluid loss (Joseph et al., 2009). Diuretics are also used to manage stable HF outside the hospital (Jessup & Brozena,

2003). Despite these therapies, stable HF can develop into ADHF for many reasons. Potential triggers for ADHF include incomplete adherence to medication protocols, untreated hypertension, cardiac arrhythmias, worsening pulmonary or kidney disease, diabetes, and medication side effects (Joseph et al., 2009). However, up to half of decompensation events do not have a discernable cause (Joseph et al., 2009). Intravenous pressure monitoring has revealed that congestion starts to increase weeks before clinical symptoms appear or worsen to the point of requiring hospitalization (Desai & Stevenson, 2012; Gheorghiade & Pang, 2009).

Most hospitalizations for ADHF involve patients with worsening chronic HF, not new HF diagnoses (Joseph et al., 2009). Patients with chronic HF, therefore, are good candidates for at-home monitoring. Monitoring body weight at home can signal impending decompensation, since increased fluid volume from worsened congestion can increase weight. However, in order for monitoring to be useful in averting hospitalization, decompensation must be detected early enough to allow time for clinical intervention. Changes in weight and other clinical symptoms are not always reliable enough to allow this early detection (Desai & Stevenson, 2012). Additionally, weight decrease does not necessarily indicate a reduction in the risk of readmission or mortality (Gheorghiade & Pang, 2009). Therefore, there is currently an unmet need for a reliable, non-invasive, early warning monitoring system for an ADHF episode. The finding that there are pre-symptomatic increases in congestion suggests an immediate use for monitoring techniques that detect sub-clinical changes in physiological status.

5.1.3 Linking voice physiology to HF and ADHF

Voice is produced when air is pushed from the lungs past the vocal folds, which are held in a closed position. The pressure difference above and below the vocal folds forces bursts of air through them, creating a periodic signal. It has previously been noted that biomechanical

properties of the vocal folds affect their readiness to vibrate in response to air pressure from the lungs (Titze, 1988). Titze (1992) defined the phonation threshold pressure (PTP) as "the minimum lung pressure required to initiate phonation." The relevant biomechanical vocal fold properties include the tissues' elasticity, thickness, and viscosity (Sivasankar & Leydon, 2010). Tissue viscosity in particular is closely linked to hydration level in the vocal folds (Finkelhor et al., 1988).

PTP and hydration level have both been linked to phonatory effort. Verdolini et al. (1994) provided healthy adult speakers with either a "hydration treatment" or "dehydration treatment." The hydration treatment consisted of high humidity, increased water intake, and the mucolytic drug guaifenesin. The dehydration treatment involved low humidity, restricted water intake, and a decongestant drug. Verdolini et al. (1994) found that PTP was inversely related to hydration level, with speakers in the dehydration condition producing the highest PTPs. Perceived vocal effort was also inversely related to hydration level. This result was expanded in Verdolini et al. (2002). In that study, healthy adults were given the diuretic Lasix (furosemide), which decreases systemic hydration. This study is particularly relevant in light of the fact that Lasix is also widely used to treat decompensated HF (Joseph et al., 2009). Verdolini et al. (2002) found that, on average, participants who received the diuretic lost 1% of their body weight over the subsequent 12 hours, indicating significant fluid loss and systemic dehydration. In this group, PTP increased by 23% after diuretic treatment. In contrast, control speakers not given the diuretic showed no weight loss and no increase in PTP during the experimental period. While diuretic-induced weight loss began within 1 to 5 hours of the drug dosage, PTP did not increase until 5 to 12 hours post-treatment. This finding suggests that PTP can remain high even after some degree of

fluid loss, so the exact time relationship between fluid status and vocal function is likely to be complex.

In a similar study to Verdolini et al. (1994), that group also found that several acoustic measures of vocal perturbation (jitter, shimmer, and signal-to-noise ratio) improved in response to hydration treatments for speakers with laryngeal nodules or polyps. These speakers also reported reduced phonatory effort after hydration treatments (Verdolini-Marston et al., 1994). Other studies have found that dehydration can reduce vocal quality as reflected by acoustic measures including increased noise-to-harmonics ratio, jitter, shimmer, and s/z ratio. Conversely, increasing hydration has been shown to improve jitter, shimmer, and maximum phonation times (Alves et al., 2019).

The findings linking hydration to PTP and vocal effort indicate a direct relationship between systemic fluid levels, vocal fold physiology, and vocal function. These results suggest that HF-related congestion and edema, if present in the vocal folds, could also affect the resulting voice signal in detectable ways. The amount of laryngeal edema required to measurably change the voice is expected to be small, especially in contrast to the large amount of systemic edema needed to produce an easily detectable increase in body weight. Therefore, voice monitoring may allow patients and clinicians to detect and track HF-related congestion at an earlier stage than weight monitoring does. This sequence of events is summarized in Figure 5.1.

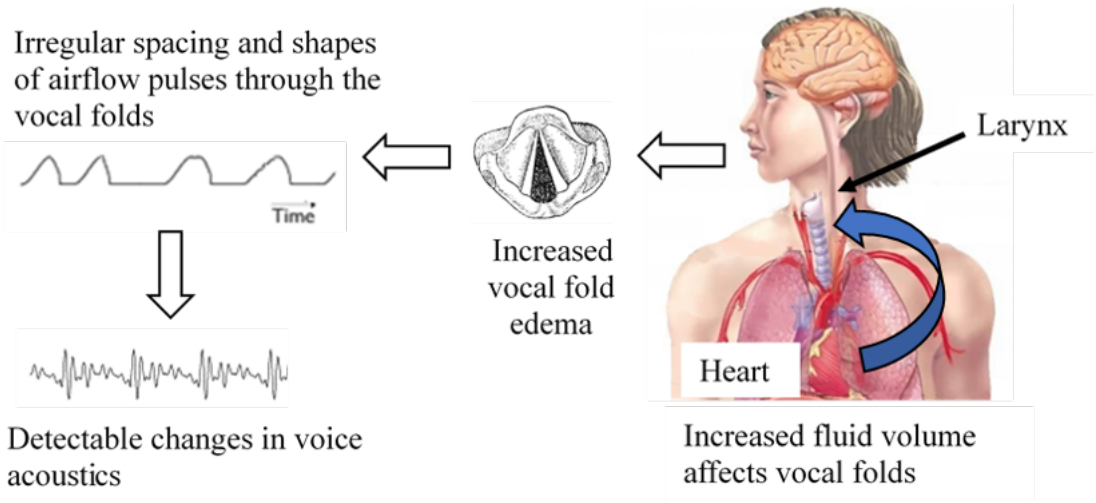

Figure 5.1. Hypothesized effects of HF-related congestion on laryngeal edema, voice airflow, and properties of the voice acoustic signal.

Recently, efforts to combat coronavirus disease 2019 (COVID-19) have correlated with significantly reduced hospitalization rates for HF in many areas, including the southeast US (Hall et al., 2020), London (Bromage et al., 2020), and Denmark (Andersson et al., 2020). These reductions in hospitalization rate have been as large as 50% (Hall et al., 2020). Patients who presented to hospitals with HF during COVID-19-related lockdown and social distancing measures tended to have more severe illness (Bromage et al., 2020), which indicates that they may be delayed in seeking care. However, it has not yet been shown whether these reductions in hospitalization rates are linked with increased mortality (Bromage et al., 2020; Andersson et al., 2020). Limiting contact with others reduces the risk of communicable respiratory infection, which is a frequent cause of HF exacerbation (Bromage et al., 2020). Nevertheless, if patients with HF are less likely to seek care in hospitals, then the need for at-home monitoring tools to predict and prevent ADHF is even greater.

5.1.4 Voice as a biomarker of general health status

Voice and speech characteristics provide an accessible, non-invasive way to monitor physiological changes throughout the body, both for voice disorders and as a biomarker of general health (Low et al., 2020; Van Stan et al., 2017). Ramig (1983) found that overall health status was significantly related to speaking and reading rate, especially in older speakers. More recently, voice biomarkers have been developed to monitor several laryngeal, neurological, and psychological disorders. These include vocal hyperfunction (Mehta et al., 2015), depression (Williamson et al., 2013), Parkinson's disease (Holmes et al., 2000), and amyotrophic lateral sclerosis (Horwitz-Martin et al., 2016). Voice has also been increasingly used as a biomarker for cardiovascular and respiratory disorders, including coronary artery disease (Maor et al., 2018), pulmonary hypertension (Sara et al., 2020), and COVID-19 (Quatieri et al., 2020).

Most relevantly, Maor et al. (2020) developed a vocal biomarker that predicted hospitalization and mortality in patients with HF. The biomarker was developed from a cohort of over 8000 speakers with chronic conditions other than HF, including cancer, lung disease, and diabetes. Acoustic features relating to the cepstrum, pitch, loudness, and perturbation were extracted from 20-second voice samples. A set of 223 more-abstract, high-level features was extracted from the time series of these acoustic features, and the resulting data set was used to train a linear model to predict all-cause mortality. When tested on a separate cohort of 2200 speakers with HF, the biomarker was associated with risk of both hospitalization and mortality. Each increase of one standard deviation in the biomarker represented an increase in risk of 48% for death and 25% for hospitalization during the 20-month follow-up period. This study strongly suggests that the pathophysiology of HF causes clinically relevant changes in the voice, and that further study of the relationship between voice and HF is likely to be fruitful.

The specific mechanisms that govern the relationship between voice and HF status are still unknown. However, in this and related work (Sara et al., 2020), changes in laryngeal perfusion, impaired vagus nerve functioning, physical compression of the left recurrent laryngeal nerve, and increased emotional distress are implicated as possible causes of voice changes from cardiopulmonary disease. Notably, Maor et al. do not hypothesize about or identify specific voice features that are well-correlated with risk of hospitalization or mortality. Because their feature set was derived from a fairly abstract representation of their acoustic voice measures, the features are not likely to be clearly relatable to specific aspects of the voice signal.

5.1.5 Current work

In this chapter, we extend the pilot study from Chapter 2 and apply findings from Chapters 3 and 4. The data set from Chapter 2 has been expanded to include voice samples from 52 patients undergoing inpatient treatment for ADHF. This data was collected from two sites: the Massachusetts General Hospital in Boston, MA, and the University of Vermont Medical Center in Burlington, VT. As in Chapter 2, patients' voices were compared at admission (pre-treatment) and discharge (post-treatment) based on a set of features relating to voice quality and breath support. Based on the findings presented in Chapter 2, two vocal tasks were added to the data collection process. First, each speaker's maximum phonation time (MPT) was assessed. This measure reflects how long a speaker can sustain /a/ vowel phonation on a single breath after inhaling as deeply as possible. MPT is related to both respiratory and laryngeal capacity for speech (Linville et al., 1989). We therefore hypothesized that MPT would increase with ADHF treatment, indicating improvements in both breath support for speech and laryngeal function. Second, each speaker read both the Rainbow Passage (as in the pilot study) and a second passage. The second passage was always new to the speaker; passages were not repeated within a

participant's hospital stay. This new second reading passage was used to control for familiarity effects that may have developed when participants read the Rainbow Passage each day.

In addition to the statistical analysis of first-to-last-day changes, these features were used to train logistic classifiers that distinguished voice samples from admission and discharge. In contrast with previous work on vocal biomarkers for HF (Maor et al., 2020), the trained classifiers were also used to identify which voice features had the greatest predictive power to differentiate pre- and post-treatment voices. In addition, we examined how participants' voices changed from day to day during their hospitalization. We computed the probability of discharge for each voice recording and examined the trajectories of these day-to-day changes.

Finally, this section presents an analysis of the data obtained from a wearable accelerometer sensor in addition to the acoustic microphone. The accelerometer is placed just below the larynx, on the anterior neck, and detects neck-skin vibrations caused by vocal activity (Mehta et al., 2012). Compared to audio microphone recordings, the accelerometer sensor provides a voice-related signal that is not filtered by the vocal tract, is less affected by environmental noise, and does not contain intelligible speech (Zañartu et al., 2009; Mehta et al., 2016). It is therefore well-suited for recording in noisy environments, including the hospital, and for preserving speaker privacy. For these reasons, the accelerometer has been widely used for ambulatory voice monitoring of speakers with voice disorders (Mehta et al., 2015). In this study, speakers' voices were recorded using both the acoustic microphone (MIC) and accelerometer (ACC), and data analysis was carried out in parallel for each sensor type.

As in Chapter 2, we hypothesized that voices from speakers at discharge (after ADHF treatment) would show increased pitch, increased vocal stability, and improved respiratory support compared to their pre-treatment baselines. We further hypothesized that (1) results from

analysis of ACC-based data would be well-correlated with MIC-based results and (2) improved performance in noisy environments would lead to increased classification accuracy from ACC-based models compared to MIC-based models.

5.2 Methods

5.2.1 Participant enrollment

We enrolled fifty-two participants (20 female and 32 male) with acute decompensated HF in Boston, MA and Burlington, VT. Forty participants were enrolled at the Massachusetts General Hospital (MGH) in Boston and twelve were enrolled at the University of Vermont Medical Center (UVM) in Burlington. Forty-nine participants were white, two African-American, and one declined to identify their race. Two participants were Hispanic or Latino, two declined to provide their ethnicity, and the remaining forty-eight participants did not identify as Hispanic or Latino. The participants' median age was 72 years, with a range of 34 – 96 years. Participants' median baseline ejection fraction (EF) prior to this study was 44% (range 12 – 78%). Each day, patients were weighed and recorded a standard speech protocol. Blood levels of N-terminal pro b-type natriuretic peptide (NT-proBNP) were tested at the beginning and end of each patient's hospitalization, since high levels of NT-proBNP are associated with HF (Dao et al., 2001). Patients also used visual analog scales to evaluate their dyspnea symptoms (DVAS) and global symptoms (GVAS) from 0 (worst) to 100 (best).

The median length of participants' hospitalizations was 7.5 (range 2 to 32) days, and the median number of days each participant was recorded was 5 (range 2 to 10) days. Median weight change (last minus first measurement) during hospitalization was −5.5 (range −24.9 to 4.8) kg. The median change in NT-proBNP level was −567 (range −9200 to 57400) pg/ml. The median

changes in participants' ratings of dyspnea and global symptoms were 11.5 (range –43 to 67) percentage points and 16 (range −20 to 68) percentage points respectively.

Table 5.1 lists the participants' demographics at the two enrollment locations.

Table 5.1. Summary of participant demographics at MGH and UVM.

Massachusetts General Hospital			
Sex Male: 24 Female: 16	*Race* White: 37 African-American: 2 Decline to answer: 1	*Ethnicity* Hispanic/Latino: 2 Non-Hispanic/Latino: 36 Decline to answer: 2	*Age (years)* Min: 34 Median: 71 Max: 89
Ejection fraction Min: 12% Median: 53% Max: 78%	*Length of stay (days)* Min: 2 Median: 9 Max: 32	*Recordings* Min: 2 Median: 5 Max: 10	
Weight change (kg) Min: −24.9 Median: −5.75 Max: 4.8	*BNP change (pg/ml)* Min: −4623 Median: −732.5 Max: 2681	*DVAS change (pp)* Min: −43 Median: 15 Max: 67	*GVAS change (pp)* Min: −20 Median: 18 Max: 68
Total: 40			

University of Vermont Medical Center			
Sex Male: 8 Female: 4	*Race* White: 12	*Ethnicity* Non-Hispanic/Latino: 12	*Age (years)* Min: 49 Median: 81 Max: 96
Ejection fraction Min: 12% Median: 40% Max: 62%	*Length of stay (days)* Min: 2 Median: 5 Max: 12	*Recordings* Min: 2 Median: 3.5 Max: 7	
Weight change (kg) Min: −15 Median: −4.05 Max: 1.4	*BNP change (pg/ml)* Min: −9200 Median: −1535 Max: 57400	*DVAS change (pp)* Min: −42 Median: 2.5 Max: 32	*GVAS change (pp)* Min: −15 Median: 15 Max: 36
Total: 12			

HF patients were considered for enrollment based on the inclusion and exclusion criteria listed in Table 5.2. Inclusion criteria called for patients to have pre-existing diagnoses of chronic HF and current diagnoses of ADHF requiring at least 48 hours of diuresis. Patients were also required to be above their "target weight," which refers to their typical body weight without HF-related extra fluid volume. Exclusion criteria included respiratory infection, pulmonary disease, kidney disease, and history of a voice disorder.

Table 5.2. Inclusion and exclusion criteria for study participants.

Inclusion	Exclusion
• 18 years old or older • Prior diagnosis of heart failure with daily diuretic use • Identified within 24 hours after admission • Diagnosis of acute HF as defined by at least one symptom or one clinical sign • Believed to be above target weight • Anticipated need for IV loop diuretic for at least 48 hours	• NT-proBNP level below 400 pg/ml • Inability to perform voice assessment • Not fluent in English • Current respiratory infection • Significant pulmonary disease requiring the use of inhaler, bronchodilators, or steroids • Anticipated need for vasoactive agent or ultrafiltration • Systolic BP below 90 mmHg • Serum creatinine above 3 on admission or hemodialysis required • Hemodynamically significant arrythmia • Complex congenital heart disease • Sepsis • Active smoking within 1 year • History of diagnosed vocal cord pathology or dysfunction

5.2.2 Data collection

At both enrollment sites, patients with acute decompensated heart failure were enrolled in the study as soon after their admission as possible, and at most 24 hours after admission. The voice recording protocol was performed at enrollment and then approximately daily (up to a maximum of 10 recordings) while the patient remained in the hospital. If a patient was expected

to be hospitalized longer than 10 days, recordings were less frequent so that their 10 recordings could span the entire course of their treatment through discharge. When possible, each patient's last recording was done on the day of, or day before, their discharge. For most patients, weight, DVAS, and GVAS were measured on every recording day. On some occasions, DVAS and GVAS were measured only on recording days near admission and discharge. NT-proBNP levels were reported at or near admission and discharge.

Both a microphone (H1 Handy Recorder, Zoom Corporation, Tokyo, Japan) and neck-mounted accelerometer (model BU-27135, Knowles Corp., Itasca, IL) were used to record participants' speech (Mehta et al., 2016). The standard voice recording protocol included a maximum phonation time on /a/, three /a/ vowels of 3 – 5 seconds each, six CAPE-V sentences (Kempster et al., 2009), the Rainbow Passage (Fairbanks, 1960), a second reading passage, and 30 seconds of spontaneous speech. The second reading passage was chosen from a set of 10 such that each participant read a new passage each day. All passages were edited to a Flesch-Kincaid reading level of approximately grade 6. The maximum phonation time and second reading passage tasks were added midway through data collection, so only a subset of participants performed those tasks. The number of participants who performed each task and the number of tokens each task generated are summarized in Table 5.3. Collectively, the 52 speakers performed 3004 vocal tasks over 255 recording sessions.

Table 5.3. For each task, counts of speakers who performed it and the number of tokens generated.

Task	Speakers	Tokens
Maximum phonation	27	116
Sustained vowels (3–5 s)	52	757
CAPE-V sentence	52	1528
Rainbow Passage	52	240
Second reading passage	27	108
Spontaneous speech	52	255

5.2.3 Signal pre-processing

Each session was recorded in full with the microphone and accelerometer, generating one MIC and one ACC recording per participant per day. MIC recordings were sampled at 44100 Hz and ACC recordings at 11025 Hz. Recordings of both types were resampled to 25000 Hz. To remove low-frequency noise artifacts, MIC recordings were high-pass filtered at 70 Hz (Hann band-pass filter, 100 Hz smoothing) using Praat (Boersma & Weenink, 2018). The resulting MIC and ACC signals for each recording session were cross-correlated using MATLAB's xcorr function (*MATLAB*, 2020). The two signals were then aligned using the lag that yielded the maximum cross-correlation, resulting in a single two-channel recording per participant per day. The various voice recording tasks that were performed in each recording session were identified and labelled by hand in Praat. Based on those labels, the MIC and ACC segments corresponding to each task were extracted from the whole-session recording separately. Following this procedure, each recording session ultimately generated two audio files (one MIC and one ACC) per task.

5.2.4 Feature extraction

Frame-by-frame fundamental frequency (F0), creak, and cepstral peak prominence (CPP) contours were computed for each MIC- or ACC-based single-task audio file. The Praat voice report was used to determine additional measures of vocal stability, and speech phrase durations were used as measures of respiratory capacity. Finally, maximum phonation time was calculated as the duration between the first and last voiced frames in each MPT task recording. Here, a "feature" refers to an acoustic measure computed on a specific vocal task. For example, computing F0 SD for a sustained vowel recording yields the "sustained vowel F0 SD" feature. A full list of the 63 features that were used in the subsequent analyses is presented in Table 5.4.

Table 5.4. Combinations of acoustic measure and speaking task that generated each feature.

Acoustic measure	Sustained vowel	CAPE-V sentences	Rainbow Passage	Spontaneous speech	2nd reading passage	Maximum phonation task
F0 mean	x	x	x	x	x	
F0 median	x	x	x	x	x	
F0 SD	x	x	x	x	x	
Maximum phonation time						x
CPP mean	x	x	x	x	x	
CPP median	x	x	x	x	x	
CPP SD	x	x	x	x	x	
Creak % (0.3 threshold)	x	x	x	x	x	
Creak % (0.02 threshold)	x	x	x	x	x	

Table 5.4 (Continued). Combinations of acoustic measure and speaking task that generated each feature.

Acoustic measure	Sustained vowel	CAPE-V sentences	Rainbow Passage	Spontaneous speech	2nd reading passage	Maximum phonation task
Jitter %	x					
Shimmer %	x					
Harmonics-to-noise ratio	x					
Low-high ratio	x					
Speech phrase count			x	x	x	
Speech phrase %			x	x	x	
Speech phrase duration total			x	x	x	
Speech phase duration mean			x	x	x	
Speech phrase duration median			x	x	x	
Speech phrase duration SD			x	x	x	

Fundamental frequency measures

F0 was extracted using Praat's cross-correlation pitch-tracking method (40-ms Hanning window every 1 ms, with pitch ceiling 400 Hz). Based only on frames with non-zero F0, the 5th and 95th percentile F0 values were determined for each task. To minimize distortion from pitch tracking artifacts, only frames with F0 within that 5th – 95th percentile range were used to compute mean F0, median F0, and F0 standard deviation (SD). Additionally, for sustained vowel tasks, the F0 SD computation was further restricted to frames with F0 within 50 Hz of the median F0. Many of Praat's pitch tracking errors for sustained vowel tasks involved doubling or halving the true F0, which substantially raised F0 SD. Using this 100-Hz range allowed the F0 SD calculation to include normal pitch fluctuations but not the inaccurate doubling or halving errors.

Creak measures

For each task, the probability of creak was computed every 10 ms using a method developed by Ishi et al.,(2008) and Kane et al. (2013). This method uses short-term power contours, intra-frame periodicity, and inter-pulse similarity as well as measures indicating secondary or widely-spaced glottal pulses as inputs to an artificial neural network that generates frame-by-frame creak probabilities. Here, frames were considered voiced if they (1) had non-zero F0 based on the Praat pitch track, (2) were separated from another voiced frame by less than 500 ms, or (3) had a creak probability above 0.8. A "breath group" was defined as a group of frames containing no unvoiced regions longer than 500 ms, and all frames within a breath group were treated as voiced for this computation. Adding high-probability creak frames additionally accounted for frames with highly irregular phonation—often on the margins of voiced regions—where Praat's pitch tracking did not identify an underlying F0. For each utterance, the creak

percent was given by (# creaky frames) / (# voiced frames) × 100. The "# creaky frames" quantity was defined as voiced frames with creak probability above either 0.02 (Murton et al., 2019), or above 0.3 (Drugman et al., 2014).

Cepstral peak prominence measures

CPP was calculated as the difference in dB between the magnitude of the highest peak and the noise floor in the power cepstrum, using a 40-ms Hamming window computed every 10 ms (Awan et al., 2010). The location of the CPP was limited to quefrencies between 3.3 and 16.7 ms, which is equivalent to F0 between 60 and 300 Hz. CPP was reported only for frames that met one of the three voicing criteria used for the F0 measures: frames with positive F0, within a breath group, or with creak probability above 0.8. For each task, frames with CPP between the 5th and 95th percentile were used to compute mean, median, and SD of CPP.

Praat (Boersma & Weenink, 2018) was also used to compute smoothed CPP (CPPS) following the procedure in Chapter 4 (Murton et al., 2020). The Praat CPPS values were well-correlated with the MATLAB-based CPP values described above, with correlation coefficients of 0.91 for both CPP mean and CPP median. Unlike the MATLAB CPP computation, the Praat CPPS could not be filtered for frames that met the voicing criteria used for the F0 measures. Therefore, the Praat CPPS values were not included in the subsequent statistical and machine learning modeling. Instead, they were used to provide context for interpreting the MATLAB CPP values relative to the norms identified in Chapter 4.

Other acoustic perturbation measures

For vowel signals only, the Praat voice report was used to compute jitter, shimmer, mean harmonics-to-noise ratio, and the low-high spectral energy ratio. The low-high energy ratio was calculated as the band energy difference in dB between 0–4 kHz and 4–10 kHz (Awan & Roy,

2006). These measures were computed only on the center 500 ms of each vowel to avoid instability from voicing onsets and offsets.

Speech phrase measures

Measures of speech phrase duration were computed for the Rainbow Passage, second reading passage, and spontaneous speech tasks. Within a task, pause durations longer than 500 ms were treated as containing an inhalation (Mehta et al., 2015). A speech phrase was defined as the speech between two successive inhalations. In other words, voiced regions separated from each other by less than 500 ms were considered to be part of the same speech phrase. Measures of breath capacity that were computed following this method included the number of speech phrases, total duration of all the speech phrases, and mean, median, and SD of the speech phrase durations in each utterance.

5.2.5 Statistical analysis

When a participant performed the same task multiple times on a single day, the measures based on that task were averaged to produce a single value per measure per day. For each participant, the change in each measure from the first session to the last was computed. Each patient served as their own control, so the pre-treatment values of each measure were compared to the post-treatment values in paired Bonferroni-corrected t-tests. Cohen's *d* for paired samples was also computed as $\frac{\bar{x}_1 - \bar{x}_2}{\sigma}$ where $\bar{x}_1$ and $\bar{x}_2$ were the means of the first- and last-day values respectively and σ was the standard deviation of the first-to-last day differences. These results were compared to the preliminary statistical results obtained in Murton et al. (2017).

5.2.6 Machine learning: Methods

Correlation matrix

To investigate whether voice features can distinguish stable vs. decompensated heart failure, we used machine learning to classify speech samples into "admission" and "discharge" classes. Patients' first voice recordings were labeled as "admission" and their last voice recordings were labeled as "discharge". We used supervised learning to train a classifier based on a logistic regression model, with the acoustic voice features described above as inputs and the admission/discharge classes as labels.

We visualized relationships between the various acoustic features with a correlation matrix. The absolute magnitude of the correlation coefficient (MATLAB `corrcoef`) was calculated for each pair of features, where a value closer to 1 indicates a stronger correlation. This correlation matrix is helpful for understanding which features tend to carry more similar information about the voice signal. Measures that are highly correlated tend to carry redundant information, whereas uncorrelated measures provide information about different properties of the signal.

Model creation

Each data point for the model input consisted of (1) the admission or discharge label and (2) all the voice features associated with a single day of recording for a single participant. There were thus 104 input points, two for each of the 52 participants. Some features could not be computed for every recording, since occasionally participants did not complete a task or the recording was unusable for some reason. Any data point with one or more features missing was excluded from the model input. The remaining input data was z-normalized so that each feature had a mean of 0 and standard deviation of 1.

The maximum phonation task was added partway through data collection, so only participants 26 and later ever completed it. If the MPT feature were an input to the model, only data from the half of participants who performed it could be used. Therefore the model creation process was run twice: once with all the participants but not the MPT feature (called "MPT–") and once with the MPT feature but only the participants who completed it (the "MPT+" model). After removing missing data, the MPT– model had 91 inputs from 51 participants, and the MPT+ model had 42 inputs from 24 participants.

MATLAB's `fitclinear` function was used to train logistic classifiers using the leave-one-participant-out cross-validation method described below. In these models, the n input features $[x_1, x_2, \dots x_n]$ are combined linearly using learned coefficients $[\beta_0, \beta_1, \beta_2, \dots \beta_n]$.

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

The learning process attempts to find β coefficients that minimize the loss function L for a prediction relative to the labels $[y_1, y_2, \dots y_n]$. These labels indicate whether a data point was produced at admission or discharge, and are represented numerically as +1 or –1 (*MATLAB Fitclinear*).

$$L\big(y, f(x)\big) = \log\big(1 + e^{-yf(x)}\big)$$

Here, y is the known label (+1 or –1) and $f(x)$ is the model output, which ranges from +1 to –1. Intuitively, this loss function penalizes large mismatches between the label y and the model output $f(x)$ for each point. It especially penalizes predictions that have opposite sign to the labels. When y and $f(x)$ have the same sign for a given data point, the model's classification is correct: the prediction correctly aligns with that point's label. In that case, the product $-y * f(x)$

is negative and $e^{-yf(x)} < 1$. If y and $f(x)$ do not have the same sign (i.e., if the classification is incorrect), then $-y * f(x) > 0$ and $e^{-yf(x)} > 1$.

Cross-validation and evaluation

These models were trained and tested using a variant of leave-one-out cross-validation. For each cross-validation fold, the model was trained on data from all but one participant and then tested on the held-out participant's data. This "leave-one-participant-out" method contrasts with the more typical leave-one-out cross-validation, in which one data point would be held out for testing. Each participant yielded two input data points (admission and discharge), so leaving out only one point means that the other point would be included in the training data. The model might then be able to learn some intrinsic characteristic of the participant who appears in the test set. The leave-one-participant-out approach maximizes the amount of training data available in each fold, but also prevents the model from both training and testing on data from the same participant.

Leave-one-participant-out cross-validation sees each participant in the test set exactly once, so a single prediction is made for every data point. The cross-validated model's performance was evaluated by calculating (1) the proportion of recordings correctly classified as first-day or last-day, (2) the area under the receiver operating characteristic (ROC) curve, (3) sensitivity, and (4) specificity. Confusion matrices were also generated to visually inspect model performance.

Regularization

This data set had many features relative to the number of data points, so there was a high risk of overfitting causing poor test accuracy. To reduce overfitting, the model fitting process used L1 (or lasso) regularization (Tibshirani, 1996). Without regularization, the learner finds

feature coefficients that minimize classification loss. In L1 regularization, a term given by the sum of the absolute values of the coefficients is added to the loss function:

$$\min_{\beta} \frac{1}{n}\sum_{i}^{n} Loss(y_i, f(x_i)) + \lambda \sum_{j}^{m} |\beta_j|$$

This regularization term penalizes large coefficient weights. L1 regularization performs a form of feature selection: as λ increases, more coefficient weights are set to zero. Adding this regularization term can improve test accuracy: with fewer features involved in predicting outputs, it is harder to overfit to the training data. If λ is too large, accuracy decreases: many features are discounted from the model and not enough remain to capture the data's patterns. Eventually, for some large enough value of λ, all weights are zero and the model simply predicts the most common class for all inputs. For each model, we used fitclinear hyperparameter optimization to sweep 30 logarithmically spaced λ values in the range [0.001, 1]. The resulting models were used to identify the λ that yielded the lowest cross-validated loss.

Odds ratio

The logistic regression model was also used to identify the voice features that best discriminated between voices at admission and discharge. The odds ratios for each feature coefficient β_i was calculated as e^{β_i}. An odds ratio further from 1 (either above or below) indicates greater predictive power for that variable.

Model comparison

This machine learning process was carried out separately for the MIC-only and ACC-only data sets. The resulting two models were compared based on their percent of correct classifications, area under the ROC curve, sensitivity, and specificity.

5.3 Results

5.3.1 Microphone recordings

T-tests and Cohen's d

Table 5.5 shows the measure, task, admission day mean across all participants, mean and range of change across all participants, paired t-test *p*-value, and paired Cohen's *d* value for each of the 63 features. Features are ordered by the absolute value of the Cohen's *d*, since greater magnitudes of Cohen's *d* indicate larger effect sizes. For F0-related measures (F0 mean, median, and SD), the changes are presented in semitones, where the difference between start and end values is given by $12 * \log_2 \frac{end\ value}{start\ value}$. P-values from the paired t-tests are presented without Bonferroni correction for ease of interpretability. Due to the large number of t-tests, none reach significance after the correction.

Table 5.5. Admission mean, mean and range of change, *p*-value, and Cohen's *d* for each MIC-based feature.

Measure	Task	Admission mean	Mean change	Range of change	p (paired t-test)	Cohen d (paired)
Total phrase duration (s)	Rainbow Passage	32.90	-2.51	[-13.72, 7.34]	1.34E-04	-0.50
Total phrase duration (s)	2nd passage	36.91	-2.59	[-18.85, 12.78]	0.18	-0.50
Phonation time (s)	Max phonation	8.28	2.04	[-3.38, 9.06]	0.0065	0.49
F0 mean (Hz)	Sustained vowel	145	1.02	[-5.61, 10.22]	0.016	0.35
Speech phrase %	2nd passage	0.75	0.006	[-0.18, 0.11]	0.75	-0.34
F0 median (Hz)	Sustained vowel	148	0.89	[-4.64, 11.12]	0.031	0.31
F0 median (Hz)	2nd passage	142	-0.38	[-2.17, 1.56]	0.27	-0.31
Speech phrase duration SD	2nd passage	1.91	-0.53	[-4.07, 0.98]	0.16	-0.29
F0 mean (Hz)	2nd passage	144	-0.45	[-4.91, 1.89]	0.33	-0.27
CPP SD (dB)	Sentences	3.57	-0.21	[-2.13, 2.26]	0.058	-0.27
Speech phrase duration mean (s)	Spontaneous	2.04	0.48	[-2.83, 6.43]	0.076	0.25
CPP mean (dB)	Sentences	18.22	-0.31	[-3.34, 2.83]	0.082	-0.25
CPP SD (dB)	2nd passage	3.46	-0.087	[-1.53, 1.72]	0.66	-0.24
CPP mean (dB)	2nd passage	17.73	-0.10	[-2.26, 2.49]	0.74	-0.24
CPP median (dB)	2nd passage	17.08	-0.041	[-1.73, 2.23]	0.88	-0.24
# of speech phrases	2nd passage	14.87	-0.20	[-7, 9]	0.89	0.23
# of speech phrases	Spontaneous	15.16	-2.66	[-32, 27]	0.059	-0.22

Table 5.5 (Continued). Admission mean, mean and range of change, *p*-value, and Cohen's *d* for each MIC-based feature.

Measure	Task	Admission mean	Mean change	Range of change	p (paired t-test)	Cohen d (paired)
CPP median (dB)	Sentences	17.71	-0.29	[-3.19, 2.57]	0.11	-0.22
Creak % (0.02 threshold)	Sentences	0.27	0.030	[-0.47, 0.39]	0.12	0.22
Phrase duration median (s)	Spontaneous	1.71	0.46	[-4, 7.27]	0.11	0.22
F0 SD (Hz)	Sustained vowel	5.22	-0.88	[-11.76, 9.52]	0.091	-0.21
CPP median (dB)	Sustained vowel	21.52	0.48	[-6.79, 9.2]	0.29	0.21
# of speech phrases	Rainbow Passage	13.83	-0.81	[-19, 35]	0.43	-0.19
Speech phrase %	Spontaneous	0.63	0.029	[-0.21, 0.35]	0.21	0.18
Creak % (0.3 threshold)	2nd passage	0.089	-0.007	[-0.21, 0.13]	0.76	-0.18
Creak % (0.02 threshold)	Spontaneous	0.25	0.025	[-0.25, 0.32]	0.16	0.18
CPP SD (dB)	Sustained vowel	2.86	-0.17	[-1.85, 1.3]	0.093	-0.17
Speech phrase duration SD	Spontaneous	1.46	0.23	[-2.76, 5.73]	0.24	0.17
Phrase duration median (s)	2nd passage	2.27	0.42	[-0.84, 3.36]	0.20	0.17
CPP mean (dB)	Sustained vowel	21.38	0.35	[-6.84, 8.95]	0.42	0.17
F0 SD (Hz)	Spontaneous	22.98	-1.78	[-32.82, 22.52]	0.26	-0.15
Low-high ratio	Sustained vowel	-27.17	-0.36	[-12.37, 10.61]	0.64	-0.14
F0 SD (Hz)	Sentences	20.98	1.25	[-27.09, 23.42]	0.31	0.14
CPP SD (dB)	Spontaneous	3.56	-0.13	[-1.83, 1.97]	0.22	-0.14

Table 5.5 (Continued). Admission mean, mean and range of change, *p*-value, and Cohen's *d* for each MIC-based feature.

Measure	Task	Admission mean	Mean change	Range of change	p (paired t-test)	Cohen d (paired)
CPP SD (dB)	Rainbow Passage	3.59	-0.11	[-1.49, 1.58]	0.26	-0.14
F0 mean (Hz)	Sentences	150	0.22	[-4.03, 4.01]	0.37	0.13
F0 SD (Hz)	2nd passage	21.98	-0.58	[-39.55, 14.39]	0.86	-0.12
Creak % (0.3 threshold)	Spontaneous	0.068	0.010	[-0.19, 0.27]	0.33	0.12
Harmonics-to-noise ratio	Sustained vowel	13.93	0.48	[-12.62, 12.99]	0.55	0.11
Phrase duration median (s)	Rainbow Passage	3.24	-0.43	[-16.61, 2.14]	0.32	-0.10
Speech phrase %	Rainbow Passage	0.75	-0.002	[-0.66, 0.26]	0.90	0.10
Creak % (0.02 threshold)	Rainbow Passage	0.23	0.010	[-0.37, 0.4]	0.61	0.10
Creak % (0.3 threshold)	Sentences	0.068	0.009	[-0.45, 0.32]	0.51	0.092
Speech phrase duration mean (s)	2nd passage	2.66	0.24	[-1.62, 3.7]	0.53	0.083
F0 median (Hz)	Sentences	149	0.16	[-3.83, 4.66]	0.55	0.078
F0 median (Hz)	Spontaneous	145	-0.19	[-6.13, 4.98]	0.52	-0.077
Creak % (0.3 threshold)	Rainbow Passage	0.062	0.006	[-0.29, 0.33]	0.65	0.077
Speech phrase duration mean (s)	Rainbow Passage	3.44	-0.35	[-16.43, 3.6]	0.43	-0.072
CPP median (dB)	Spontaneous	16.77	0.053	[-2.08, 4.07]	0.72	0.065
Jitter (%)	Sustained vowel	0.012	-0.001	[-0.04, 0.04]	0.67	-0.065
F0 mean (Hz)	Rainbow Passage	146	0.075	[-4.77, 3.4]	0.76	0.063

Table 5.5 (Continued). Admission mean, mean and range of change, *p*-value, and Cohen's *d* for each MIC-based feature.

Measure	Task	Admission mean	Mean change	Range of change	p (paired t-test)	Cohen d (paired)
Creak % (0.02 threshold)	2nd passage	0.25	0.006	[-0.31, 0.23]	0.88	0.060
Creak % (0.3 threshold)	Sustained vowel	0.069	-0.001	[-0.26, 0.36]	0.94	0.050
CPP mean (dB)	Rainbow Passage	17.84	-0.055	[-2.37, 4.3]	0.74	-0.042
Speech phrase duration SD (s)	Rainbow Passage	1.87	-0.15	[-14.2, 5.07]	0.66	-0.035
F0 mean (Hz)	Spontaneous	147	-0.10	[-4.62, 4.74]	0.71	-0.031
Shimmer (%)	Sustained vowel	0.083	3.4E-04	[-0.1, 0.1]	0.95	-0.030
Creak % (0.02 threshold)	Sustained vowel	0.19	-0.013	[-0.59, 0.46]	0.67	-0.030
F0 median (Hz)	Rainbow Passage	144	0.017	[-3.16, 3.56]	0.94	0.028
CPP median (dB)	Rainbow Passage	17.16	0.034	[-2.78, 5.85]	0.86	0.022
CPP mean (dB)	Spontaneous	17.61	-0.044	[-2.43, 3.38]	0.77	-0.017
F0 SD (Hz)	Rainbow Passage	19.98	-0.41	[-36.57, 24.95]	0.79	-0.017
Total phrase duration (s)	Spontaneous	25.54	-1.12	[-34.55, 45.01]	0.61	-0.009

Feature correlation

Figure 5.2 visualizes the absolute magnitude of the correlation coefficient for each pair of MIC-based features. A value closer to 1 indicates a stronger correlation between a feature pair. The measures are labeled from left to right on the x-axis and are presented in the same order (top to bottom) on the y-axis.

Figure 5.2. MIC-based correlation coefficients for each pair of features.

F0 mean: vowel
F0 mean: sentences
F0 mean: RP
F0 mean: free
F0 mean: 2nd'
F0 median: vowel
F0 median: sentences
F0 median: RP
F0 median: free
F0 median: 2nd
F0 SD: vowel
F0 SD: sentences
F0 SD: RP
F0 SD: free
F0 SD: 2nd
MPT: max phonation task
CPP mean: vowel
CPP mean: sentences
CPP mean: RP
CPP mean: free
CPP mean: 2nd
CPP median: vowel
CPP median: sentences
CPP median: RP
CPP median: free
CPP median: 2nd
CPP SD: vowel
CPP SD: sentences
CPP SD: RP
CPP SD: free
CPP SD: 2nd
Creak % (0.3): vowel
Creak % (0.3): sentences
Creak % (0.3): RP
Creak % (0.3): free
Creak % (0.3): 2nd
Creak % (0.02): vowel
Creak % (0.02): sentences
Creak % (0.02): sentences
Creak % (0.02): free
Creak % (0.02): 2nd
Jitter %: vowel
Shimmer %: vowel
HNR: vowel
LHR: vowel
Phrase count: RP
Phrase count: free
Phrase count: 2nd
Phrase time: RP
Phrase time: free
Phrase time: 2nd
Phrase %: RP
Phrase %: free
Phrase %: 2nd
Phrase duration mean: RP
Phrase duration mean: free
Phrase duration mean: 2nd
Phrase duration median: RP
Phrase duration median: free
Phrase duration median: 2nd
Phrase duration SD: RP
Phrase duration SD: free
Phrase duration SD: 2nd
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1

Logistic classification: no regularization

Figure 5.3 shows confusion matrices for the MPT– and MPT+ models trained with no regularization. Figure 5.4 additionally shows a confusion matrix for a non-regularized model trained only on the MPT feature. Quantities in blue-shaded squares indicate the number of data points correctly classified as admission or discharge. Grey-shaded squares indicate incorrectly classified points.

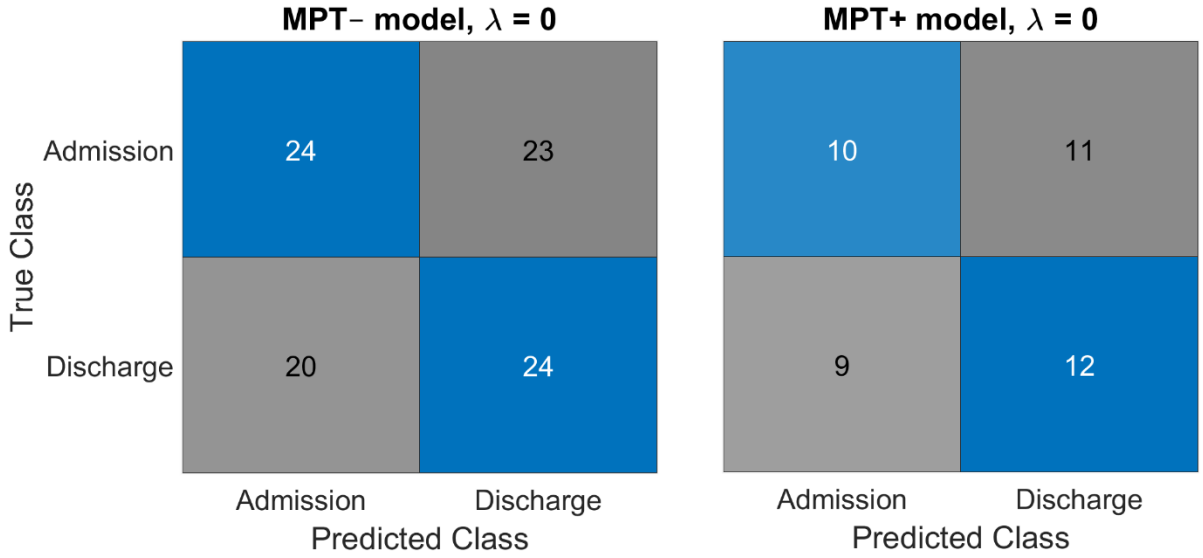

Figure 5.3. Confusion matrices for the MIC-based MPT+ and MPT– models without regularization.

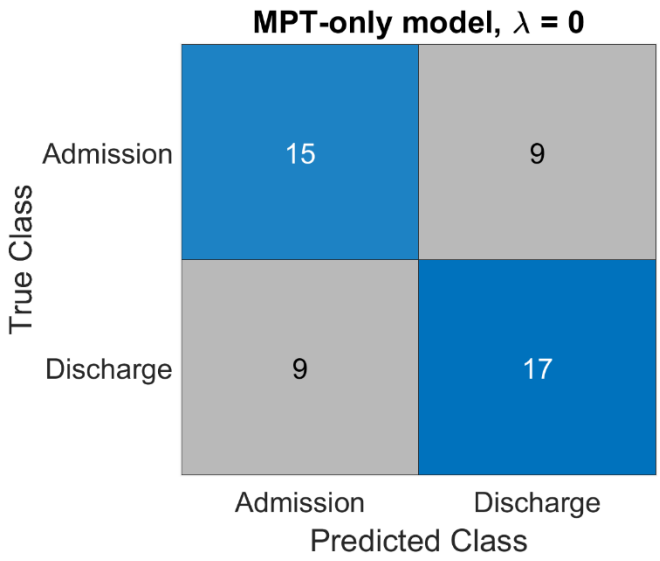

Figure 5.4. Confusion matrix for the MIC-based non-regularized MPT-only model.

Logistic classification: L1 regularization

Thirty λ values between 0.001 and 0.3 were tested to identify the regularization parameter that would yield the lowest regularized loss. That process is visualized for the MPT– and MPT+ models in Figure 5.5 and Figure 5.6 respectively. L1 regularization sets increasingly many coefficients to zero as λ increases. The stable area in the rightmost region of both plots, where the objective function no longer varies with λ, indicates that λ was high enough to set all coefficients to 0. Therefore, no higher λ values needed to be inspected.

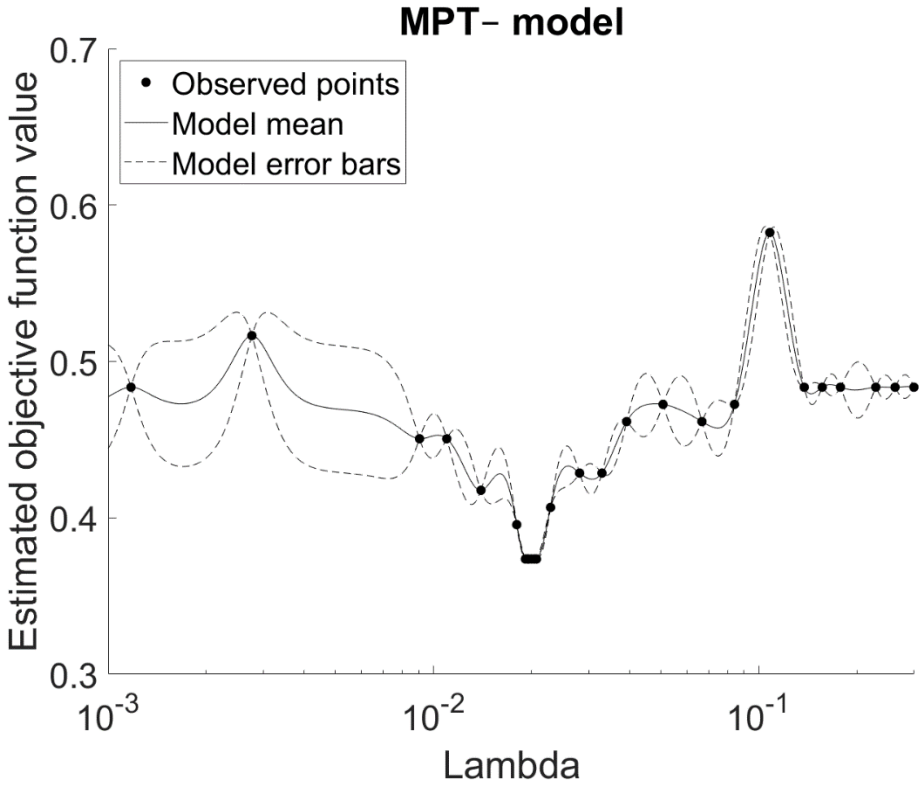

Figure 5.5. Estimated objective function value for each lambda value that was tested during optimization of the MIC-based MPT– model. Lower objective function values indicate a better fit.

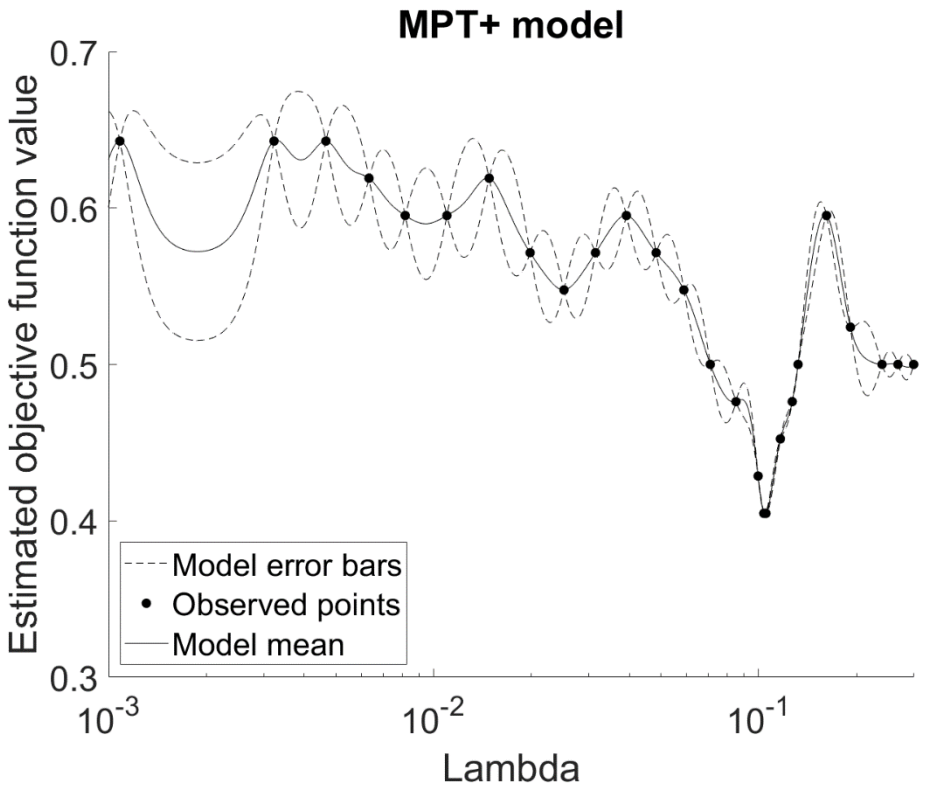

Figure 5.6. Estimated objective function values for each lambda value that was tested during optimization of the MIC-based MPT+ model. Lower objective function values indicate a better fit.

Figure 5.7 shows confusion matrices for the MPT– and MPT+ models trained with no regularization. As in Figure 5.3 and Figure 5.4, quantities in blue-shaded squares indicate the number of data points correctly classified as admission or discharge. Grey-shaded squares indicate incorrectly classified points.

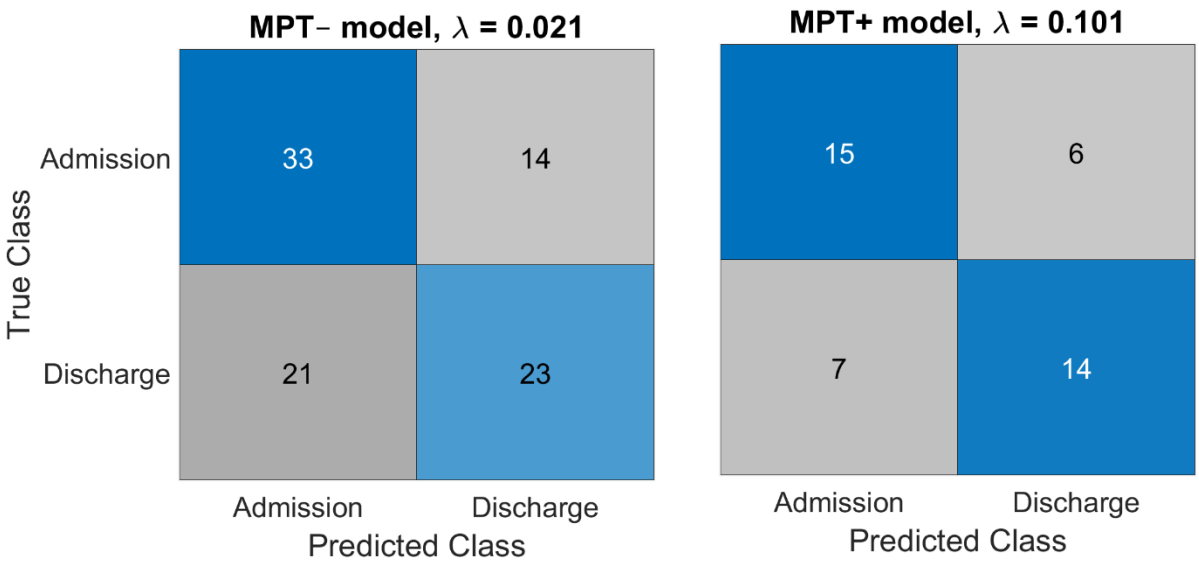

Figure 5.7. Confusion matrices and optimized lambda values for the MIC-based regularized MPT– and MPT+ models.

Model performance

The models' performances were evaluated based on their accuracy, area under the receiver operating characteristic curve (AUC), true admission rate (TAR), true discharge rate (TDR), admission predictive value (APV), and discharge predictive value (DPV). "Admission" and "discharge" are used in place of "positive" and "negative" in the names of these metrics to increase clarity. As used here, the TAR and TDR metrics correspond to sensitivity and specificity. Performance metrics for the five models are summarized in Table 5.6.

Table 5.6. Performance metrics for each MIC-based logistic classifier model.

λ	Model	Accuracy	AUC	TAR	TDR	APV	DPV
none	MPT–	0.53	0.49	0.51	0.55	0.55	0.51
none	MPT+	0.52	0.46	0.48	0.57	0.53	0.52
none	MPT-only	0.64	0.61	0.63	0.65	0.63	0.65
L1	MPT–	0.62	0.66	0.70	0.52	0.61	0.62
L1	MPT+	0.69	0.65	0.71	0.67	0.68	0.70

Coefficient weights and odds ratios

In the cross-validation process, many models were created and their performances were averaged. However, it is also desirable to identify a single set of feature weights. These weights can provide information about each feature's predictive power. To identify a single set of feature weights, we trained new MPT– and MPT+ models without cross-validation, using all the training data. These models used L1 regularization with the optimized λ parameters shown in Figure 5.7, so only a subset of the input features ended up with non-zero β weights. The resulting β feature weights and odds ratios (given by e^{β}) are listed in Table 5.7 and Table 5.8.

The models' training data was z-normalized to have mean of 0 and SD of 1. Therefore, the output β weights are also in these normalized units. If a feature's SD in the training data was σ, then its normalized values are related to its original values by $\frac{1}{\sigma}$. A normalized feature weight β can be then converted to its original units by dividing by σ. Both the normalized and original-unit β weights and odds ratios are listed in these tables.

Because they are all on the same scale, the normalized β weights and odds ratios are interpretable relative to each other. The normalized odds ratios for each feature indicate the change in the probability of discharge when that feature increases by 1 standard deviation. A larger absolute value for a β weight, or an odds ratio farther from 1, indicates that a feature makes a greater contribution to the classification decision. In contrast, the original-unit β weights and odds ratios are interpretable in terms of the features themselves. In this case, the odds ratio for each feature represents the change in the probability of discharge when that feature increases by 1 unit.

Table 5.7. Feature weights and odds ratios for the MIC-based regularized MPT– model.

Feature	Normalized units		Original units	
	β weight	**Odds ratio**	**β weight**	**Odds ratio**
Creak %: 0.3 threshold (sentences)	0.65	1.92	8.82	6.7×10^3
CPP SD (sentences)	–0.50	0.61	–0.67	0.51
Total phrase duration (rainbow)	–0.35	0.70	–0.065	0.94
F0 mean (vowel)	0.34	1.40	0.008	1.008
CPP median (spontaneous)	0.31	1.37	0.32	1.38
Phrase duration median (spontaneous)	0.24	1.27	0.16	1.18
F0 SD (spontaneous)	–0.24	0.79	–0.021	0.98
Phrase duration mean (spontaneous)	0.22	1.25	0.14	1.15
Phrase % (rainbow)	0.16	1.18	1.30	3.67
F0 SD (vowel)	–0.16	0.85	–0.047	0.96
Phrase duration SD (rainbow)	–0.14	0.87	–0.079	0.92
CPP SD (rainbow)	–0.086	0.92	–0.11	0.89
CPP median (vowel)	0.074	1.08	0.024	1.02

Table 5.8. Feature weights and odds ratios for the MIC-based regularized MPT+ model.

Feature	Normalized units		Original units	
	β	**Odds ratio**	**β**	**Odds ratio**
Max phonation time (max phonation)	0.30	1.34	0.055	1.06
F0 SD (sentence)	0.22	1.25	0.022	1.02
Phrase duration median (spontaneous)	0.11	1.11	0.063	1.06

Within-stay trajectories

The MPT– and MPT+ models, which were trained on all of the participants' data, were also used to generate the probability of discharge for every daily voice recording rather than simply the ones from admission and discharge. This process generated a series of day-by-day discharge probabilities for each participant. These trajectories are shown in Figure 5.8 and Figure 5.9. In Figure 5.8, the 52 participants are ordered from the greatest to least first-to-last-day change in discharge probability. Within each plot, a participant's voice recordings are ordered sequentially from admission to discharge. The equivalent plots for the 27 participants who were included in the MPT+ model are shown in Figure 5.9. If the probability of discharge for a point was below 0.5, the point was classified as "admission" and plotted in red. Black points indicate recordings for which the discharge probability was above 0.5.

For most participants, the probability of discharge increased over the course of the hospital stay. At the same time, most participants showed day-to-day fluctuations outside of that overall trajectory. Trajectories for the MPT+ model were generally flatter than those for the MPT– model.

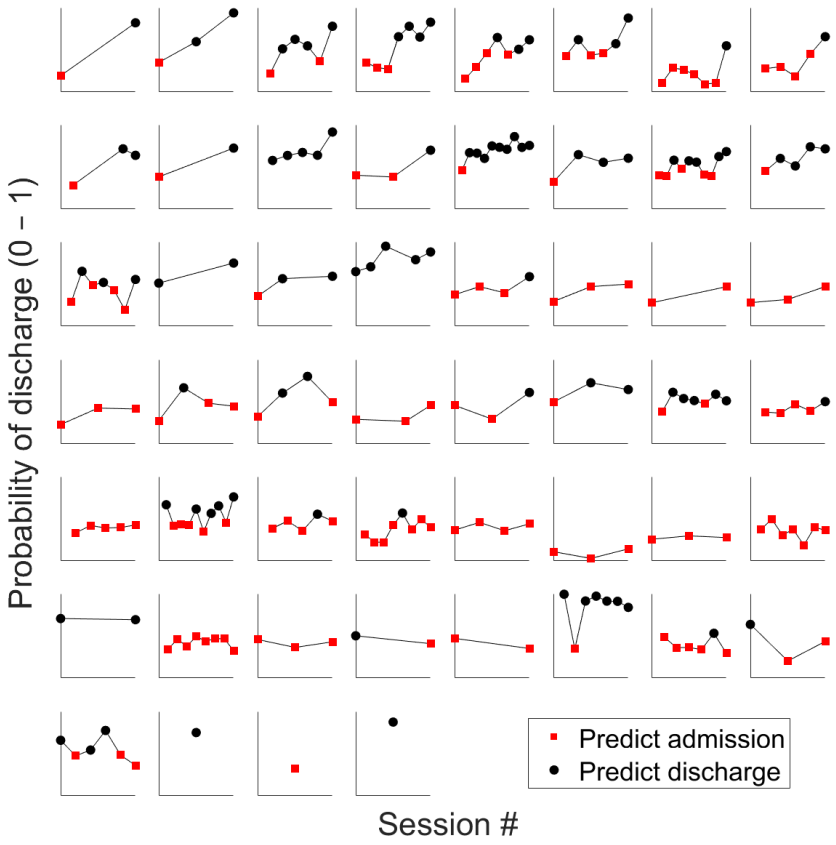

Figure 5.8. Day-to-day discharge probabilities for each speaker based on the MIC-based MPT– model. Red squares indicate admission predictions (discharge probability < 0.5) and black dots indicate discharge predictions.

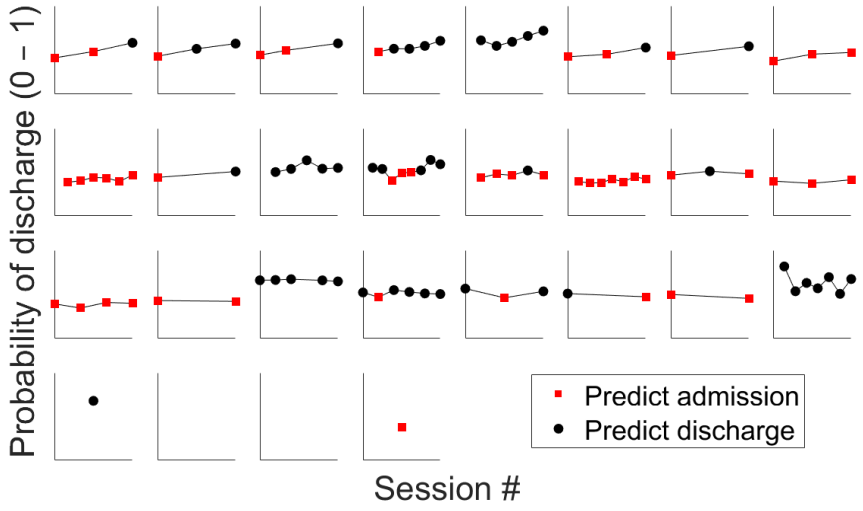

Figure 5.9 Day-to-day discharge probabilities for each speaker based on the MIC-based MPT+ model. Red squares indicate admission predictions (discharge probability < 0.5) and black dots indicate discharge predictions.

5.3.2 Accelerometer recordings

T-tests and Cohen's d

Table 5.9 shows the measure, task, admission day mean across all participants, mean and range of change across all participants, paired t-test *p*-value, and paired Cohen's *d* value for each of the 63 features. Features are ordered by the absolute value of the Cohen's *d*, since greater magnitudes of Cohen's *d* indicate larger effect sizes. For F0-related measures (F0 mean, median, and SD), the changes are presented in semitones, where the difference between start and end values is given by $12 * \log_2 \frac{end\ value}{start\ value}$. P-values from the paired t-tests are presented without Bonferroni correction for ease of interpretability. Due to the large number of t-tests, none reach significance after the correction.

Table 5.9. Admission mean, mean and range of change, *p*-value, and Cohhen's *d* for each ACC-based feature.

Measure	Task	Admission mean	Mean change	Range of change	p (paired t-test)	Cohen d (paired)
Phonation time (s)	Max phonation	8.27	1.85	[-4.42, 9.12]	0.019	0.45
# of speech phrases	Rainbow passage	14.59	-1.62	[-17, 13]	0.054	-0.41
F0 mean (Hz)	Sustained vowel	147	1.06	[-3.67, 10.22]	0.012	0.35
Total phrase duration (s)	Rainbow passage	32.30	-1.50	[-26.82, 17.17]	0.082	-0.33
Creak % (0.3 threshold)	2nd passage	0.10	-0.026	[-0.2, 0.15]	0.27	-0.32
CPP mean (dB)	Sentences	18.52	-0.41	[-4.11, 4.47]	0.085	-0.31
F0 median (Hz)	Sustained vowel	149	0.99	[-3.67, 11.12]	0.025	0.30
CPP SD (dB)	Sentences	3.42	-0.25	[-2.52, 2.83]	0.082	-0.30
CPP SD (dB)	Sustained vowel	2.50	-0.24	[-2.13, 0.99]	8.29E-03	-0.29
Creak % (0.02 threshold)	2nd passage	0.31	-0.075	[-0.56, 0.25]	0.17	-0.29
CPP median (dB)	Sentences	18.21	-0.40	[-4.81, 5.64]	0.12	-0.28
Harmonics-to-noise ratio	Sustained vowel	21.30	2.55	[-15.02, 22.87]	0.025	0.22
Total phrase duration (s)	2nd passage	35.43	-1.75	[-29.51, 23.72]	0.59	-0.22
CPP SD (dB)	Spontaneous	3.35	-0.15	[-1.99, 3.39]	0.28	-0.21
CPP SD (dB)	Rainbow passage	3.43	-0.14	[-2.72, 3.55]	0.34	-0.19
Speech phrase %	Spontaneous	0.61	0.040	[-0.37, 0.52]	0.14	0.19
F0 SD (Hz)	Sentences	20.52	1.00	[-18.75, 19.93]	0.34	0.18

Table 5.9 (Continued). Admission mean, mean and range of change, *p*-value, and Cohen's *d* for each ACC-based feature.

Measure	Task	Admission mean	Mean change	Range of change	p (paired t-test)	Cohen d (paired)
CPP SD (dB)	2nd passage	3.08	0.31	[-0.86, 3.73]	0.31	0.17
Phrase duration mean (s)	Spontaneous	2.02	0.32	[-2.83, 6.43]	0.15	0.17
F0 mean (Hz)	Sentences	149	0.27	[-4.02, 4.01]	0.29	0.16
CPP mean (dB)	Rainbow passage	18.13	-0.17	[-4.79, 4.95]	0.49	-0.16
CPP median (dB)	Rainbow passage	17.75	-0.22	[-6.62, 5.39]	0.45	-0.16
Phrase duration mean (s)	2nd passage	2.58	0.35	[-2.11, 4]	0.40	0.16
Phrase duration median (s)	2nd passage	2.28	0.68	[-2.19, 7.38]	0.28	0.15
# of speech phrases	2nd passage	15.47	-0.71	[-6, 6]	0.47	0.14
# of speech phrases	Spontaneous	14.02	-1.62	[-31, 27]	0.28	-0.13
F0 mean (Hz)	2nd passage	140	0.025	[-2.5, 3.78]	0.96	0.12
Speech phrase %	Rainbow passage	0.73	0.014	[-0.43, 0.26]	0.40	0.12
Phrase duration median (s)	Spontaneous	1.75	0.26	[-4, 7.27]	0.32	0.12
F0 median (Hz)	Sentences	148	0.26	[-3.57, 4.23]	0.32	0.11
F0 median (Hz)	Spontaneous	144	-0.26	[-6.13, 3.36]	0.40	-0.11
F0 SD (Hz)	Sustained vowel	5.01	-0.74	[-11.81, 9.52]	0.15	-0.11
F0 SD (Hz)	2nd passage	20.09	1.15	[-10.45, 12.03]	0.49	0.11
CPP mean (dB)	Spontaneous	17.75	-0.074	[-2.85, 5.2]	0.73	-0.10

Table 5.9 (Continued). Admission mean, mean and range of change, *p*-value, and Cohen's *d* for each ACC-based feature.

Measure	Task	Admission mean	Mean change	Range of change	p (paired t-test)	Cohen d (paired)
Jitter (%)	Sustained vowel	0.010	-2.63E-	[-0.06, 0.04]	0.20	-0.10
Speech phrase %	2nd passage	0.72	-1.81E-	[-0.59, 0.3]	0.97	-0.10
Creak % (0.3 threshold)	Rainbow passage	0.074	-7.04E-	[-0.45, 0.2]	0.63	0.088
Creak % (0.02 threshold)	Spontaneous	0.27	-5.20E-	[-0.58, 0.36]	0.84	0.087
CPP mean (dB)	2nd passage	17.62	0.60	[-1.52, 5.02]	0.17	0.085
F0 mean (Hz)	Rainbow passage	144	0.22	[-3.47, 5.43]	0.41	0.084
Creak % (0.02 threshold)	Sentences	0.29	-6.33E-	[-0.37, 0.54]	0.78	0.075
F0 mean (Hz)	Spontaneous	146	-0.21	[-4.62, 3.37]	0.47	-0.071
CPP median (dB)	Spontaneous	17.21	-0.018	[-4.36, 5.07]	0.94	-0.059
F0 SD (Hz)	Rainbow passage	18.31	0.53	[-18.76, 24.95]	0.58	0.058
Creak % (0.3 threshold)	Sustained vowel	0.080	1.11E-03	[-0.38, 0.8]	0.97	0.054
Speech phrase duration SD	Rainbow passage	1.48	0.053	[-6.1, 3.61]	0.77	0.052
Low-high ratio	Sustained vowel	-35.78	-0.97	[-19.99, 11.38]	0.40	-0.047
F0 SD (Hz)	Spontaneous	21.11	-1.06	[-22.18, 22.52]	0.40	-0.045
Shimmer (%)	Sustained vowel	0.053	-5.07E-	[-0.1, 0.07]	0.33	-0.041
Phrase duration median (s)	Rainbow passage	2.71	-0.11	[-10.14, 3.8]	0.71	-0.040
Speech phrase duration SD (s)	Spontaneous	1.31	0.062	[-2.39, 2.76]	0.62	0.040

Table 5.9 (Continued). Admission mean, mean and range of change, *p*-value, and Cohen's *d* for each ACC-based feature.

Measure	Task	Admission mean	Mean change	Range of change	p (paired t-test)	Cohen d (paired)
CPP mean (dB)	Sustained vowel	20.23	-0.17	[-9.36, 6.06]	0.70	-0.036
CPP median (dB)	2nd passage	17.21	0.65	[-1.61, 4.32]	0.12	0.036
Phrase duration mean (s)	Rainbow passage	2.81	0.044	[-9.91, 4.14]	0.89	0.033
Creak % (0.02 threshold)	Sustained vowel	0.19	-0.026	[-0.69, 0.73]	0.40	-0.019
CPP median (dB)	Sustained vowel	20.36	-0.15	[-9.6, 6.39]	0.75	-0.018
Speech phrase duration SD (s)	2nd passage	1.68	-0.093	[-4.07, 1.89]	0.80	-0.016
Creak % (0.3 threshold)	Sentences	0.079	-0.011	[-0.31, 0.23]	0.40	-0.013
Creak % (0.02 threshold)	Rainbow passage	0.27	-0.030	[-0.56, 0.28]	0.20	-0.010
F0 median (Hz)	Rainbow passage	144	0.07	[-3.34, 4.62]	0.77	-0.007
Total phrase duration (s)	Spontaneous	24.83	-0.58	[-37.29, 38.9]	0.80	-0.006
Creak % (0.3 threshold)	Spontaneous	0.082	-3.96E-	[-0.55, 0.27]	0.82	0.005
F0 median (Hz)	2nd passage	140	-0.19	[-2.26, 3.61]	0.70	-0.003

Feature correlation

Figure 5.10 visualizes the absolute magnitude of the correlation coefficient for each pair of accelerometer-based features. As in Figure 5.2, a value closer to 1 indicates a stronger correlation between a feature pair. The measures are labeled from left to right on the x-axis and are presented in the same order (top to bottom) on the y-axis.

Figure 5.10. ACC-based correlation coefficients for each pair of features.

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
F0 mean: vowel
F0 mean: sentences
F0 mean: RP
F0 mean: free
F0 mean: 2nd'
F0 median: vowel
F0 median: sentences
F0 median: RP
F0 median: free
F0 median: 2nd
F0 SD: vowel
F0 SD: sentences
F0 SD: RP
F0 SD: free
F0 SD: 2nd
MPT: max phonation task
CPP mean: vowel
CPP mean: sentences
CPP mean: RP
CPP mean: free
CPP mean: 2nd
CPP median: vowel
CPP median: sentences
CPP median: RP
CPP median: free
CPP median: 2nd
CPP SD: vowel
CPP SD: sentences
CPP SD: RP
CPP SD: free
CPP SD: 2nd
Creak % (0.3): vowel
Creak % (0.3): sentences
Creak % (0.3): RP
Creak % (0.3): free
Creak % (0.3): 2nd
Creak % (0.02): vowel
Creak % (0.02): sentences
Creak % (0.02): sentences
Creak % (0.02): free
Creak % (0.02): 2nd
Jitter %: vowel
Shimmer %: vowel
HNR: vowel
LHR: vowel
Phrase count: RP
Phrase count: free
Phrase count: 2nd
Phrase time: RP
Phrase time: free
Phrase time: 2nd
Phrase %: RP
Phrase %: free
Phrase %: 2nd
Phrase duration mean: RP
Phrase duration mean: free
Phrase duration mean: 2nd
Phrase duration median: RP
Phrase duration median: free
Phrase duration median: 2nd
Phrase duration SD: RP
Phrase duration SD: free
Phrase duration SD: 2nd

Logistic classification: no regularization

Figure 5.11 shows confusion matrices for the MPT– and MPT+ models trained with no regularization. Figure 5.12 additionally shows a confusion matrix for a non-regularized model trained only on the MPT feature. Quantities in blue-shaded squares indicate the number of correctly classified data points. Grey-shaded squares indicate incorrectly classified points.

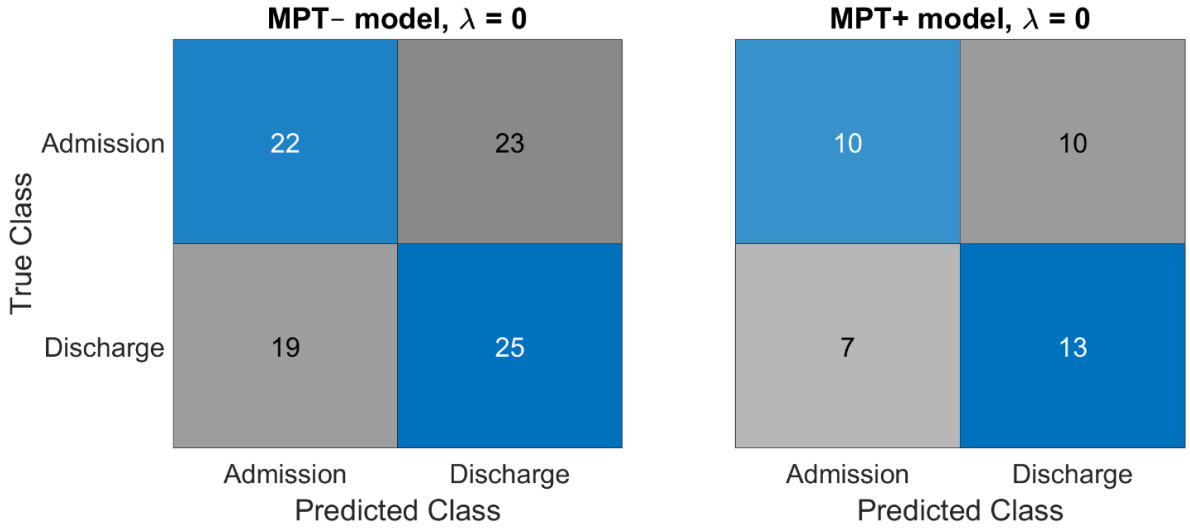

Figure 5.11. Confusion matrices for the ACC-based MPT– and MPT+ models without regularization.

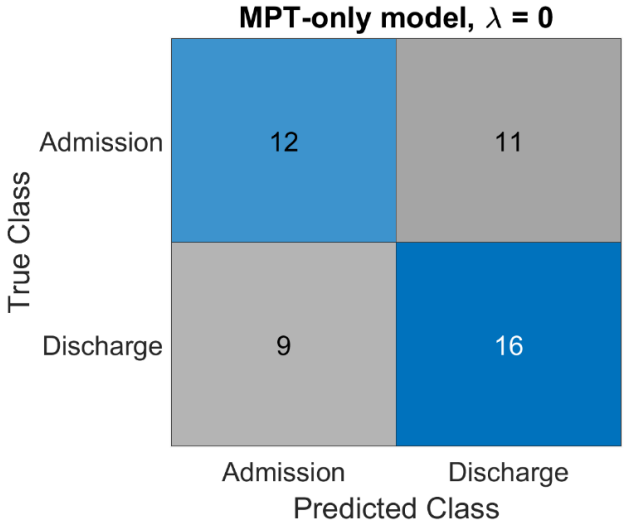

Figure 5.12. Confusion matrix for the ACC-based non-regularized MPT-only model.

Logistic classification: L1 regularization

Thirty λ values between 0.001 and 0.3 (for the MPT+ model) or between 0.001 and 0.1 (for the MPT– model) were tested to identify the regularization parameter that would yield the lowest regularized loss. That process is visualized for the MPT– and MPT+ models in Figure 5.13 and Figure 5.14 respectively.

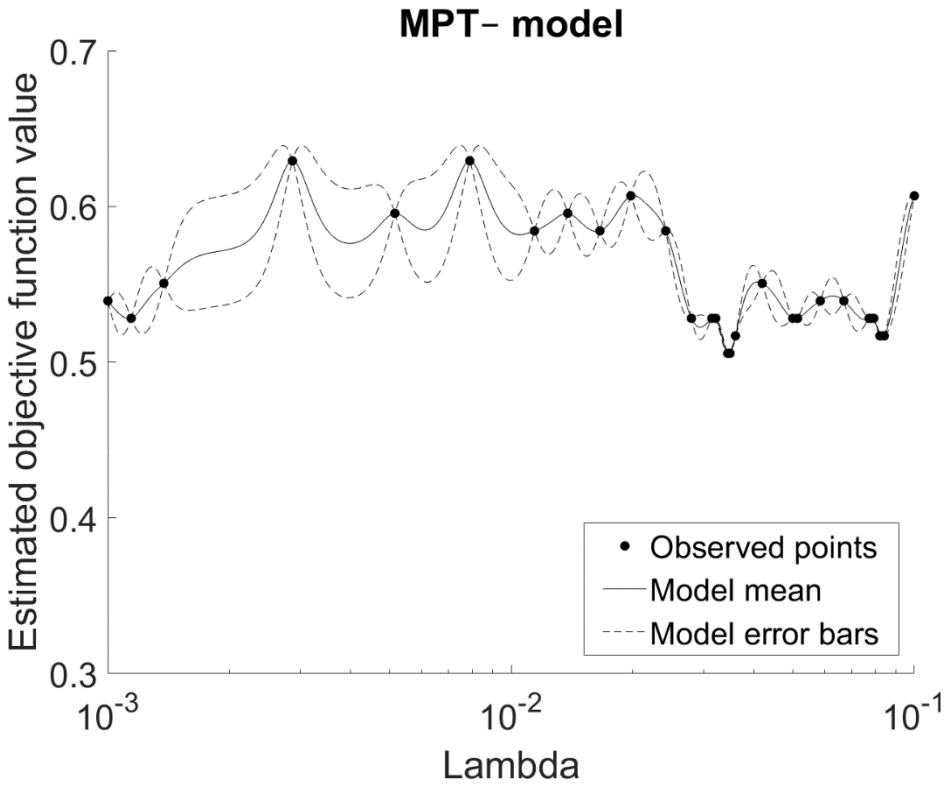

Figure 5.13. Estimated objective function value for each lambda value that was tested during optimization of the ACC-based MPT– model. Lower objective function values indicate a better fit.

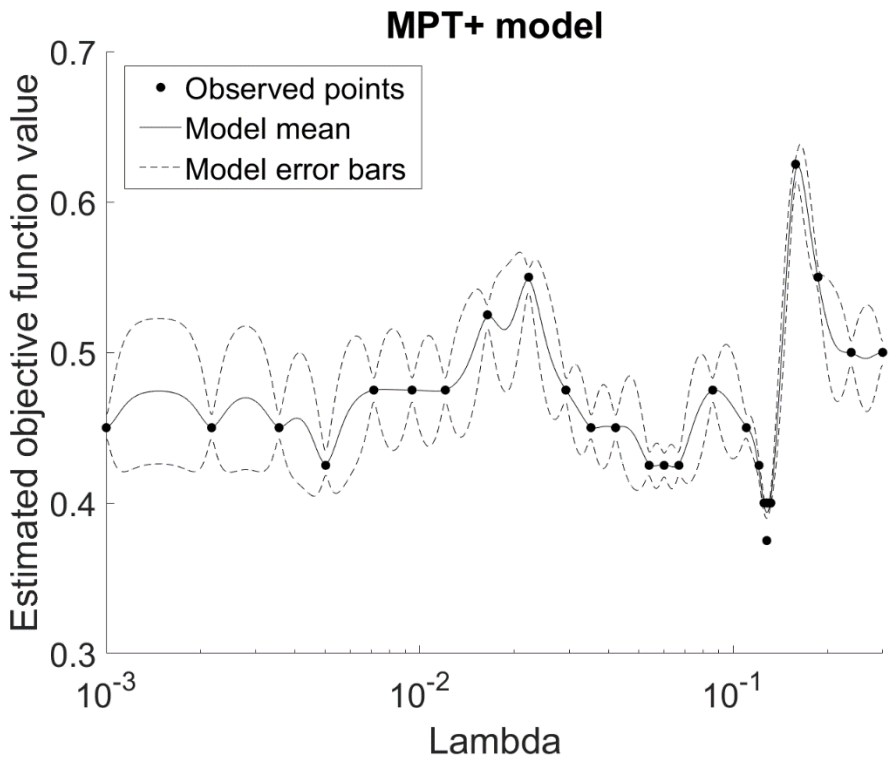

Figure 5.14. Estimated objective function value for each lambda value that was tested during optimization of the ACC-based MPT+ model. Lower objective function values indicate a better fit.

Figure 5.15 shows confusion matrices for the MPT– and MPT+ models trained with no regularization. Quantities in blue-shaded squares indicate the number of data points correctly classified as admission or discharge. Grey-shaded squares indicate incorrectly classified points.

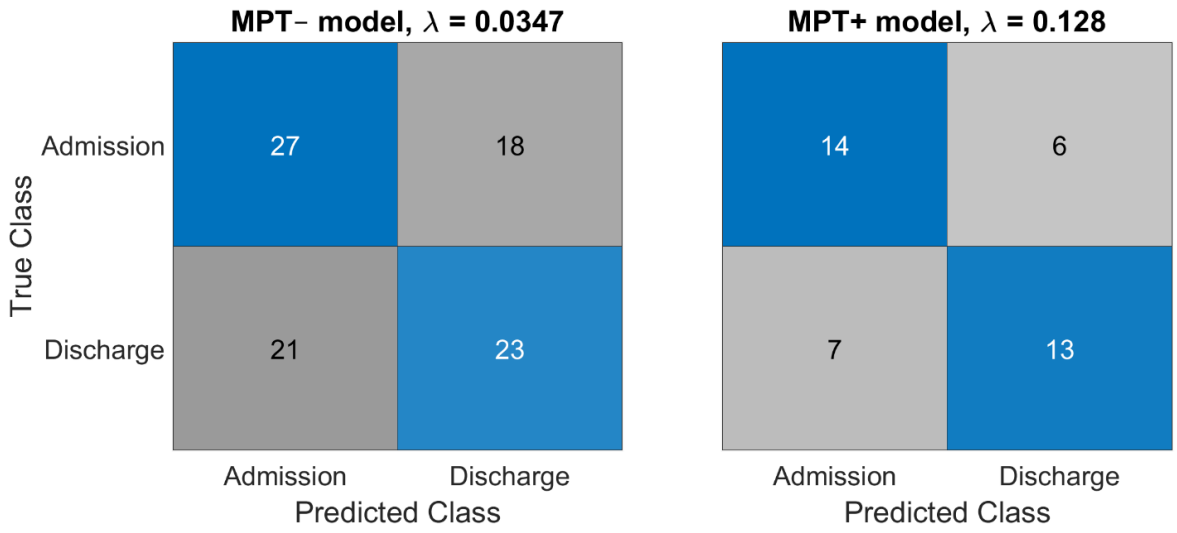

Figure 5.15. Confusion matrices and optimized lambda values for the ACC-based regularized MPT– and MPT+ models.

Model performance

The models' performances were evaluated based on their accuracy, area under the receiver operating characteristic curve (AUC), true admission rate (TAR), true discharge rate (TDR), admission predictive value (APV), and discharge predictive value (DPV). "Admission" and "discharge" are used in place of "positive" and "negative" in the names of these metrics to increase clarity. As used here, the TAR and TDR metrics correspond to sensitivity and specificity. Performance metrics for the five ACC-based models are summarized in Table 5.10.

Table 5.10. Performance metrics for each ACC-based logistic classifier model.

λ	Model	Accuracy	AUC	TAR	TDR	APV	DPV
none	MPT–	0.53	0.53	0.49	0.57	0.54	0.52
none	MPT+	0.58	0.54	0.50	0.65	0.59	0.57
none	MPT-only	0.58	0.60	0.52	0.64	0.57	0.59
L1	MPT–	0.56	0.51	0.60	0.52	0.56	0.56
L1	MPT+	0.68	0.63	0.70	0.65	0.67	0.68

Coefficient weights and odds ratios

Table 5.11 and Table 5.12 show β feature weights and odds ratios for MPT– and MPT+ models trained on the full data set, following the procedure used to generate Table 5.7 and Table 5.8.

Table 5.11. Feature weights and odds ratios for the ACC-based regularized MPT– model.

Feature	Normalized units		Original units	
	β weight	Odds ratio	β weight	Odds ratio
CPP SD (sentences)	–0.16	0.85	–0.19	0.83
Phrase duration mean (spontaneous)	0.12	1.13	0.10	1.11
Creak %: 0.02 threshold (sentences)	0.11	1.12	0.64	1.89
CPP mean (sentences)	–0.11	0.90	–0.069	0.93
Phrase count (rainbow)	–0.10	0.91	–0.014	0.99
Total phrase duration (rainbow)	–0.088	0.92	–0.014	0.99
F0 mean (vowel)	0.079	1.08	0.0019	1.002
Phrase % (spontaneous)	0.075	1.078	0.45	1.58
HNR (vowel)	0.066	1.07	0.0082	1.008

Table 5.12. Feature weights and odds ratios for the ACC-based regularized MPT+ model.

Feature	Normalized units		Original units	
	β	**Odds ratio**	**β**	**Odds ratio**
F0 SD (sentence)	0.18	1.20	0.019	1.02
Max phonation time (max phonation)	0.18	1.20	0.034	1.03

Within-stay trajectories

Day-by-day discharge probabilities for each participant, generated by the ACC-based MPT– and MPT+ models, are shown in Figure 5.16 **Error! Reference source not found.**and Figure 5.17. These figures were generated following the same procedures as for Figure 5.8 and Figure 5.9. In each figure, participants are presented in order from the greatest to least first-to-last-day change in discharge probability. Within each plot, a participant's voice recordings are ordered sequentially from admission to discharge. If the probability of discharge for a point was below 0.5, the point was classified as "admission" and plotted in red. Black points indicate recordings for which the discharge probability was above 0.5.

For most participants, the probability of discharge increased over the course of the hospital stay. At the same time, most participants showed day-to-day fluctuations outside of that overall trajectory. Trajectories for the MPT+ model were generally flatter than those for the MPT– model.

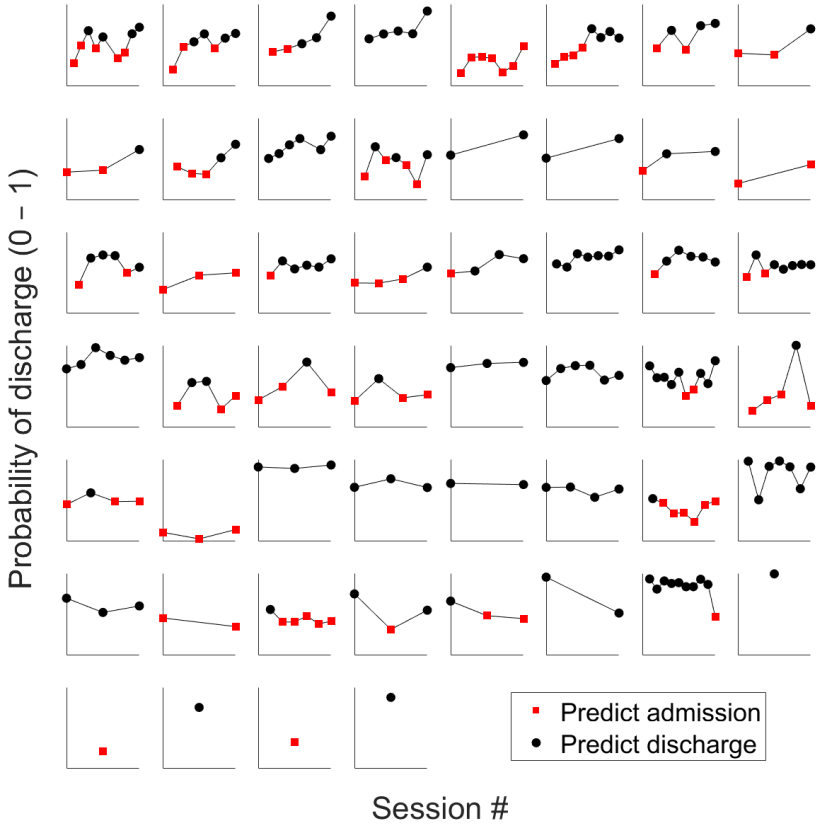

Figure 5.16. Day-to-day discharge probabilities for each speaker based on the ACC-based MPT– model. Red squares indicate admission predictions (discharge probability < 0.5) and black dots indicate discharge predictions.

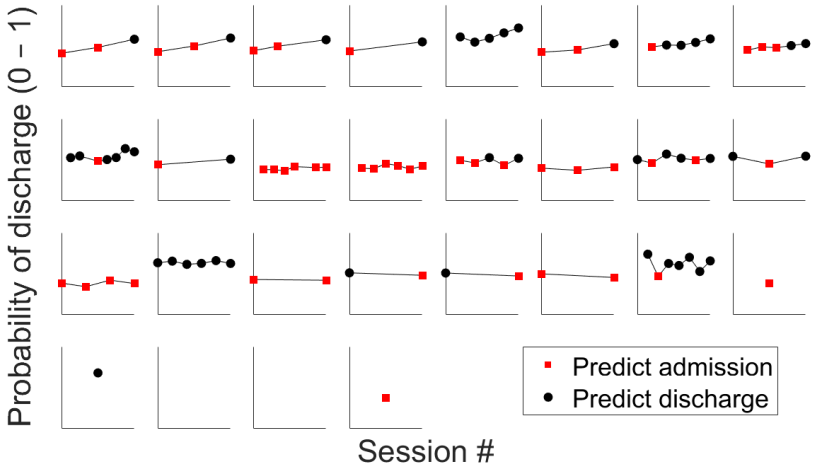

Figure 5.17. Day-to-day discharge probabilities for each speaker based on the ACC-based MPT+ model. Red squares indicate admission predictions (discharge probability < 0.5) and black dots indicate discharge predictions.

5.4 Discussion

The following sections discuss the microphone-based analysis unless otherwise stated. Discussion of the accelerometer-based data can be found in Section 5.4.6.

5.4.1 Effect sizes

Total phrase duration

For a continuous speech task, the "total phrase duration" feature was given by the sum of the durations of all the speech phrases in that utterance. In other words, this measure represents the total duration of the recording minus any pause time. As shown in Table 5.5, the effect size (Cohen's *d*) for this measure was –0.5 for both the Rainbow Passage and the second reading passage. This value indicates that total phrase duration tended to decrease from admission to discharge, with a moderate effect size (Sawilowsky, 2009). The Rainbow Passage text did not vary from day to day, and the second reading passages were all approximately equal length and presented in random order. A decrease in the amount of time needed to read these texts, then, indicates that participants were able to read faster at discharge compared to admission.

In our pilot study (Murton et al., 2017), total phrase duration in the Rainbow Passage decreased from admission to discharge for most participants. In that study, participants read the same text each day. Therefore we initially hypothesized that the decrease in total phrase duration may have been a practice effect: participants were able to read the text more quickly because they were more familiar with it. We later added the second passage-reading task to control for any familiarity effect. Since there were 10 additional reading texts, participants read an unfamiliar one each day. Total phrase duration also decreased for this second reading task, with an equally large effect size (–0.5) as for the Rainbow Passage task. This finding indicates that participants' ability to read faster at discharge is not due only to familiarity with the reading text. In future work, it may also be useful to look at speech rate in the spontaneous speech tasks, since that task clearly does not depend on the speaker's reading ability.

In a relevant finding, Ramig (1983) found that, especially among older speakers, speaking and reading rates were correlated with overall level of health. Similarly, Linville et al. (1989) examined relationships between various speech production measures in women aged 67 and older. They found that reduced reading rate correlated with increases in phonatory instability, including jitter and F0 SD. They suggest that "generalized loss of physiological control" could alter both reading rate and phonatory stability. In their case, the suggested change in physiological control was primarily related to age, but HF-related changes in physiological status might similarly affect vocal control. In particular, Linville et al. (1989) found a "significant correlation" between reduced reading rate and increased vocal fold edema on laryngoscopic exam. The causes of their speakers' vocal fold edema were not specified, but this finding strongly suggests that HF-related vocal fold edema may have direct physiological effects on voice measures, including speaking and reading rate.

Maximum phonation time

The effect size for MPT was also moderate, at 0.49 (Sawilowsky, 2009). This is a promising finding: MPT is related to phonatory ability and breath support, both of which we hypothesized would be affected by HF-related edema. It is important to note that the mean MPT even at discharge, when speakers were at their healthiest, was only approximately 10 seconds. In contrast, Maslan et al. (2011) measured MPT in healthy adults aged 65 years or older and found that MPT did not decrease with age alone in the absence of comorbidities: healthy older adults had similar MPTs as younger adults. In fact, they found a mean MPT of 22 seconds in healthy speakers age 60-90 years, which is substantially longer than the mean 10-second MPT for our speakers at discharge. It is important to keep in mind that the admission vs. discharge classification presented in this analysis is not the same as identifying whether someone is healthy or has HF. This study is only aimed at differentiating degrees of severity within already-diagnosed speakers.

Figure 5.18 shows box-and-whisker plots for admission and discharge MPTs, as well as the admission-to-discharge change in MPT. Notably, one participant had very high MPT at both admission (31 seconds) and discharge (28 seconds). Those MPTs are even longer than the mean MPT for this participant's age group in Maslan et al. (2011), which was approximately 23 seconds. The speaker's voice was severely dysphonic, with a Praat CPPS of 4.9 dB at admission and 7.4 dB at discharge. These CPPS values are substantially below the clinical threshold of 14.45 dB identified in Murton et al. (2020) and are very likely to indicate a voice disorder. Perceptually, the speaker had substantial strain and low airflow, which likely allowed them to conserve breath and produce a long MPT. For this participant, and any other speakers with similar voices, long MPT likely does not indicate a low decompensation risk. This study did

exclude participants with a history of voice disorders. However, many voice issues are undiagnosed, and elderly speakers may be particularly unlikely to seek diagnosis or treatment (Cohen et al., 2012). In future data collection, it may be desirable to actively screen participants for voice disorders using some acoustic or auditory-perceptual criteria. Data from speakers with non-HF-related dysphonia should be interpreted with caution.

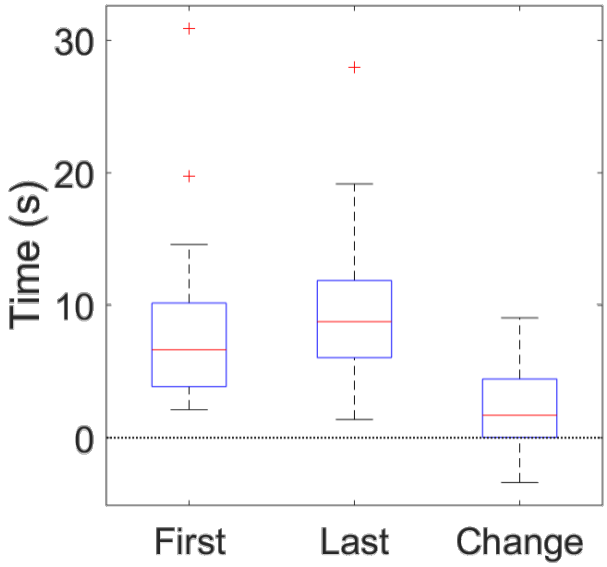

Figure 5.18. Distributions of MIC-based maximum phonation times for speakers' first days, last days, and first-to-last day changes.

Cepstral peak prominence

The effect size for CPP in sustained vowels was comparatively small, at approximately 0.2 for CPP mean and median. However, inspecting the distribution of CPP values reveals that CPP was fairly high at admission for many participants, with a mean CPP of 21.4 dB at admission. The mean Praat CPPS for sustained vowels at admission was 11.8 dB, which is below the 14.45 dB norm identified in Chapter 4 (Murton et al., 2020). However, the Chapter 4 norms were computed on younger, comparatively healthy speakers relative to the HF population here. Future work is needed to identify CPP norms for older adults. Regardless, speakers whose CPPS

was low at admission were much more likely to show increases in CPPS with treatment. Of speakers with CPPS below the 11.8 dB mean at admission, 76% (13/17) had increased CPPS at discharge. In contrast, of speakers with admission CPPS above 11.8 dB, only 47% (14/30) had increased CPPS at discharge. The difference between the CPPS changes in the low-CPPS and high-CPPS groups was nearly significant ($p = 0.055$) and is visualized in Figure 5.19. Future work could investigate any other characteristics common to this low-CPP group, and potentially identify speakers whose voices are likely to respond to changes in HF status.

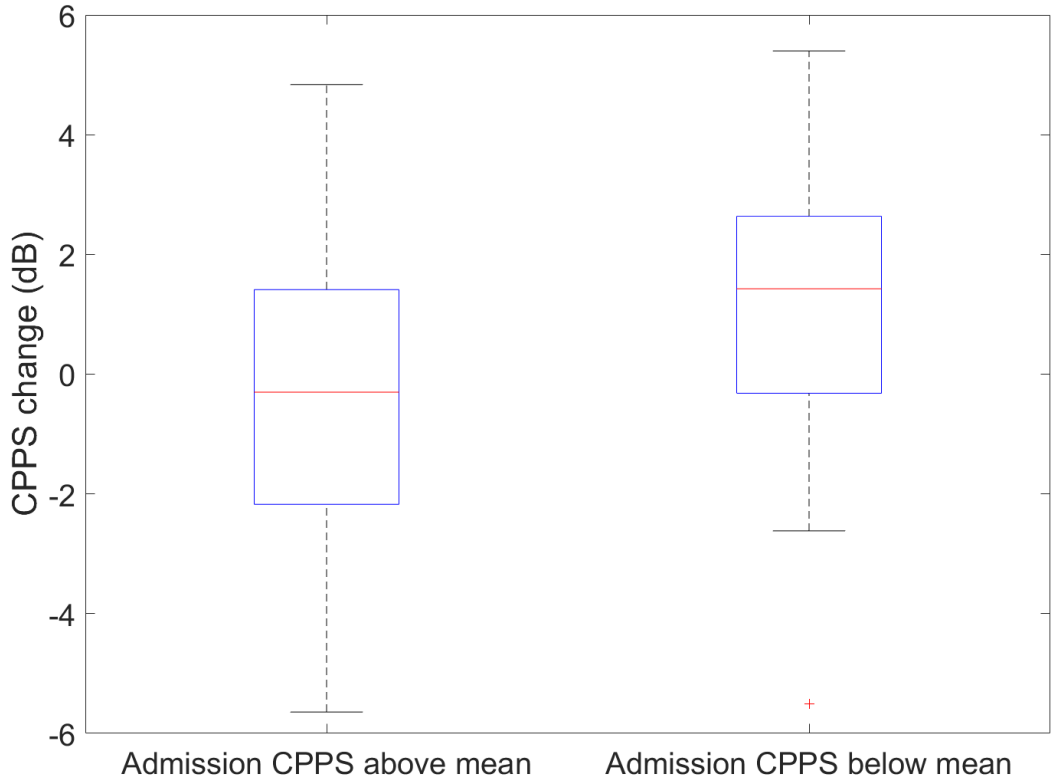

Figure 5.19. Distribution of changes in CPPS from admission to discharge, for speakers with CPPS below the group mean at admission (left) and above the group mean at admission (right).

Interestingly, CPP tended to decrease for continuous speech tasks. This finding is unexpected, especially in light of the increases in CPP for sustained vowels. Increases in the

usage of creaky voice, which tended to occur for continuous speech but not sustained vowels, may contribute to the decline in CPP at discharge relative to admission.

5.4.2 Feature correlations

For most measures, the features calculated on the various continuous speaking tasks were highly correlated with each other. For example, the correlation coefficient between the Rainbow Passage F0 mean and the CAPE-V sentence F0 mean was 0.96. In some but not all cases, the features for continuous speech tasks were also well-correlated with the equivalent sustained vowel features. For example, the correlation between the Rainbow Passage creak percent and the sustained vowel creak percent was 0.68, but the correlation between the Rainbow Passage CPP SD and the sustained vowel CPP SD was only 0.22. In general, the creak percent features were moderately correlated with the CPP, F0 SD, jitter, shimmer, and HNR features. The mean and median F0 features tended to be less correlated with other acoustic measures. The MPT, CPP SD, and total speech time features were largely uncorrelated with any other feature.

5.4.3 Logistic regression classification

Classification accuracy from the MPT– and MPT+ models without regularization was poor. Accuracy (percent of correct classifications) was 53% for the MPT– model and 52% for MPT+. These results are unsurprising given the large number of features relative to the number of data points. The models are likely overfit to the training data leading to poor test accuracy. L1 regularization, which yielded markedly better accuracy, performs a kind of feature selection: increasing λ sends increasingly many feature weights to zero, removing them from the model. The MPT-only model, which had 64% accuracy, was at much less risk of overfitting even without regularization since it used only one feature. The finding that MPT alone classified

correctly 64% of recordings further indicates that MPT is a promising feature for additional analysis.

The MPT– model with L1 regularization found its optimized λ at λ = 0.021. This model had higher test accuracy (62%) than the corresponding non-regularized model. However, its predictions tended to be biased towards admission (54 admission predictions vs. 37 discharge predictions). The input data set was slightly biased towards admission recordings, with 47 recordings from admission and 44 from discharge, so that input bias may have partially caused the output bias. To adjust this bias in future work, it may be fruitful to adjust the probability threshold needed to predict discharge. Currently any data point with discharge probability over 0.5 received a discharge prediction, but that 0.5 threshold could be optimized with more data.

The MPT+ model with L1 regularization found an optimized λ at λ = 0.101. At 69%, this model's test accuracy also compared favorably to the non-regularized MPT+ model. This model's predictions were more balanced than those of the MPT– model, with 22 admission and 20 discharge predictions. Its predictions were also more accurate overall. This comparison further suggests that MPT is a useful feature to identify decompensated heart failure and is worth investigating more closely in future work.

5.4.4 Odds ratios and predictive power

The feature weights and odds ratios from the regularized models provide information about each feature's predictive power. Features with weights of 0 did not have enough predictive power to be preserved by the L1 regularization process. Of the remaining features, weights further from 0 (i.e., those with absolute value further from 0) indicate that the model's prediction is based relatively more on that feature. For a weight β, the odds ratio is given by e^{β}. For each feature, the odds ratio represents the change in the prediction that results from a 1-unit increase

in that feature (if the feature weight is in the original units), or from a 1-standard-deviation increase (if the weight is normalized). The odds ratios for the MPT– model indicate that a discharge prediction was associated with increased mean F0 in vowels; greater phonatory stability (decreased F0 SD, increased CPP, and decreased CPP SD) in vowels; increased creak percent in sentences; faster speech rate in the Rainbow Passage; and reduced, more regular pauses in the Rainbow Passage and spontaneous speech. More specifically, in the Rainbow Passage, the percent of time devoted to speaking rather than pausing increased and the standard deviation of phrase durations decreased. The mean and median phrase duration both increased for spontaneous speech. These results broadly support our hypotheses that voices at discharge would have reduced instability, increased F0, longer phrase durations, and less pausing.

The MPT+ model's optimized λ was higher, so fewer features had non-zero weights after regularization. In this model, discharge probability was associated with increased F0 SD in sentences, increased MPT, and longer phrase durations in spontaneous speech. Notably, the MPT– model showed a *decrease* in F0 SD for sustained vowels, while this model showed an *increase* in F0 SD in CAPE-V sentences. For sustained vowels, F0 SD is a measure of the stability of the phonation: it is not expected that the F0 would vary during the vowel production. In contrast, F0 during continuous speech is not expected to remain stable. Increased F0 SD during continuous speech is more likely related to increased use of prosody than to phonatory instability.

5.4.5 Within-stay trajectories

The day-to-day trajectories of discharge probability are shown in Figure 5.8 and Figure 5.9. They reveal that, in general, most participants showed increased probability of discharge on their last day compared to their first. This increase occurred for most participants even though it

did not always cross the category boundary from an admission prediction to a discharge prediction. The MPT– model showed more day-to-day variability than the MPT+ one did, possibly because it incorporated more features. In the MIC-based MPT– model, 49 participants had usable data from more than one recording session. Of those 49 participants, 39 (80%) had increased discharge probability on their last day compared to their first day. Similarly, 74% (35/47) participants had increased discharge probability from the first to last day in the ACC-based MPT– model. This result suggests that future work could investigate appropriate thresholds for distinguishing voices with and without decompensation. In theory, the classification could be related to the change in discharge or admission probability rather than an absolute threshold. In other words, for a patient being monitored at home, a certain level of increase in the probability of admission could trigger an alert to the care provider.

Notably, many day-to-day trajectories were not monotonic: they had repeated fluctuations in discharge probability. Alternately, for some speakers, the trajectory involved several stable days followed by a dramatic change in discharge probability and then more stable days. In future work, these fluctuations could be correlated with weight changes or medication dosages to determine a more exact relationship between fluid level and voice quality.

5.4.6 Accelerometer findings

For acoustic measures (e.g. F0- and CPP-related measures), mean changes and effect sizes tended to be similar for the ACC-based data (Table 5.9) and MIC-based data (Table 5.5). In contrast, several of the large effect sizes for speech-phrase-related measures in the MIC data were not comparably large for the ACC data. For example, total phrase duration had an effect size of –0.5 for both the MIC-based Rainbow Passage and second reading passage. For the ACC-based Rainbow Passage and second reading passage, though, the effect sizes were only –0.33

and –0.22 respectively. The speech phrase measures were computed by identifying voiced and unvoiced regions in the speech signals, so differences in the behavior of the pitch-tracking algorithms for MIC and ACC signals could have contributed to this difference.

Logistic classifier performance was broadly similar for the MIC-based (Table 5.6) and ACC-based (Table 5.10). In particular, the regularized MPT+ model performed very similarly for with both MIC and ACC data. The MPT-only and regularized MPT– models had somewhat lower test accuracy with ACC data compared to MIC. Both regularized models preserved similar features for MIC and ACC input. As for the MIC model in Section 5.4.4, the features most predictive features in the ACC-based MPT– model related to increased mean F0 mean; greater phonatory stability (reduced CPP SD, and increased HNR); increased creaky voice percent; faster speech; and less-frequent pauses. Like the MIC-based model, the ACC-based MPT+ model showed that increased MPT and F0 SD in CAPE-V sentences were related to increased discharge probability.

5.4.7 Towards a clinical implementation

The findings presented above indicate that voice signals alone can separate ADHF patient's pre- and post-treatment voice samples with accuracy up to 69%. The acoustic features that best predicted that a sample was post-treatment were longer MPT, more stable phonation (decreased F0 SD, increased CPP, and decreased CPP SD), increased creak percent, faster speech rate, and reduced pausing compared to pre-treatment voices. These features are hypothesized to reflect physiological changes in the larynx and lungs due to fluid accumulation from ADHF. Importantly, these results are based on speakers who were recovering from ADHF and were receiving inpatient treatment. The long-term goal of this work is to monitor stable HF patients at home to predict and prevent ADHF episodes before they require hospitalization (see

"Future work", below). Therefore, it is important to determine whether these voice changes (or similar ones) are also present in speakers who are developing ADHF. If they are, then patients and clinicians could use at-home voice monitoring to assess ADHF risk and guide outpatient treatment. This work also showed that models based on data from the accelerometer (positioned on the neck surface) performed similarly to models based on the acoustic microphone. These results are also based on a short set of speech tasks, most of which did not involve spontaneous speech (i.e., reading and sustained vowel tasks). Future work might reveal that longer or more natural speech samples provide better information about vocal function and ADHF risk. In that case, the neck-surface accelerometer could be used for ambulatory voice monitoring to evaluate ADHF risk based on real-world voice use. The accelerometer preserves speaker privacy and is more robust to background noise than the acoustic microphone, making it particularly useful for ambulatory recordings.

5.4.8 Future work

Suggestions for future work that are raised by specific results are discussed above. More broadly, future work could look at subgroups of the already-enrolled participants to identify whether there are non-voice factors that affect how voice responds to HF status. For example, any effects of age, sex, HF subtype, or previous medical history could all be relevant. Additionally, data from the intermediate recordings (between admission and discharge) could be used to train a linear model that predicts, for example, how many days from admission a sample was recorded on. In contrast to the logistic models presented here, a linear model could provide more information about the trajectory of voice change during treatment for HF.

As discussed in Chapter 2, this study examined voice changes in patients recovering from decompensated HF. However, the eventual goal of this work is to predict and prevent episodes of

decompensation by monitoring stable HF patients who are at risk. In future data collection, it would be useful to collect data from HF patients who are not decompensated but are at risk of developing ADHF. For example, the 30-day readmission rate after an episode of ADHF is approximately 24% (Desai & Stevenson, 2012). Analyzing voice from patients who had just recovered from ADHF could be used to predict readmission. This prospective monitoring could also test our hypotheses in the forward direction (starting with stable patients and attempting to predict decompensation) rather than following improving patients as we did here. Monitoring discharged patients may also provide a clearer idea of how closely voice changes leading up to decompensation actually mirror voice changes during treatment.

Chapter 6. Conclusion

6.1 Summary

This thesis examined three interrelated applications of voice analysis for health monitoring. The voice signal can convey information about a speaker's identity, language, emotional state, cognitive status, and vocal, mental, and physical health. In addition to being so information-rich, voice samples are also non-invasive and relatively easy to collect. For these reasons, voice monitoring shows considerable promise as a tool for clinicians and patients to manage illness and promote health.

Chapter 2 presented a pilot study of vocal biomarkers for acute decompensated heart failure. ADHF was theorized to affect voice production by increasing the amount of edema in the vocal folds and lungs. Patients who were undergoing inpatient treatment for ADHF provided daily voice samples, and their voices at admission (pre-treatment) were compared to their voices at discharge (post-treatment). At discharge, speakers with ADHF produced voices with more creaky voice, higher fundamental frequency, decreased cepstral peak prominence variation, and faster speech rate compared to their own voices at admission. These findings suggest that speech biomarkers can indicate ADHF status, and may be useful for proactive monitoring of patients at risk of developing ADHF.

Two findings from Chapter 2 prompted additional inquiry in separate studies. First, as noted above, speakers produced more creaky voice at discharge, after ADHF treatment, compared to their voices at admission. Creaky voice is a type of non-modal phonation that is regularly produced by healthy speakers, but can indicate a voice disorder. Its comparative increase in post-treatment voices was somewhat unusual because most other acoustic measures tended to indicate improved vocal quality with ADHF treatment. However, because creaky voice

production is governed in part by a speaker's linguistic system as well as vocal tract physiology, it is harder to predict how physiological changes like AHDF would affect creaky voice. For example, it is possible that pre-treatment speakers with AHDF were less able than normal to generate creaky voice, and that the higher post-treatment amount of creaky voice reflects their normal phonation. An early investigation of this finding was presented in Chapter 3.

Chapter 3 evaluated an automatic creaky voice detection algorithm against a set of hand labels to determine a perceptually relevant threshold probability for the algorithm's output. IPPs are grammatically, pragmatically, and clinically significant but are time-consuming and difficult to detect by hand. Improving their automatic detection can therefore support a wide range of voice- and speech-related work. The automatic detection algorithm evaluated in Chapter 3 divides a signal into short frames and produces a probability that each frame contains creaky voice. To identify whether a specific frame contains creaky voice, these probabilities need to be converted into a binary decision using an appropriate threshold. We compared these frame-by-frame creak probabilities to a set of hand labels in recordings of American English conversational speech from healthy speakers without voice disorders. Results suggested that the algorithm's output aligned well with the hand labels, but at a much lower threshold probability than had been previously reported.

Second, cepstral peak prominence improved with ADHF treatment primarily for the subset of speakers who began treatment with comparatively low CPP. In other words, some speakers maintained a high CPP despite having severe enough ADHF to require hospitalization, and CPP in those speakers did not tend to increase after treatment. This finding suggests that CPP may be a relevant biomarker for ADHF only in speakers whose CPP drops below some clinically significant threshold during decompensation. In addition to this application for ADHF

monitoring, clinically relevant CPP thresholds would also aid voice clinicians who wish to use CPP to evaluate and monitor voice disorders. However, previous work on identifying perceptually relevant threshold values for CPP, especially in older speakers, was limited. Chapter 4 presents an analysis of voices with and without voice disorders to identify these CPP thresholds.

Chapter 4 identified cepstral peak prominence values that aid in clinical voice evaluation by distinguishing speakers with and without voice disorders. Previous work indicated that different voice analysis programs produce CPP values in different ranges, and that CPP values also fall into different ranges for sustained vowels and continuous speech. Therefore, this study's first experiment employed two widely used voice analysis programs to determine CPP thresholds separately for sustained /a/ vowels and for the Rainbow Passage. Voice samples from 295 speakers with voice disorders were compared to voice samples from 50 vocally healthy controls. This dataset was used to identify CPP threshold values that best separated the speakers with voice disorders from the control speakers. The CPP thresholds we identified varied with speech task and analysis program, but classification accuracy was high for all thresholds. This study's second experiment used linear regression models to relate CPP values from both programs to auditory-perceptual ratings of dysphonia severity. CPP provided good estimates of dysphonia severity ratings, and this second experiment's findings also validated the thresholds from the first experiment.

Chapter 5 presents a continuation of the pilot study from Chapter 2. This study enrolled 52 total patients receiving inpatient ADHF treatment, carried out additional statistical and machine learning analyses, and compared data collected from an acoustic microphone to data collected from a neck-skin acceleration sensor. Unlike the acoustic microphone, which records

sound waves from the air, the acceleration sensor records vibrations of the neck skin around the larynx during phonation. Therefore, the acceleration sensor's signal does not include information about the user's speech (which is shaped by the vocal tract above the larynx) or about other sounds in the vicinity. In other words, the acceleration sensor better preserves a speaker's privacy and is more robust to background noise compared to the acoustic microphone. These properties mean that the acceleration sensor is particularly useful for real-world applications of voice monitoring, since it can be worn more readily in daily life. In Chapter 5, logistic regression models were trained to classify voice samples as pre-treatment (i.e., at admission) or post-treatment (at discharge). For each voice sample, these models output a probability that the sample was produced at discharge rather than at admission. The models distinguished pre-treatment from post-treatment voices with accuracy up to 69%. Additionally, the probability of discharge increased from the pre-treatment voice sample to the post-treatment sample for up to 80% of participants. We also analyzed the predictive power of various acoustic features in these models. Evaluating predictive power allowed us to understand which features contributed most to the predicted discharge probability, and to evaluate our results in the context of our hypotheses about the physiology of ADHF and voice production. That analysis revealed that voices at discharge were associated with longer maximum phonation times, more stable phonation, increased creak percent, faster speech rate, and reduced pausing compared to voices at admission, broadly supporting our hypotheses. Overall, models based on data from the neck-skin acceleration sensor performed similarly to models based on the acoustic microphone, suggesting that the accelerometer could be used to evaluate ADHF risk in environments that are not conducive to microphone recordings. These findings indicate that ADHF-related fluid

accumulation in the lungs and larynx is likely to measurably affect voice, and that voice may therefore provide a useful biomarker for proactive at-home monitoring of ADHF risk.

6.2 Future work

Previous work on the algorithm used in Chapter 3 has identified three acoustic realizations of creaky voice and shown that these patterns are detected with varying accuracy. The data set we used in Chapter 3 contains hand labels for prosodic structure in addition to labels for IPPs. Analyzing the algorithm's performance in different prosodic contexts could help elucidate how IPP production is influenced by prosody and syntactic structure. For example, in American English, IPPs tend to occur at specific locations including phrase beginnings, phrase endings, and accented syllables (Pierrehumbert & Talkin, 1992). However, IPPs may not be produced the same way in all locations where they occur. Future work could be aimed at understanding how the algorithm evaluates IPPs in all these different contexts, so that its results can be used with more confidence on a wide range of speech tasks. Additionally, future work could examine the distribution of IPPs that speakers produce before and after ADHF treatment. Understanding where these speakers produce IPPs relative to the linguistically-expected locations could help explain why creaky voice production tended to increase in response to ADHF treatment.

Chapter 4 found CPP thresholds that accurately distinguished speakers with and without voice disorders. However, certain types of non-modal phonation appear to affect CPP in ways that merit additional study. For example, CPP from one voice with significant strain was substantially higher for Praat than for ADSV, although both programs' CPPs for this voice fell below the clinical threshold. Other work suggests that a typical upper bound for CPP should also be established: in some hyperfunctional or rough voices, very high CPP values high might also indicate a voice disorder. The effects of voicing detection algorithm, vowel quality, and speaker

age on appropriate CPP thresholds are also not well-understood and warrant future investigation. In particular, work on voice monitoring for ADHF would benefit from a better understanding of typical CPP thresholds in older speakers.

The findings from Chapter 5 suggested future work to identify factors that affect how voice changes with ADHF status. For example, pre-treatment CPP was low for some speakers; in these speakers, but not others, CPP tended to increase after treatment. Grouping patients by age, sex, HF type, previous medical history, or other factors could help reveal whether there are specific patient populations who would benefit most from voice monitoring for ADHF. The trajectories of voice change during ADHF treatment would also benefit from further study. The logistic classifiers showed that discharge probability increased in response to treatment for up to 80% of participants. Rather than classifying voice samples into "admission" or "discharge" categories, it could be useful to identify a threshold for the increase in discharge probability that would signal a shift from the admission to discharge class. For stable patients being monitored at home, this technique could establish a baseline for each patient individually and then look for sufficiently large voice changes to indicate an increased risk of decompensation.

Finally, this work's ultimate goal is to develop voice monitoring tools that predict and prevent episodes of ADHF. Future data collection could focus not on already-hospitalized ADHF patients, but on stable HF patients who are at risk but not hospitalized. Prospective monitoring of this type is needed to test whether voice changes during ADHF treatment (as presented in Chapters 2 and 5) are similar to voice changes as an ADHF episode approaches. Patients who recently received inpatient ADHF treatment are at high risk of readmission over the subsequent several months. Using voice monitoring to predict readmission for ADHF could be a stepping stone to predicting ADHF episodes more generally in at-risk patients. Additionally, recent work

by other groups suggests that vocal biomarkers can help to diagnose and monitor COVID-19 (Adans-Dester et al., 2020; Quatieri et al., 2020). The work on ADHF presented here may be useful for monitoring changes to vocal and respiratory function that are caused by COVID-19 and other health conditions.

Appendix: Reading texts

CAPE-V sentences

1. The blue spot is on the key again.

2. How hard did he hit him?

3. We were away a year ago.

4. We eat eggs every Easter.

5. My momma makes lemon muffins.

6. Peter will keep at the peak.

Rainbow Passage

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.

Rainbow Passage (part 2)

Over the centuries people have explained the rainbow in many ways. To the Hebrews it was a sign that there would be no more big floods. The Greeks used to say that it was a sign from the gods that there would be war or heavy rain. The Norse thought the rainbow was a bridge that the gods used to travel from earth to their home in the sky. Others have tried to explain rainbows

using science. Aristotle thought that they came from the sun's rays reflecting the rain. Since then, physicists have found out that refraction by the raindrops actually causes the rainbows.

Rainbow Passage (part 3)

There are many different ideas about rainbows. The colors in the rainbow depend on the size of the raindrops. Bigger raindrops make a wider colored band. The actual primary rainbow that we see comes from adding up many bows. For example, mixing red and green light makes yellow. If the red part of one rainbow lines up with the green part of another, that makes a rainbow with a very wide yellow band. It is very common for rainbows to look like this: mostly red and yellow, with not very much green or blue.

Bamboo Passage

Bamboo walls are getting to be very popular. They are strong, easy to use, and good looking. They provide a good background and create the mood in Japanese gardens. Bamboo is a grass, and is one of the most rapidly growing grasses in the world. Many varieties of bamboo are grown in Asia, although it is also grown in America. Last year we bought a new home and have been working on the flower gardens. In a few more days, we will be done with the bamboo wall in one of our gardens. We have really enjoyed the project.

The North Wind and the Sun

The North Wind and the Sun were arguing about who was stronger, when a traveler came along wearing a warm cloak. They agreed that whoever made the traveler take his cloak off first would be considered the strongest. The North Wind blew as hard as he could, but the traveler just

folded his cloak around him more closely. Finally, the North Wind gave up. Then the Sun shone out warmly, and the traveler took off his cloak right away. And so, the North Wind had to confess that the Sun was the stronger of the two.

Wolf Passage (part 1)

There was once a poor shepherd boy who used to watch his flocks in the fields next to a dark forest. One hot afternoon, he thought up a good plan to get some company for himself and also have a little fun. Raising his fist in the air, he ran down to the village shouting 'Wolf, Wolf.' As soon as they heard him, the villagers all rushed from their homes. They were full of concern for his safety, and two of his cousins even stayed with him for a short time.

Wolf Passage (part 2)

A few days later the boy tried exactly the same trick again, and once more he was successful. However, not long after, a wolf that had just escaped from the zoo was looking for a change from its usual diet of chicken and duck. It came out from the forest and began to threaten the sheep. Racing down to the village, the boy of course cried out even louder than before. Unfortunately, all the villagers were convinced that he was trying to fool them a third time. They told him, 'Go away and don't bother us again.' And so the wolf had a feast.

Caterpillar Passage (part 1)

Do you like amusement parks? Well, I sure do. To amuse myself, I went twice last spring. My most memorable moment was riding on the Caterpillar, which is a gigantic rollercoaster high above the ground. When I saw how high the Caterpillar rose into the bright blue sky I knew it

was for me. After waiting in line for thirty minutes, I made it to the front where the man measured my height to see if I was tall enough. I gave the man my coins, asked for change, and jumped on the cart.

Caterpillar Passage (part 2)

Tick, tick, tick, the Caterpillar climbed slowly up the tracks. It went so high I could see the parking lot. Boy, was I scared!\ I thought to myself, “There’s no turning back now.” People were so scared they screamed as we swiftly zoomed fast, fast, and faster along the tracks. As quickly as it started, the Caterpillar came to a stop. Unfortunately, it was time to pack the car and drive home. That night I dreamed of the wild ride on the Caterpillar. Taking a trip to the amusement park and riding on the Caterpillar was my favorite moment ever!

Arthur the Rat (part 1)

Once upon a time there was a rat who couldn't make up his mind. Whenever the other rats asked him if he would like to come out hunting with them, he would answer in a hoarse voice, "I don't know." And when they said, "Would you rather stay inside?" he wouldn't say yes, or no either. He'd always avoid making a choice. One fine day his aunt Josephine said to him, "Now look here! You can’t carry on like this. You have no more mind of your own than a blade of grass!" The young rat coughed and looked wise, but said nothing.

Arthur the Rat (part 2)

One night the rats heard a loud noise in the loft. It was a very dreary old place. The roof let the rain come washing in, the beams and rafters had all rotted through, so that the whole thing was

quite unsafe. One of the beams fell with one edge on the floor. "This won't do," said their leader. "We can't stay cooped up here any longer." So they sent out scouts to search for a new home. A little later on that evening, the scouts came back and said they had found a barn where there would be room for all of them.

References

Abraham, W. T., Adamson, P. B., Bourge, R. C., Aaron, M. F., Costanzo, M. R., Stevenson, L. W., Strickland, W., Neelagaru, S., Raval, N., Krueger, S., Weiner, S., Shavelle, D., Jeffries, B., & Yadav, J. S. (2011). Wireless pulmonary artery haemodynamic monitoring in chronic heart failure: A randomised controlled trial. *The Lancet*, *377*(9766), 658–666. https://doi.org/10.1016/S0140-6736(11)60101-3

Adans-Dester, C. P., Bamberg, S., Bertacchi, F. P., Caulfield, B., Chappie, K., Demarchi, D., Erb, M. K., Estrada, J., Fabara, E. E., Freni, M., Friedl, K. E., Ghaffari, R., Gill, G., Greenberg, M. S., Hoyt, R. W., Jovanov, E., Kanzler, C. M., Katabi, D., Kernan, M., … Bonato, P. (2020). Can mHealth Technology Help Mitigate the Effects of the COVID-19 Pandemic? *IEEE Open Journal of Engineering in Medicine and Biology*, *1*, 243–248. https://doi.org/10.1109/OJEMB.2020.3015141

ADSV (3.4.2). (2019). [Computer software]. PENTAX Medical.

Alves, M., Krüger, E., Pillay, B., van Lierde, K., & van der Linde, J. (2019). The Effect of Hydration on Voice Quality in Adults: A Systematic Review. *Journal of Voice*, *33*(1), 125.e13-125.e28. https://doi.org/10.1016/j.jvoice.2017.10.001mx

American English Map Task. (1999). https://dspace.mit.edu/bitstream/handle/1721.1/32533/README_AmEngMapTask.pdf

Andersson, C., Gerds, T., Fosbøl, E., Phelps, M., Andersen, J., Lamberts, M., Holt, A., Butt, J. H., Madelaire, C., Gislason, G., Torp-Pedersen, C., Køber, L., & Schou, M. (2020). Incidence of New-Onset and Worsening Heart Failure Before and After the COVID-19 Epidemic Lockdown in Denmark: A Nationwide Cohort Study. *Circulation: Heart Failure*, *13*(6). https://doi.org/10.1161/CIRCHEARTFAILURE.120.007274

Awan, S. N., & Awan, J. A. (2020). A Two-Stage Cepstral Analysis Procedure for the Classification of Rough Voices. *Journal of Voice*, *34*(1), 9–19. https://doi.org/10.1016/j.jvoice.2018.07.003

Awan, S. N., Giovinco, A., & Owens, J. (2012). Effects of Vocal Intensity and Vowel Type on Cepstral Analysis of Voice. *Journal of Voice*, *26*(5), 670.e15-670.e20. https://doi.org/10.1016/j.jvoice.2011.12.001

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

Awan, S. N., Roy, N., Jetté, M. E., Meltzner, G. S., & Hillman, R. E. (2010). Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V. *Clinical Linguistics & Phonetics*, *24*(9), 742–758. https://doi.org/10.3109/02699206.2010.492446

Awan, S. N., Roy, N., Zhang, D., & Cohen, S. M. (2016). Validation of the Cepstral Spectral Index of Dysphonia (CSID) as a Screening Tool for Voice Disorders: Development of Clinical Cutoff Scores. *Journal of Voice*, *30*(2), 130–144. https://doi.org/10.1016/j.jvoice.2015.04.009

Aydinli, F. E., Özcebe, E., & İncebay, Ö. (2019). Use of cepstral analysis for differentiating dysphonic from normal voices in children. *International Journal of Pediatric Otorhinolaryngology*, 7.

Beckman, Hirschberg, & Shattuck-Hufnagel. (2006). The original ToBI system and the evolution of the ToBI framework. In *Prosodic typology* (Vol. 1). Oxford University Press.

Boersma, P., & Weenink, D. (2018). *PRAAT* (5.0.40) [Computer software]. University of Amsterdam. http://www.praat.org

Bőhm, T., & Shattuck-Hufnagel, S. (2009). Do Listeners Store in Memory a Speaker's Habitual Utterance-Final Phonation Type? *Phonetica*, *66*(3), 150–168. https://doi.org/10.1159/000235658

Boorsma, E. M., ter Maaten, J. M., Damman, K., Dinh, W., Gustafsson, F., Goldsmith, S., Burkhoff, D., Zannad, F., Udelson, J. E., & Voors, A. A. (2020). Congestion in heart failure: A contemporary look at physiology, diagnosis and treatment. *Nature Reviews Cardiology*. https://doi.org/10.1038/s41569-020-0379-7

Borlaug, B. A., & Paulus, W. J. (2011). Heart failure with preserved ejection fraction: Pathophysiology, diagnosis, and treatment. *European Heart Journal*, *32*(6), 670–679. https://doi.org/10.1093/eurheartj/ehq426

Brockmann-Bauser, M., Van Stan, J. H., Carvalho Sampaio, M., Bohlender, J. E., Hillman, R. E., & Mehta, D. D. (2019). Effects of Vocal Intensity and Fundamental Frequency on Cepstral Peak Prominence in Patients with Voice Disorders and Vocally Healthy Controls. *Journal of Voice*, S0892199719304552. https://doi.org/10.1016/j.jvoice.2019.11.015

Bromage, D. I., Cannatà, A., Rind, I. A., Gregorio, C., Piper, S., Shah, A. M., & McDonagh, T. A. (2020). The impact of COVID-19 on heart failure hospitalization and management: Report from a Heart Failure Unit in London during the peak of the pandemic. *European Journal of Heart Failure*, *22*(6), 978–984. https://doi.org/10.1002/ejhf.1925

Cohen, S. M., Kim, J., Roy, N., Asche, C., & Courey, M. (2012). Prevalence and causes of dysphonia in a large treatment-seeking population: Prevalence and Causes of Dysphonia. *The Laryngoscope*, *122*(2), 343–348. https://doi.org/10.1002/lary.22426

Dao, Q., Krishnaswamy, P., Kazanegra, R., Harrison, A., Amirnovin, R., Lenert, L., Clopton, P., Alberto, J., Hlavin, P., & Maisel, A. (2001). Utility of B-type natriuretic peptide in the

diagnosis of congestive heart failure in an urgent-care setting. *Journal of the American College of Cardiology*, *37*(2), 379–385. https://doi.org/10.1016/S0735-1097(00)01156-6

Delgado-Hernández, J., León-Gómez, N., & Jiménez-Álvarez, A. (2019). Diagnostic accuracy of the Smoothed Cepstral Peak Prominence (CPPS) in the detection of dysphonia in the Spanish language. *Loquens*, *6*(1), 058. https://doi.org/10.3989/loquens.2019.058

Desai, A. S., & Stevenson, L. W. (2012). Rehospitalization for Heart Failure: Predict or Prevent? *Circulation*, *126*(4), 501–506. https://doi.org/10.1161/CIRCULATIONAHA.112.125435

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, *24*(4), 423–444. https://doi.org/10.1006/jpho.1996.0023

Drugman, T., Kane, J., & Gobl, C. (2014). Data-driven detection and analysis of the patterns of creaky voice. *Computer Speech & Language*, *28*(5), 1233–1253. https://doi.org/10.1016/j.csl.2014.03.002

Fairbanks, G. (1960). The Rainbow Passage. In *Voice and articluation drillbook* (2nd ed., pp. 124–139). Harper & Row.

Fang, J. (2016). Heart failure with preserved ejection fraction: A kidney disorder? *Circulation*, *134*(6), 435–437. https://doi.org/10.1161/CIRCULATIONAHA.116.022249

Finkelhor, B. K., Titze, I. R., & Durham, P. L. (1988). The effect of viscosity changes in the vocal folds on the range of oscillation. *Journal of Voice*, *1*(4), 320–325. https://doi.org/10.1016/S0892-1997(88)80005-5

Fit linear classification model to high-dimensional data—MATLAB fitclinear. (n.d.). Retrieved August 4, 2020, from https://www.mathworks.com/help/stats/fitclinear.html

Fraile, R., & Godino-Llorente, J. I. (2014). Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control*, *14*, 42–54. https://doi.org/10.1016/j.bspc.2014.07.001

Gerratt, B. R., & Kreiman, J. (2001). Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, *29*(4), 365–381. https://doi.org/10.1006/jpho.2001.0149

Gheorghiade, M., & Pang, P. S. (2009). Acute Heart Failure Syndromes. *Journal of the American College of Cardiology*, *53*(7), 557–573. https://doi.org/10.1016/j.jacc.2008.10.041

Gobl, C. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*(1–2), 189–212. https://doi.org/10.1016/S0167-6393(02)00082-1

Gordon, M., & Ladefoged, P. (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics*, *29*(4), 383–406. https://doi.org/10.1006/jpho.2001.0147

Habibzadeh, F., Habibzadeh, P., & Yadollahie, M. (2016). On determining the most appropriate test cut-off value: The case of tests with continuous results. *Biochemia Medica*, 297–307. https://doi.org/10.11613/BM.2016.034

Hall, M. E., Vaduganathan, M., Khan, M. S., Papadimitriou, L., Long, R. C., Hernandez, G. A., Moore, C. K., Lennep, B. W., Mcmullan, M. R., & Butler, J. (2020). Reductions in Heart Failure Hospitalizations During the COVID-19 Pandemic. *Journal of Cardiac Failure*, *26*(6), 462–463. https://doi.org/10.1016/j.cardfail.2020.05.005

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

Heman-Ackah, Y. D., Sataloff, R. T., Laureyns, G., Lurie, D., Michael, D. D., Heuer, R., Rubin, A., Eller, R., Chandran, S., Abaza, M., Lyons, K., Divi, V., Lott, J., Johnson, J., & Hillenbrand, J. (2014). Quantifying the Cepstral Peak Prominence, a Measure of Dysphonia. *Journal of Voice*, *28*(6), 783–788. https://doi.org/10.1016/j.jvoice.2014.05.005

Henton, C. G. (1986). Creak as a sociophonetic marker. *The Journal of the Acoustical Society of America*, *80*(S1), S50–S50. https://doi.org/10.1121/1.2023837

Hillenbrand, J., & Houde, R. A. (1996). Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech. *Journal of Speech, Language, and Hearing Research*, *39*(2), 311–321. https://doi.org/10.1044/jshr.3902.311

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

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

Holmes, R. J., Oates, J. M., Phyland, D. J., & Hughes, A. J. (2000). Voice characteristics in the progression of Parkinson's disease. *International Journal of Language & Communication Disorders*, *35*(3), 407–418.

Horwitz-Martin, R. L., Quatieri, T. F., Lammert, A. C., Williamson, J. R., Yunusova, Y., Godoy, E., Mehta, D. D., & Green, J. R. (2016). *Relation of Automatically Extracted Formant Trajectories with Intelligibility Loss and Speaking Rate Decline in Amyotrophic Lateral Sclerosis*. 1205–1209. https://doi.org/10.21437/Interspeech.2016-403

Ishi, C. T., Sakakibara, K.-I., Ishiguro, H., & Hagita, N. (2008). A Method for Automatic Detection of Vocal Fry. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 47–56. https://doi.org/10.1109/TASL.2007.910791

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. https://doi.org/10.1044/1058-0360.0603.66

Jessup, M., & Brozena, S. (2003). Medical progress: Heart failure. *The New England Journal of Medicine*, *348*, 2007–2018.

Joseph, S., Cedars, A., Gregory, E., Geltman, E., & Mann, D. (2009). Acute decompensated heart failure. *Texas Heart Institute Journal*, *36*(5), 510.

Kane, J., Drugman, T., & Gobl, C. (2013). Improved automatic detection of creak. *Computer Speech & Language*, *27*(4), 1028–1047. https://doi.org/10.1016/j.csl.2012.11.002

Keating, P., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. *Proceedings of the International Congress of Phonetic Sciences*.

Kemp, C. D., & Conte, J. V. (2012). The pathophysiology of heart failure. *Cardiovascular Pathology*, *21*(5), 365–371. https://doi.org/10.1016/j.carpath.2011.11.007

Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus Auditory-Perceptual Evaluation of Voice: Development of a Standardized Clinical Protocol. *American Journal of Speech-Language Pathology*, *18*(2), 124–132. https://doi.org/10.1044/1058-0360(2008/08-0017)

Lee, Y., Kim, G., & Kwon, S. (2019). The Usefulness of Auditory Perceptual Assessment and Acoustic Analysis for Classifying the Voice Severity. *Journal of Voice*, S0892199719300876. https://doi.org/10.1016/j.jvoice.2019.04.013

Linville, S. E., Skarin, B. D., & Fornatto, E. (1989). The Interrelationship of Measures Related to Vocal Function, Speech Rate, and Laryngeal Appearance in Elderly Women. *Journal of Speech, Language, and Hearing Research*, *32*(2), 323–330. https://doi.org/10.1044/jshr.3202.323

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, *5*(1), 96–116. https://doi.org/10.1002/lio2.354

Maor, E., Perry, D., Mevorach, D., Taiblum, N., Luz, Y., Mazin, I., Lerman, A., Koren, G., & Shalev, V. (2020). Vocal Biomarker Is Associated With Hospitalization and Mortality Among Heart Failure Patients. *Journal of the American Heart Association*, *9*(7). https://doi.org/10.1161/JAHA.119.013359

Maor, E., Sara, J. D., Orbelo, D. M., Lerman, L. O., Levanon, Y., & Lerman, A. (2018). Voice Signal Characteristics Are Independently Associated With Coronary Artery Disease. *Mayo Clinic Proceedings*, *93*(7), 840–847. https://doi.org/10.1016/j.mayocp.2017.12.025

Maslan, J., Leng, X., Rees, C., Blalock, D., & Butler, S. G. (2011). Maximum Phonation Time in Healthy Older Adults. *Journal of Voice*, *25*(6), 709–713. https://doi.org/10.1016/j.jvoice.2010.10.002

Massachusetts Eye and Ear Infirmary. (1994). *Voice disorders database (CD-ROM)* (1.03) [Computer software]. Kay Elemetrics Corporation.

MATLAB (R2020a: 9.8.0.1396136). (2020). [Computer software]. The MathWorks, Inc.

Mehta, D. D., Zañartu, M., Feng, S. W., Cheyne, H. A., & Hillman, R. E. (2012). Mobile Voice Health Monitoring Using a Wearable Accelerometer Sensor and a Smartphone Platform. *IEEE Transactions on Biomedical Engineering*, *59*(11), 3090–3096. https://doi.org/10.1109/TBME.2012.2207896

Mehta, Daryush D., Van Stan, J. H., & Hillman, R. E. (2016). Relationships Between Vocal Function Measures Derived from an Acoustic Microphone and a Subglottal Neck-Surface Accelerometer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(4), 659–668. https://doi.org/10.1109/TASLP.2016.2516647

Mehta, Daryush D., Van Stan, J. H., Zañartu, M., Ghassemi, M., Guttag, J. V., Espinoza, V. M., Cortés, J. P., Cheyne, H. A., & Hillman, R. E. (2015). Using Ambulatory Voice Monitoring to Investigate Common Voice Disorders: Research Update. *Frontiers in Bioengineering and Biotechnology*, *3*. https://doi.org/10.3389/fbioe.2015.00155

Miranda, D., Lewis, G., & Fifer, M. (2016). Heart failure. In *L. S. Lilly (Ed.), Pathophysiology of heart disease: A collaborative project of medical students and faculty* (pp. 220–248). Lippincott Williams & Wilkins.

Mueller, P. B. (1997). The aging voice. *Seminars in Speech and Language*, *18*(2), 159–169.

Murton, O., Hillman, R., & Mehta, D. (2020). Cepstral Peak Prominence Values for Clinical Voice Evaluation. *American Journal of Speech-Language Pathology*, *29*(3), 1596–1607. https://doi.org/10.1044/2020_AJSLP-20-00001

Murton, O. M., Hillman, R. E., Mehta, D. D., Semigran, M., Daher, M., Cunningham, T., Verkouw, K., Tabtabai, S., Steiner, J., Dec, G. W., & Ausiello, D. (2017). Acoustic speech analysis of patients with decompensated heart failure: A pilot study. *The Journal of the Acoustical Society of America*, *142*(4), EL401–EL407. https://doi.org/10.1121/1.5007092

Murton, O., Shattuck-Hufnagel, S., Choi, J.-Y., & Mehta, D. D. (2019). Identifying a creak probability threshold for an irregular pitch period detection algorithm. *The Journal of the Acoustical Society of America*, *145*(5), EL379–EL385. https://doi.org/10.1121/1.5100911

Núñez-Batalla, F., Cartón-Corona, N., Vasile, G., García-Cabo, P., Fernández-Vañes, L., & Llorente-Pendás, J. L. (2019). Validez de las medidas del pico cepstral para la valoración objetiva de la disfonía en sujetos de habla hispana. *Acta Otorrinolaringológica Española*, *70*(4), 222–228. https://doi.org/10.1016/j.otorri.2018.04.008

O'Connor, C. M., Stough, W. G., Gallup, D. S., Hasselblad, V., & Gheorghiade, M. (2005). Demographics, Clinical Characteristics, and Outcomes of Patients Hospitalized for Decompensated Heart Failure: Observations From the IMPACT-HF Registry. *Journal of Cardiac Failure*, *11*(3), 200–205. https://doi.org/10.1016/j.cardfail.2004.08.160

Oppenheim, A. V., & Schafer, R. W. (2004). Dsp history - From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, *21*(5), 95–106. https://doi.org/10.1109/MSP.2004.1328092

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., Paul, D., Švec, J. G., & Hillman, R. (2018). Recommended Protocols for Instrumental Assessment of Voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. *American Journal of Speech-Language Pathology*, *27*(3), 887–905. https://doi.org/10.1044/2018_AJSLP-17-0009

Pierrehumbert, J., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In *Papers in Laboratory Phonology II* (pp. 90–116). Cambridge University Press.

Quatieri, T. F., Talkar, T., & Palmer, J. S. (2020). A Framework for Biomarkers of COVID-19 Based on Coordination of Speech-Production Subsystems. *IEEE Open Journal of Engineering in Medicine and Biology*, *1*, 203–206. https://doi.org/10.1109/OJEMB.2020.2998051

Ramig, L. A. (1983). Effects of physiological aging on speaking and reading rates. *Journal of Communication Disorders*, *16*(3), 217–226. https://doi.org/10.1016/0021-9924(83)90035-7

Redi, L., & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics*, *29*(4), 407–429. https://doi.org/10.1006/jpho.2001.0145

Sara, J. D. S., Maor, E., Borlaug, B., Lewis, B. R., Orbelo, D., Lerman, L. O., & Lerman, A. (2020). Non-invasive vocal biomarker is associated with pulmonary hypertension. *PLOS ONE*, *15*(4), e0231441. https://doi.org/10.1371/journal.pone.0231441

Sawilowsky, S. S. (2009). New Effect Size Rules of Thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 597–599. https://doi.org/10.22237/jmasm/1257035100

Scherer, S., Schwenker, F., Campbell, N., & Palm, G. (2009). Multimodal laughter detection in natural discourses. In *Human Centered Robot Systems* (Vol. 6, pp. 111–120). Springer. https://doi.org/10.1007/978-3-642-10403-9_12

Sharma, K., & Kass, D. A. (2014). Heart Failure With Preserved Ejection Fraction: Mechanisms, Clinical Features, and Therapies. *Circulation Research*, *115*(1), 79–96. https://doi.org/10.1161/CIRCRESAHA.115.302922

Sivasankar, M., & Leydon, C. (2010). The role of hydration in vocal fold physiology: *Current Opinion in Otolaryngology & Head and Neck Surgery*, *18*(3), 171–175. https://doi.org/10.1097/MOO.0b013e3283393784

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Titze, I. R. (1988). The physics of small-amplitude oscillation of the vocal folds. *The Journal of the Acoustical Society of America*, *83*(4), 1536–1552. https://doi.org/10.1121/1.395910

Titze, I. R. (1992). Phonation threshold pressure: A missing link in glottal aerodynamics. *The Journal of the Acoustical Society of America*, *91*(5), 2926–2935.

Unal, I. (2017). Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach. *Computational and Mathematical Methods in Medicine*, *2017*, 1–14. https://doi.org/10.1155/2017/3762651

Van Stan, J. H., Mehta, D. D., & Hillman, R. E. (2017). Recent Innovations in Voice Assessment Expected to Impact the Clinical Management of Voice Disorders. *Perspectives of the ASHA Special Interest Groups*, *2*(3), 4–13. https://doi.org/10.1044/persp2.SIG3.4

Verdolini, K., Min, Y., Titze, I. R., Lemke, J., Brown, K., Mersbergen, M. van, Jiang, J., & Fisher, K. (2002). Biological Mechanisms Underlying Voice Changes Due to Dehydration. *Journal of Speech, Language, and Hearing Research*, *45*(2), 268–281. https://doi.org/10.1044/1092-4388(2002/021)

Verdolini, K., Titze, I. R., & Fennell, A. (1994). Dependence of Phonatory Effort on Hydration Level. *Journal of Speech, Language, and Hearing Research*, *37*(5), 1001–1007. https://doi.org/10.1044/jshr.3705.1001

Verdolini-Marston, K., Sandage, M., & Titze, I. R. (1994). Effect of hydration treatments on laryngeal nodules and polyps and related voice measures. *Journal of Voice*, *8*(1), 30–47. https://doi.org/10.1016/S0892-1997(05)80317-0

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

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Horwitz, R., Yu, B., & Mehta, D. D. (2013). Vocal biomarkers of depression based on motor incoordination. *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge - AVEC '13*, 41–48. https://doi.org/10.1145/2512530.2512531

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35.

Yu, M., Choi, S. H., Choi, C.-H., & Choi, B. (2018). Predicting Normal and Pathological Voice using a Cepstral Based Acoustic Index in Sustained Vowels versus Connected Speech. *Communication Sciences & Disorders*, *23*(4), 1055–1064. https://doi.org/10.12963/csd.18550

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *The Journal of the Acoustical Society of America*, *123*(5), 3878–3878. https://doi.org/10.1121/1.2935783

Zañartu, M., Ho, J. C., Kraman, S. S., Pasterkamp, H., Huber, J. E., & Wodicka, G. R. (2009). Air-Borne and Tissue-Borne Sensitivities of Bioacoustic Sensors Used on the Skin Surface. *IEEE Transactions on Biomedical Engineering*, *56*(2), 443–451. https://doi.org/10.1109/TBME.2008.2008165