Voice Biomarkers of Recovery From Acute Respiratory Illness

Brian Tracey, *Senior Member, IEEE*, Shyamal Patel, *Member, IEEE*, Yao Zhang, Kara Chappie, Dmitri Volfson, Federico Parisi, Catherine Adans-Dester, Francesco Bertacchi, Paolo Bonato, *Senior Member, IEEE*, and Paul Wacnik, *Member, IEEE*

***Abstract*—Voice analysis is an emerging technology which has the potential to provide low-cost, at-home monitoring of symptoms associated with a variety of health conditions. While voice has received significant attention for monitoring neurological disease, few studies have focused on voice changes related to flu-like symptoms. Herein, we investigate the relationship between changes in acoustic features of voice and self-reported symptoms during recovery from a flu-like illness in a cohort of 29 subjects. Acoustic features were automatically extracted from "sick" and "well" visit data collected in the laboratory setting, and feature down-selection was used to identify those that change significantly between visits. The selected acoustic features were extracted from at-home data and used to construct a combined distance metric that correlated with self-reported symptoms (0.63 rank correlation). Changes in self-reported symptoms corresponding to 10% of the ordinal scale used in the study were detected with an area under the curve of 0.72. The results show that acoustic features derived from voice recordings may provide an objective measure for diagnosing and monitoring symptoms of respiratory illnesses.**

***Index Terms*—Biomedical signal processing, biomedical acoustics, speech analysis, wearable sensors.**

I. Introduction

IN RECENT years we have seen a growing interest in using voice characteristics to monitor a variety of health conditions [1]. This has been driven by the low-cost and non-invasive nature of voice recordings, as well as advances in machine learning and audio signal processing. Significant attention has been focused on identifying acoustic features associated with voice changes in disorders such as Parkinson's disease [2], [3], depression [4], dementia [5], hypertension [6], post-traumatic stress disorder [7], and COVID-19 [8]. However, changes in voice characteristics during influenza-like illnesses are not well understood.

During speech production, air flows through the oral tract and the nasal passages, particularly when pronouncing nasal consonants or other nasal sounds. The nasal passages produce resonances at distinct frequencies, but also anti-resonances (sound blockages), generally all at higher frequencies. These resonances and anti-resonances are attenuated during congestion. Thus, during decongestion, the higher-frequency spectra become both more peaked and more attenuated due to anti-resonances [9]. Several previous studies examined voice changes after acute decongestion induced by nasal decongestant sprays. These studies demonstrated changes in spectra during vocalization of Chinese phonemes [9], and also demonstrated changes in the Voice Low tone to High tone Ratio (VLHR) metric [10]. Prior work on capturing nasality of speech is also relevant, as congestion may be comparable to hyponosality. Several studies have used nasality metrics such as VLHR or analysis of power third-octave band powers to characterize hypernasality, the sound associated with cleft palate and other conditions [11], [12]. In third-octave analysis, the acoustic spectra are divided into log-spaced frequency bands, under the assumption that the distinctive formants (resonances) for each vowel will be in the same third-octave band for most subjects. These studies generally found third-octave analysis to be more sensitive than VLHR to hypernasality [11], [12].

In this work, we tracked the recovery of study participants from symptoms associated with acute influenza-like illness (ILI) using acoustic features extracted during the performance of a series of sustained phoneme tasks. We recruited participants with acute ILI who recorded their voice over a two-week period during the sick to well transition. Voice recording was performed in the lab at the beginning (sick visit) and at the end (well visit) of the study. In addition, participants recorded their symptoms as well as voice twice daily (morning and evening) at home using a smartphone app for a period of approximately 14 days in between the two lab visits.

Manuscript received July 20, 2021; revised November 17, 2021; accepted December 12, 2021. Date of publication December 21, 2021; date of current version June 6, 2022. This work was supported in part by Spaulding Rehabilitation Hospital and in part by Pfizer. *(Corresponding author: Brian Tracey.)*

Brian Tracey was with Signal Insights LLC, Arlington, MA 02474 USA. He is now with Takeda Pharmaceuticals, Cambridge, MA 02139 USA (e-mail: brian.tracey@takeda.com).

Shyamal Patel was with Pfizer Inc., Cambridge MA 02139 USA. He is now with Oura Health Ltd., San Francisco, CA 94133 USA (e-mail: shyamal.patel@gmail.com).

Yao Zhang, Kara Chappie and Paul Wacnik are with Pfizer Inc., Cambridge MA 02139 USA (e-mail: yao.zhang@pfizer.com; kara.chappie@pfizer.com; paul.wacnik@pfizer.com).

Dmitri Volfson was with Pfizer Inc., Cambridge MA 02139 USA and is now with Takeda Pharmaceuticals, Cambridge, MA 02139 USA.

Federico Parisi, Catherine Adans-Dester, Francesco Bertacchi, and Paolo Bonato are with the Harvard Medical School/Spaulding Rehabilitation Hospital, Charlestown MA 02129 USA (e-mail: fbertacchi@g.harvard.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2021.3137050.

Digital Object Identifier 10.1109/JBHI.2021.3137050

To our knowledge, this is the first study to report that acoustic features associated with respiratory symptoms change during natural patient recovery from a flu-like illness. In addition to previously studied acoustic features such as VLHR [10] and third-octave band metrics [13], we explored a wider set of acoustic features and showed that features capturing spectral structure correlate well with self-reported change in symptoms. Furthermore, we use at-home recordings to create a distance metric which combines information from a selected set of acoustic features extracted from only three sustained phoneme tasks, and show this metric can capture patient trajectories over time with good correlation to self-reported symptoms. These results suggest that at-home, smartphone-based monitoring of changes in the acoustic features of voice may be an effective, objective approach for tracking patient recovery and may merit further investigation for early detection of respiratory infections.

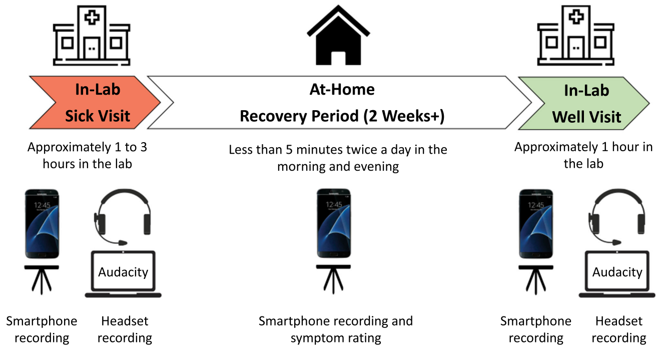

In-Lab
Sick Visit
At-Home
Recovery Period (2 Weeks+)
In-Lab
Well Visit
Approximately 1 to 3 hours in the lab
Less than 5 minutes twice a day in the morning and evening
Approximately 1 hour in the lab
Audacity
Audacity
Smartphone recording
Headset recording
Smartphone recording and symptom rating
Smartphone recording
Headset recording

Fig. 1. Overview of data collection. Participants performed voice exercises both during sick and well visits, recorded simultaneously on a smartphone and a headset system. Between visits, participants recorded voice exercises using a smartphone app every morning and evening and rated their status on symptoms of respiratory illness.

TABLE I
SELF-REPORTED SYMPTOMS FOR PROPOSED ILLNESS PHENOTYPES

Phenotype	Symptoms included
Congestion-specific	Need to blow nose, Nasal obstruction, Post-nasal discharge
Non-congestion	Runny nose, Cough, Sore throat, Thick Nasal Discharge

TABLE II
SUMMARY OF CANDIDATE ACOUSTIC ENDPOINTS FOUND FROM DOWN-SELECTION BASED ON IN-LAB MEASUREMENTS

Category	Acoustic features
Pitch variability	Pitch interquartile range / median (LG)
	Jitter, local (LG)
Amplitude variability	Shimmer, dB
	200 Hz third-octave band power standard deviation (LG)
Spectral structure	Harmonicity
	Voice low/high ratio (VLHR)
	Spectral entropy (1.5-2.5 kHz, 1.6-3.2 kHz)
	Spectral contrast (1.7-3.2 kHz (LG), 3.2-6.4 kHz (LG))
	Spectral flatness (1.5-2.5 kHz band)
Formant 1 (F1)	F1 frequency (LG)
	F1 bandwidth standard deviation
MFCC	mean MFCC: MFCC 6, 8
	Standard deviation of MFCC (LG): MFCC 2, 3, 8, 9, 10, 11, 12

Endpoints are grouped by general category of voice characteristics. LG denotes log-transformed variables; other variables are not transformed.

II. MATERIALS AND METHODS

Before discussing our methods in detail, we give a brief overview. Section A describes the experimental protocol. Sections B-C describe preprocessing and extraction of acoustic features. Section D describes analysis of self-reported symptoms, and Section E describes how in-lab and at-home data were analyzed to identify the most informative phonemes and acoustic features.

A. Experimental Protocol

Individuals with flu-like symptoms who met the inclusion/exclusion criteria were recruited and studied during their recovery phase. The study was approved by the Partners Human Research Committee IRB (protocol #2017P002684 / PHS, 2/7/2018). The study protocol was explained to all subjects and written consent was obtained. A sick-to-well study design was chosen to make it easier to recruit participants as opposed to a well-to-sick design, which would significantly increase the duration and complexity of the study in addition to requiring a much larger sample size. We collected self-reported symptoms and voice data both in-lab and at-home to capture their sick and well states as well as day-to-day changes during the recovery period. As shown in Fig. 1, participants were asked to perform voice tasks during both lab visits (sick visit and well visit) as well as at-home during the minimum two-week period between these visits. Accelerometer-based systems [14] offer advantages in terms of environmental noise and privacy, but we focused on smartphone-based recordings (scalable to large-scale clinical trials). Thus we created a custom-built app running on a Samsung Galaxy S7 Edge smart phone with an Android operating system and used it to record both at-home and in-lab data. As a check, in-lab data were simultaneously recorded using a Shure SM10 A headset microphone with an ART USB Dual Pre preamp (35 dB channel gain) and Audacity software version 2.2.2 running on a laptop. Both systems used non-lossy compression and had a bit depth of 16; Samsung data were sampled at 44.1 kHz, and Audacity data were sampled at 48 kHz. The protocol did not specify mouth-to-microphone distance. Participants recorded sustained phonations of both nasal consonants and cardinal vowels for 5-10 seconds. The four vowels using the International Phonetic Alphabet (IPA) were /a/, /i/, /u/, and /ae/ (participants were prompted to pronounce sounds using the more vernacular cues “o,” “E,” “OO,” and “a”). The three nasal consonants were the sounds /n/, /m/ and /ng/ (note that the English /ng/ sound does not map directly to the IPA). During the home monitoring period and concurrent with the voice tasks, participants were asked to rate their perceived symptom severity (0-5, 0 being lowest and 5 being highest) for 19 symptoms in the morning and 16 symptoms in the evening related to respiratory tract illness (see Supplementary Materials Table III for details). A Composite Symptom Score (CSS) was formed by summing of morning

TABLE III
FINAL SET OF ACOUSTIC FEATURES FOR THE 3 SELECTED PHONEMES

Phoneme	Acoustic Feature
/m/	Harmonicity
	Pitch interquartile range (IQR) (LG)
	F1 bandwidth standard deviation
	Spectral Entropy: 1.5-2.5kHz, 1.6-3.2 kHz
	Standard Deviation of MFCC (LG): MFCC 2, 10
	Spectral Flatness 1.5-2.5 kHz
	Mean MFCC: MFCC 8
	Shimmer (local, dB)
	Spectral Contrast 3.2-6.4 kHz (LG)
	200 Hz TOB (third-octave band) standard deviation (LG)
/n/	Harmonicity
	F1 bandwidth standard deviation
	Pitch interquartile range (IQR) (LG)
	Spectral Entropy: 1.5-2.5kHz, 1.6-3.2 kHz
	Spectral Flatness 1.5-2.5 kHz
	Standard Deviation of MFCC (LG): MFCC 1, 2, 3, 11
	Mean MFCC: MFCC 8
	Spectral Contrast 1.6-3.2 kHz (LG)
/a/	F1 bandwidth standard deviation
	Pitch interquartile range (IQR) (LG)
	Spectral Entropy: 1.6-3.2 kHz
	Jitter (local) (LG)
	Standard Deviation of MFCC (LG): MFCC 9, 12
	Mean MFCC: MFCC 6
	Spectral Contrast 3.2-6.4 kHz (LG)

plus evening scores on 7 symptoms (‘post-nasal discharge’, ‘nasal obstruction’, ‘runny nose’, ‘thick nasal discharge with mucus’ ‘cough’, ‘sore throat’, and ‘need to blow nose’, yielding a CSS range 0-70). After completing two weeks of recordings, participants were instructed to return to the lab for their well visit when all symptoms had resolved (i.e. for some participants the well visit took place several days after the end of the 2-week monitoring period).

B. Voice Signal Preprocessing and Phoneme Segmentation

Voice recordings were subject to several screening and preprocessing steps. Recording lengths were checked to identify and remove any incomplete recordings. When multiple complete recordings exist for a given time point, the final recording was used for analysis, under the assumption that participants had re-recorded their speech due to technical problems encountered during previous attempts.

Segmentation was carried out to identify the time periods during which individual phonemes were present in the audio recordings. Segmentation was made easier by the fact that the volume of each phoneme is relatively constant and the expected number of vocalizations in each recording is known. The acoustic intensity throughout the recording was first computed using Praat [15], and Otsu’s method [16] was used to find a threshold separating background noise from more energetic events representing speech segments. A morphological ‘fill’ operation was then used to fill short (<0.2 sec) gaps in the detected segments. Finally, the first 0.75 sec of each detected phoneme was discarded to avoid transient effects and a time segment of 2 sec was retained for analysis, similar to other studies [10]. See the Supplementary Materials for more discussion.

This automated segmentation approach was validated by comparison to manual segmentation (manual annotations were made using [17]) for a portion of the in-lab recordings. In these recordings, the automated quality checks discarded 26 of 228 automatically segmented vowels; for the remaining vowels, we observed good agreement between the manual and automated vowel segment start times (mean difference: 0.015 sec, standard deviation: 0.04 sec).

C. Extraction of Acoustic Features

For each segmented phoneme, multiple acoustic features were computed using the Parselmouth interface [18] to Praat as well as custom Python code. Features were computed at the default Praat frame rate (generally 10 msec) and statistics were computed across frames in each phoneme. Samsung and Audacity processing was identical except for pitch floor parameters noted below.

Power and power variability features: The root-mean-square (RMS) acoustic amplitude for each segment was computed and used to normalize sound amplitude, following Lee *et al.* [9]. RMS was converted to dB for consideration as a feature. Power variability was measured using shimmer, which captures the rapid variability in waveform amplitudes measured at glottal pulse intervals [6], [19]. We examined several variants of shimmer [15]. We expected shimmer to reduce as participants recover, inflammation decreases, and vocal fold function returns to its normal state.

Motivated by previous work on third-octave band analysis for hypernasality [13], [20] we analyzed power fluctuations in the output of 1/3 octave band filters at various frequencies. Exploratory data analysis revealed little change in average 1/3 octave output, but significant reductions in the power variability (measured as the standard deviation) of 1/3 octave filters for several bands. Fluctuations in the 200 Hz third-octave band (passband 178-224 Hz) were significant for all sustained nasal and vowel phonemes. Because this frequency band captures the pitch fundamental for much of our cohort, we interpret this output to be a measure of power variability at frequencies near the pitch fundamental. Similar to shimmer, we expect variability in the 200 Hz third-octave band to decrease as participants recover.

Pitch and pitch variability features: Pitch characteristics of each segment were extracted using the autocorrelation method from Praat [15]. For Samsung files, manual review indicated that the presence of low-frequency periodic background noise in laboratory measurements led to some false pitch detections, so the pitch floor was increased to 80 Hz for males and 100 Hz for females (for Audacity data, a pitch floor of 50 Hz was acceptable). Within each vocalization, estimated pitch values were processed to compute the pitch statistics. In some noisier recordings, pitch estimation was difficult so that the algorithm temporarily locked onto the wrong frequency. Thus, we used robust measures of location and spread (median and interquartile range) and computed the quartile variation (IQR/median), denoted `coviqr_pitch` below. Unreliable pitch estimates were screened by requiring that this quartile variation be less than 10% (threshold determined empirically). Values greater than this were treated as missing data. Pitch variability on shorter time scales was captured using jitter [18]. We expect that jitter and other measures of pitch fluctuation should decrease as participants recover.

Spectral structure features: Congestion is known to change the spectral structure of phonemes that involve the nasal passages. We compute multiple metrics to capture this effect:

- Harmonicity [21] captures the power ratio of harmonic to non-harmonic components. Because resonances and anti-resonances at high frequencies become more pronounced in decongested participants, harmonicity should increase with decongestion.
- Spectral entropy [22] computes the entropy of the spectrum in a desired frequency band. During decongestion, the spectrum should become more peaked (particularly at higher frequencies), so entropy should decrease.
- Spectral contrast [23] sorts power spectrum values in a desired frequency band by intensity and computes the ratio of the highest quartile of values (peaks) to the lowest quartile of values (troughs). During decongestion, contrast in higher-frequency bands should increase as the spectrum becomes more peaked.
- Spectral flatness [24] is computed as the ratio of the geometric to arithmetic mean of spectrum values in a given frequency band, and (like entropy) seeks to differentiate flat vs. peaked spectra. During decongestion, flatness should decrease.
- Voice low-to-high ratio (VLHR) [10] is the ratio of low-to-high frequency energy. We used a separation between low and high of 600 Hz, following [12]. In decongested participants, anti-resonances decrease high-frequency energy, so VLHR should increase as congestion decreases.
- Mel-frequency Cepstral Coefficients (MFCC) are widely used features in speech processing and are useful in analysis of Parkinsonian speech [25]. These coefficients, which represent the discrete cosine transform of a scaled power spectrum, have less straightforward physical interpretation, but are sensitive to changes in the spectrum and are also robust to environmental noise. We computed 21 MFCC values using librosa [26].

Except for harmonicity and MFCC, the above spectral features are computed over specific sub-bands of the spectrum where changes were expected to be more pronounced (for example, spectral contrast was computed between 1.6-3.2 kHz, and also between 3.2-6.4 kHz). These sub-band frequencies were manually selected based on exploratory data analysis. Praat Parselmouth routines were used to estimate properties of the acoustic formants, which represent resonances of the vocal tract. For each segmented sound, the mean formant frequency, standard deviation of formant frequency, and formant bandwidth were computed. Manual review revealed that estimates of formants 2 and 3 were sometimes unreliable. Therefore, we only used estimates derived from formant 1, denoted as F1.

D. Analysis of Self-Reported Symptoms

To describe changes in self-reported symptoms vs. time, we summed the self-reported scores in the various symptom categories (all, congestion-related, and non-congestion related), with symptoms grouped as shown in Table I. The sum of all symptoms is referred to below as the Composite Symptom Score (CSS). We then fit an exponential decay model ($score \sim a\exp(-b(day-1))+\epsilon$) using a nonlinear mixed effect models with subject as a random effect, using R version 3.6.3 and nlme library version 3.1.144. Model fits and residuals were examined for goodness of fit. The model parameter of primary interest is b, the decay rate which captures changes in symptoms. Separate analyses were done for each symptom group so we could explore acoustic markers of different types of respiratory illness (phenotypes).

E. Selection of Informative Acoustic Features and Phonemes

We next analyzed the acoustic features described above, with the goal of defining a distance metric that would characterize the changes in acoustic features over time. Down-selection procedures described below were used for identifying a subset of phonemes, and a set of acoustic features for each phoneme, which would optimize the correlation between this distance metric and self-reported symptom recovery. Supplementary Materials Fig. 1 shows a detailed outline of the process; italicized text in the discussion below refers to rows of the figure.

Analysis of in-lab data: After computing features, we transformed them for normality and tested to identify features that may change during recovery. As this was only the first round of feature selection, our selection criteria were permissive in order to retain all potentially useful features. The Shapiro-Wilk test [27] was used to select the transformation (square root, log, or none) yielding the most normally-distributed data (see Supp. Mat. Table 1 for selected transformations). Paired T-tests were used to identify the transformed features that changed between sick and well visits. Because not all participants reported full symptom recovery, we identified a “high-recovery” subgroup of participants with above-median decrease in self-reported symptoms (i.e. $b > median(b)$, from the decay model above), and accepted features with a significant ($p < 0.05$) change in either the full or high-recovery group, for any phoneme. To reduce sensitivity to recording device, we required changes to be detected in both Samsung and Audacity data. Finally, for highly correlated (Pearson $\rho > 0.9$) features, we picked the feature with the most significant change.

At-home data, preprocessing: The candidate acoustic features identified from the step above were then computed for at-home Samsung recordings. All acoustic features (23 features for 7 phonemes) were computed for all phonemes and standardized across the dataset, so that each feature had zero mean and unit variance. For consistency and because of a higher rate of missing values for recordings made during the morning hours, only recordings made during the evening hours were used. Participants were dropped if more than 20% of the data was missing. Missing values for the remaining participants were imputed as shown in Supplementary Materials Fig. 1.

At-home data, phoneme and feature selection: Identification of the best phoneme and feature combination was done in two steps. We first performed an exhaustive search to identify the best phoneme combination (out of 127 possible combinations of seven phonemes). To do this, we used principal component analysis (PCA) to compute the first 6 principal components (capturing $> 50\%$ of variance) for each phoneme combination and calculated the Euclidean distance between the vectors representing acoustic features on each pair of days. We then computed Spearman's rank correlation between the distance metric for each day relative to the final at-home day (representing the well state) and self-reported symptom ratings. Favorable phoneme combinations were those for which a) the median Spearman's correlation was high and b) the 25th percentile correlation score was also high. Thus, the phoneme combination with the lowest coefficient of quartile variation (IQR/median) was selected as the best phoneme combination. Next, we performed unsupervised feature selection by applying Sparse PCA [28] to the feature space of the best phoneme combination to further reduce dimensionality of the dataset. The choice of unsupervised feature selection was motivated by the fact that while self-reported symptoms can be useful for understanding the direction of change, they are not very reliable as an absolute measure of illness. Features in the first 30 principal components (determined empirically) with a non-zero weight were selected for further analysis and represented the best phoneme and feature combination.

Finally (last row of Supplementary Materials Fig. 1) we assessed the sensitivity of distance metric calculated on the feature space represented by the best phoneme and feature combination to detect changes in self-reported symptom severity. We calculated area under the curve (AUC) of the receiver operating characteristic (ROC) for a Gaussian naïve Bayes classifier trained to differentiate between all pairs of days with a score of zero (i.e. well state) and all pairs of days that represent a change from zero (i.e. well state to sick state). We hypothesized that change in the distance metric would be greater for pairs of days that had a large change in symptoms.

III. Results

109 individuals were screened by telephone based on self-reported respiratory symptoms prior to the first lab visit. Of these, 56 were deemed ineligible based on inclusion/exclusion criteria. 21 were subsequently deemed ineligible for the following reasons: three declined participation because of data collection procedures, five recovered between contact and time they came to the lab, two lived too far away from the study site, three had no prior experience with smart-phones and no caretaker to help, one had a BMI above the required range, four had symptoms most likely due to seasonal allergies/other pathology, two had very mild symptoms and one declined to provide information necessary for enrolling in the study. 32 individuals were enrolled (24 female) in the study, carried out between March 8th, 2018 and May 21st, 2018. The median age of participants was 33 years (range 21-80). At screening, enrolled participants self-reported being sick for a median of 3.5 days (range 1-30 days) with upper respiratory symptom frequency as follows; 32 reported a cough, 30 reported sniffling, 28 reported needing to clear their throat, and 25 reported sneezing. Three participants were excluded from analysis: one participant dropped out of the study and two did not capture any voice data at home.

A. Selected Acoustic Features From In-Lab Data

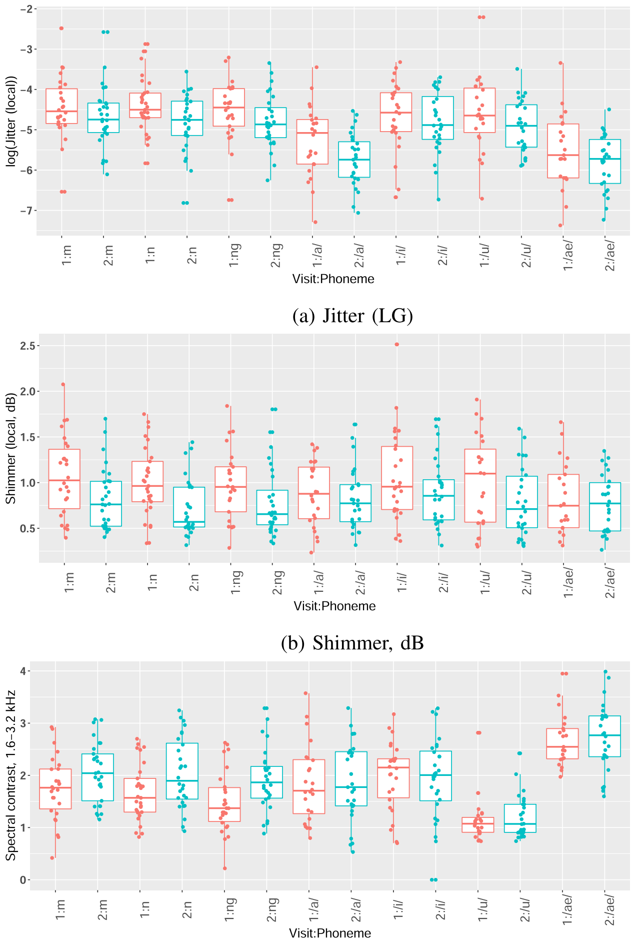
Multiple facets of voice were observed to change during recovery. In addition to those shown in Table II, the log-transformed RMS amplitude demonstrated a significant change between lab visits. However, amplitude was not selected for further analysis because recordings were uncalibrated and mouth-to-phone distance was likely to be variable during at-home recordings. Further, VLHR narrowly missed our inclusion criteria ($p = 0.07$ in Audacity data, $p < 0.05$ in Samsung), but was included as a candidate because past studies had demonstrated its usefulness [9], [10]. We also tested whether changes in acoustic features depended on age, sex or BMI, and did not find any dependence, suggesting that our strategy of comparing changes within each subject helps remove dependence on these factors.

Fig. 2 shows changes in three example acoustic features, chosen to represent several aspects of voice. Jitter (a measure of pitch instability) and shimmer (a measure of amplitude instability) decreased during recovery across all phonemes, indicating that participants have better voice stability after recovery. Spectral contrast at higher frequencies increases for nasal sounds, consistent with nasal resonances becoming more pronounced as congestion reduces in recovery.

B. Correlation Between Voice Features and Symptom-Based Illness Phenotypes In-Lab Recordings

While all individuals were screened for flu-like symptoms prior to enrollment, their baseline condition and their degree of recovery was variable. We hypothesized that participants may fall into different illness phenotypes, and also that congestion should have a noticeable effect on voice. As discussed in Methods and shown in Table I, we grouped self-reported symptoms based on clinical input into two illness phenotypes (congestion-specific and all others) and fitted decay models to symptom phenotype scores for each participant. Histograms of decay constants fitted to the Composite Symptom Score (CSS) and other subscores (congested, non-congested) are shown in Fig. 3. While the overall cohort clearly reports improvement, there is noticeable variability in the recovery profiles of self-reported symptoms.

Fig. 4 shows the correlation between acoustic features and self-reported symptom recovery, with separate decay constants (b values from Fig. 3 computed for the CSS (i.e. all symptoms), the sum of all congestion-related symptoms, and the sum of all non-congestion-related symptoms. Spearman correlation coefficients were computed, and all correlation values with a trend towards significance ($p < 0.1$) are shown as a function of symptom group. Absolute values of correlation are plotted; for



(a) Jitter (LG)

(b) Shimmer, dB

(c) Spectral contrast, 1.6-3.2 kHz

Fig. 2. Representative acoustic feature changes from visit 1 (sick, in red) to visit 2 (well, in blue) for in-lab data based on Samsung recordings for all phonemes. Jitter (local) (a) and shimmer (local, dB) (b) metrics decrease in recovery for all phonemes, while spectral contrast at higher frequencies (c) increases for nasal sounds. The tick labels for X-axis are formatted as visit:phoneme (e.g. 1:m).

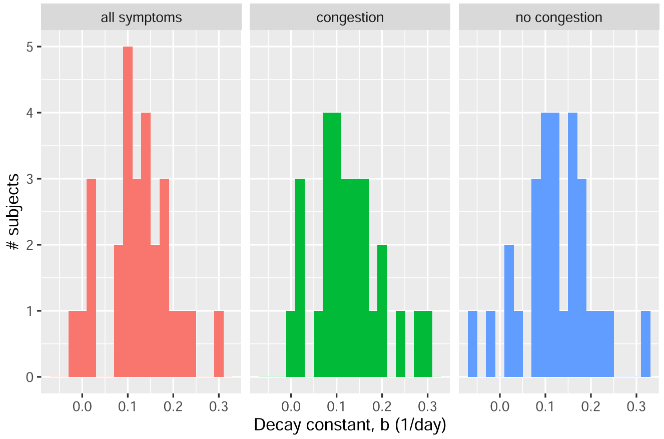


Fig. 3. Histograms of decay constant values for the three symptom phenotypes (i.e. all symptoms, congestion symptoms and non-congestion symptoms) listed in Table I. Positive values correspond to a decrease in symptoms; zero values correspond to no change; and negative values correspond to a worsening of symptoms.

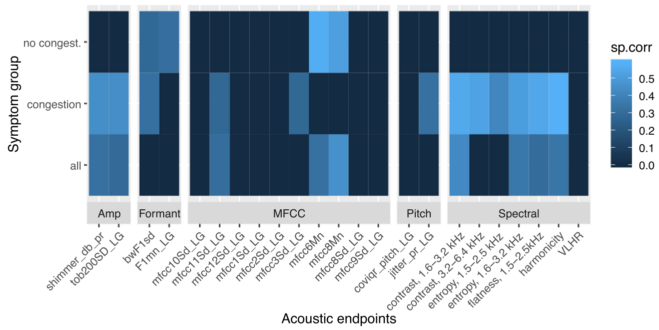


Fig. 4. Patterns of acoustic feature correlation with self-reported symptom change. Correlation is computed between decay constants from exponential models fit to symptoms vs. change in lab measurement for each acoustic feature (visit 2 – visit 1). Changes in self-reported congestion correlate most strongly with spectral features, but also with amplitude instability features (shimmer and 1/3 octave power standard deviation). Formant-related metrics and several MFCC metrics appear correlated with other (non-congestion) aspects of recovery.

most features the direction of correlation is the same between symptom groups. An exception is the standard deviation of formant 1 bandwidth (`bw1sd`), which is positively correlated with non-congestion symptoms but negatively correlated with congestion symptoms (and thus uncorrelated with all summed symptoms). We observe a stronger correlation between changes in higher-frequency spectral structure and changes in self-reported symptoms associated with the congestion phenotype compared to the non-congestion phenotype. A full set of tabulated correlation coefficients can be found in the Supplementary Materials.

C. Recovery-Related Acoustic Changes in At-Home Data

As described in Methods, we extracted the candidate acoustic features in Table II from at-home recordings and derived a distance metric which quantifies the change in the acoustic feature space between any pair of days for a given participant. For analyzing changes in acoustic features in response to recovery, we used the distance metric between the last at-home recording (assumed to be a “well day” because symptoms for all but two participants had resolved significantly by the last at-home day) and each preceding day, which was computed as a measure of change from the ‘well’ state. An exhaustive search was performed to determine which subset of phonemes resulted in the best correlation. The phoneme combination /n/, /m/ and /a/ gave the lowest value (0.34) of the coefficient of quartile variation (for comparison, the median value was 0.72 and max was 1.3) and were therefore selected for further analysis. A second round of down-selection was performed using Sparse PCA to identify a subset of acoustic features for each of the three phonemes, which resulted in a total of 32 features (listed in Table III). The final feature set included 12 features from /n/, 12 features from /m/ and eight features from /a/.

Fig. 5 shows the rank correlation for each participant between the distance metric (computed using 32 features derived from 3 selected phonemes) and the CSS. Because participants varied in their degree of recovery, they are sorted in order of increasing

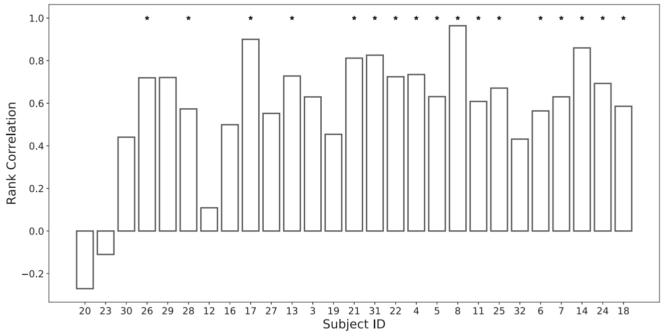
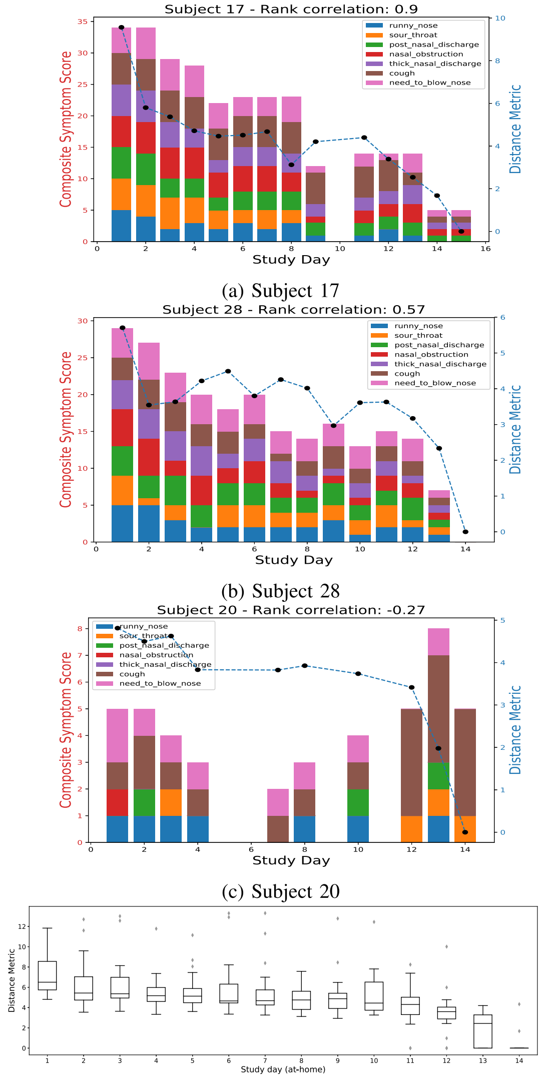


Fig. 5. Per-participant rank correlations between distance measure and self-reported symptoms. Participants are sorted left to right in order of greatest change in symptoms. * Indicates that rank correlation was statistically significant ($p < 0.05$).

value of b using the decay model fit to CSS as described earlier. Correlations are generally higher for participants who exhibited a more rapid recovery (higher values of b). The average rank correlation for participants with b value above median was $0.7(\pm0.13)$ compared to $0.46(\pm0.33)$ for participants with a b value below median. Overall, the correlation between distance metric and CSS was moderate (median: 0.63) with significant variability between participants.

Fig. 6(a)–(c) shows representative time courses of three participants with different recovery profiles. Panels a-c compare self-reported symptoms per day (bars) to the distance metric relative to the last day (dashed blue line). Participant 17 (Fig. 6(a)) is an example of a participant who showed a significant and relatively monotonic reduction in symptoms over the course of the study, which was reflected in the distance metric. For Participant 28 (Fig. 6(b)), the reduction in symptoms was more gradual and less monotonic compared to Participant 17, and the recovery appears to stabilize around day 7-12 before a slight drop on day 13. Agreement with the distance metric is moderate, but we can still observe a transition from illness to recovery. In contrast to these two participants, the self-reported symptoms for Participant 20 (Fig. 6(c)) were mild (CSS $=$ 5 on day 1) at study onset, and non-congestion symptoms (cough and sore throat) worsened over time. Consequently, the agreement with the distance metric was poor. Fig. 6(d)) shows a boxplot (across participants) of the distance metric from the last day (assumed to be a "well day"). In this case a roughly monotonic decrease in distance is seen across the population, reflecting an overall trend of participants recovering from respiratory illness and improvement in symptoms.

In practice, an important goal for monitoring health status would be the detection of changes from sick to well, or vice versa. Fig. 7 shows an approach for characterizing the ability of the distance metric to detect the magnitude of change in self-reported symptoms. Fig. 7(a)) shows the distance metric as a function of score difference (thus a change from 14 to 15, or 5 to 4, or 0 to 1 all correspond to a score difference of 1) across all pair of days for all participants. Fig. 7(b)) shows receiver operating characteristic (ROC) curves for detecting each change (as compared to no change or a score difference of 0). The area under the curve (AUC) metric is 0.89 for a 7-point change and



(a) Subject 17

(b) Subject 28

(c) Subject 20

(d) All subjects

Fig. 6. Distance metric for individual participants (a)–(c), comparing the distance from the final day (lines) vs. CSS (bars). Participant 17 and 28 had excellent and moderate rank correlation coefficients respectively, while participant 20 had poor correlation. (d) Distance metric vs. day, across all participants. Note that distance decreases as participants approach 'well' state, typically around 14 days.

a more moderate 0.72 for detecting a 3-4 point change. While performance should be checked on a larger dataset, this result suggests the distance metric is better able to detect changes in illness as the magnitude of change in symptoms increases.

IV. Discussion

In this paper, we analyzed changes in acoustic features of sustained phonation during the course of recovery from acute respiratory illness. Data collected under two conditions were analyzed; in-lab voice recordings from sick and well lab visits, and at-home voice recordings made between the lab visits. Both in-lab and at-home data shows that sustained phonations of nasal consonants and cardinal vowel appear to carry useful cues of recovery. In both vowels and consonants, measures of

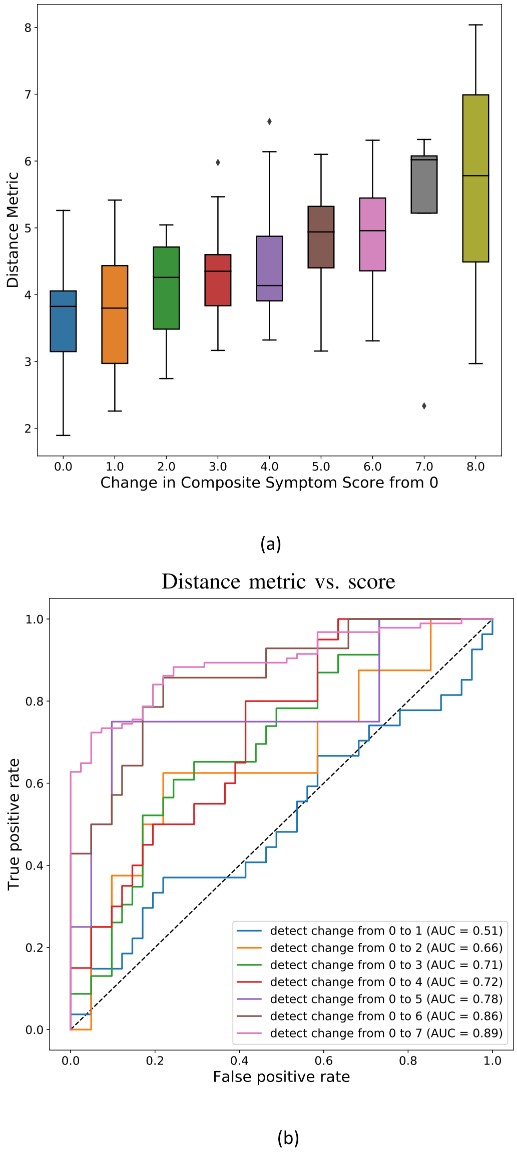


Fig. 7. Quantification of ability to detect changes in the self-reported CSS. (a) distance metric changes vs. change in self-report score, showing that as the difference in self-reported status on a given day increase, the distance between acoustic measures also increases. (b) ROC curves and associated AUC values for detecting changes of different magnitude in the CSS (vs. no change).

pitch variability reduce in recovery, as does amplitude variability as measured shimmer and other metrics. Spectral cues appear to capture congestion-induced changes in nasal resonances, as shown in Fig. 4.

Furthermore, we derived a distance metric for tracking voice changes that tracks well (AUC > 0.7 for detecting a 3 to 4-point change in CSS) with self-reported symptom scores. This is especially encouraging because self-report of congestion is known to be quite variable, leading to difficulties in establishing a statistically significant relationship between objective (e.g. imaging-derived) and self-reported congestion measures [20]. Research suggests that the *changes* in subjective congestion ratings may be more reliable than the absolute congestion ratings [20]. In a highly powered study (N $>$ 2000), Kjaergaard [29] reported correlation between objective measures of congestion and a score based on changes in subjective measurements. This greater reliability of changes in self-reported symptom ratings (rather than absolute ratings) supports our use of Spearman's ranked correlation coefficient, which measures ordering of rankings instead of absolute values.

We carried out feature down-selection to help understand which phonemes and acoustic features were most informative. Identifying the most useful phonemes has an important practical benefit, as recording fewer phonemes simplifies data collection and reduces participant burden by requiring them to complete fewer voice tasks. In addition, while machine learning approaches can employ very large numbers of features [3], [30], this can reduce interpretability. Therefore, we were motivated to reduce the number of features in part to gain insight into which features were most important.

While levels of background noise were acceptable in our experiment, future lab-based studies would benefit from careful experimental characterization of the noise environment (and when possible, reduction of ambient noise) before data collection, as well as segmentation approaches for separating phonations from background speech or noise. Particularly for home-based studies (where the environment is less controlled), it is important to select acoustic features that are robust to environmental noise as well as variability during the data capture process (e.g. distance between mouth and microphone). As noted above, we focused on higher-frequency bands (1.5 – 6.4 kHz) for many of the spectral structure metrics used (spectral contrast, entropy, etc.); this increases robustness by removing lower-frequency background noise, while also focusing on frequencies that are expected to be most impacted by congestion changes [9], [10]. However, further investigation of robust features is warranted in a larger and more diverse population.

Shimmer and jitter are well understood to be robust metrics as participants are not able to consciously control these voice characteristics. In contrast, we chose not to include overall signal power (RMS amplitude) in our at-home data analysis, even though it appeared to increase significantly between lab visits. Because our protocol did not control for mouth-to-microphone distance, we felt amplitude would not be a robust feature. Future studies would benefit from protocols for at-home recording that would ensure consistent speaker-microphone distance during recording, allowing use of voice amplitude as a feature. Similarly, we noted that several participants were clearly controlling the pitch of their phonemes in a musical fashion (for example, 'singing' the phonemes in an ascending scale). Thus, while voice pitch has been shown to be an important cue for conditions such as Parkinson's [25], we felt that it might not be very robust in our sustained phonation study; a limitation of our work is that we did not explore data from scripted speech where pitch and other cues may be more robust.

Our study has several limitations. Our subjects' disease status was established through self-report instead of clinical exams. Subject recruitment was not balanced by age or sex, making it difficult to assess impacts of these factors (although our approach analyzes changes in each subject relative to the subject's baseline data, which should reduce sensitivity to these effects).

Our sample size was small, so performance estimates reported here should be checked in a larger study. While the results here suggest that acoustic features derived from voice could be useful for tracking recovery from respiratory illness, in many cases (for example vaccine clinical trials or pandemics) an key goal would be detecting the onset of illness. These weaknesses could be addressed by studies in larger cohorts to capture variability of voice during healthy baseline, which may lead to approaches for detecting the onset of illness as well as monitoring the progression of symptoms.

Acknowledgment

The authors would like to thank Adan Rivas, Hao Zhang, and Mar Santamaria of the Pfizer Digital Medicine and Translational Imaging Group, Cambridge, MA, USA, for their contributions to study design and data preparation. They would also like to thank Anne O'Brien, Gregory Schiurring, Eric Fabara, as well as members of the Motion Analysis Laboratory at Spaulding Rehabilitation Hospital, Boston, MA, USA, for their contributions to data collection. They would also like to thank Lukas Adamowicz and Isik Karahanoglu for reviewing the code and data analysis steps. They would also like to thank Xuemei Cai, David Caouette, Joe Mather, and Timothy J. McCarthy of the Pfizer Digital Medicine and Translational Imaging Group, Cambridge, MA, USA, for reviewing and providing feedback on the manuscript.

References

[1] D. D. Matías, M. Zañartu, S. W. Feng, H. A. I. Cheyne, and R. E. Hillman, "Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform," *IEEE Trans. Biomed. Eng*., vol. 59, no. 11, pp. 3090–3096, Nov. 2012.

[2] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinsons disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng*., vol. 57, no. 4, pp. 884–893, Apr. 2010.

[3] J. M. Tracy, Y. Özkanca, D. C. Atkins, and R. H. Ghomi, "Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease," *J. Biomed. Informat*., vol. 104, Apr. 2020, Art. no. 103362.

[4] M. Faurholt-Jepsen *et al.*, "Voice analysis as an objective state marker in bipolar disorder," *Transl. Psychiatry*, vol. 6, no. 7, Jul. 2016, Art. no. e856.

[5] J. C. Hailstone *et al.*, "Voice processing in dementia: A neuropsychological and neuroanatomical analysis," *Brain*, vol. 134, pp. 2535–2547, 2011.

[6] J. D. S. Sara *et al.*, "Non-invasive vocal biomarker is associated with pulmonary hypertension," *PLoS One*, vol. 15, no. 4, Apr. 2020, Art. no. e0231441.

[7] C. R. Marmar *et al.*, "Speech-based markers for posttraumatic stress disorder in US veterans," *Depression Anxiety*, vol. 36, no. 7, pp. 607–616, Jul. 2019.

[8] T. F. Quatieri, T. Talkar, and J. S. Palmer, "A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 203–206, May 2020.

[9] G.S. Lee, C. C. H.Yang, C. P. Wang, and T. B. J. Kuo, "Effect of nasal decongestion on voice spectrum of a nasal consonant-vowel," *J. Voice*, vol. 19, no. 1, pp. 71–77, Mar. 2005.

[10] G. Lee, C. H. C.Yang, and T. B. J. Kuo, "Voice low tone to high tone ratio-a new index for nasal airway assessment," *Chin. J. Physiol*., vol. 46, no. 3, pp. 123–7, Sep. 2003.

[11] N. Attuluri and M. Pushpavathi, "Correlation of perceived nasality with the acoustic measures (one third octave spectral analysis & voice low tone to high tone ratio)," *Glob J. Otolaryngol.*, vol. 8 no. 3, 2017, doi: 10.19080/GJO.2017.08.555737.

[12] A. P. Vogel, H. M. Ibrahim, S. Reilly, and N. Kilpatrick, "A comparative study of two acoustic measures of hypernasality," *J. Speech, Lang. Hear. Res*., vol. 52, no. 6, pp. 1640–1651, 2009.

[13] M. Novotny, J. Rusz, R. Čmejla, H. Růžičková, J. Klempíř, and E. Růžička, "Hypernasality associated with basal ganglia dysfunction: Evidence from Parkinson's disease and Huntington's disease," *PeerJ*, vol. 2016, no. 9, Sept. 2016, doi: 10.7717/peerj.2530.

[14] R. E. Hillman, J. T. Heaton, A. Masaki, S. M. Zeitels, and H. A. Cheyne, "Ambulatory monitoring of disordered voices," *Ann. Otol., Rhinol. Laryngol.*, vol. 115, no. 11, pp. 795–801, 2006.

[15] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.,* vol. 5, no. 9, pp. 341–345, 2001.

[16] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern*., vol. 9, no. 1, pp. 62–66, Jan. 1979.

[17] B. Meléndez-Catalán, E. Molina, and E. Gómez, "BAT: An open-source, web-based audio events annotation tool," in *Proc. Web Audio Conf.*, 2017, pp. 1–4.

[18] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A Python interface to praat," *J. Phonetics*, vol. 71, pp. 1–15, Nov. 2018.

[19] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *J. Neurolinguistics*, vol. 20, no. 1, pp. 50–64, Jan. 2007.

[20] M. J. Schumacher, "Nasal congestion and airway obstruction: The validity of available objective and subjective measures," *Curr. Allergy Asthma Rep.*, vol. 2, no. 3, pp. 245–251, 2002.

[21] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Inst. Phonetic Sci.*, vol. 17, no. 1193, Tech. Rep., 1993, pp. 97–110.

[22] J.-L. Shen, J.-W. Hung, and L.-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proc. 5th Int. Conf. Spoken Lang.*, 1998, pp. 1–4.

[23] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, "Music type classification by spectral contrast feature," in *Proc. IEEE Int. Conf. Multimedia Expo.*, vol. 1, 2002, pp. 113–116.

[24] N. Madhu, "Note on measures for spectral flatness," *Electron. Lett.*, vol. 45, no. 23, pp. 1195–1196, 2009.

[25] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," *J. Acoust. Soc. Amer*., vol. 129, no. 1, pp. 350–367, 2011.

[26] B. Mcfee *et al.*, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, vol. 8, 2015, pp. 18–25.

[27] J. P. Royston, "An extension of Shapiro and Wilk's W. test for normality to large samples," *Appl. Statist*., vol. 31, no. 2, pp. 115–124, 1982.

[28] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Statist*., vol. 15, no. 2, pp. 265–286, Jun. 2006.

[29] T. Kjaergaard, M. Cvancarova, and S. K. Steinsvåg, "Nasal congestion index: A measure for nasal obstruction," *Laryngoscope*, vol. 119, no. 8, pp. 1628–1632, Aug. 2009.

[30] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2016.