Learning to Detect Vocal Hyperfunction From Ambulatory Neck-Surface Acceleration Features: Initial Results for Vocal Fold Nodules

Marzyeh Ghassemi*, Jarrad H. Van Stan, Daryush D. Mehta, *Member, IEEE*, Matías Zañartu, *Member, IEEE*, Harold A. Cheyne II, Robert E. Hillman, and John V. Guttag

***Abstract*—Voice disorders are medical conditions that often result from vocal abuse/misuse which is referred to generically as vocal hyperfunction. Standard voice assessment approaches cannot accurately determine the actual nature, prevalence, and pathological impact of hyperfunctional vocal behaviors because such behaviors can vary greatly across the course of an individual's typical day and may not be clearly demonstrated during a brief clinical encounter. Thus, it would be clinically valuable to develop noninvasive ambulatory measures that can reliably differentiate vocal hyperfunction from normal patterns of vocal behavior. As an initial step toward this goal we used an accelerometer taped to the neck surface to provide a continuous, noninvasive acceleration signal designed to capture some aspects of vocal behavior related to vocal cord nodules, a common manifestation of vocal hyperfunction. We gathered data from 12 female adult patients diagnosed with vocal fold nodules and 12 control speakers matched for age and occupation. We derived features from weeklong neck-surface acceleration recordings by using distributions of sound pressure level and fundamental frequency over 5-min windows of the acceleration signal and normalized these features so that intersubject comparisons were meaningful. We then used supervised machine learning to show that the two groups exhibit distinct vocal behaviors that can be detected using the acceleration signal. We were able to correctly classify 22 of the 24 subjects, suggesting that in the future measures of the acceleration signal could be used to detect patients with the types of aberrant vocal behaviors that are associated with hyperfunctional voice disorders.**

***Index Terms*—Ambulatory voice monitoring, clinical detection, machine learning, vocal cord, vocal fold nodules.**

I. Introduction

THIS paper presents initial results from an ongoing project that is aimed at developing ambulatory monitoring of laryngeal voice production (phonation) into an effective clinical tool. In particular, we hope to improve the assessment of voice disorders associated with daily vocal abuse/misuse (e.g., yelling) referred to generically as vocal hyperfunction. Vocal nodules (depicted in Fig. 1) are one of the common manifestations of vocal hyperfunction that arise secondarily to chronic tissue trauma on the surface of the vocal cords (folds).

The primary role of vocal hyperfunction in this diagnosis is difficult to determine because the associated vocal fold pathology is most probably the result of long-standing and inconsistent behaviors whose effect is cumulative over time. In addition, once pathological changes in vocal fold tissue have taken place, their very presence alters vocal function in ways that require additional effort to maintain phonation. Thus, separating primary hyperfunction (initial cause of the disorder) from reactive hyperfunction (reaction to the presence of pathology) is not easily achieved.

The work reported here is part of an ongoing project intended to gain insight into these complex relationships by analyzing data collected from an accelerometer (ACC) placed on the neck. In this study, we analyzed data from patients suffering from vocal hyperfunction with associated vocal fold nodules. In future studies, we intend to also examine subjects who have had nodules surgically removed, but before they have undergone voice therapy (vocal retraining). Clinical experience and previous work [2] suggest that these subjects will display persistent postsurgical hyperfunctional behaviors that are not confounded by the presence of vocal fold pathology and would eventually result in a recurrence of vocal nodules if the behaviors are not ameliorated by voice therapy. In this paper, we show that by using supervised machine learning techniques we can build a

Manuscript received October 10, 2013; revised December 19, 2013; accepted December 21, 2013. Date of publication January 2, 2014; date of current version May 15, 2014. This work was supported in part by the National Library of Medicine's university-based Biomedical Informatics Research Training Program, in part by the Intel Science and Technology Center, and in part by the National Institutes of Health National Institute on Deafness and Other Communication Disorders under Grant R33DC011588, the Chilean CONICYT under Grant FONDECYT 11110147, and the MIT International Science and Technology Initiatives MIT-Chile Seed Fund under Grant 2745333. *Asterisk indicates Corresponding author.*

*M. Ghassemi is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: mghassem@mit.edu).

J. H. Van Stan is with the Center for Laryngeal Surgery and Voice Rehabilitation and Institute of Health Professions, Massachusetts General Hospital, Boston, MA 02114 USA (e-mail: JVanStan@mghihp.edu).

D. D. Mehta is with the Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, Boston, MA 02114 USA and also with the Department of Surgery, Harvard Medical School, Boston, MA 02115 USA (e-mail: daryush.mehta@alum.mit.edu).

M. Zañartu is with the Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso 2390123, Chile (e-mail: matias.zanartu@usm.cl).

H. A. Cheyne II is with the Bioacoustics Research Program, Laboratory of Ornithology, Cornell University, Ithaca, NY 14850 USA (e-mail: hac68@cornell.edu).

R. E. Hillman is with the Center for Laryngeal Surgery and Voice Rehabilitation and Institute of Health Professions, Massachusetts General Hospital, Boston, MA 02114 USA, and also with the Department of Surgery, Harvard Medical School, Boston, MA 02115 USA (e-mail: hillman.robert@mgh.harvard.edu).

J. V. Guttag is with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: guttag@csail.mit.edu).

Digital Object Identifier 10.1109/TBME.2013.2297372

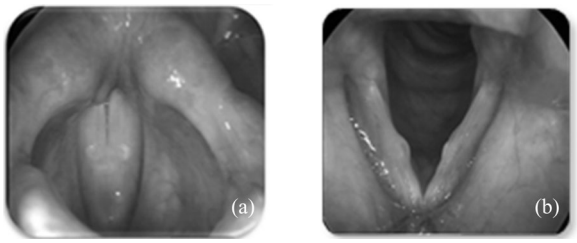

Fig. 1. Endoscopic images of the larynx of (a) a normal subject and (b) a subject with nodules. The vocal folds for (b) are shown in an open state to illustrate the location of vocal fold nodules that develop due to tissue trauma.

classifier to distinguish between patients with vocal nodules and matched control subjects.

Section II presents clinical background relevant to the problem of tracking voice use–related measures. Section III describes our data collection methods, novel features, and classification algorithms. Section IV summarizes the classification performance of a logistic regression model and a support vector machine (SVM). We discuss the contributions of particular features for potential clinical application in Section V, list limitations in Section VI, and summarize our work in Section VII.

II. Background

A. Clinical Background on Vocal Hyperfunction

An estimated 6.6% of the United States' working-age population suffers from a voice disorder at any given point in time [1]. Such disorders are caused by a malfunctioning of the vocal folds in the larynx and can have a devastating effect on the ability of individuals to speak, sing, and/or project their voice. Associated economic losses can be significant because of increases in healthcare costs and reductions in occupational productivity.

Many voice disorders are chronic or recurring conditions that are likely to result from faulty and/or abusive daily patterns of vocal behavior, referred to generically as *vocal hyperfunction*. While the general impact of hyperfunctional disorders on vocal function has been described based on laboratory studies [2], [3] and clinical evaluations, critical aspects of voice use may not be captured in such brief assessment periods. Clinicians employ an assessment protocol that includes a patient's self-report and self-monitoring, which are subjective and prone to unreliability when assessing the prevalence and persistence of vocal behaviors during diagnosis and management [4], [5]. The diagnosis of some common voice disorders could be enhanced by the ability to unobtrusively monitor and quantify hyperfunctional vocal behaviors as individuals go about their normal daily activities.

There are two types of vocal hyperfunction that can be differentiated from each other and from normal voice production: 1) *adducted hyperfunction*, which contributes to chronic vocal fold tissue trauma and the formation of growths, such as nodules, that may cause a disordered voice quality [see Fig. 1(b)]; and 2) *nonadducted hyperfunction*, which contributes to vocal fatigue and a disordered voice quality in the absence of vocal fold tissue damage [2].

B. Smartphone-Based Voice Health Monitoring

Our group recently reported on the development of a platform for unobtrusive, noninvasive ambulatory voice monitoring that uses a neck-placed miniature ACC as the phonation sensor and a smartphone as the data acquisition platform [6]. This device collects the unprocessed ACC signal and daily calibration recordings from speakers. The raw ACC signal is collected at an 11 025-Hz sampling rate, 16-bit quantization, and 80-dB dynamic range to obtain frequency content of neck surface vibrations up to 5000 Hz.

ACC data are preferable to acoustic recordings because 1) continuous daily recording of the acoustic signal raises privacy concerns, 2) the ACC signal is less affected by external acoustic noise sources [7], and 3) the ACC signal captured below the larynx is easier to analyze than the oral signal because the resonances of the respiratory system are relatively time-invariant compared to the vocal tract resonances that are continuously altered during speech production by movements of the articulators (tongue, lips, and jaw) [8].

C. Acoustically Inspired Measures of Vocal Hyperfunction

Previous ACC-based ambulatory monitoring approaches have estimated basic acoustic parameters to quantify voice use, including fundamental frequency (F0, related to pitch), sound pressure level (SPL, related to loudness), and phonation time [9], [10]. Analogous to noise dosimetry for hearing health, vocal dose measures incorporate F0 and SPL over long periods of time to estimate the total exposure of vocal fold tissue to potentially damaging forces (e.g., shear stresses) associated with vibration [10].

The prevailing view is that vocal fold tissue can be damaged by accumulating cycles of vibration and repeated collision between the two vocal folds without sufficient recovery time [11]. However, prior preliminary results suggest that long-term averages of these voice use measures do not capture the difference between high-voice users with disorders such as nodules and high-voice users with normal voices [12]. In a similar way, average measures were not significantly different between the paired subjects in the current study. Although average vocal measures have been found to distinguish various types of teachers in occupational and nonoccupational contexts [13]–[16], such differences have not been investigated between individuals with and without voice disorders. Thus, to differentiate patients with nodules from their control subjects in this study, we augmented ACC-derived features to capture both average and extreme characteristics of vocal behavior.

III. Methods

A. Data Collection

All subjects were monitored over the course of one week using the voice health monitor [6]. Data were gathered from 12 pairs of trained adult singers. All subjects were female with an average age of 21.6 years (SD = 2.7 years). The subjects were instructed to wear the device during all waking hours; strict compliance was not a precondition for data inclusion. For

example, if a subject wore the device for only 4 h on one day, we did not exclude data from that day from analysis. Each subject pair comprised a patient diagnosed with vocal fold nodules and a vocally normal subject matched for age (within 5 years). The severity of the nodules and associated abnormalities in voice quality varied across patients. Diagnoses were based on a team evaluation (laryngologist and speech-language pathologist) at the Center for Laryngeal Surgery and Voice Rehabilitation at the Massachusetts General Hospital that included 1) a complete case history, 2) endoscopic imaging of the larynx, 3) voice-related quality of life questionnaire [17], 4) consensus auditory-perceptual evaluation of voice assessment [18], and 5) aerodynamic and acoustic assessments of vocal function.

B. Feature Extraction

After each week of recording, data were downloaded from the smartphone and preprocessed to yield voice use features. Daily accumulation of voice use was quantified by F0, SPL, and three vocal dose measures: phonation time, cycle dose, and distance dose [10]. Phonation time reflects the total duration of voiced frames and is computed in terms of time and percentage of total time. Cycle dose is the number of vocal fold oscillations during a period of time. Distance dose estimates the total distance traveled by the vocal folds, combining cycle dose with F0 and estimates of vibratory amplitude based on SPL.

We took a similar approach to prior work in preprocessing the ACC signal by extracting F0 (in Hz) and SPL (in dB SPL) from nonoverlapping frames of 50 ms in duration [6], [19]. For each frame, SPL was computed using calibration factors for each sensor to map ACC level to acoustic SPL. F0 was estimated from the first peak in the time-based autocorrelation function. From these values, we used a simple voice activity detection method that defined voiced frames as consisting of plausible ranges of SPL (62–130 dB SPL) and F0 (50–500 Hz) [20].

Given SPL and F0 point estimates for each voiced frame each day, we segmented these time series into nonoverlapping 5-min windows. We calculated the three vocal dose measures (phonation time, cycle dose, and distance dose) over each window using the F0 and SPL from voiced frames. We then treated the SPL and F0 time series within each window as a distribution and calculated their mean, skewness, kurtosis, 5th percentile, and 95th percentile values. These values (five descriptive statistics for F0 and SPL distributions and three vocal dose measures) formed the 13 "basis features."

In addition, we created 13 "normalized features" to indicate how far subjects strayed from their baseline behavior. Each basis feature was converted into units of standard deviation based on that feature's baseline distribution over an average hour in the first half of the day. For example, the cycle dose distribution in a particular 5-min window would be converted to a normalized cycle dose by subtracting the mean and dividing by the standard deviation of the cycle dose time series over the first half of the day. The 13 normalized features are similar in content to the basis features, but can be compared meaningfully across subjects and days.

FEATURE EXTRACTION ALGORITHM

```
for each subject s:
  for each day d:
    Compute SPL and F0 for each 50 ms frame
    Unvoiced frames receive NaN values
    Group SPL/F0 vectors into 5 min windows w,
    /   each w comprising 1,200 frames
    for each w:
      Compute 13 basis features
      Compute average and SD for each feature
      /  over first half of day
      Compute 13 normalized features
```

Fig. 2. Pseudocode describing the feature extraction procedure. Variables are in italics. The NaN indicators allow unvoiced frames to be ignored in the calculation of distributional features. Features are not computed for windows with less than 10% voiced frames.

TABLE I
DISTRIBUTION OF DATASET

Subject	*Total Windows*	*Percentage of Data*	*Percentage of Data Per Pair*
P01	310	1.84%	7.06%
N01	878	5.22%	
P02	906	5.39%	11.23%
N02	983	5.84%	
P03	677	4.02%	8.54%
N03	761	4.52%	
P04	749	4.45%	7.50%
N04	513	3.05%	
P05	751	4.46%	7.72%
N05	548	3.26%	
P06	775	4.61%	9.20%
N06	772	4.59%	
P07	472	2.81%	7.09%
N07	720	4.28%	
P08	706	4.20%	8.37%
N08	701	4.17%	
P09	803	4.77%	8.13%
N09	565	3.36%	
P10	777	4.62%	8.32%
N10	623	3.70%	
P11	684	4.07%	8.06%
N11	672	3.99%	
P12	706	4.20%	8.78%
N12	771	4.58%	

Fig. 2 summarizes the feature extraction algorithm in pseudocode. The 13 baseline features and 13 normalized features were calculated in every 5-min window with more than 10% of the frames voiced (15 345 windows total) and were combined into a single 26-D feature vector (time-ordering was ignored). See Table I for a detailed listing of the proportion of data available from each subject after frames were discarded. As shown, the retained data tended to be evenly balanced across subject pairs.

C. Discrimination Using Hypothesis Testing

We examined feature correlations across all subjects to determine whether any feature had a Pearson's correlation coefficient

TABLE II
DESCRIPTION OF FEATURES USED IN MODEL FITTING

Feature Name	*Description*	*K–S Statistic (p-value)*
Distance Dose	The distance travelled by the vocal folds within a 5-minute frame.	0.03 ($p < 0.001$)
% Phon	Percent phonation time, calculated within a 5-minute frame.	0.02 ($p = 0.04$)
SPL Skew	Distributional skew of the sound pressure level within a 5-minute frame.	0.10 ($p < 0.001$)
SPL Kurtosis	Distributional kurtosis of the sound pressure level within a 5-minute frame.	0.04 ($p < 0.001$)
SPL 5th Percentile	5th Percentile of the sound pressure level within a 5-minute frame.	0.09 ($p < 0.001$)
SPL 95th Percentile	95th Percentile of the sound pressure level within a 5-minute frame.	0.05 ($p < 0.001$)
F0 Skew	Distributional skew of the fundamental frequency within a 5-minute frame.	0.06 ($p < 0.001$)
F0 Kurtosis	Distributional kurtosis of the fundamental frequency l within a 5-minute frame.	0.06 ($p < 0.001$)
F0 5th Percentile	5th Percentile of the fundamental frequency within a 5-minute frame.	0.10 ($p < 0.001$)
F0 95th Percentile	95th Percentile of the fundamental frequency within a 5-minute frame.	0.10 ($p < 0.001$)
Normalized Distance Dose	Normalized Distance Dose	0.04 ($p < 0.001$)
Normalized Mean F0	Normalized Mean Fundamental Frequency	0.07 ($p < 0.001$)
Normalized SPL Skew	Normalized Sound Pressure Level Skew	0.05 ($p < 0.001$)
Normalized SPL Kurtosis	Normalized Sound Pressure Level Kurtosis	0.06 ($p < 0.001$)
Normalized SPL 5th Percentile	Normalized Sound Pressure Level 5th Percentile	0.05 ($p < 0.001$)
Normalized F0 Skew	Normalized Fundamental Frequency Skew	0.05 ($p < 0.001$)
Normalized F0 Kurtosis	Normalized Fundamental Frequency Kurtosis	0.05 ($p < 0.001$)

higher than 0.9 with at least one other feature. If a pair of correlated features were identified, the feature that was most correlated with the remaining features was eliminated, leaving 17 features (see Table II). We did not use orthonormal dimensionality reduction techniques because we believe that the interpretability of the features is critical for clinical application. Note that the resulting 17-feature vector is not designed to detect underlying physiological problems, but rather to capture hyperfunctional vocal behaviors that are typically associated with vocal fold nodules.

Statistical differences between distributions were tested using the Kolmogorov–Smirnov (K–S) statistic [21] and with a p-value modified with the Bonferroni correction for multiple hypothesis tests ($p < 0.0019$) [22]. The K–S statistic converges to zero for large datasets if the samples come from the same empirical distribution.

D. Classification Using Machine Learning Techniques

We approached our task as a binary classification problem. Each of the 15 345 windows was labeled as positive or negative depending on whether it was associated with a patient or control subject, respectively (51% of windows came from the patient group). This approach ignored the fact that subjects with vocal fold nodules exhibit inconsistent degrees of hyperfunctional behaviors during the day. There may also be instances where vocally normal subjects exhibit some degree of hyperfunctional behavior; but we expect fewer of these based on the lack of vocal pathology.

L1-norm regularized logistic regression [23] and SVM [24] models were trained for the binary classification problem. We used a neutral cost function for classifier training (i.e., the cost of a misclassification is the same regardless of the underlying label). We first divided data using leave-one-out cross-validation to generate 12 datasets, each consisting of 11 training pairs and one test pair. All windows from the 11 training pairs (22 subjects total) were then subdivided using fivefold cross-validation (1/5th validation and 4/5ths training in each fold). The training data in each fold were used to select optimal beta values for the logistic regression model and slack parameters for the soft-margin linear kernel SVMs. From these five trained models (one per fold), the best model was selected based on the highest area under the ROC curve (AUC) performance on the validation set. Pseudocode describing this procedure is given in Fig. 3.

The chosen models were used to classify windows in the test set; in Section IV, we report the test set AUC, F-score, accuracy, sensitivity (Sens), specificity (Spec), positive predictive value (PPV), and negative predictive value (NPV). We also used the models to classify all windows from all subjects with a classification threshold of 0.5. We then examined the proportion of windows classified as positives for each subject to classify each subject as a positive or negative case.

MODEL TRAINING PROCEDURE

for each *subject pair p*:
  All windows from *control*[*p*] and *positive*[*p*] are
  / *test_data[p]*
  Windows from all other subjects are *train_data[p]*
  Separate *train_data[p]* into 5 folds
  for each *fold f*:
    Optimize model parameters on *train_data[p]* in *f*
    Calculate AUC on *train_data[p]* not in *f*
  Choose model from fold with best AUC
  Report model performance on *test_data[p]*

Fig. 3. Pseudocode describing the training and model selection procedure. Variables are in italics and matrix indexing uses brackets.

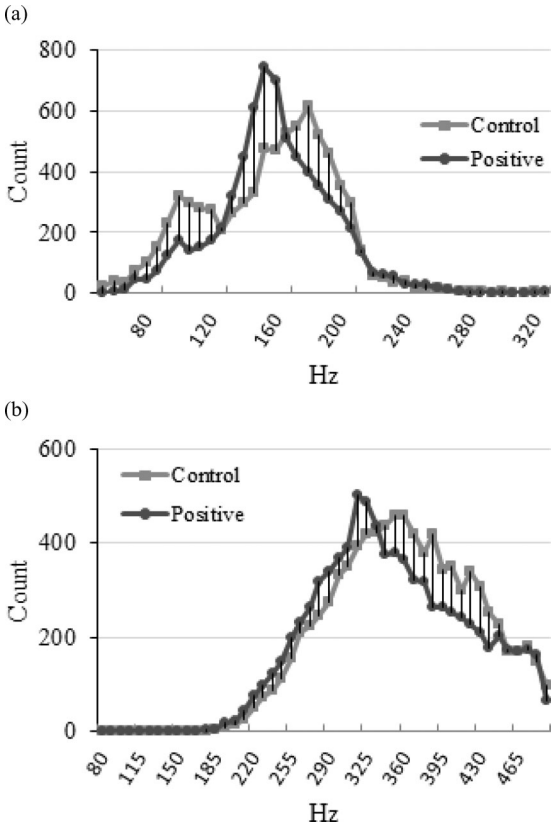

Fig. 4. Distributions of 5-min derived values of (a) F0 5th percentile and (b) F0 95th percentile over all days from all subjects. These features had a probability less than 0.001 of the two populations being drawn from the same distribution (K–S test).

IV. RESULTS

Fig. 4 illustrates differences between the distributions of measures of extreme vocal behavior: F0 5th percentile and F0 95th percentile. There are statistically significant differences in how much the nodule and normal group vary from each other, which is reflected in the K–S statistic reported for the figure. There were distributional differences for the basis features and normalized features between subjects in the nodules group and those in the control group. In our dataset, there was no statistically significant difference between the phonation time of subjects with nodules and those without. It also was not a discriminative feature (using regularization techniques) when building any of the discriminative classifiers used.

Table III summarizes the performance of each measure to classify hyperfunction in the 12 subject pairs. The maximum number that the "association count" field can have is 12. This occurs when that particular variable (row) has a statistically significant effect (p-value < 0.05, absolute average odds ratios ≥ 1.05) in each subject pair during the testing phase. Many associations persisted across all subject pairs rather than in only a few pairs and also tended to agree well on the magnitude of the association. The 95% confidence interval is from the lowest bound across pairs to the highest bound across pairs.

Logistic regression performance had an average AUC of 0.705, F-score of 0.630, and accuracy of 0.660 (Sens. 0.499, Spec. 0.806, PPV 0.719, NPV 0.621) across the twelve subject pairs. The performance of the linear SVM was not substantially different (AUC of 0.708, F-score of 0.650).

After all test data predictions, we applied a classification threshold of 0.5 to the logistic regression model output to determine whether a window was positive or negative. We then examined the proportion of 5-min windows classified as positives in each subject. Applying this method, there is a clear separation between patients and controls as seen in Fig. 5.

To clarify whether a single day of data could be used, we selected only the first day of data available from all subjects and trained classifiers in the same manner as above. We applied the same classification threshold of 0.5 to the logistic regression model output, and only 17 of the 24 were correctly classified. From this result, we believe that subjects do not necessarily produce their most distinguishing behaviors on any given single day a device was used. A primary reason we have chosen a week is to gain a full understanding of subject behaviors and habits as they go about their routines.

V. DISCUSSION

We found several features that could be used for patient identification using our long-term mobile monitoring approach. Model performance indicates that features were more important than the learning technique in decision boundary determination. Many significant differences between the patients and their matched normals appear to be related to extreme, rather than average, behaviors. This is supported by normalized features accounting for several of the most important features, indicating that behavior straying from an individuals' baseline behavior may be more important than their absolute behavior. This is an important difference from previous work that focused on averages to differentiate subjects with and without voice disorders [12].

In general, the PPV was higher than the sensitivity in all subject pairs. As mentioned in Section III-D, our approach to labeling all windows from a patient as positive ignored the fact that subjects with vocal fold nodules may exhibit inconsistent degrees of hyperfunctional behaviors during the day. Based on our low sensitivity and high PPV, we believe that there are roughly three clusters of data: one with data mostly from controls, one with data mostly from patients, and one mixed. A mixed cluster "near" the control cluster with more patients than

TABLE III
MODEL PERFORMANCE ACROSS THE 12 SUBJECT PAIRS, SORTED BY DECREASING MEAN ODDS RATIO

Variable	*Association Counts*		*Mean Multivariate Associations*		
	Hyperfunction	Normal	Beta Mean (SD)	Odds Ratio Mean (95% CI)	SVM Weight Mean (SD)
Normalized Mean F0	12	0	0.37 (0.09)	1.45 (1.21–1.89)	2.83 (0.78)
Normalized SPL Skew	12	0	0.36 (0.07)	1.44 (1.30–1.80)	11.87 (1.30)
SPL Kurtosis	12	0	0.27 (0.09)	1.31 (1.05–1.62)	9.79 (3.25)
F0 Kurtosis	12	0	0.13 (0.03)	1.14 (1.06–1.27)	15.09 (3.22)
Normalized Distance Dose	12	0	0.13 (0.04)	1.14 (1.08–1.32)	3.88 (0.93)
Normalized SPL 5^{th} Percentile	0	12	−0.34 (0.08)	0.71 (0.54–0.86)	−2.61 (0.41)
Normalized SPL Kurtosis	0	12	−0.38 (0.09)	0.68 (0.52–0.84)	−8.35 (2.49)
Normalized F0 Kurtosis	0	12	−0.64 (0.06)	0.53 (0.44–0.63)	−21.46 (1.90)
SPL Skew	0	12	−1.28 (0.21)	0.28 (0.17–0.44)	−6.16 (0.80)

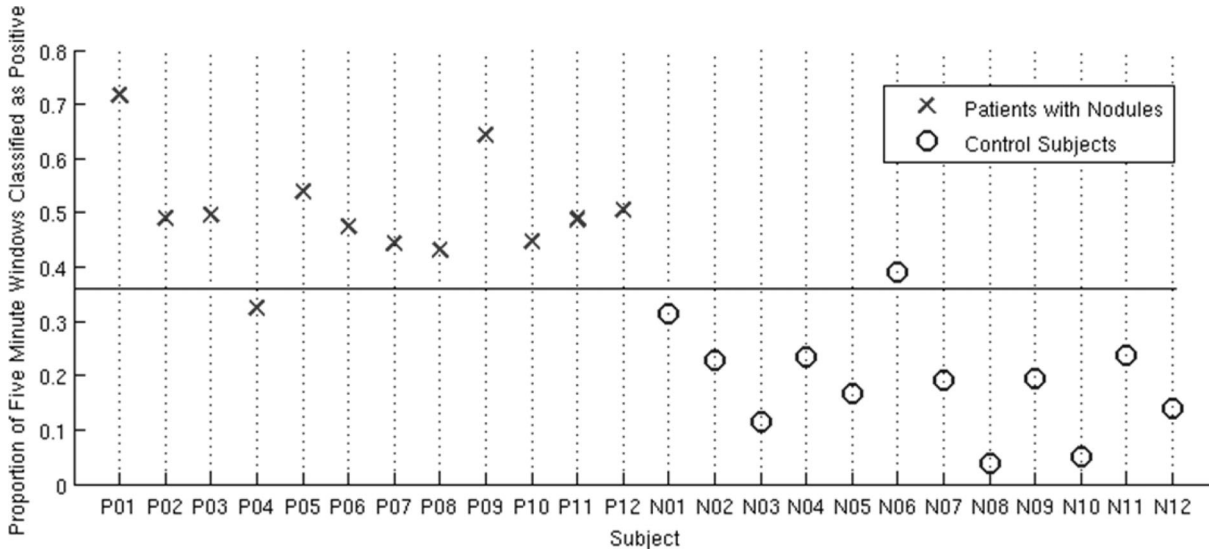

Fig. 5. Proportion of 5-min windows classified as positive by the best logistic regression model. We obtained separation for 22 out of the 24 subjects.

controls would pull the classifier decision boundary toward the positive class.

The simplest reason a feature would be a good predictor of either class is a correlation with the outcome. However, some measures were predictive of vocal fold nodules but were not correlated with the outcome on their own. This indicates that these measures did not have a direct linear relationship with the class label, but instead were important biasing factors once other variables were taken into account.

Subjects with vocal fold nodules, on average, had a normalized F0 mean that was higher relative to the first half of their day. Positive subjects also had a peaky un-normalized F0 distribution (higher F0 kurtosis), and a normalized SPL with a heavier right side tail (higher normalized SPL skew). These observations could be interpreted to mean that subjects with vocal fold nodules tend to deviate from their baseline F0 and SPL as their days progressed, possibly reflecting increased difficulty in producing phonation.

The vocally normal group had an SPL distribution with a heavier tail on the right side (higher SPL skew), and their SPL and F0 levels normalized to the first half of their day had longer, heavier right-hand tails (normalized SPL kurtosis, normalized F0 kurtosis, respectively). This observation could be interpreted to mean that control subjects tended to have brief F0 deviations that were mostly bringing their lower pitches higher, rather than their higher pitches even higher. It also suggests that even when control subjects exhibited higher SPL ranges, they tended to stay within their baseline.

It was striking that some of the most heavily weighted features in predicting the presence of nodules (SPL kurtosis and F0 kurtosis) were mirrored by the corresponding normalized feature being strongly weighted toward the opposite prediction. The opposite was true of SPL skew: the feature itself is associated with vocal normalcy, but the normalized feature was associated with vocal fold nodules. One possible explanation for these results is that patients with vocal fold pathology such as nodules progressively increase muscle activation levels during the course of the day to maintain a functional loudness level in the face of increasing fatigue and dysphonia. This corresponds to the daily "vicious cycle" that has been observed clinically and described in the literature for hyperfunctional voice disorders [2]. Under these conditions, the effort to support inefficient voicing leads to fatigue and a progressive increase in muscle activity that tenses the vocal folds and is reflected by an increase in F0 [25]. The increasing trend of F0 over the day has also been reported previously during controlled, laboratory recordings of

teachers before and after a work day [26] and is hypothesized to be related to vocal fatigue.

Our assessment provides an approach to address the fragility of traditional clinical assessments. Instead of a single vocal snapshot, we are able to capture potentially hyperfunctional behavior that is representative of at-work and/or at-home voice use. Model performance on classifying windows was reasonable (AUC 0.705, accuracy 0.660), and we were also able to obtain separation between the classes for 22 of 24 subjects. A window-based classification would be most useful in a real-time biofeedback application designed to reduce nodule-associated behaviors, whereas a subject-based classification would be relevant in a screening test for vocal nodules. An important question for exploration is whether individuals with nodules who exhibit associated behavior early in the day are at greater risk for increased damage. This should become clearer as additional speaker data are obtained for analysis.

VI. Limitations

As noted previously, it is not possible to determine the extent to which differences observed in the vocal behavior of patients with vocal nodules preceded (primary hyperfunction) or followed (reactive hyperfunction) the formation of the nodules.

Many of the subjects in our study were students of performing arts programs, and our collection periods may have included substantial episodes of singing. It is possible that periods of singing played a factor in our ability to successfully classify subjects.

We performed analysis on 24 subjects, which is a small sample size. Our findings must be replicated on larger datasets once they become available.

VII. Conclusion

In this paper, we used distributional features of SPL and F0 derived from a noninvasive ACC signal to classify 5-min windows as belonging to a subject with normal voice or to a subject diagnosed with vocal fold nodules. We evaluated these features on 12 patients with vocal fold nodules and 12 matched control subjects. We identified several correlations related to subject class and were able to separate 22 of the 24 subjects based on the proportion of 5-min windows classified as positive.

Wearable voice monitoring systems have the potential to provide relevant, real-time information about speaker vocal status by providing reliable and objective measures of voice use during an individual's day. Large-sample data collection from patients and subjects with normal voices over long periods of time is warranted to provide further opportunity to explore potential behavioral targets.

Acknowledgment

The authors are grateful to S. W. Feng for signal processing contributions and to R. Petit for smartphone application programming.

References

[1] N. Roy, R. M. Merrill, S. D. Gray, and E. M. Smith, "Voice disorders in the general population: Prevalence, risk factors, and occupational impact," *Laryngoscope*, vol. 115, no. 11, pp. 1988–1995, 2005.

[2] R. E. Hillman, E. B. Holmberg, J. S. Perkell, M. Walsh, and C. Vaughan, "Objective assessment of vocal hyperfunction: An experimental framework and initial results," *J. Speech, Lang. Hear. Res.*, vol. 32, no. 2, pp. 373–392, 1989.

[3] R. E. Hillman, E. B. Holmberg, J. S. Perkell, M. Walsh, and C. Vaughan, "Phonatory function associated with hyperfunctionally related vocal fold lesions," *J. Voice*, vol. 4, no. 1, pp. 52–63, 1990.

[4] N. Roy, J. Barkmeier-Kraemer, T. Eadie, M. P. Sivasankar, D. Mehta, D. Paul, and R. Hillman, "Evidence-based clinical voice assessment: A systematic review," *Amer. J. Speech-Lang. Pathology*, vol. 22, no. 2, pp. 212–226, 2013.

[5] M. P. Karnell, S. D. Melton, J. M. Childes, T. C. Coleman, S. A. Dailey, and H. T. Hoffman, "Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders," *J. Voice*, vol. 21, no. 5, pp. 576–590, 2007.

[6] D. D. Mehta, M. Zañartu, S. W. Feng, H. A. 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.

[7] M. Zañartu, J. C. Ho, S. S. Kraman, H. Pasterkamp, J. E. Huber, and G. R. Wodicka, "Air-borne and tissue-borne sensitivities of bioacoustic sensors used on the skin surface," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 2, pp. 443–451, Feb. 2009.

[8] M. Zañartu, J. C. Ho, D. D. Mehta, R. E. Hillman, and G. R. Wodicka, "Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 9, pp. 1929–1939, Sep. 2012.

[9] H. A. Cheyne, H. M. Hanson, R. P. Genereux, K. N. Stevens, and R. E. Hillman, "Development and testing of a portable vocal accumulator," *J. Speech, Lang. Hear. Res.*, vol. 46, no. 6, pp. 1457–1467, 2003.

[10] I. R. Titze, J. G. Švec, and P. S. Popolo, "Vocal dose measures: Quantifying accumulated vibration exposure in vocal fold tissues," *J. Speech, Lang. Hear. Res.*, vol. 46, no. 4, pp. 919–932, 2003.

[11] I. R. Titze, E. J. Hunter, and J. G. Švec, "Voicing and silence periods in daily and weekly vocalizations of teachers," *J. Acoustical Soc. Amer.*, vol. 121, pp. 469–478, 2007.

[12] D. D. Mehta, R. W. Listfield, H. A. Cheyne II, J. T. Heaton, S. W. Feng, M. Zañartu, and R. E. Hillman, "Duration of ambulatory monitoring needed to accurately estimate voice use," in *Proc. InterSpeech: Annu. Conf. Int. Speech Commun. Assoc.*, Portland, OR, USA, 2012, pp. 1–4.

[13] E. J. Hunter and I. R. Titze, "Variations in intensity, fundamental frequency, and voicing for teachers in occupational versus nonoccupational settings," *J. Speech, Lang. Hear. Res.*, vol. 53, no. 4, pp. 862–875, 2010.

[14] S. L. Morrow and N. P. Connor, "Comparison of voice-use profiles between elementary classrooms and music teachers," *J. Voice*, vol. 25, no. 3, pp. 367–372, 2011.

[15] P. Bottalico and A. Astolfi, "Investigations into vocal doses and parameters pertaining to primary school teachers in classrooms," *J. Acoustical Soc. Amer.*, vol. 131, no. 4, pp. 2817–2827, 2012.

[16] A. Remacle, D. Morsomme, and C. Finck, "Comparison of vocal loading parameters in kindergarten and elementary school teachers," *J. Speech, Lang. Hear. Res.*, to be published, 2013. DOI: 10.1044/2013_JSLHR-S-12-0351.

[17] N. D. Hogikyan and G. Sethuraman, "Validation of an instrument to measure voice-related quality of life (V-RQOL)," *J. Voice*, vol. 13, no. 4, pp. 557–569, 1999.

[18] G. B. Kempster, B. R. Gerratt, K. Verdolini Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," *Amer. J. Speech-Lang. Pathology*, vol. 18, no. 2, pp. 124–132, 2009.

[19] M. Ghassemi, E. Shih, D. D. Mehta, S. W. Feng, J. Van Stan, R. E. Hillman, and J. Guttag, "Detecting voice modes for vocal hyperfunction prevention," in *Proc. 7th Annu. Workshop Women Mach. Learning/Neural Inform. Process. Syst. Conf.*, Lake Tahoe, NV, USA, 2012, poster presentation.

[20] D. D. Mehta, M. Zañartu, J. H. Van Stan, S. W. Feng, H. A. Cheyne II, and R. E. Hillman, "Smartphone-based detection of voice disorders by long-term monitoring of neck acceleration features," in *Proc. The 10th Annu. Body Sensor Netw. Conf.*, Cambridge, MA, USA, 2013, pp. 1–4.

[21] F. J. MasseyJr., "The Kolmogorov–Smirnov test for goodness of fit," *J. Amer. Statist. Assoc.*, vol. 46, no. 253, pp. 68–78, 1951.
[22] W. R. Rice, "Analyzing tables of statistical tests," *Evolution*, vol. 43, no. 1, pp. 223–225, 1989.
[23] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006, vol.1.
[24] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, 2004.
[25] N. V. Welham and M. A. Maclagan, "Vocal fatigue: Current knowledge and future directions," *J. Voice*, vol. 17, no. 1, pp. 21–30, 2003.
[26] A.-M. Laukkanen and E. Kankare, "Vocal loading-related changes in male teachers voices investigated before and after a working day," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 4, pp. 229–239, 2006.

Marzyeh Ghassemi received the B.S. degree in computer science and electrical engineering with a minor in applied mathematics from New Mexico State University, Las Cruces, USA, in 2005 as a Goldwater Scholar, and the M.Sc. degree in biomedical engineering from Oxford University, U.K., as a Marshall Scholar. She is currently working toward the Ph.D. degree at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA.

Jarrad H. Van Stan received the B.M. degree in applied voice from the University of Delaware, Newark, USA, in 2001, and the M.A. degree in speech pathology from Temple University, Philadelphia, PA, USA, in 2005.

He is currently a Speech Language Pathologist and a Senior Clinical Research Coordinator at the MGH Center for Laryngeal Surgery and Voice Rehabilitation and the Ph.D. degree at the MGH Institute of Health Professions, Boston, MA, USA. He is a Board Recognized Specialist in swallowing disorders and his research interests include voice and swallowing assessment and rehabilitation.

Daryush D. Mehta (S'01–M'11) received the B.S. degree in electrical engineering (*summa cum laude*) from the University of Florida, Gainesville, USA, in 2003, the S.M. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, USA, in 2006, and the Ph.D. degree from the MIT in speech and hearing bioscience and technology, Harvard–MIT Division of Health Sciences and Technology, Cambridge, MA, USA, in 2010.

He currently holds appointments at Massachusetts General Hospital (Assistant Biomedical Engineer in the Department of Surgery) and Harvard Medical School (Instructor in surgery), Boston, MA, USA. He is also an Honorary Senior Fellow in the Department of Otolaryngology, University of Melbourne, Australia.

Matías Zañartu (S'08–M'11) received the Ph.D. and M.S. degrees in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2010 and 2006, respectively, and his professional title and B.S. degree in acoustical engineering from the Universidad Tecnológica Vicente Pérez Rosales, Santiago, Chile, in 1999 and 1996, respectively.

He is currently an Academic Research Associate at the Department of Electronic Engineering, Universidad Técnica Federico Santa Maria, Valparaiso, Chile. His research interests include digital signal processing, nonlinear dynamic systems, acoustic modeling, speech/audio/biomedical signal processing, speech recognition, and acoustic biosensors.

Dr. Zañartu was the recipient of a Fulbright Scholarship, an Institute of International Education IIE-Barsa Scholarship, a Qualcomm Q Award of Excellence, and the Best Student Paper in Speech Communication in the 157th meeting of the Acoustical Society of America.

Harold A. Cheyne II received the B.S. degree in electrical engineering (*summa cum laude*) from Tufts University in 1993, and the Ph.D. degree from MIT in speech and hearing bioscience and technology in the Harvard -MIT Division of Health Sciences and Technology, Cambridge, in 2002.

He is currently the Director of Technology for the Bioacoustics Research Program at the Cornell Lab of Ornithology, and consults in audio electronics design and audio forensics through his firm L.A.S.E.R., LLC.

Robert E. Hillman received the B.S. and M.S. degrees in speech pathology from Pennsylvania State University, University Park, USA, in 1974 and 1975, respectively, and the Ph.D. degree in speech science from Purdue University, West Lafayette, IN, USA, in 1980.

He is currently Codirector/Research Director of the Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, a Professor of Surgery at Harvard Medical School, and a Director of Research Programs, MGH Institute of Health Professions, Boston, MA, USA. His research has been funded by both governmental and private agencies since 1981, and he has more than 100 publications on normal and disordered voice.

Prof. Hillman is a fellow of the American Laryngological Association and has received the Honors of the American Speech-Language-Hearing Association (ASHA's highest honor).

John V. Guttag received the Bachelor's degree in english and the Master's degree in applied mathematics both from Brown University, Providence, RI, USA, in 1971 and 1972, respectively, and the Doctorate degree in computer science from the University of Toronto, Toronto, ON, Canada, in 1975.

He is the Dugald Jackson Professor at the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, USA. He was a member of the faculty at the University of Southern California, Los Angeles, USA, from 1975 to 1978, and joined the MIT faculty in 1979. From 1993 to 1998, he served as an Associate Department Head for computer science of MIT's Electrical Engineering and the Computer Science Department. From January 1999 through August 2004, he served as a Head of that department. He also coheads the MIT Computer Science and Artificial Intelligence Laboratory's Networks and Mobile Systems Group. This group studies issues related to computer networks, applications of networked and mobile systems, and advanced software-based medical instrumentation and decision systems.

Prof. Guttag is a fellow of the ACM and a member of the American Academy of Arts and Sciences.