Identifying bias in models that detect vocal fold paralysis from audio recordings using explainable machine learning and clinician ratings

Daniel M. Low [1,2], Vishwanatha Rao [3,4], Gregory Randolph [4,5], Phillip C. Song [4,5] *, Satrajit S. Ghosh, PhD[1,2,5] *

[1] Program in Speech and Hearing Bioscience and Technology, Harvard Medical School, Boston, MA, USA

[2] McGovern Institute for Brain Research, MIT, Cambridge, MA, USA

[3] Department of Biomedical Engineering, Columbia University, New York, NY, USA

[4] Department of Otolaryngology–Head and Neck Surgery, Massachusetts Eye and Ear Infirmary, Boston, MA, USA

[5] Department of Otolaryngology–Head and Neck Surgery, Harvard Medical School, Boston, MA, USA

* Equal contribution

Corresponding author

Correspondence can be addressed to Daniel M. Low, Office: 46-4033F, 43 Vassar St, Cambridge, MA 02139, USA. E-mail: dlow@g.harvard.edu.

Abstract

Introduction. Detecting voice disorders from recordings of voice could allow for frequent, remote, and low-cost screening before costly clinical visits and a more invasive laryngoscopy examination. The goals of this study were to detect unilateral vocal fold paralysis (UVFP) from voice recordings using machine learning, to identify which acoustic variables were important for prediction to increase trust by relating such variables with known pathophysiology, and to determine model performance relative to clinician performance on the same recordings.

Methods. Patients with confirmed UVFP through endoscopic examination (N=77) and controls with normal voices matched for age and sex (N=77) were included. Voice samples were elicited by reading the Rainbow Passage and sustaining phonation of the vowel "a". The 88 extended Geneva Minimalistic Acoustic Parameter Set features were extracted as inputs for four machine learning models of differing complexity. SHAP was used to identify important features.

Results. The highest median bootstrapped Area Under the Receiver Operating Characteristic Curve (ROC AUC) score was 0.87, which varied depending on model and task, and beat the top clinician performance (range: 0.74 – 0.81). When training machine learning models on clinician's CAPE-V estimates, the maximum performance was 0.84. The most important variables included some related to UVFP pathophysiology such as mean MFCC1, mean F1 amplitude and frequency, and

shimmer variability depending on model and task. Surprisingly, many UVFP recordings had higher intensity than controls, which was the most important variable for prediction. This was confirmed by having clinician's rate the background noise as a proxy for an increase in microphone gain, which showed to be higher for UVFP than controls. This and UVFP patients compensating their soft voices with extra vocal effort in order to be heard are likely causes of a systematic bias that allowed the models to detect UVFP.

Conclusion. Using the largest dataset studying UVFP to date, we achieve high performance from just a few seconds of voice recordings, surpassing expert clinicians' performance. However, we uncovered bias which may occur in voice biomarker research any time individuals have a soft voice. Explainable machine learning and clinician ratings thus provide a mechanism to detect UVFP, uncover how acoustic variables characterize a specific pathophysiology, and reveal bias.

Keywords: vocal fold paralysis, acoustic analysis, voice, speech, explainability, interpretability, machine learning, bias

INTRODUCTION

Voice recordings provide a rich source of information related to vocal tract physiology and human physical and mental health. Given advances in smartphones and wearables, these recordings can be made anytime and anywhere. Thus, the search for disorder-specific acoustic biomarkers has been gaining momentum. Voice biomarkers have been reported for detecting Parkinson's diseases[1] as well as psychiatric disorders including depression, schizophrenia, and bipolar disorder (for a systematic review, see Low et al, 2020[2]). Given our scientific understanding of the complexity of speech production, multiple acoustic features have been devised for use in machine learning models.

In Figure 1, we describe a schematic of speech production and the process of extracting certain acoustic features from an audio signal (see also Quatieri, 2008[3]). Panel (A) depicts speech as the result of the neural coordination of three subsystems: the respiratory system (lungs), the laryngeal system (vocal folds), and the resonatory system of the vocal tract (pharynx, oral cavity, nasal cavity, articulators, and subglottal effects). Speech production requires air flow from the lungs to generate sound sources that are filtered by the vocal tract. Panel (B) captures the fact that environmental, microphone, and digital sampling characteristics (e.g., background noise, microphone gain, sampling rate) can affect acoustic features. Panel (C) Waveform of the audio signal, which is the 2D representation of the contraction (positive amplitude) and rarefaction (negative amplitude) of air particles. Higher amplitudes can lead to higher

perceived loudness. Prosodic features arise from changes over longer segments of time, which is perceived in the rhythm, stress, and intonation of speech. A segment of the waveform is shown in the right panel, indicating a periodic signal from the vocal folds. Panel (D) shows that for a given time window, a spectrum (right panel) can be obtained through a Fast Fourier Transform (FFT) which represents the magnitude of the frequencies in the signal with peaks (formants F1–F3) due to vocal tract filtering of the source signal produced by the vocal folds. The spectrogram (left panel) is a representation of the spectrum as it varies over time. The approximate location of the F0 and first formants are displayed. Finally, panel (E) demonstrates that it is possible to separate source and filter components by computing the inverse FFT of the log of the magnitude of the spectrum, called the cepstrum (right panel). The peak in the cepstrum reflects the periodic glottal fold vibration while lower quefrency components reflect properties of the resonatory subsystem. For speech recognition, Mel filters are applied to the spectrum to better approximate human hearing. A conversion of the Mel-spectrum to a cepstrum using a Discrete Cosine Transform (DCT) generates mel-frequency cepstral coefficients (MFCCs). Similar to the cepstrum, lower MFCCs track vocal-tract filter information. Despite these advances, robust applications to detect vocal fold paralysis disorders remain limited[4–9] (see Supplementary Table S1 for a summary of prior machine learning studies).

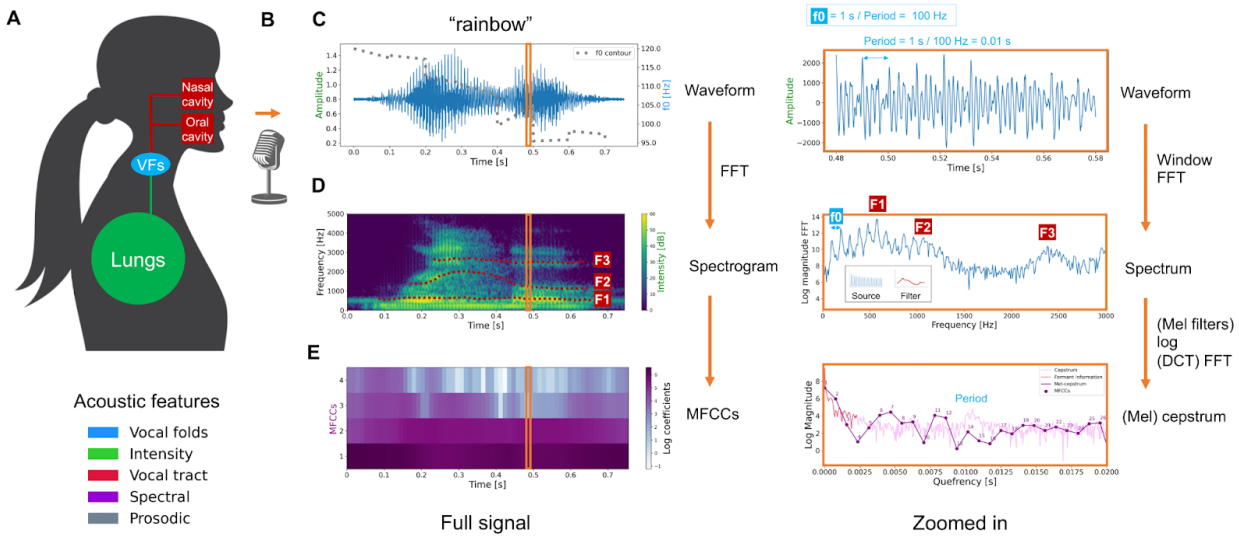

Figure 1. Schematic of speech production and the process of extracting certain acoustic features from an audio signal. (A) Speech production, (B) recording characteristics, (C) waveform (f0), (D) spectrogram (F1-F3, intensity), (E) mel-frequency cepstral coefficients (MFCCs).

Furthermore, while machine learning (ML) can be a powerful and successful approach for diagnostics, they are often treated as "black-boxes". It can be difficult to determine how the model is making a decision, that is, how it is combining input features from a given patient to generate a prediction. This is particularly worrisome given ML algorithms can detect and associate unintended or clinically irrelevant relationships and introduce bias that may be difficult to anticipate. Explainable ML refers to a series of methods and quantitative analyses for uncovering and "explaining" the rationale behind the decision made by complex algorithms, which is particularly critical in the high-stake decisions of medicine to increase trust among clinicians and patients[10].

There are many challenges for applying acoustic analysis to detect specific disorders. Voice characteristics are highly varied and change over time. Laryngeal pathology, age, gender, size, weight, general state of health, smoking/vaping, and medications can

impact vocal acoustic characteristics. Diseases in the larynx and phonatory system (i.e., larynx, resonating structures, lungs) and/or neurological system, will also affect voice. Compensatory production strategies and environmental conditions can also change the vocal signal. Furthermore, because hoarseness is such a frequent occurrence and specialty voice centers are rare, vocal fold disorders are often undiagnosed, under-reported, or misdiagnosed[11].

Unilateral vocal fold paralysis (UVFP) is an ideal model for demonstrating the explainability of ML for several reasons. UVFP occurs when the mobility of a single vocal fold is impaired as a consequence of neurological injury and diagnosis is consistently verified through routine laryngoscopy; therefore, ground truth labels are available. Second, the clinical signs of UVFP are well-described and acknowledged. These characteristics include a weak, breathy voice quality, early vocal fatigue, reduced cough strength, and aspiration with thin liquids[12,13]. Therefore, the expected acoustic differences between UVFP patients and healthy controls can be interpreted with regards to perceptual symptoms and a well-understood pathophysiology. In contrast, explaining important variables to predict a disorder which is hard to diagnose (e.g., has low inter-rater reliability) and has an unclear pathophysiology would ironically result in a poor explanation, because it would be puzzling how or even if the disorder could modulate the important acoustic variables.

We also chose to examine UVFP because it is clinically important. Vocal fold paralysis may occur due to iatrogenic injury, malignancy, idiopathic, and neurological disease,

and impacts quality of life. Overall, surgical iatrogenic injury accounts for 46% of all UVFP in adults and thyroid and parathyroid surgeries are responsible for 32% of postsurgical UVFP[14]. There is a significant need for a screening tool for the diagnosis and tracking of UVFP because of the high impact of this condition on productivity and quality of life. Screening could be done remotely and frequently, especially when surgical specialists and laryngeal exams are not readily accessible due to geographical, financial, and other barriers[15]. Using an explainable ML model as a screening tool for UVFP can provide greater clarity as to who most needs laryngoscopy and provides insight in the key voice characteristics related to the pathophysiology[16–20].

The objectives of our study were: (1) to detect UVFP using ML; (2) to evaluate the effectiveness of different models in differentiating the acoustic signals between patients with UVFP and patients with normal functioning vocal folds (i.e., controls); and (3) to explain which features are most important to the diagnostic models and examine the pathophysiological relevance. To achieve these objectives, we evaluated statistical dependencies across voice features in the data, used four different classes of machine learning algorithms to assess classification performance, evaluated the minimal set of features necessary for detection, and identified the most important features for model construction. Ultimately, we wanted to see if the most important features identified by the machine learning models matched clinically-known relevant acoustic changes.

MATERIALS AND METHODS

This study was approved by the Institutional Review Board at Massachusetts Eye and Ear Infirmary and Partners Healthcare (IRB 2019002711).

Participants and voice samples

Through retrospective chart analysis from 2009 to 2019, a total of 1043 patient charts were reviewed from a tertiary care laryngology practice who underwent endoscopic evaluation and voice testing. Of those, 53 patients with confirmed UVFP were identified. They had documented vocal fold paralysis by endoscopic examination and had undergone acoustic analysis as part of routine clinical care. Each patient had four acoustic recordings. These included three sustained vocalizations of the "a" vowel sound (ɑ in the International Phonetic Alphabet) and a reading of the introductory paragraph of the rainbow passage[21]. The acoustic recordings were all taken in an acoustically shielded room. For each of these 53 patients, a board-certified otolaryngologist reviewed their clinical history, video laryngoscopy as well as their audio samples to confirm that they were correctly classified to have UVFP. Voice samples from an additional 24 patients were collected prospectively using a mobile software, OperaVOX™ on an iPad, who were being treated for UVFP. These patients also had the same four acoustic recordings as the patients from retrospective chart review. This combination of data collection yielded a total of 77 UVFP patients for analysis, of which 48 had left UVFP and 29 right UVFP.

All of the patients were then matched with control samples from a database of patients without UVFP who had also undergone acoustic analysis. Each control was the same sex and had the same smoking status as the UVFP patient and within three years of age, and had documented laryngeal examinations that verified the absence of vocal fold mucosal pathology. The controls were excluded if they had established laryngeal surgery, vocal fold lesions, radiation, head and neck cancer, or neurological disease. The controls had recorded the same four acoustic recordings as the retrospectively gathered UVFP group. A board-certified otolaryngologist confirmed that the voice recordings and video laryngoscopies of these controls matched normal expectancies. The reading samples were divided in thirds to match the amount of vowel production samples. Reading recordings were not available for three patients and three patient vowel samples were removed due to containing multiple vowel productions or a cough. The final dataset that was analyzed is described in Table 1. Reading+vowel refers to including all samples (i.e., ~6 samples) from the same participant with the goal of either obtaining higher performance or discovering features that show variation in relation to diagnosis consistently across tasks. Mean (SD) audio lengths were 6.81s (5.47) for reading samples and 3.95s (1.00) for vowel samples. The audio samples were processed using OpenSmile with the eGeMAPS configuration file (article[22], source code[23]) which applies different summarization statistics to the time series depending on the feature resulting in 88 features per sample covering information related to the vocal folds (F0, jitter, shimmer), intensity (loudness, HNR), vocal tract (F1–3 frequency, bandwidth, amplitude), spectral balance (alpha ratio, Hammamberg index, spectral

slope, MFCC 1–4, spectral flux), and prosody (voice and unvoiced segments, loudness peaks per second).

Table 1. Sample sizes and demographic information.

	UVFP	Controls	Total
N	77	77	154
Mean age (SD)	56.4 (18.7)	56.6 (18.8)	56.5 (18.7)
Sex (F/M)	39/38	39/38	78/76
Reading	222	231	453
Vowel	227	231	458
Reading+vowel (total)	449	462	911

SD: standard deviation; F: female; M: male.

Machine learning models of increasing complexity

With the goal of classifying voices recording into either UVFP or controls, we used four machine learning algorithms of increasing complexity from the *scikit-learn* (v0.21.3) using the *pydra-ml* (v=0.3.1) toolbox[24] (default parameters were used unless otherwise specified):

(1) Logistic Regression: a simple linear model that is constrained to use few features due to an L1 penalty making it the simplest model ("liblinear" solver was used which is ideal for smaller datasets).

(2) Stochastic Gradient Descent (SGD) Classifier: it is also a linear model but tends to use more features due to an elastic net penalty that was chosen making it slightly more complex (the max_iter parameter was set to 5000 and early_stopping was set to True).

(3) Random Forest: it is an algorithm that uses simpler decision trees (i.e., weak learners) on feature subsets but then averages the trees' predictions to create a stronger learner, making it harder to interpret which features are important across trees.

(4) Multi-Layer Perceptron: it is a neural network classifier which incorporates, in our case, 100 instances of perceptrons (artificial neurons), which are connected to each input feature through weights with an added ReLU activation function to capture nonlinear structures in the data. It is not possible to know exactly how the hundreds of internal weights interact to determine feature importance, making the model difficult to interpret directly from its parameters (the max_iter parameter was set to 1000; alpha or the L2 penalty parameter was set to 1).

To generate independent test and train data splits, a bootstrapped group shuffle split sampling scheme was used. For each iteration of bootstrapping, a random selection of 20% of the participants was used to create a held-out test set. The remaining 80% of participants were used for training. This process was repeated 50 times, and the four classifiers were fitted and tested for each test/train split. The Area Under the Receiver Operating Characteristic Curve (ROC AUC; perfect classification = 1; chance = 0.5) was computed to evaluate the performance of the models on each iteration, resulting in a distribution of 50 ROC AUC scores for each classifier. For each iteration, each

classifier was trained with randomized patient/control labelings to generate a null distribution of ROC AUC scores (i.e., a permutation test). Each model's performance was statistically compared to other models and to the null distributions using an empirical p-value, a common and effective measure for evaluating classifier performance (see Definition 1 in Ojala & Garriga, 2010[25]). The significance level was set to alpha = 0.05.

Assessing feature importance

Kernel SHAP (SHapley Additive exPlanations) was used to determine which acoustic features were most important for each model to detect UVFP. This method is model agnostic in that it can take any trained target model (even “black box” neural networks) and compute feature importance[26]. It does so by performing regression with L1 penalty between different sets of input features and a single prediction made by the target model. It then uses the coefficients of the additional regression model as a measure of feature importance for a single prediction. We took the average of the absolute SHAP values across all test predictions (positive and negative values are both important for classification). We then weighted the average values by the model’s median performance since an important feature for a bad model could be a less important feature for a good model and vice versa. Since we trained each model 50 times (i.e., one for each bootstrapping split), we computed the mean SHAP values across splits for each model. This pipeline (i.e., machine learning models, bootstrapping scheme, SHAP analysis) was done using the *pydra-ml* package.

Reducing redundant features for more explainable models

Highly correlated features (i.e., multicollinearity) can influence model generation and interpretation. Two models may obtain similar performance while using different features or placing different weights on the same features (i.e., underspecification). This makes it difficult to compare algorithmic explanations across models. For instance, mean F1 frequency may be less important to a given model because the model uses mean F2 frequency which happens to capture very similar information in a particular dataset, whereas a different model may use F1 instead of F2 or use both but assign less importance to each. To enforce models to use the same features that capture very similar information and be able to compare feature importance across models, we kept a single feature out of the sets of features that share similar information above a given threshold. We used a custom algorithm we call Independence Factor (see "Reducing redundant features for more explainable models" section in Supplementary Materials).

Clinician ratings

One otorhinolaryngologist and two speech-language pathologists rated each audio recording of reading task (one per participant, not split in three) for the following variables, in order: background noise (None, Some, High); UVFP (yes, no), background noise [1,2,3], CAPE-V severity (0 to 100), CAPE-V roughness (0 to 100), CAPE-V breathiness (0 to 100), CAPE-V strain (0 to 100), CAPE-V pitch (0 to 100), CAPE-V

loudness (0 to 100; estimated loudness as if the rater were in the recording room), recording loudness (low, medium, high; loudness of the recording). Inter-rater agreement was assessed using intra-class correlation for all numerical variables and Light's k for the binary presence of UVFP . The entire reading task was provided instead of the task split in three to make assignment easier for clinicians. The reading task was chosen over the sustained vowel because we expected it to be easier to detect UVFP.

RESULTS

Performance of models using acoustic features and clinician ratings

Given dependent features provide similar information (see Supplementary Figures S1, S2, S3, S4, S5, S6, S7, S8, and S9) and distort feature importance analyses, we then tested performance after removing redundant features using the Independence Factor method previously described. Supplementary Figure S10 shows performance for different feature set sizes with increasing amounts of redundant features. From this analysis, we selected the feature-set size that resulted in best performance using the least amount of features for subsequent analyses: 39 features (reading), 13 (vowel), 19 (reading+vowel). After removing related features (i.e., reducing multicollinearity) from the original 88 features, similar performance was obtained (median ROC AUC = 0.84–0.87) using fewer features. Supplementary Materials "Feature selection" section

describes an analysis of how this method compares to removing features across each train set (see Sup. Mat. Table S2).

Performance was found to be high across most models both with and without redundant features. The bootstrapped ROC AUC distributions and permutation tests for the reduced (parsimonious) models using the non-redundant feature set are shown in Figure 2. The figure reports a one tailed statistical comparison (row > column) of models using an empirical p-value, which represents the fraction of column-model scores where the row-model classifier had a higher mean performance (e.g., a p-value of 0.02 indicates that the mean score of a row model is higher than 98% of column-model scores). Table 2 shows performance using all features and a subset of features selected by either removing redundant features while maintaining performance (as in Supplementary Figure S10) or using the top 5 most important features.

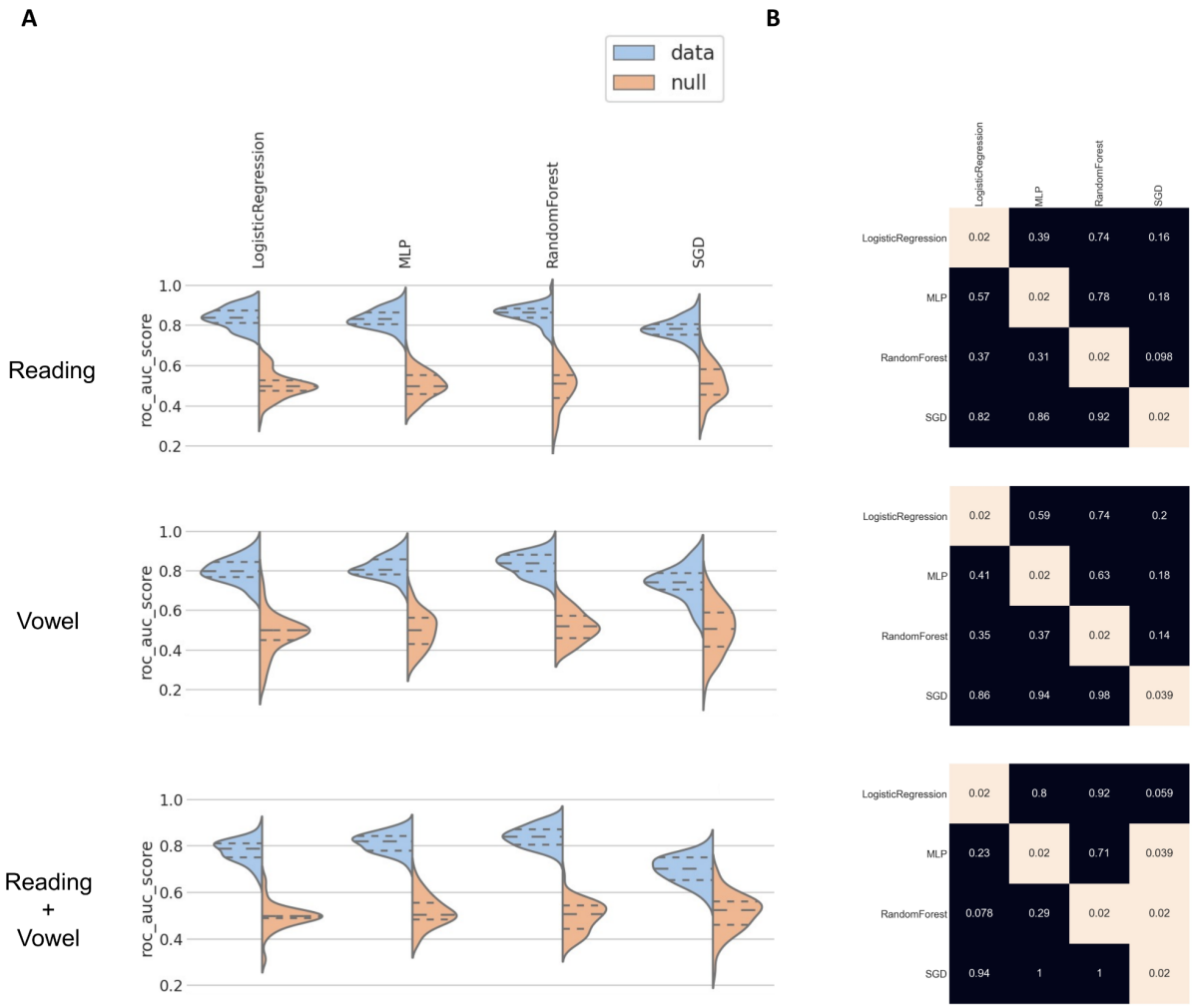

Figure 2. Model performance comparison using a permutation test. (A) Scores from models trained on true labels (blue) and trained on permuted labels (orange) over bootstrapping splits. **(B)** Statistical comparison between models (annotation = p-value, highlighted = significant results).

Table 2. Model performance.

	Features	LogisticRegression	MLP	RandomForest	SGD
Reading	88	.87 (.78–.93; .50)	.87 (.80–.93; .50)	.87 (.76–.91; .49)	.83 (.76–.89; .50)
Vowel	88	.84 (.77–.89; .50)	.86 (.79–.91; .50)	.86 (.79–.91; .51)	.80 (.72–.87; .50)
Reading+Vowel	88	.84 (.76–.91; .50)	.86 (.74–.92; .48)	.85 (.77–.92; .49)	.79 (.72–.86; .51)
Reading	39	.84 (.76–.92; .50)	.83 (.76–.91; .50)	.87 (.77–.91; .51)	.78 (.71–.86; .51)
Vowel	13	.80 (.70–.90; .50)	.81 (.74–.91; .50)	.84 (.75–.90; .52)	.74 (.58–.87; .51)
Reading+Vowel	19	.79 (.70–.84; .50)	.82 (.75–.88; .51)	.84 (.77–.91; .51)	.70 (.61–.77; .52)
Reading	Top 5	.81 (.73–.89; .50)	.86 (.78–.92; .47)	.85 (.77–.90; .50)	.75 (.56–.87; .57)
Vowel	Top 5	.78 (.67–.87; .50)	.82 (.74–.92; .53)	.81 (.72–.87; .50)	.72 (.57–.82; .49)
Reading+Vowel	Top 5	.80 (.70–.86; .50)	.82 (.74–.88; .50)	.81 (.74–.89; .53)	.72 (.55–.83; .52)
Reading	88 - Top 5	.85 (.76–.92; .50)	.87 (.77–.92; .49)	.85 (.77–.90; .52)	.82 (.71–.89; .51)
Vowel	88 - Top 5	.84 (.75–.93; .50)	.86 (.72–.93; .51)	.84 (.74–.94; .52)	.80 (.70–.90; .48)
Reading+Vowel	88 - Top 5	.84 (.74–.89; .50)	.85 (.76–.91; .50)	.85 (.76–.91; .50)	.79 (.71–.87; .50)

Performance of models using either all 88 features, non-redundant features (39, 13, 19), top five most important features, all 88 features minus top 5 most important features, and clinician ratings of CAPE-V and noise (i.e., background noise) and loudness (i.e., loudness of the recording). Median ROC AUC score from 50 bootstrapping splits (90% confidence interval; median score of null model). For full distributions of scores see Figure S10 in Supplementary Materials. Removing features is a post-hoc analysis because features were selected based on observing performance on the test sets, and therefore performance might be slightly overly optimistic and would need to be tested on an independent test set for further validation. MLP: Multi-Layer Perceptron; SGD: Stochastic Gradient Descent Classifier.

How important are the most important features?

Studies tend to report and describe the top N features, but it is not clear what performance the model would obtain for those features when used alone since measurement is usually based on models that use additional features with multiple interactions. In contrast, in our study we ran models on the top 5 features together (Table 2), which allowed us to actually demonstrate their predictive capability. The lower performance of these top 5 features relative to a richer feature set helps demonstrate that model performance is dependent on interactions across multiple additional features. We also ran models without top 5 features to demonstrate that leaving features that are redundant with these top features results in almost equivalent high performance to using all 88 features since the redundant features share information. Given 24 UVFP patients were recorded with a different device, we trained models without their samples to make sure these differences in recordings were not driving performance. There was a small drop in performance, which could be due to a bias (the full, original model using information of the recording device), but could also be due to removing training samples. The drop in performance is not large enough to suspect that differences in recording are driving the full original model's performance (see Sup. Mat. Table S3, Table S4, and analysis in Supplementary section "Performance removing participants that used other recording system").

Assessing feature importance

Figure 3 reports feature importance using SHAP for all models. For further description of features and the chosen classification of features, see Eyben et al. (2015)[22] and Low

et al. (2020)[2]. When reviewing important features, it is key to note that any of the features with which it is codependent could be a reasonable important feature (see clusters of redundant features in Supplementary Figures S1-S9). To understand the role of the most important features we ran a post-hoc analysis with the top 5 features for each data type (reading, vowel, reading+vowel), performance is shown in Table 2 and we further display the distribution of each top feature and its individual performance in Figure 4. Figure 5 reports similarity between top 5 features and all original 88 eGeMAPS features. Features that have a high distance correlation (i.e., cluster) with top 5 features were not used in models to avoid redundancy, but still share similar information and can therefore be considered important features as well. Hierarchically-clustered heatmaps for other data types (vowel, reading, both) and groups (UVFP patients, controls, both) are displayed in Supplementary Figures S1, S2, S3, S4, S5, S6, S7, S8, and S9.

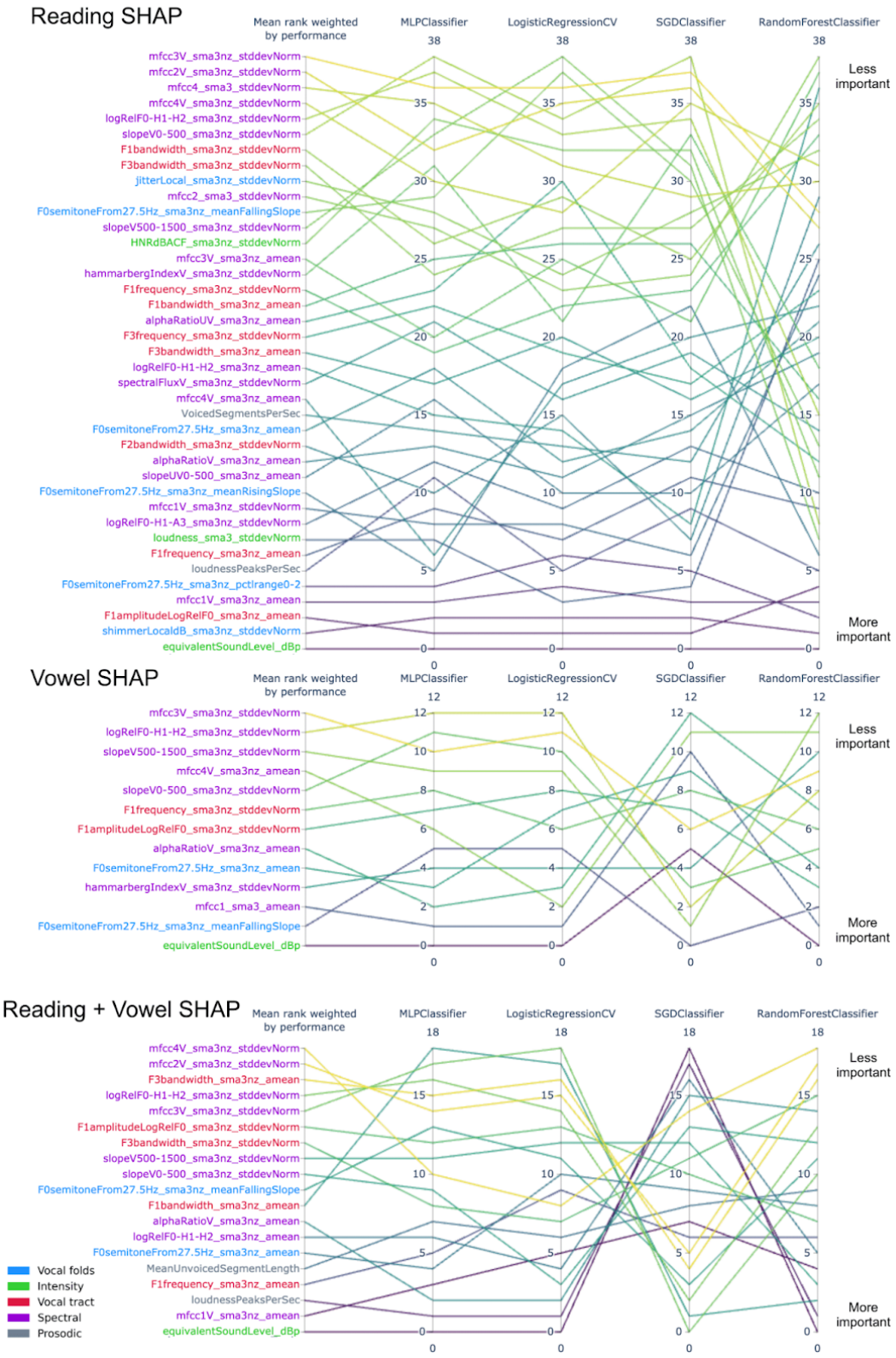

Figure 3. Feature importance parallel coordinate plot. Rank reads from bottom (most important) to top (least important). Mean rank is weighted by performance of each model to avoid a lower performing model biasing the mean rank

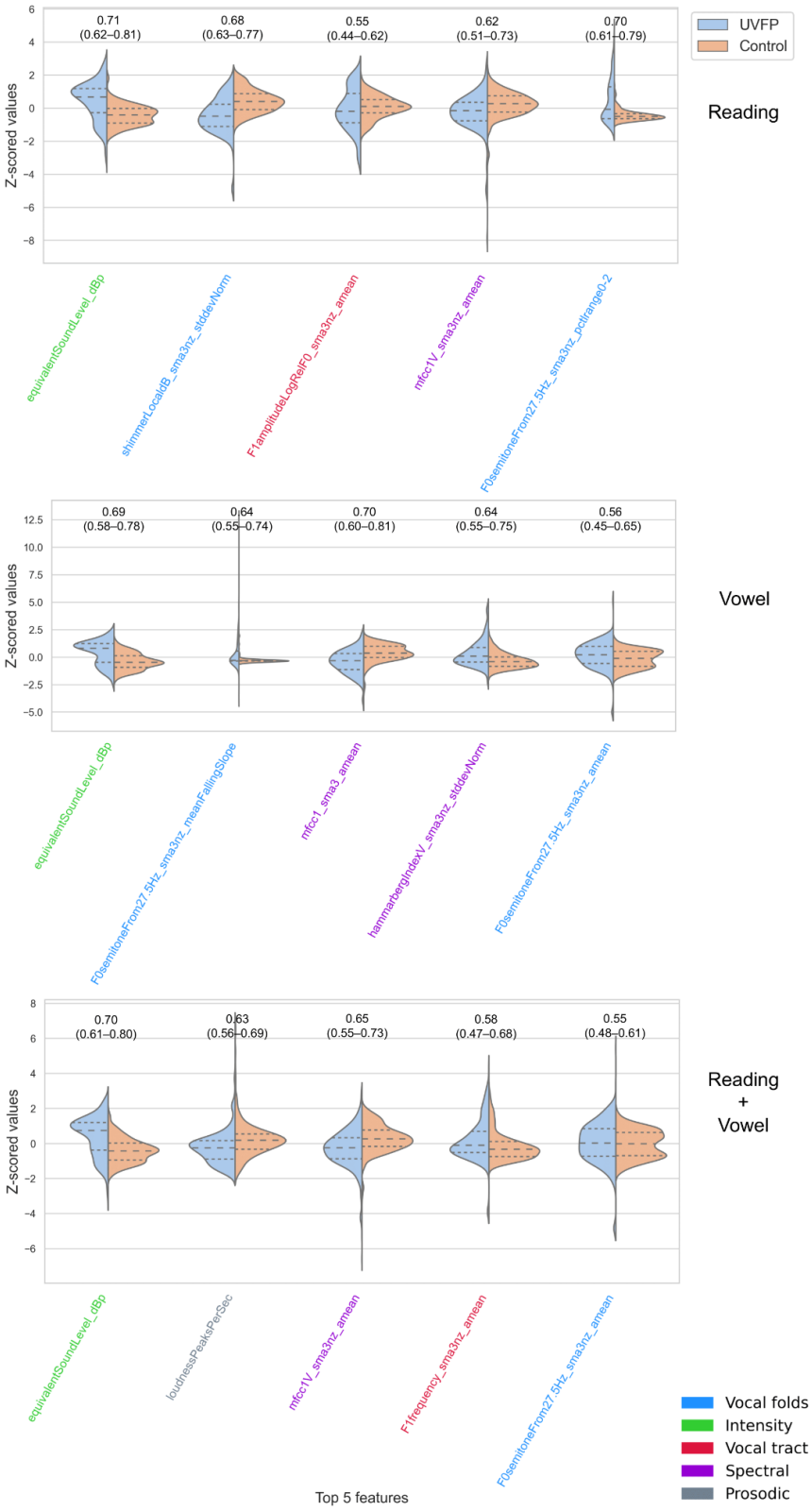

Figure. 4. Distributions for top 5 features and corresponding performance for single features. Logistic Regression with L1 penalty was used. No single feature is enough to dissociate groups with high performance. Null models' median performance was 0.5.

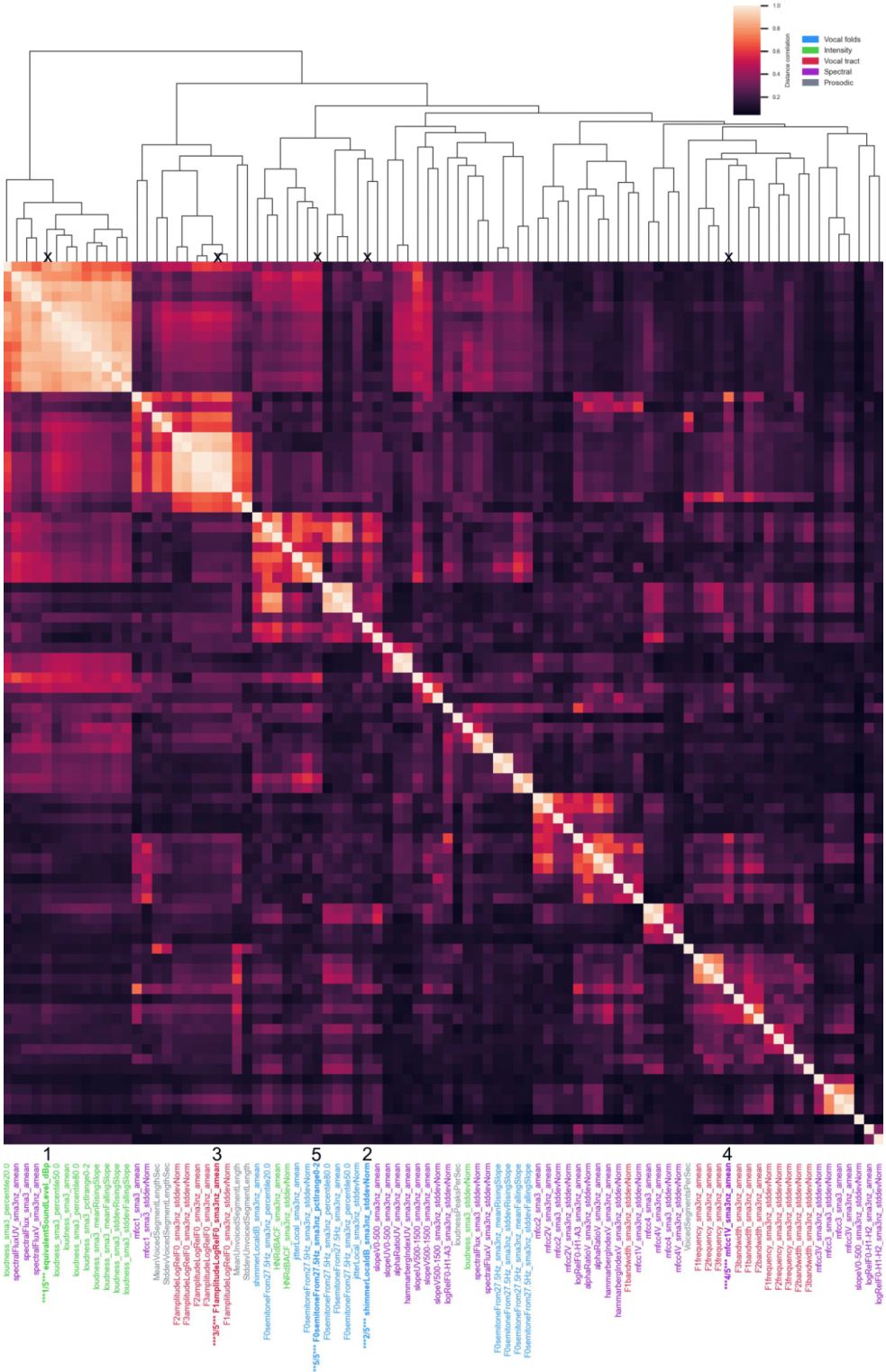

Figure 5. Feature redundancy with top 5 features highlighted. Top 5 features are highlighted in bold and their rank is displayed. Squares are clusters of redundant features. Computed with all participants on the reading task.

Clinician ratings

The median ROC AUC for humans was 0.78 (min. = 0.74 to max. = 0.81) meaning the machine learning models performed better than the highest performing clinician. Interestingly, using the average clinician's CAPE-V ratings within machine learning models was able to obtain a median ROC AUC of 0.81 with the Random Forest model (Table 3). Using clinicians perceptions of background noise and recording loudness achieved a maximum median ROC AUC of 0.77.

Table 3. Performance using clinician ratings as variables for machine learning models

	Features	LogisticRegression	MLP	RandomForest	SGD
CAPE-V	6	.80 (.69–.88)	.81 (.71– .90)	.84 (.71– .94)	.80 (.48– .91)
Noise+loudness	2	.76 (.59 –.86)	.77 (.61– .87)	.74 (.63 –.81)	.68 (.43– .81)

Common cutoffs for inter-rater agreement are poor for values less than .40, fair for values between .40 and .59, good for values between .60 and .74, and excellent for values between .75 and 1.0. In Figures 6 and 7 we report the inter-rater reliability (Flight's kappa and ICC) along with the distribution of the ratings. Background noise had poor reliability across rater, UVFP and recording loudness had fair reliability (see Figure 6) and CAPE-V-inspired ratings scored good to excellent except for pitch which was fair (see Figure 7).

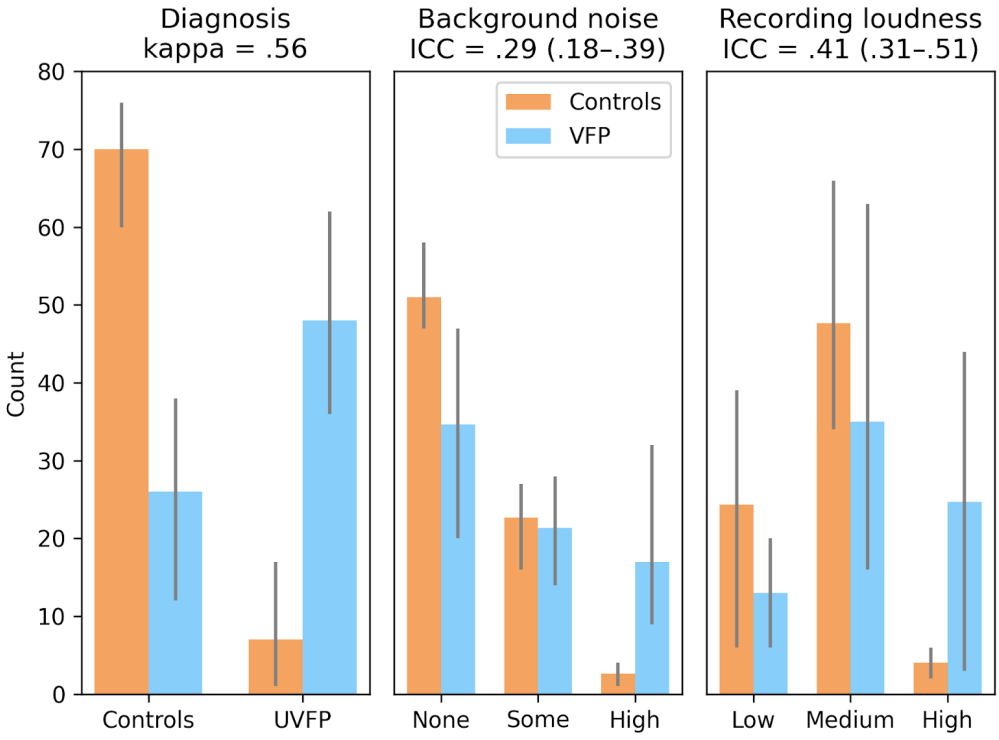

Figure 6. Clinician ratings for unilateral vocal fold paralysis (UVFP), background noise, and recording loudness indicating likely bias. Error bars indicate maximum and minimum count across the three raters. The disproportionate amount of UVFP samples rated as having high background noise and high loudness indicates likely bias, where the gain might have been raised for some UVFP patients and they may have phonated more intensely. kappa: Light's kappa; ICC: intra-class correlation coefficient.

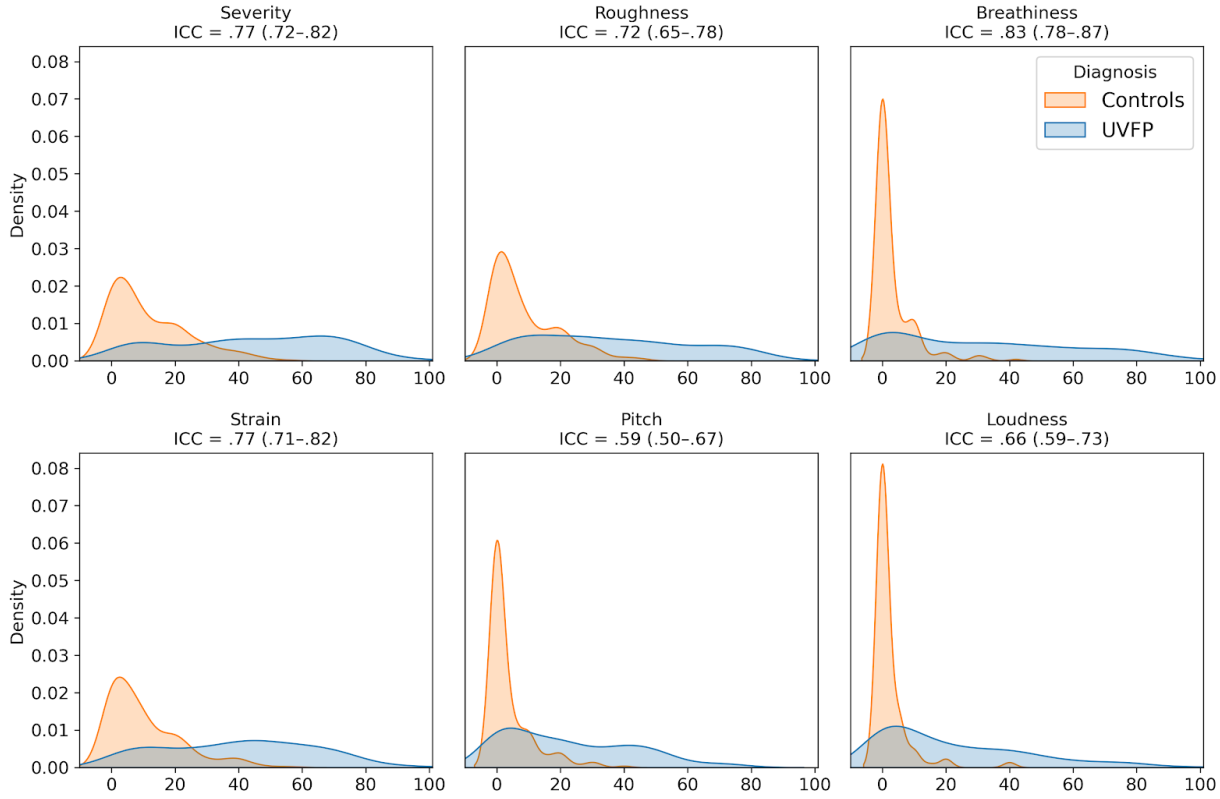

Figure 7. Distribution of CAPE-V ratings on audio recordings of read speech. ICC: intra-class correlation coefficient.

Discussion

This study achieves high performance in detecting UVFP from healthy voices which could have several clinical applications: (1) postoperative screening for thyroid surgery-related UVFP since after thyroid surgery, UVFP is common, occurring in up to 5 to 10% of cases[27]. Furthermore, laryngoscopy is not readily available to all postoperative populations and symptomatic changes are notoriously variable. An ML-based screening could help identify patients needing further workup and treatment, and earlier diagnosis is essential to optimize long-term outcomes [28,29]. (2) Monitoring

voice during speech therapy and after surgical treatment for confirmed UVFP to measure when and if the patient's voice is approximating a healthy voice. (3) Preoperative screening prior to surgeries that are at high risk for developing UVFP such as thyroid, head and neck, cardiac, thoracic, esophageal, and cervical spine operations.

We achieve robust classification performance and associate this performance with relevant acoustic features. Critically, we demonstrate that interpreting performance accuracy has to be contextualized with respect to the type of the ML model used and the voice-eliciting task.

The need for automated assessments of vocal fold paralysis

We chose vocal fold paralysis as the study cohort for several reasons. The acoustic changes associated with vocal fold paralysis are relatively reliable and consistent. Application of objective acoustic measurements towards differentiating between voice conditions have been limited[4–9] (see Supplementary Table S1 for a summary of prior machine learning studies). Our study expands on prior studies which have used pre-existing commercial databases, smaller sample sizes, fewer features, and/or methods for model evaluation that can be biased in small datasets given the test sets may not be representative (for a discussion on bootstrapping for clinical datasets, see 5[2]). As a clinical entity, UVFP can have detrimental effects on voice, vocation, and quality of life, with resultant morbidity related to respiration, swallowing and aspiration. The costs associated with UVFP not only relate to patient morbidity and diminished quality of life but also to the economic burden placed on our healthcare system. Greater

lengths of hospitalization and increased hospital costs have been associated with postsurgical VFP[27,30]. Access to specialists for diagnosis is limited and early detection and management of UVFP appear to improve length of stay and surgical outcomes[31].

Explaining acoustic features relevant to detecting vocal fold paralysis

Objective acoustic measurement changes associated with vocal fold paralysis have been described and these changes include reduced loudness and maximum phonation time, higher perturbation measurements such as jitter and shimmer, and increased signal to noise ratio[13,32,33]; however these were univariate models, and we have demonstrated that using single variables does not seem to provide high predictive performance. While other multivariate machine learning models have been used (see Supplementary Table S1), these used few features and small or undefined samples and only report feature importance results for one model; therefore it is not clear whether the important features reported would hold using larger feature sets or how other models would perform. Using a much larger initial set of acoustic features for analysis, we demonstrate that several machine learning algorithms of increasing complexity (using more parameters) successfully identify vocal fold paralysis from healthy voices. We also report that these models can use different features to achieve similar performance. Different models emphasize different features not simply because of its relevance to a disorder, but because of the mathematics associated with the model[34,35]. The variability of the ranking of features used by our individual models also illustrates the potential

danger of using the single highest performing model, which is commonly seen in published literature.

Instead of simply reporting the important features from the highest performing model, we analyzed the models to find common features. Some of the most important features across models were: intensity (especially equivalent sound pressure level which was redundant with multiple loudness features and seems to be due to some patients trying to use more breath for projection), Mel Frequency Cepstral Coefficients (especially the first coefficient, which captures spectral envelope or slope), mean F0 semitones (given F0 originates from vocal-fold oscillation, a vocal-fold paralysis is expected to alter F0), mean F1 amplitude and frequency (influenced by how the vocal tract filters F0 and the shape of the glottal pulse which would be affected by UVFP), and voiced and unvoiced segments (prosodic and speech articulation features which may be altered due to changes in the periodicity of F0). Shimmer variability was important just for reading, and it captures variability in glottal pulses and pressure patterns which ultimately affect F0. When we removed these top 5 features from the full feature set, performance is practically equivalent to using 88 features, as expected, since there are features that are redundant with these top 5 features. Therefore, it is not that only these 5 specific features drive performance, but rather the information they contain, which in this dataset is also captured by other features as shown in Figure 5.

These acoustic features corroborate our clinical understanding of glottal incompetence from UVFP and with common patient complaints of reduced loudness, vocal instability,

hoarseness, and rough voice. Uncovering and understanding the basic mechanisms and features that models use to generate predictions and outcomes are important as these tools become part of the clinical decision making process.

Identifying bias

Equivalent Sound Level was higher in UVFP patients than controls. This is counter-intuitive because UVFP patients are known to have softer voices as already described; however, clinicians rated most UVFP samples as being louder than controls. The bias discovered was likely due to increasing the gain on the microphone for some UVFP patients, which would explain the increased background noise in UVFP patients' recordings. A second source of bias may have occurred from requesting UVFP patients to speak louder in order to meet the minimum intensity threshold on the recording softwares Computerized Speech Lab™ and OperaVOX, or patients could have tried this on their own knowing they were being recorded. This behavioral compensation is likely to occur in biomarker research when the participant has a soft voice, especially in retrospective studies like ours where the study goal is not known at the time of recording or when certain software properties lead individuals with weak voices to speak louder. Even though the current models perform better than the clinicians, a systematic comparison would require more clinician and model assessments across datasets where the model's training is done on a single dataset. It is possible such a model might learn intrinsic characteristics of a dataset that do not generalize as well as

clinical expertise. Having said this, this procedure would help us understand the extent to which UVFP detection is generalizable from acoustic data alone.

Comparing tasks, model complexity, and feature set sizes

In addition to getting a better understanding of features, we explored performance in the context of different vocal tasks. Participants carried out two different tasks to elicit voice, *reading*, which captures more complex speech dynamics, and *sustaining vowels*, which is a simpler measure of vocalization and the respiratory subsystem. Overall, these dynamics from the speech task may have improved model performance as was observed. Comparing simpler and more complex models is important because simpler models such as Logistic Regression could be preferred because they tend to generalize better given they are less at risk for overfitting the training set and they are more interpretable and thus biases can be assessed more directly[36].

By removing redundant features, we can concentrate on finding the most useful features for further analysis. Performance decreased only slightly while we made models more parsimonious and explainable. Performance using the top 5 features dropped performance in most cases, with the exception of samples from the reading task which obtained an AUC of 0.85 using just the 5 features (see Figure 4). Using the individual features from within these top 5 one at a time (univariate models) reduced performance significantly to 0.55-0.71. This indicates the need for these models to combine multiple features to achieve high performance and any model evaluation

should not focus on only the common or top features without testing their predictive performance.

Limitations and future directions

We cannot determine how the bias will affect the model's performance on future samples, but it will likely underperform in samples where gain cannot be changed and where participants are instructed to not overproject their voice; however, it is possible the model underperforms for other reasons including dataset shift (e.g., the distribution of voice characteristics or demographics is different in a new sample). It is not clear whether these models could detect UVFP from other voice disorders or just healthier voices; however, a model that generalizes well in classifying UVFP from controls could be used to monitor UVFP patients remotely and affordably during treatment or detect risk for UVFP when it is the most likely cause (e.g., dysphonia after thyroid surgery). Larger sample sizes with curated examinations can help increase diverse representation across voice quality and thereby potentially reduce bias in classifier performance. Additional datasets will also help confirm the generalizability of these findings beyond the cross-validation approach used here. Our choice of a standardized feature set worked well in this setting, but may fail to work for differential voice disorder diagnosis or when generalizing to larger datasets, which may bring in additional sources of variance unaccounted for in this dataset. With the availability of more data, additional features could be extracted that better capture changes in coordination (e.g., XCORR[37]), vocal fold characteristics (e.g., cepstral peak prominence[38]) or speech rate

(i.e., given UVFP patients may speak slower). While our feature importance evaluation method, SHAP, shows a certain amount of robustness across models, alternative model-agnostic feature-importance methods (e.g., LIME, permutation importance) as well as model-specific methods (coefficient values for linear models, mean decrease in impurity for Random Forest) could be compared. Model understandability –how easily are the explanations understood by a speech scientist or clinician– could be assessed rigorously. Finally, debiasing the models by removing features correlated with the biased ones could be attempted although it is not clear how exactly intensity may influence certain features. Therefore, the effect of the bias can be assessed by testing the model's generalizability to new unbiased datasets.

Conclusion

Using the largest UVFP dataset to date, our study demonstrates the importance of checking for biases using explainable machine learning and clinician perceptual ratings. We also demonstrate the feasibility and value of testing multiple ML algorithms on data obtained from different voice tasks to better understand the process that models use to predict vocal changes associated with laryngeal disease, since analyzing a single algorithm will result in a biased view of how predictions are achieved. During this process, we discovered a likely bias resulting from the soft voice that characterizes many UVFP patients: the microphone gain was likely raised, possibly generating higher background noise and patients were prompted by the software used to speak louder. Deciphering how these models work, being able to understand strengths and

weaknesses of different algorithms, and making sure the training sets are representative of the intended uses are all aspects of ML that clinicians need to understand prior to application. We believe that establishing reliable ML tools should involve controlling audio recording, providing instructions so participants do not overproject their voice, identifying appropriate methods for feature extraction and performance evaluation (e.g., bootstrapping), explaining feature importance which may require understanding redundancy across features (i.e., addressing multicollinearity), and applying multiple models of varying complexity to understand how much feature importance can vary to then make inferences from the features that are important *across* models. With these considerations, ML applications could aid in vocal fold paralysis diagnosis, allowing for the potential development of in-home screening assessments and continuous pre- and post-treatment monitoring.

Acknowledgments

We would like to thank Daryush Mehta, Robert Hillman, John Guttag for their feedback on an earlier version of this study and to Cody Sullivan and Carolyn Hsu for their help in rating the audio samples. DML was supported by a National Institute on Deafness and Other Communication Disorders T32 training grant [5T32DC000038-28] and is supported by a RallyPoint Fellowship and an Amelia Peabody Professional Development Award. The work was supported by a gift to the McGovern Institute for Brain Research at MIT. SSG was partially supported by National Institutes of Health

grants for the development of pydra-ml [R01 EB020740], for reproducible practices [P41 EB019936], and the Bridge2AI voice data generation project [1OT2OD032720-01]. The authors declare that there is no conflict of interest.

Data Availability Statement

All data and code are available through Github (https://github.com/danielmlow/vfp) and Zenodo (https://doi.org/10.5281/zenodo.5009208).

Author Contributions

Daniel M. Low: Data curation, Methodology, Formal analysis, Software, Writing - Original Draft; Vishwanatha Rao: Data Curation, Formal analysis, Writing - Original Draft; Gregory Randolph: Writing - Review & Editing; Philip C. Song: Conceptualization, Methodology, Writing - Original Draft, Supervision, Data curation; Satrajit S. Ghosh: Conceptualization, Methodology, Writing - Original Draft, Supervision, Software;

References

1. Tracy JM, Özkanca Y, Atkins DC, Hosseini Ghomi R. Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. *J Biomed Inform*. 2020;104:103362. doi:10.1016/j.jbi.2019.103362
2. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investig Otolaryngol*. 2020;5(1):96-116. doi:10.1002/lio2.354
3. Quatieri TF. *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education; 2008.
4. Schönweiler R, Hess M, Wübbelt P, Ptok M. Novel approach to acoustical voice analysis using artificial neural networks. *Journal of the Association for Research in Otolaryngology*. 2000;1(4):270--282.
5. Godino-Llorente JI, Gomez-Vilda P. Automatic Detection of Voice Impairments by Means of

Short-Term Cepstral Parameters and Neural Network Based Detectors. *IEEE Trans Biomed Eng*. 2004;51(2):380-384. doi:10.1109/TBME.2003.820386

6. Fraile R, Sáenz-Lechón N, Godino-Llorente JI, Osma-Ruiz V, Fredouille C. Automatic Detection of Laryngeal Pathologies in Records of Sustained Vowels by Means of Mel-Frequency Cepstral Coefficient Parameters and Differentiation of Patients by Sex. *Folia Phoniatr Logop*. 2009;61(3):146-152. doi:10.1159/000219950
7. Voigt D, Döllinger M, Yang A, Eysholdt U, Lohscheller J. Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods. *Comput Methods Programs Biomed*. 2010;99(3):275-288. doi:10.1016/j.cmpb.2010.01.004
8. Lopes LW, Simões LB, da Silva JD, et al. Accuracy of acoustic analysis measurements in the evaluation of patients with different laryngeal diagnoses. *J Voice*. 2017;31(3):382-e15.
9. Powell ME, Rodriguez Cancio M, Young D, et al. Decoding phonation with artificial intelligence (D E P AI): Proof of concept. *Laryngoscope Investig Otolaryngol*. 2019;4(3):328-334. doi:10.1002/lio2.259
10. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub; 2020.
11. Stachler RJ, Francis DO, Schwartz SR, et al. Clinical Practice Guideline: Hoarseness (Dysphonia) (Update). *Otolaryngol Neck Surg*. 2018;158(1_suppl):S1-S42. doi:10.1177/0194599817751030
12. Pinho CMR, Jesus LMT, Barney A. Aerodynamic measures of speech in unilateral vocal fold paralysis (UVFP) patients. *Logoped Phoniatr Vocol*. 2013;38(1):19-34. doi:10.3109/14015439.2012.696138
13. Hartl DM, Crevier-Buchman L, Vaissière J, Brasnu DF. Phonetic Effects of Paralytic Dysphonia. *Ann Otol Rhinol Laryngol*. 2005;114(10):792-798. doi:10.1177/000348940511401009
14. Sritharan N, Chase M, Kamani D, Randolph M, Randolph GW. The vagus nerve, recurrent laryngeal nerve, and external branch of the superior laryngeal nerve have unique latencies allowing for intraoperative documentation of intact neural function during thyroid surgery: IONM Normative Range During Thyroid Surgery. *The Laryngoscope*. 2015;125(2):E84-E89. doi:10.1002/lary.24781
15. Randolph GW, Kamani D. The importance of preoperative laryngoscopy in patients undergoing thyroidectomy: Voice, vocal cord function, and the preoperative detection of invasive thyroid malignancy. *Surgery*. 2006;139(3):357-362. doi:10.1016/j.surg.2005.08.009
16. Colton RH, Paseman A, Kelley RT, Stepp D, Casper JK. Spectral Moment Analysis of Unilateral Vocal Fold Paralysis. *J Voice*. 2011;25(3):330-336. doi:10.1016/j.jvoice.2010.03.006
17. Balasubramanium RK, Bhat JS, Fahim S, Raju R. Cepstral Analysis of Voice in Unilateral Adductor Vocal Fold Palsy. *J Voice*. 2011;25(3):326-329. doi:10.1016/j.jvoice.2009.12.010
18. Little MA, Costello DAE, Harries ML. Objective Dysphonia Quantification in Vocal Fold Paralysis: Comparing Nonlinear With Classical Measures. *J Voice*. 2011;25(1):21-31. doi:10.1016/j.jvoice.2009.04.004
19. Bielamowicz S, Stager SV. Diagnosis of unilateral recurrent laryngeal nerve paralysis: laryngeal electromyography, subjective rating scales, acoustic and aerodynamic measures. *The Laryngoscope*. 2006;116(3):359-364. doi:10.1097/01.MLG.0000199743.99527.9F
20. Hartl DAM, Hans S, Vaissière J, Brasnu DAMF. Objective acoustic and aerodynamic measures of breathiness in paralytic dysphonia. *Eur Arch Oto-Rhino-Laryngol Off J Eur Fed Oto-Rhino-Laryngol Soc EUFOS Affil Ger Soc Oto-Rhino-Laryngol - Head Neck Surg*.

2003;260(4):175-182. doi:10.1007/s00405-002-0542-2
21. Fairbanks G. *Voice and Articulation Drillbook*. Harper; 1960. https://books.google.com/books?id=qN1ZAAAAMAAJ
22. Eyben F, Scherer KR, Schuller BW, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans Affect Comput*. 2016;7(2):190-202. doi:10.1109/TAFFC.2015.2457417
23. audEERING GmbH. openSMILE (Version 2.3) [https://github.com/naxingyu/opensmile/blob/3a0968e7b36c1b730a4ffd2977031091ee9abf7f/config/gemaps/eGeMAPSv01a.conf]. Published online 2017.
24. Satrajit Ghosh, Low DM, Hoda1394, et al. *Nipype/Pydra-Ml: Zenodo Release for Doi*. Zenodo; 2020. doi:10.5281/ZENODO.4170850
25. Ojala M, Garriga GC. Permutation tests for studying classifier performance. *J Mach Learn Res*. 2010;11(6).
26. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. *ArXiv170507874 Cs Stat*. Published online November 24, 2017. Accessed October 19, 2020. http://arxiv.org/abs/1705.07874
27. Francis DO, Pearce EC, Ni S, Garrett CG, Penson DF. Epidemiology of Vocal Fold Paralyses after Total Thyroidectomy for Well-Differentiated Thyroid Cancer in a Medicare Population. *Otolaryngol Neck Surg*. 2014;150(4):548-557. doi:10.1177/0194599814521381
28. Vila PM, Bhatt NK, Paniello RC. Early-injection laryngoplasty may lower risk of thyroplasty: A systematic review and meta-analysis: Early Injection for Unilateral Vocal Fold Paralysis. *The Laryngoscope*. 2018;128(4):935-940. doi:10.1002/lary.26894
29. Dhillon VK, Randolph GW, Stack BC, et al. Immediate and partial neural dysfunction after thyroid and parathyroid surgery: Need for recognition, laryngeal exam, and early treatment. *Head Neck*. 2020;42(12):3779-3794. doi:10.1002/hed.26472
30. Jeannon J-P, Orabi AA, Bruch GA, Abdalsalam HA, Simo R. Diagnosis of recurrent laryngeal nerve palsy after thyroidectomy: a systematic review. *Int J Clin Pract*. 2009;63(4):624-629. doi:10.1111/j.1742-1241.2008.01875.x
31. Bhattacharyya N, Kotz T, Shapiro J. Dysphagia and Aspiration with Unilateral Vocal Cord Immobility: Incidence, Characterization, and Response to Surgical Treatment. *Ann Otol Rhinol Laryngol*. 2002;111(8):672-679. doi:10.1177/000348940211100803
32. Ramig LA, Titze IR, Scherer RC, Ringel SP. Acoustic Analysis of Voices of Patients with Neurologic Disease: Rationale and Preliminary Data. *Ann Otol Rhinol Laryngol*. 1988;97(2):164-172. doi:10.1177/000348948809700214
33. Morsomme D, Jamart J, Wéry C, Giovanni A, Remacle M. Comparison between the GIRBAS Scale and the Acoustic and Aerodynamic Measures Provided by EVA for the Assessment of Dysphonia following Unilateral Vocal Fold Paralysis. *Folia Phoniatr Logop*. 2001;53(6):317-325. doi:10.1159/000052685
34. Kriegeskorte N, Douglas PK. Interpreting encoding and decoding models. *Curr Opin Neurobiol*. 2019;55:167-179. doi:10.1016/j.conb.2019.04.002
35. Jacobucci R, Littlefield AK, Millner AJ, Kleiman EM, Steinley D. Evidence of Inflated Prediction Performance: A Commentary on Machine Learning and Suicide Research. *Clin Psychol Sci*. Published online January 7, 2021:216770262095421. doi:10.1177/2167702620954216
36. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x
37. Williamson JR, Quatieri TF, Helfer BS, Ciccarelli G, Mehta DD. Vocal and Facial

Biomarkers of Depression based on Motor Incoordination and Timing. In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14*. ACM Press; 2014:65-72. doi:10.1145/2661806.2661809

38. Murton O, Hillman R, Mehta D. Cepstral Peak Prominence Values for Clinical Voice Evaluation. *Am J Speech Lang Pathol*. 2020;29(3):1596-1607. doi:10.1044/2020_AJSLP-20-00001

Supplementary Table legends

Table S1. Prior studies on voice disorders. N: sample size; MEEI: Kay Electronics Mass. Eye and Ear Infirmary (MEEI) CD-ROM dataset; UVFP: Unilateral Vocal Fold Paralysis; SD: standard deviation; f0: fundamental frequency.

Table S2. Comparison of selecting features on the entire dataset (useful for explainability) versus selecting on 50 bootstrap (80–20) train splits. Original total features are 88. CI = Confidence Interval.

Table S3. Performance of models without 24 patients recorded on iPad. Median ROC AUC score from 50 bootstrapping splits (90% confidence interval; median score of null model). The control group represents 60% of the training samples. MLP: Multi-Layer Perceptron; SGD: Stochastic Gradient Descent Classifier.

Table S4. False negative rate (FNR) of training on one recording device and testing on 24 UVFP patients that used iPad. FNR is generally quite low. Performance can also be influenced by having a smaller training set in order to balance the classes.

Supplementary Figure legends

Figure S1. All participants, reading task: Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

Figure S2. All participants, vowel task: Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

Figure S3. All participants, reading+vowel tasks: Visualization of features with shared information using pairwise distance correlation across the 88 features. Squares are clusters of redundant features.

Figure S4. Patients, reading task: Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

Figure S5. Patients, vowel task: Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

Figure S6. Patients, reading+vowel tasks: Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

Figure S7. Controls, reading task: Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

Figure S8. Controls, vowel task: Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features extracted. Squares are clusters of redundant features.

Figure S9. Controls, reading+vowel tasks: Visualization of features with shared information using pairwise distance correlation across the 88 features. Squares are clusters of redundant features.

Figure S10. Performance as a function of feature set size using Independence Factor method for reducing feature redundancy. The feature sets remove features with distance correlation ≥ 0.2 up to 1.0 (i.e., keeping all features) in increments of 0.1.