Uncovering the important acoustic features for detecting vocal fold paralysis with explainable machine learning

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

[1] Program in Speech and Hearing Bioscience and Technology, Harvard Medical School, Boston, MA, USA
[2] Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
[3] McGovern Institute for Brain Research, MIT, Cambridge, MA, 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

Correspondence can be addressed to Philip C. Song and Daniel M. Low. Massachusetts Eye and Ear Infirmary, 243 Charles St, Boston, MA 02114, USA. E-mails: phillip_song@meei.harvard.edu, dlow@mit.edu.

Abstract

Objectives: To detect unilateral vocal fold paralysis (UVFP) from voice recordings using an explainable model of machine learning.
Study Design: Case series - retrospective with a control group.
Methods: Patients with confirmed UVFP through endoscopic examination (N=77) and controls with normal voices matched for age and sex (N=77) were included. Two tasks were used to elicit voice samples: reading the Rainbow Passage and sustaining phonation of the vowel /a/. The eighty-eight extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features were extracted as inputs for four machine learning models of differing complexity. Training and testing were performed using bootstrapped cross-validation. SHAP was used to identify important features.
Results: The median Area Under the Receiver Operating Characteristic Curve (ROC AUC) score ranged from 0.79 to 0.87 depending on model and task. After removing redundant features for explainability, the highest median ROC AUC score was 0.84 using only 13 features for the vowel task and 0.87 using 39 features for the reading task. The most important features included intensity measures, mean MFCC1, mean F1 amplitude and frequency, and shimmer variability depending on model and task.
Conclusion: Using the largest dataset studying UVFP to date, we achieve high performance from just a few seconds of voice recordings while discovering which acoustic features are important across models. Notably, we demonstrate that the models use different combinations of features to achieve similar effect sizes. Overall the categories of features related to vocal fold physiology were conserved across the models. Machine learning thus provides a mechanism to detect UVFP and contextualize the accuracy relative to both model architecture and pathophysiology.

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

Level of Evidence: Type 3

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[2]. Despite these advances, robust applications to detect specific voice disorders remain limited[3,4].

There are multiple challenges for applying acoustic analysis to detect specific disorders. Voice characteristics can be highly varied and change over time. Laryngeal pathology, language, age, gender, size, weight, general state of health, smoking/vaping, and medications all represent variables that can impact the vocal acoustic characteristics. Diseases in the larynx and phonatory system (encompassing the larynx, resonating structures, and lungs) and neurological system, will also affect voice. Patient compensation 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.

Unilateral vocal fold paralysis (UVFP) occurs when the mobility of a single vocal fold is impaired as a consequence of neurological injury. The clinical features of UVFP are weak, breathy voice quality, early vocal fatigue, reduced cough strength and aspiration with thin liquids[5,6]. Diagnosis is made based on a laryngeal examination, most commonly nasopharyngoscopy, which is performed by an otolaryngologist. UVFP is commonly associated with surgery or malignancy, idiopathic, and neurological disease, and impacts quality of life. Overall, surgical iatrogenic injury accounts for 46% of all UVFP in adults. Of these, thyroid and parathyroid surgeries are responsible for 32% of postsurgical UVFP[8]. 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, as well as association with malignancies and neoplasms, and surgical complications, especially in regions where surgical specialists are not readily accessible.

Machine learning algorithms are capable of capturing complex nonlinear relationships in data. They are naturally suited for applications using high dimensional physiological measures such as voice samples, cardiac abnormalities, skin examinations, and radiographs. Using a machine learning model as a screening tool for UVFP can reduce the need for laryngoscopy and provide an understanding of voice characteristics related to the pathophysiology[9–13]. Because machine learning algorithms can work with high-dimensional data, they can detect and associate unintended or clinically irrelevant relationships that may go unrecognized to the user. This can introduce new knowledge or introduce bias. To elucidate how a model works, several approaches help quantitatively estimate important input characteristics[14]. The objectives of our study were: (1) to detect UVFP using machine learning; (2) to evaluate the effectiveness of different models in differentiating the acoustic signals between patients with UVFP and patients with working vocal folds (i.e., controls); and (3) to explain which features are most important to the diagnostic models and their 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 evaluate detection performance, evaluated the minimal set of

features necessary for detection, and estimated the most relevant features for model construction.

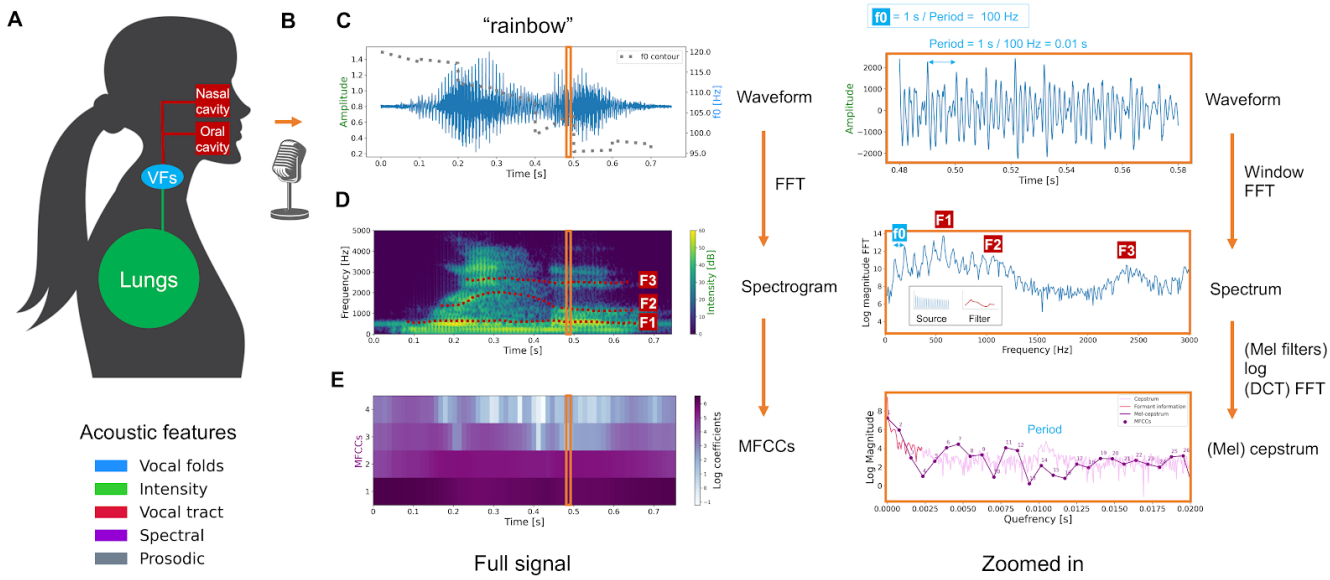

Figure 1. Schematic of speech production and different ways of displaying an audio signal along with associated acoustic features. (A) Speech is a result of the neural coordination of three subsystems: the respiratory system (lungs), the laryngeal system (vocal folds), and the resonatory system (pharynx, oral cavity, nasal cavity, and subglottal effects). Speech production requires air flow from the lungs to generate sound sources that are filtered by the vocal tract. **(B)** Environmental, microphone, and digital sampling characteristics (e.g., background noise, microphone gain, sampling rate) can affect acoustic features. **(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. **(D)** 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 at 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. **(E)** To separate source and filter components one can compute 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. For more in depth discussion see, Chapter 3 of Quatieri (2008)[7].

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 vocalizations of the /a/ vowel sound and a reading of the introductory paragraph of the rainbow passage[15]. 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. A separate 24 samples were collected prospectively using a mobile software, OperaVOX™ on an iPad from patients who were currently 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 as the UVFP patient and within three years of age. The controls had recorded the same four vocal files 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. 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[16], source code[17]) 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).

Machine Learning Models of Increasing Complexity

Using the pydra-ml toolbox18, four machine learning algorithms of increasing complexity from the scikit-learn package (sklearn) were used (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 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 nonlinear 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).

	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
Total	449	462	911

Table 1. Sample sizes and demographic information.
SD: standard deviation; F: female; M: male.

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 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 = 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 a Wilcoxon signed-rank test.

Finally, we used Kernel SHAP to determine which features were important for each model. This method is model agnostic in that it can take any trained target model (even "black box" neural networks) and compute feature importance[19]. 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 absolute SHAP value across all test predictions (positive and negative values are 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 parsimonious and explainable models

Highly correlated features can influence model generation and interpretation. Two models may obtain similar performance while using different features or placing different weights on the same features. This makes it difficult to compare algorithmic explanations across models. For instance, mean F1 frequency may be less important to a given model because it uses mean F2 frequency which happens to capture very similar information, 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 whereby for each feature in alphabetical (i.e., arbitrary) order, we removed features that show strong dependence above a given threshold. The step was repeated for remaining features. We used distance correlation through the Python dcor package to capture linear and nonlinear relationships[20]. We used the following threshold values for the distance correlation [1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2] to compute the Independence Factor, which removed increasingly more features (i.e., 1.0 keeps all features and 0.2 removes features that have a distance correlation above 0.2). We chose the feature size which contains at least one model that scores within three percentage points of the performance using all features, with the goal of obtaining a more parsimonious model for subsequent explanation while maintaining high accuracy. Thus, removing redundant features makes the models easier to interpret for clinical relevance. To visualize the original redundancy across features, we computed clustermaps using seaborn package performing hierarchical clustering with the average-linkage method and Euclidean distance. This was performed on the pairwise distance correlation, computed separately on data from UVFP, controls, UVFP+controls and on reading, vowel, and reading+vowel.

RESULTS

Performance with and without redundant features

Figure 2 shows dependence across features using samples from all participants for reading+vowel (see Supplementary Materials Figures S1–9 for separation between UVFP and controls separated by tasks). For further description of features and the chosen classification, see Eyben et al. (2015)[16] and Low et al. (2020)[2]. The chosen classification of features appears to be empirically replicated in this dataset given most low-level clusters (i.e., have higher dependency) are for the most part homogenous (i.e., of the same color). This also allows us to observe exceptions (e.g., mean spectral flux clusters with loudness features) which could otherwise be missed if using only a priori theoretical knowledge.

Given dependent features provide similar information and distort feature importance analyses, we then tested performance after removing redundant features using the Independence Factor method previously described. See Figure 3 for performance for different feature set sizes with increasing amounts of redundant features. From this analysis, we used 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). By removing redundant features (i.e., reducing multicollinearity) from the original 88 features, similar performance was obtained (median ROC AUC = 0.84–0.87) using fewer features. See Supplementary Materials "Feature selection" section for an analysis of how this method compares to removing features across each train set.

The cross-validated ROC AUC distributions and permutation tests for the parsimonious models are shown in Figure 4. See Table 2 for performance using all features and a subset of features selected by either removing redundant features while maintaining performance (as in Figure 3) or using the top 5 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, 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 features (see Supplementary Materials Table S2).

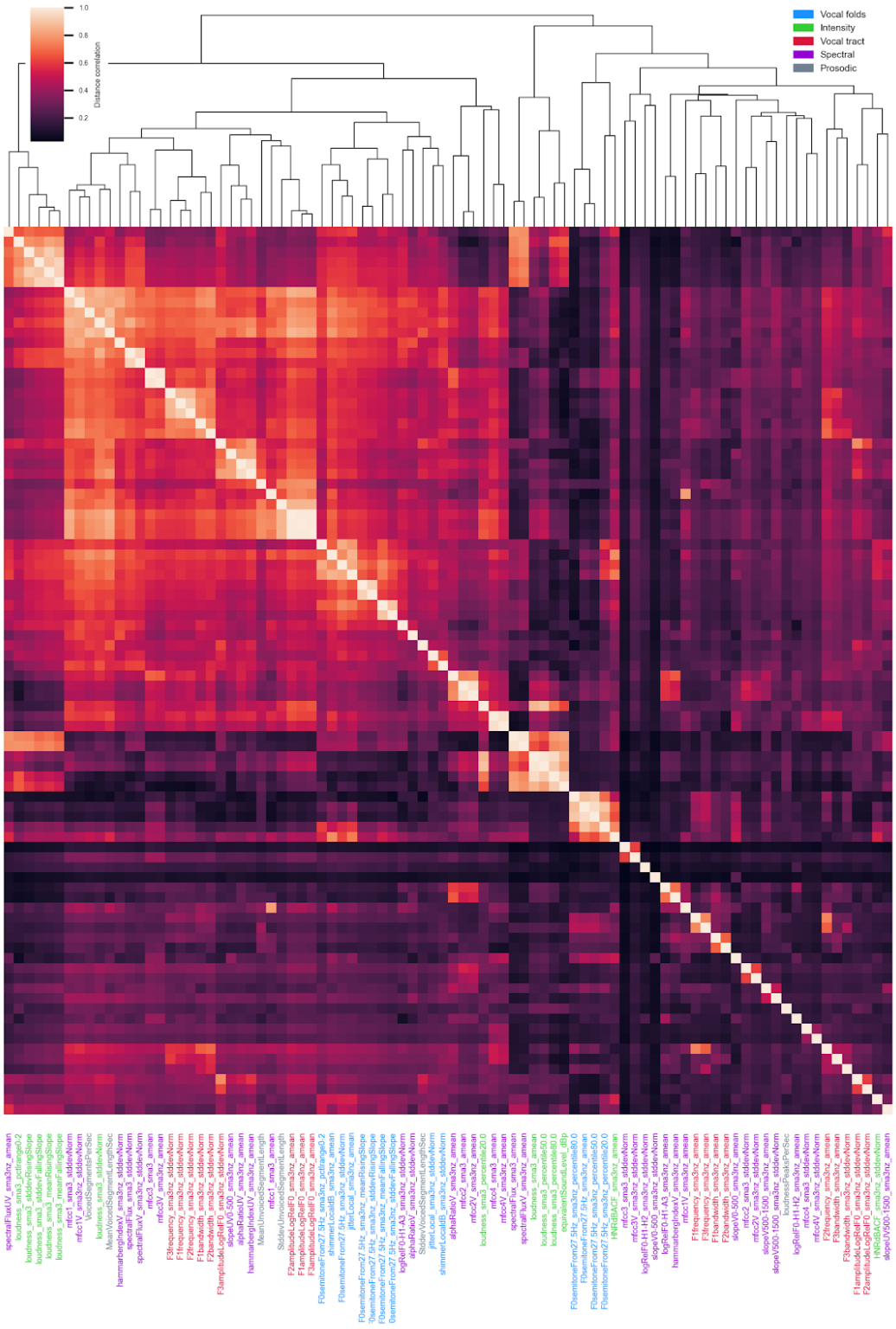

Figure 2. Visualization of features with redundant information. Pairwise distance correlation across the 88 eGeMAPs features extracted from both reading and vowel recordings. Squares are clusters of redundant features. A value of 1 indicates complete dependence, while a value of 0 indicates independence. The features are organized according to a hierarchical clustering algorithm.

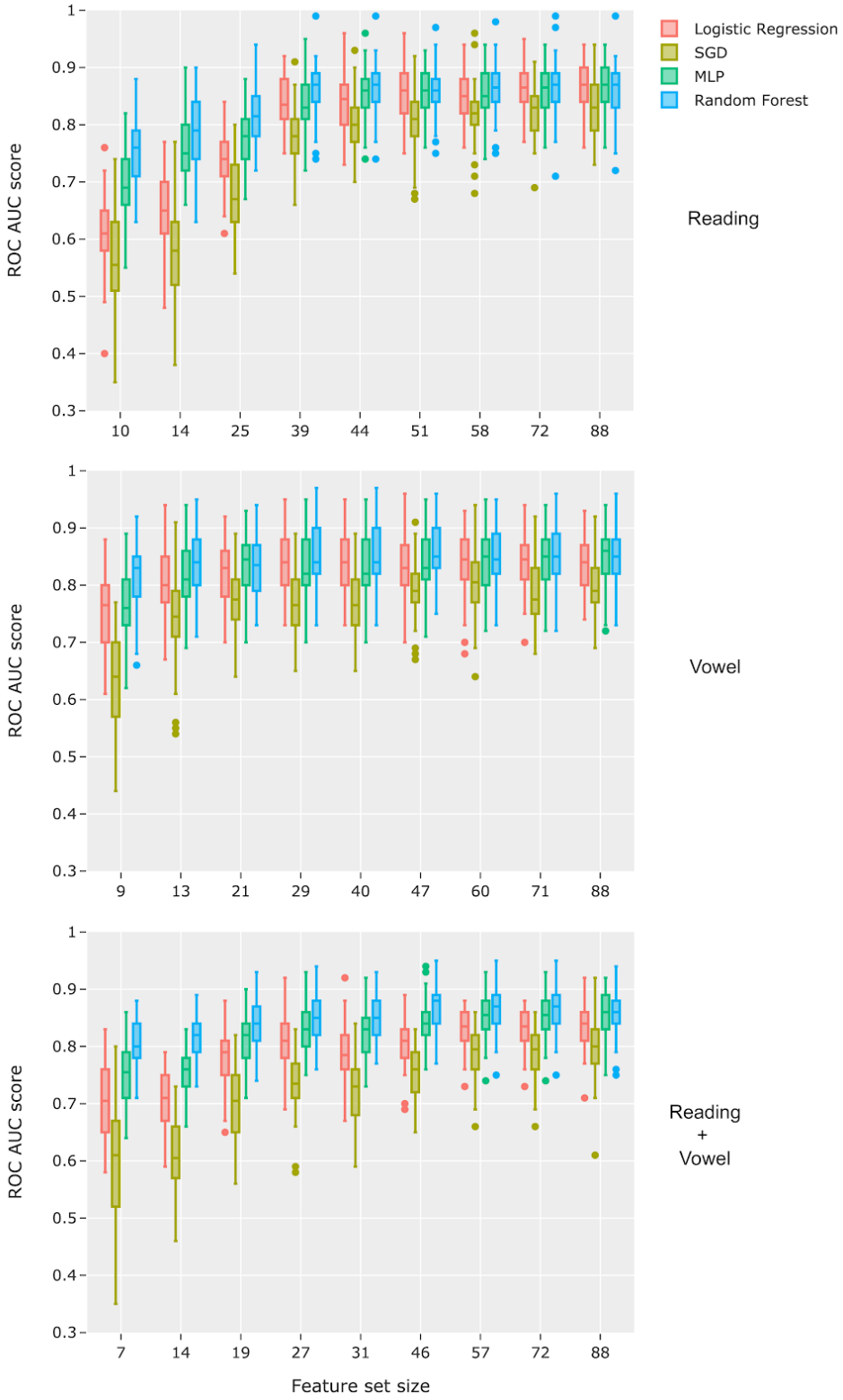

Figure 3. Performance as a function of feature set size using Independence Factor method for reducing feature redundancy. The feature sets remove features with distance correlation ≥ 0.2 up to 1.0 (i.e., keeping all features) in increments of 0.1.

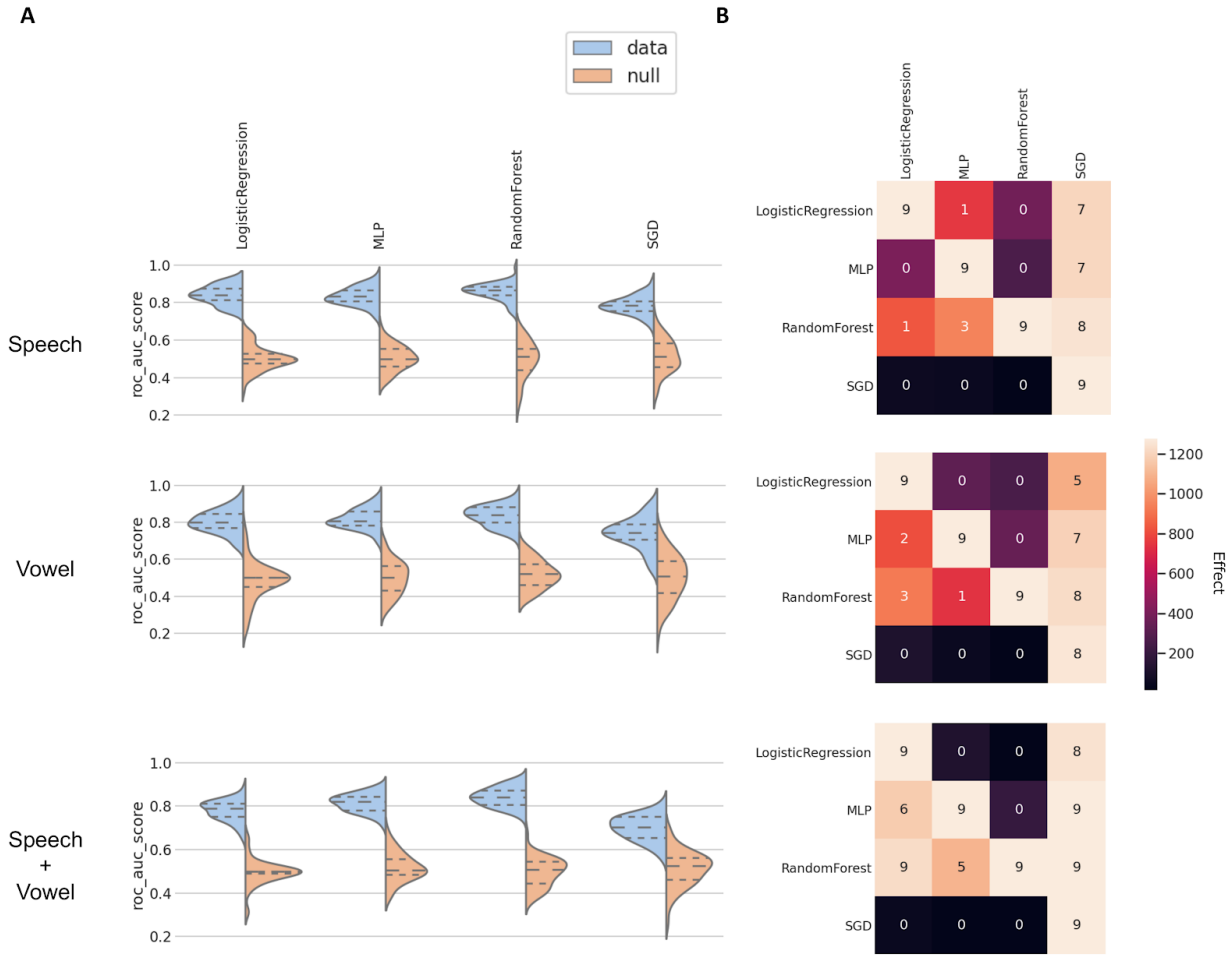

Figure 4. Model performance comparison using non-redundant feature sets. (A) The distribution of performance from models trained on true labels (blue) and trained on permuted labels (orange) over 50 bootstrapping splits. **(B)** One tailed statistical comparison (row > column) of models using a Wilcoxon signed rank test (non-parametric paired t-test). The diagonal represents a comparison of the model with a null model. The figures report statistical values as the color of each cell and the corresponding -log10(P-value) as the annotation; higher numbers indicate stronger effects (color) and lower P-values (annotations). For instance, -log10(0.01) = 2 and -log10(0.001) = 3. All of these tests between data and null distributions were significant compared to the alpha value of 0.05.

Task	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	5	.81 (.73–.89; .50)	.86 (.78–.92; .47)	.85 (.77–.90; .50)	.75 (.56–.87; .57)
Vowel	5	.78 (.67–.87; .50)	.82 (.74–.92; .53)	.81 (.72–.87; .50)	.72 (.57–.82; .49)
Reading+Vowel	5	.80 (.70–.86; .50)	.82 (.74–.88; .50)	.81 (.74–.89; .53)	.72 (.55–.83; .52)

Table 2. Performance of models using either all 88 features, non-redundant features (39, 13, 19), and top five features.
Median ROC AUC score from 50 bootstrapping splits (95% confidence interval; median score of null model). For full distributions of scores see Figure 3. 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.

Feature Importance

See Figure 5 for feature importance using SHAP for all models. 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 3. We further show the distributions for each top feature for both groups (UVFP and controls) and test the classification performance using single features (Figure 6).

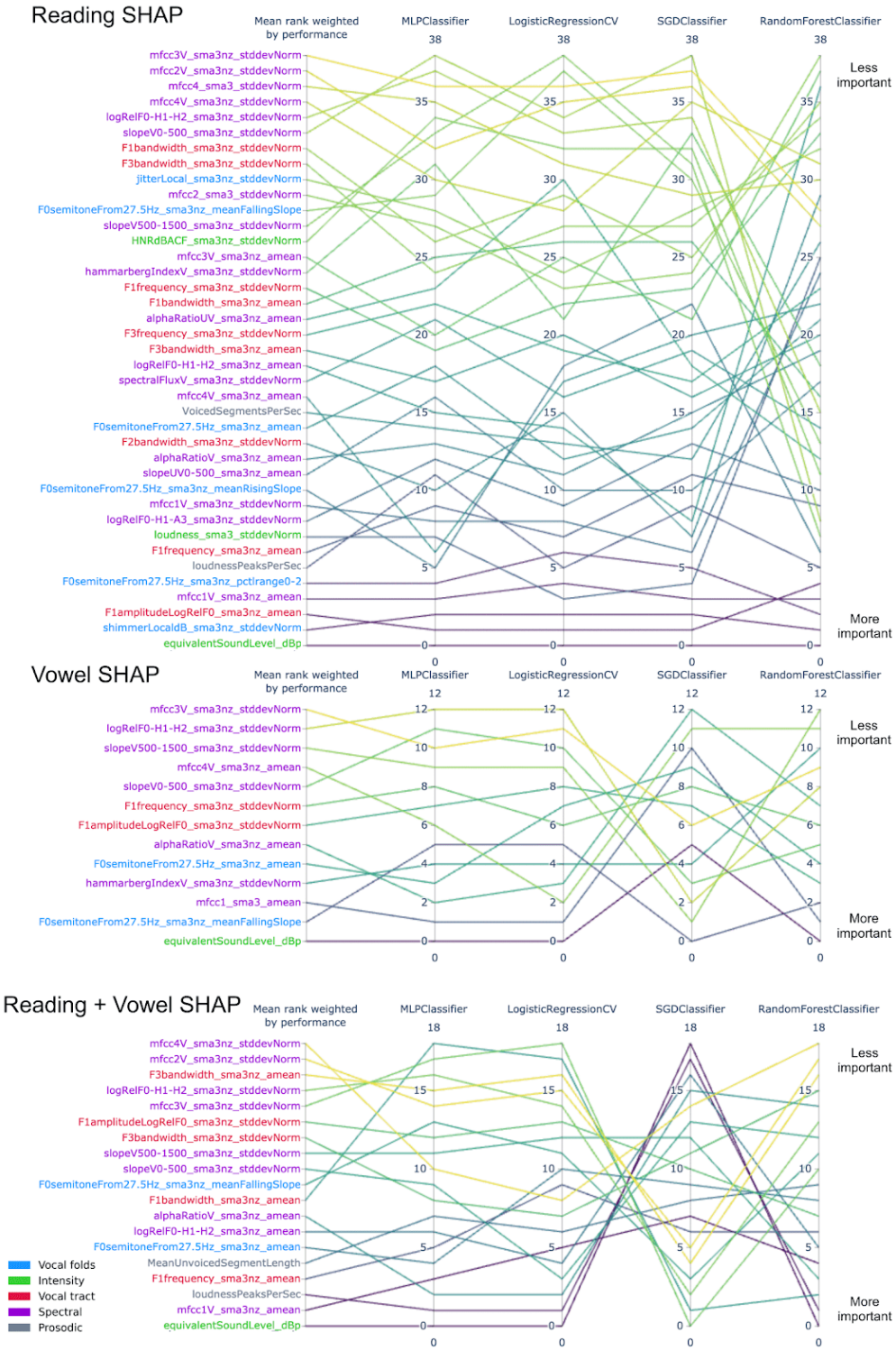

Figure 5. Feature importance parallel coordinate plot. Rank reads from bottom (most important) to top (least important). Mean rank is weighed by performance of each model to avoid a lower performing model biasing the mean rank. When reviewing important features, it is key to note that any of the features with which it is codependent could be a reasonable important feature (see Figure 2).

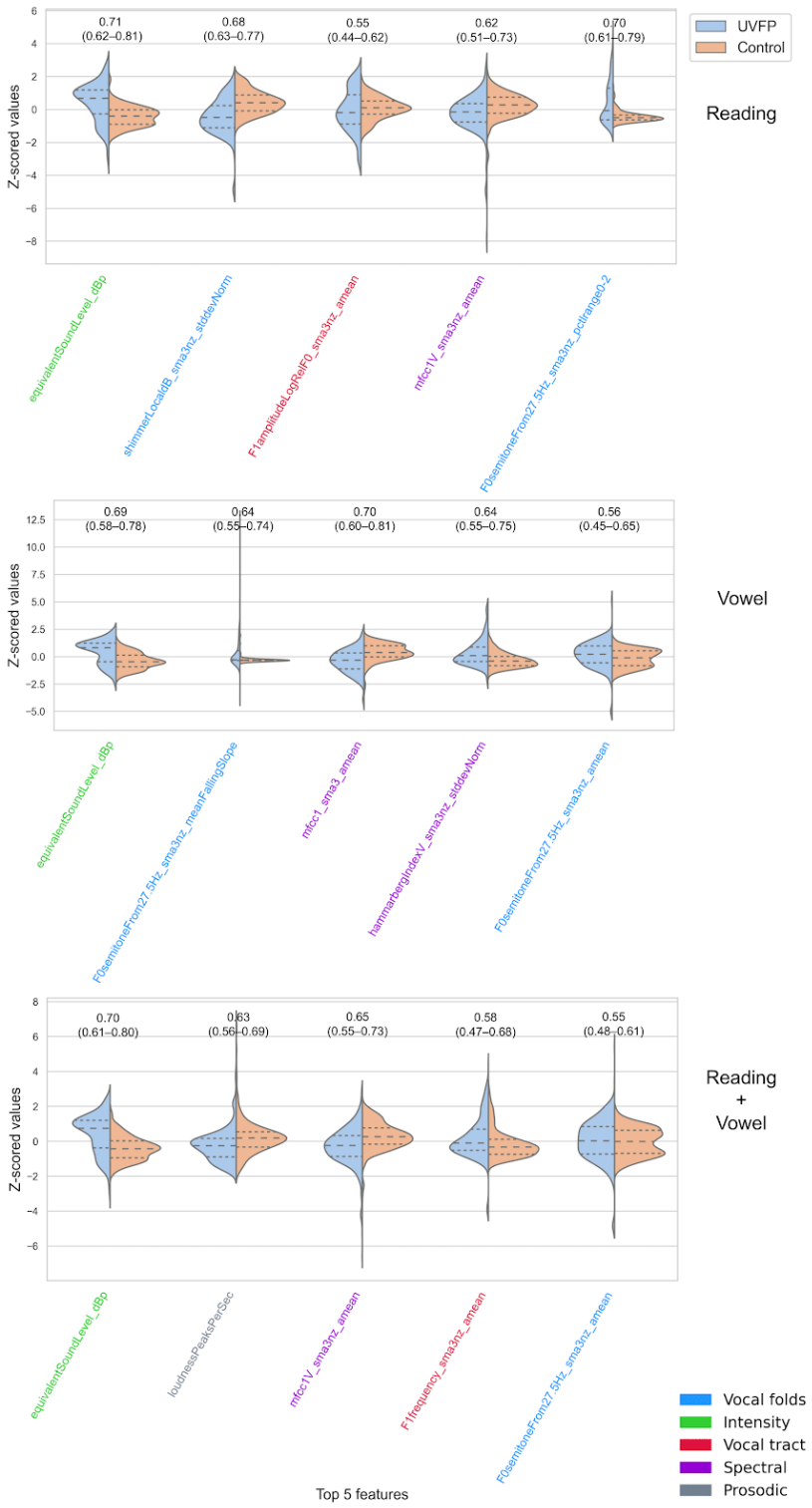

Figure. 6. Distributions for top 5 features and corresponding performance for single features using a Logistic Regression with L1 penalty. No single feature is enough to dissociate groups with high performance; high performance is obtained by combining features as in Figure 5. The median performance of all null models was 0.5.

Discussion

This study presents the results of machine learning on one of the largest datasets of vocal fold paralysis voice samples to detect UVFP. We achieve high performance while explaining which acoustic features are important. Critically, we demonstrate that interpreting detection performance has to be contextualized with respect to the type of the machine learning model, the features considered important, and the voice-eliciting task.

The need for automated assessments of vocal fold paralysis

Vocal fold pathology is associated with measurable voice changes, but application of objective acoustic measurements towards differentiating between voice conditions have been inconsistent[21–25]. Prior studies have used pre-existing commercial databases, smaller sample sizes, fewer features, and/or methods for model evaluation that can be biased in small datasets (for a discussion, see [2]). 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. 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[26,27]. Access to specialists for diagnosis is limited and early detection and management of UVFP appear to improve length of stay and surgical outcomes[28].

Our approach: comparing tasks, model complexity, and feature set sizes

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, speech performed slightly better. Using all features, all of the models except the SGD classifier demonstrated strong and consistent classification performance and correctly discriminated between UVFP and controls 84–87% of the time depending on the voice-eliciting task. 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[29].

We then removed redundant features with the goal of forcing models to use the same subset of features instead of one of multiple redundant features to understand how important the subset was for each model. Performance decreased only slightly while we made models more parsimonious. Notably, it was possible to detect UVFP using only 13 features from the vowel task. These features corresponded to information about energy expended and fidelity of vocal fold vibration. While the smallest number of features varied per task, this post hoc analysis provided an intuition behind the most useful features and can help tune data collection to leverage certain classes of information.

When using only the top 5 features, performance remained high only mainly for the reading dataset, since the top 5 rank was more consistent across models. When using only one of the top 5 features (see Figure 6), performance decreased considerably, and demonstrates that high performance requires combining multiple features.

Explaining important features related to vocal fold paralysis

Our study provides a computational explanation for how the algorithms perform the task of identifying vocal fold paralysis from normal voices and show that models can achieve similar performance using different features. This should raise watchfulness when evaluating feature importance only using the single highest performing model, which is commonly seen in the field of detecting medical disorders using machine learning.

Objective acoustic measurement changes associated with vocal fold paralysis have been described[30–32]. These changes include reduced loudness and maximum phonation time, higher perturbation measurements such as jitter and shimmer, and increased signal to noise ratio. In our study, using a much larger pool of possible acoustic features for analysis, the methodology and processing suggest that certain acoustic features can –in combination– show predictive capability using different models and different voice-eliciting tasks. The analysis of four models of increasing complexity using three datasets generated 12 separate ranks lists of feature importance.

Instead of measuring feature importance on the highest performing model which often is only slightly better than the second best model, we find the features that are important across models. 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 enunciate louder), Mel Frequency Cepstral Coefficients (especially the first coefficient, which captures spectral envelope or slope), mean F0 semitones (which should be altered in UVFP), mean F1 amplitude and frequency (influenced by how the vocal tract filters f0), and voiced and unvoiced segments (prosodic 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. These acoustic features track 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.

Limitations and future directions

The development of a machine learning screening tool for vocal fold paralysis is favorable and accuracy will improve with larger sample size and additional curated examinations. A major advantage of machine learning algorithms is the ability to merge multiple forms of data, including electrophysiological, image, video, motion capture, and acoustic into a comprehensive data set. However, the development and training of these algorithms rely on the accurate introduction of training data that are relevant to the given task. The accuracy will

improve with larger numbers, but even with our sample size, the algorithms were able to discriminate with reasonable fidelity.

While we chose a standardized feature set, additional features could be extracted that better capture changes in coordination (e.g., XCORR[33]) or vocal fold characteristics (e.g., cepstral peak prominence[34]) . Furthermore, models that capture sequential dynamics could be used (e.g., recurrent neural networks). Regarding performance, future studies could do hyperparameter tuning on larger datasets, which could increase performance.

We removed features on the entire dataset instead of a nested cross-validation approach which is preferable because this could result in different features being selected. In an empirical evaluation of the alternate approach, we found that performance did not drop considerably, did not change our interpretation, while saving significant computation time. Moreover, while 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.

Conclusion

Using the largest dataset to date, our study demonstrates the feasibility and value of testing multiple machine learning algorithms on data obtained from different voice tasks to better understand the process that models use to predict vocal changes associated with laryngeal disease. However, 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 machine learning that clinicians need to understand prior to application. We believe that establishing reliable machine learning tools should involve using expertly-curated clinical data, identifying appropriate methods for feature extraction and performance evaluation, explaining feature importance which may require removing redundant features, and applying multiple models of various complexity to understand how much feature importance can vary to then make inferences from the features that are important across models. With these considerations, machine learning applications can aid in vocal fold paralysis diagnosis, allowing for the potential development of in-home screening assessments and continuous pre- and post-treatment monitoring.

All data and code has been publicly released[35].

Acknowledgements

DML was supported by a National Institutes of Health (NIH) training grant (NIDCD 5T32DC000038). The work was supported by a gift to the McGovern Institute for Brain Research at MIT. SSG was partially supported by NIH grant R01 EB020740 (development of pydra-ml) and P41 EB019936 (reproducible practices). None of the authors has a conflict of interest.

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. 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.
4. Gillespie AI, Dastolfo C, Magid N, Gartner-Schmidt J. Acoustic analysis of four common voice diagnoses: moving toward disorder-specific assessment. *J Voice*. 2014;28(5):582–588.
5. 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
6. 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
7. Quatieri TF. *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education; 2008.
8. 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
9. 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
10. 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
11. 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
12. 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
13. 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
14. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub; 2020.
15. Fairbanks G. *Voice and Articulation Drillbook*. Harper; 1960. https://books.google.com/books?id=qN1ZAAAAMAAJ
16. 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

17. audEERING GmbH. openSMILE (Version 2.3) [https://github.com/naxingyu/opensmile/blob/3a0968e7b36c1b730a4ffd2977031091ee9abf7f/config/gemaps/eGeMAPSv01a.conf]. Published online 2017.
18. Satrajit Ghosh, Low DM, Hoda1394, et al. *Nipype/Pydra-Ml: Zenodo Release for Doi*. Zenodo; 2020. doi:10.5281/ZENODO.4170850
19. 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
20. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *Ann Stat*. 2007;35(6):2769-2794. doi:10.1214/009053607000000505
21. 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
22. 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.
23. 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
24. 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
25. 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
26. 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
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. 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
29. 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
30. Hartl DM, Hans S, Vaissière J, Brasnu DF. Objective acoustic and aerodynamic measures of breathiness in paralytic dysphonia. *Eur Arch Otorhinolaryngol*. 2003;260(4):175-182. doi:10.1007/s00405-002-0542-2
31. 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
32. 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

33. 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
34. 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
35. Low DM. *https://github.com/danielmlow/vfp*. Zenodo; 2020. Accessed November 23, 2020. doi.org/10.5281/zenodo.4287654

Supplementary Materials

Visualization of Redundant Features

See Figure S1–S9 for a visualization of redundant features for all participants, patients, and controls and for reading, vowel, and reading+vowel tasks. When stratifying samples by disorder and task, clustering becomes more homogenous (clusters tend to contain a single feature type) in comparison to when all participants or both tasks are included.

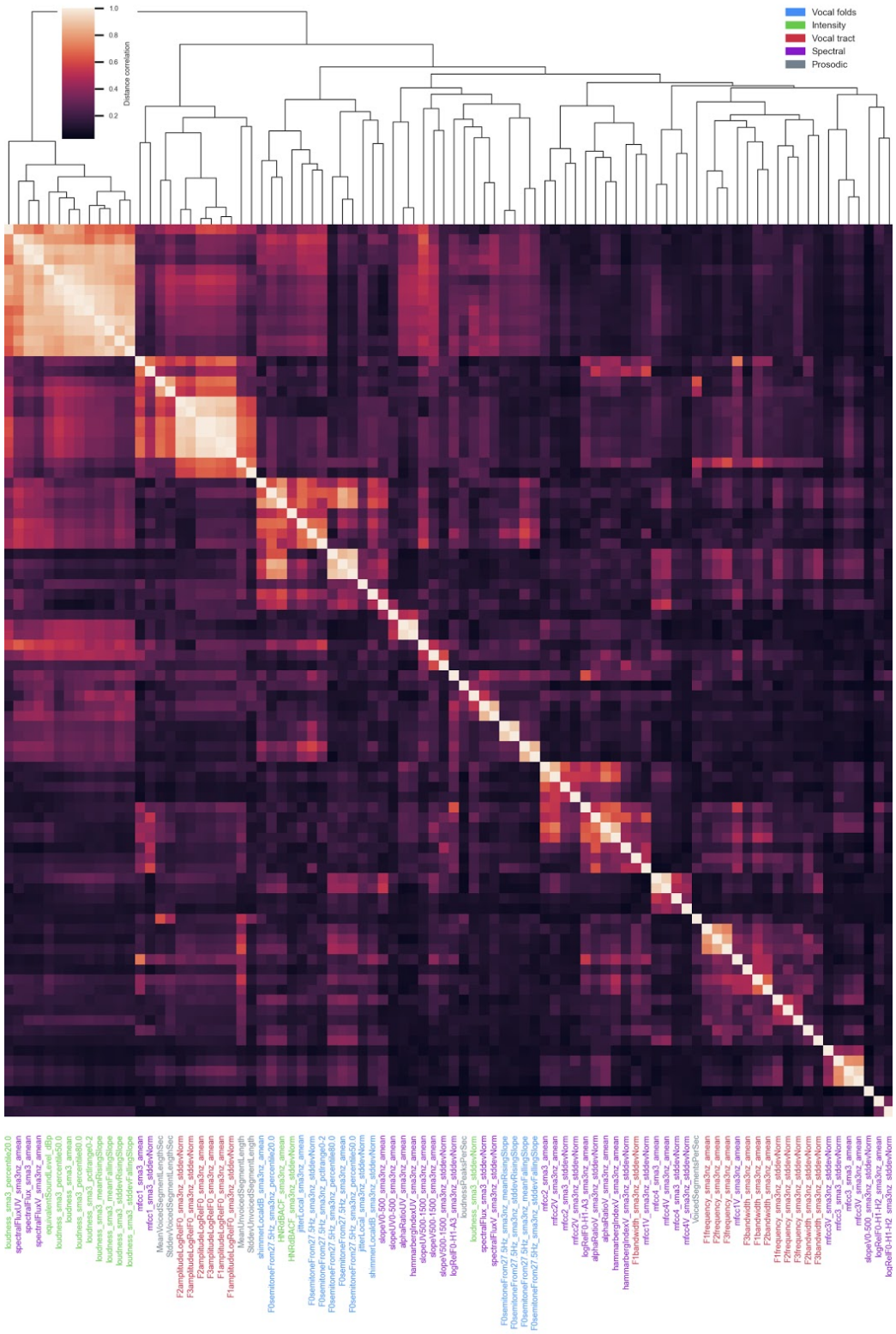

Figure S1. All participants, reading task. Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

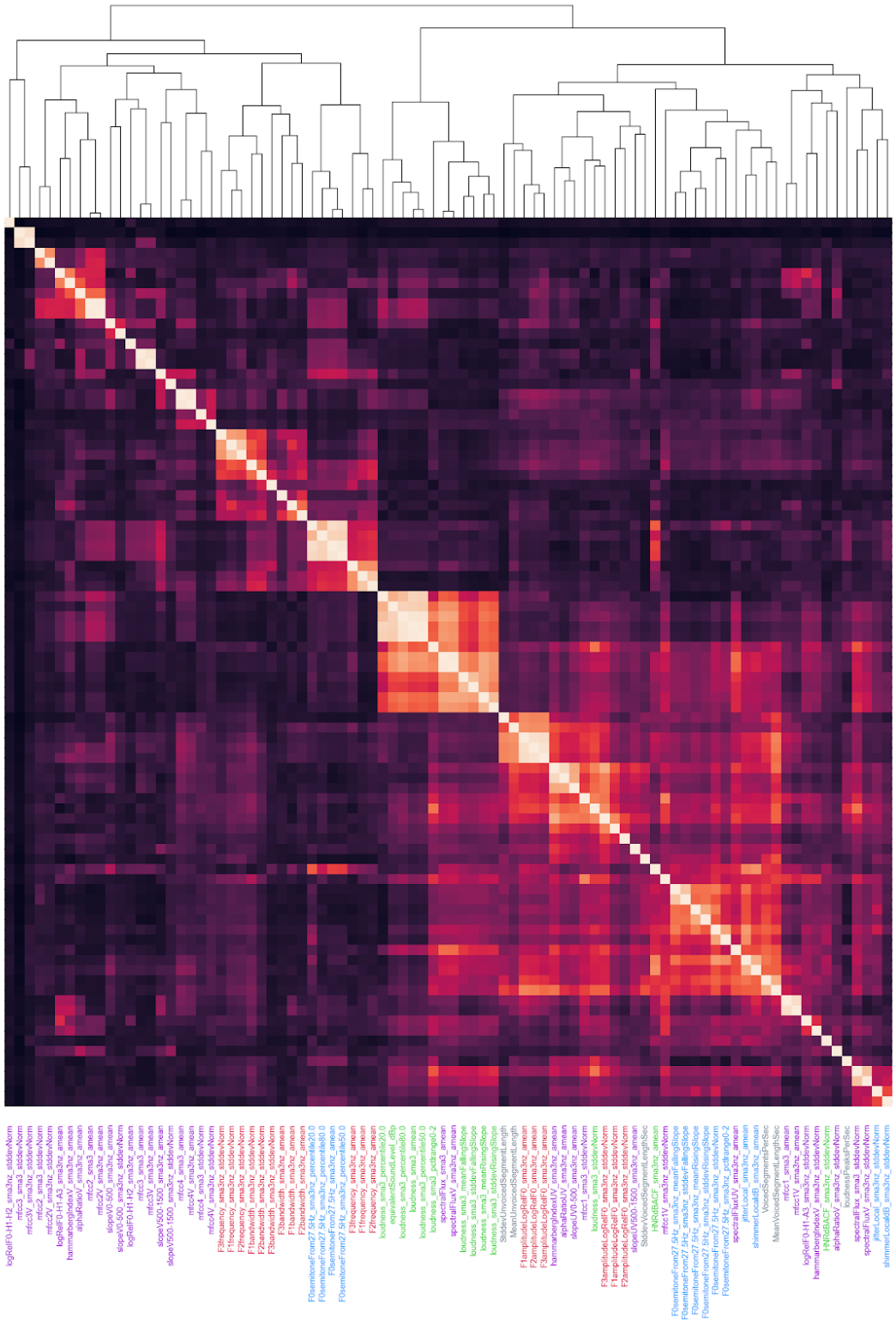

Figure S2. All participants, vowel task. Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

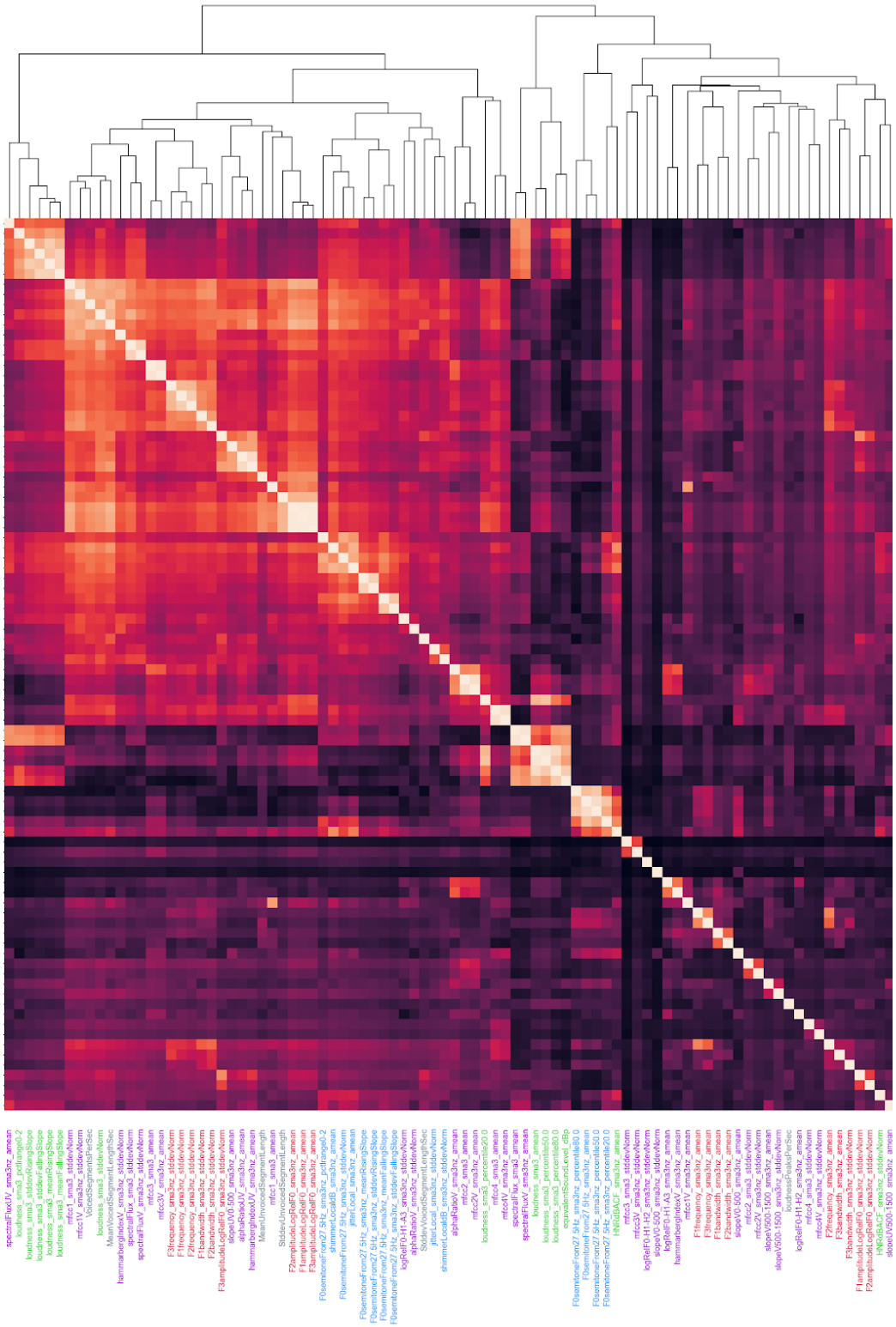

Figure S3. All participants, reading+vowel tasks. Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

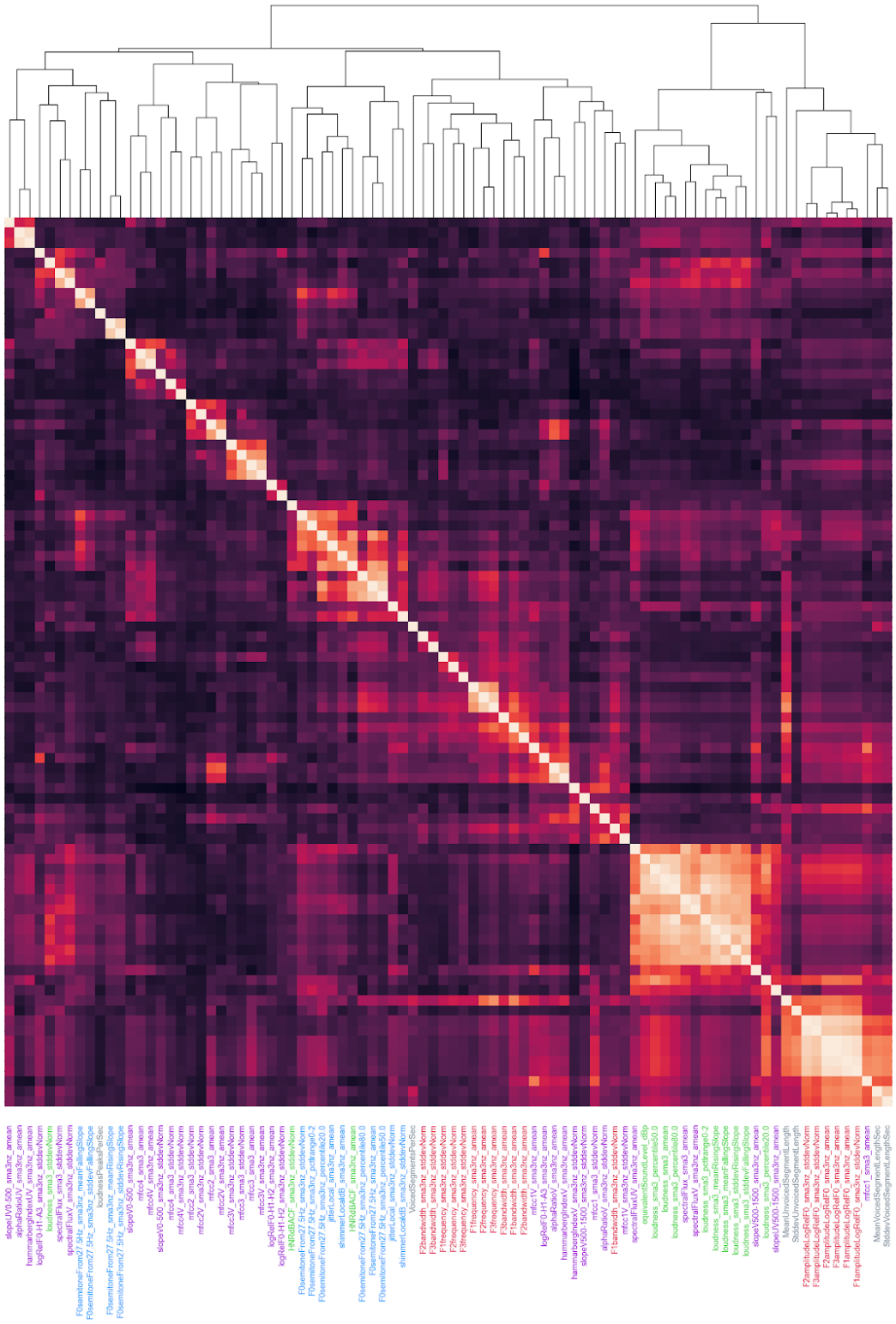

Figure S4. Patients, reading task. Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

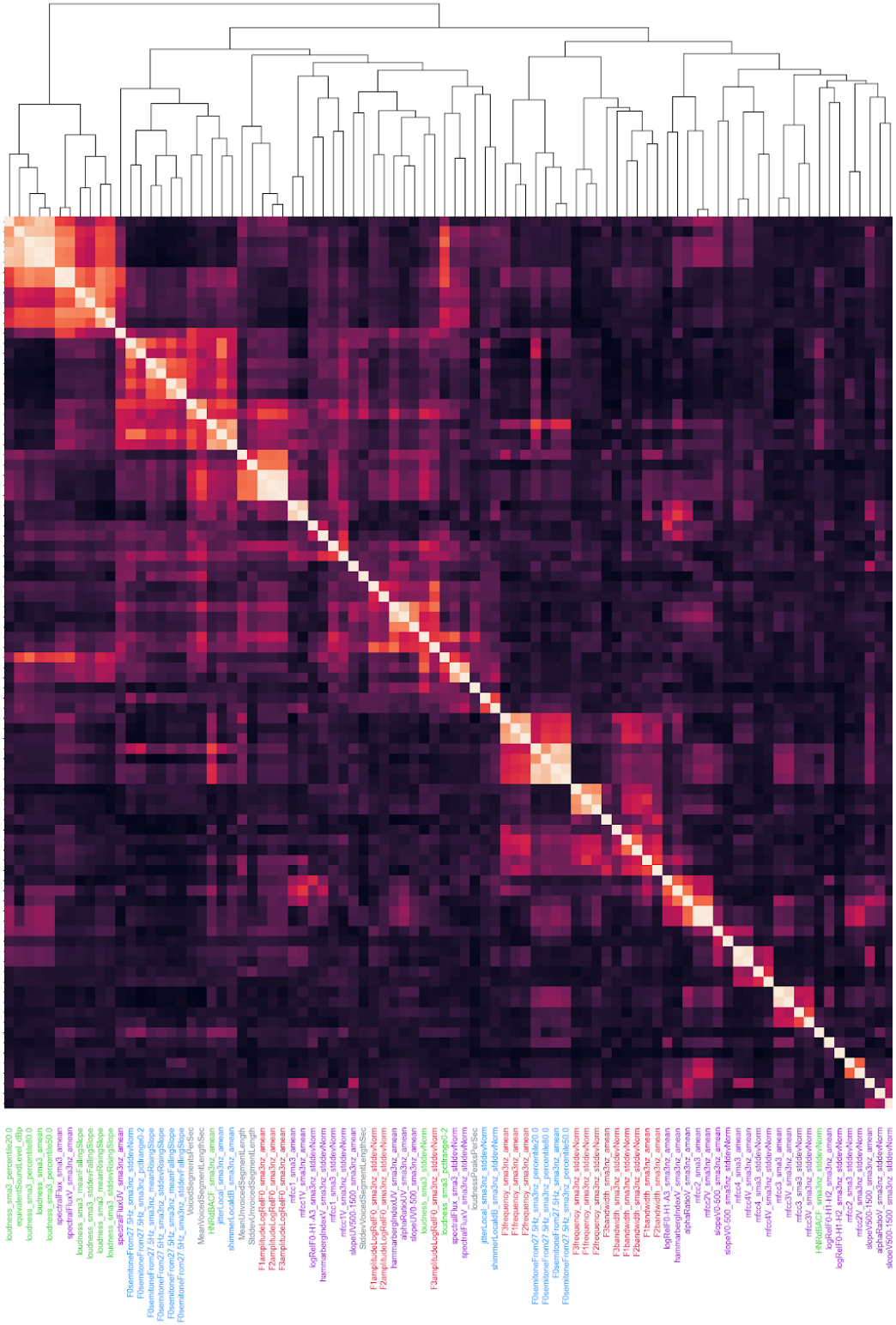

Figure S5. Patients, vowel task. Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

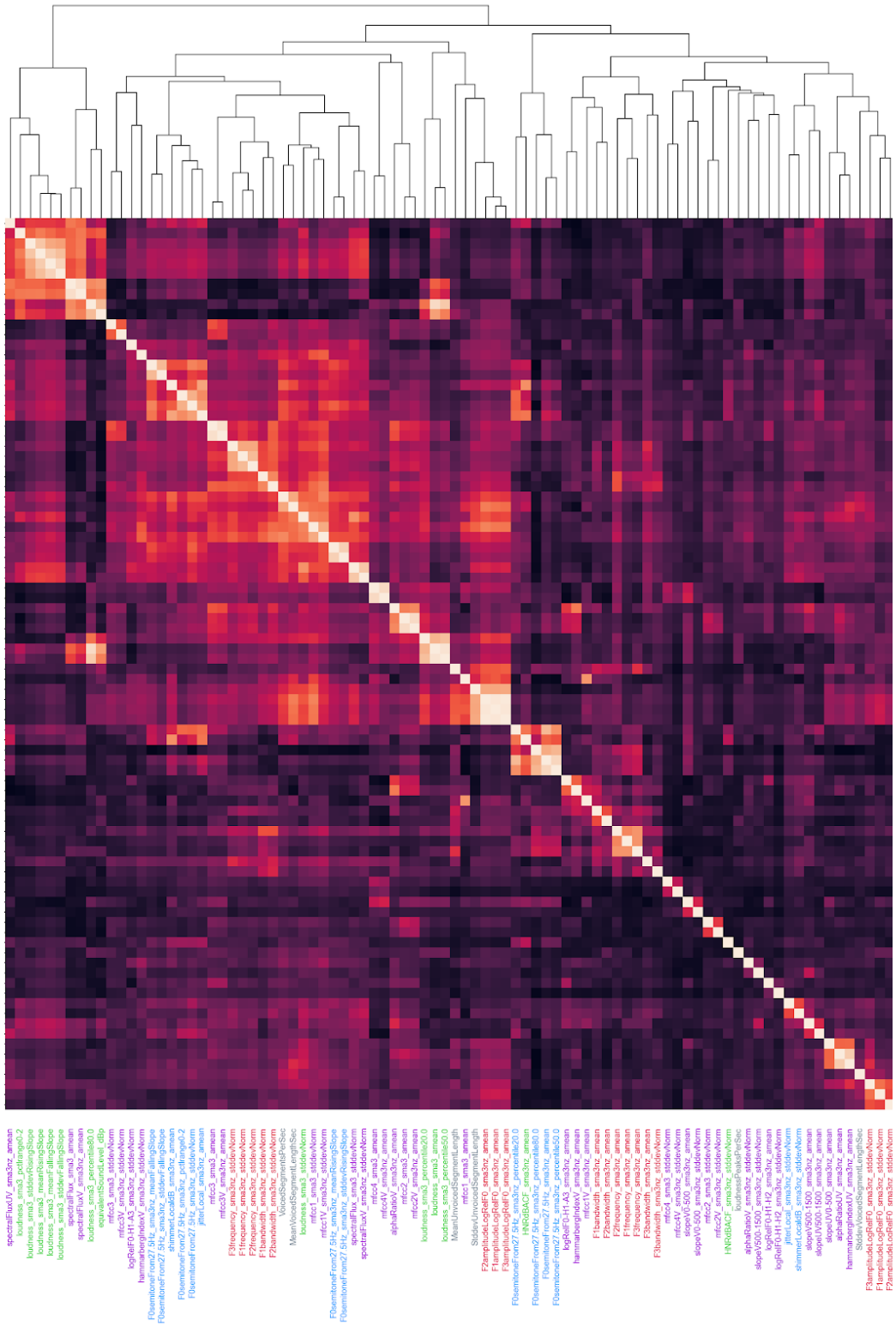

Figure S6. Patients, reading+vowel tasks. Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

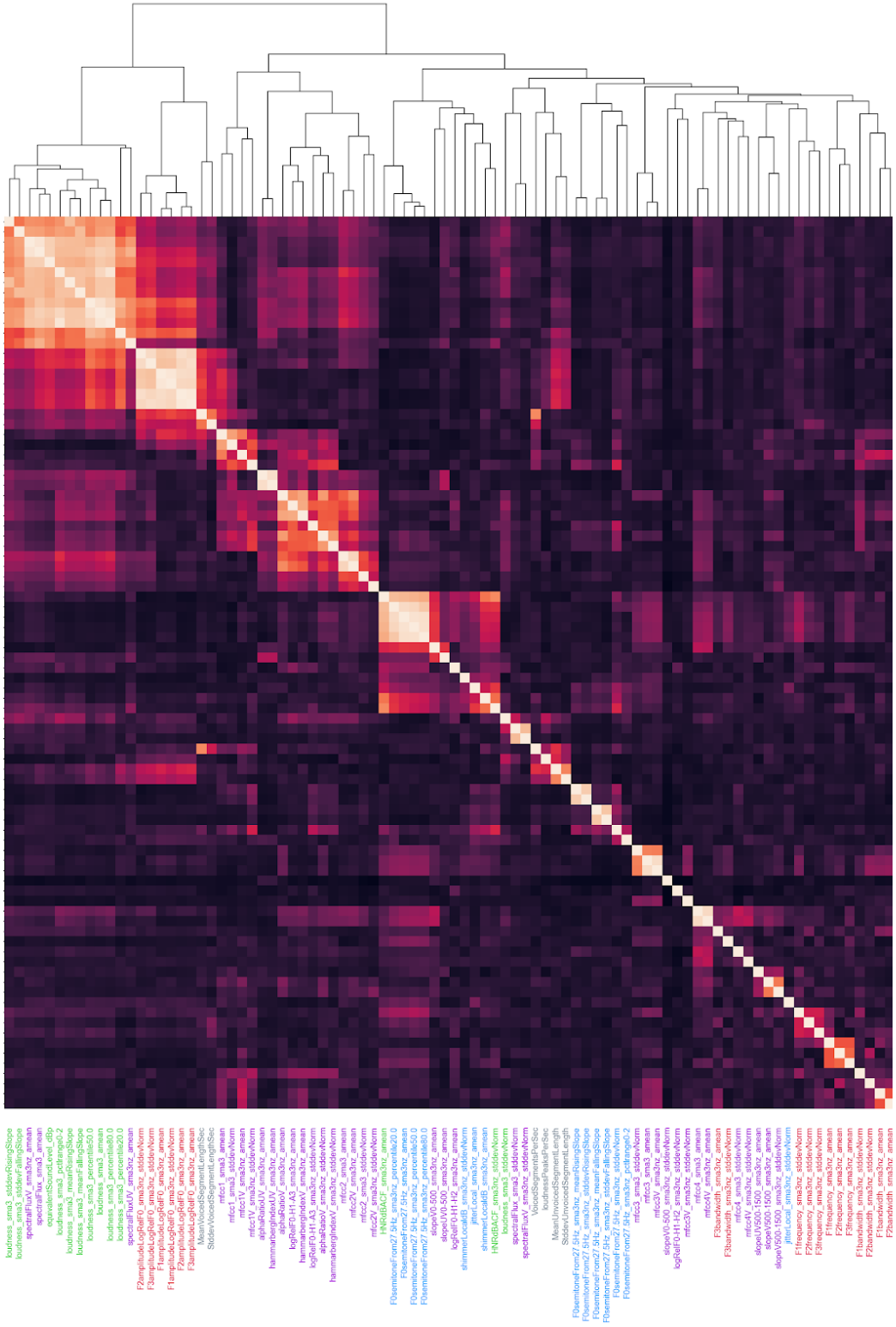

Figure S7. Controls, reading task. Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

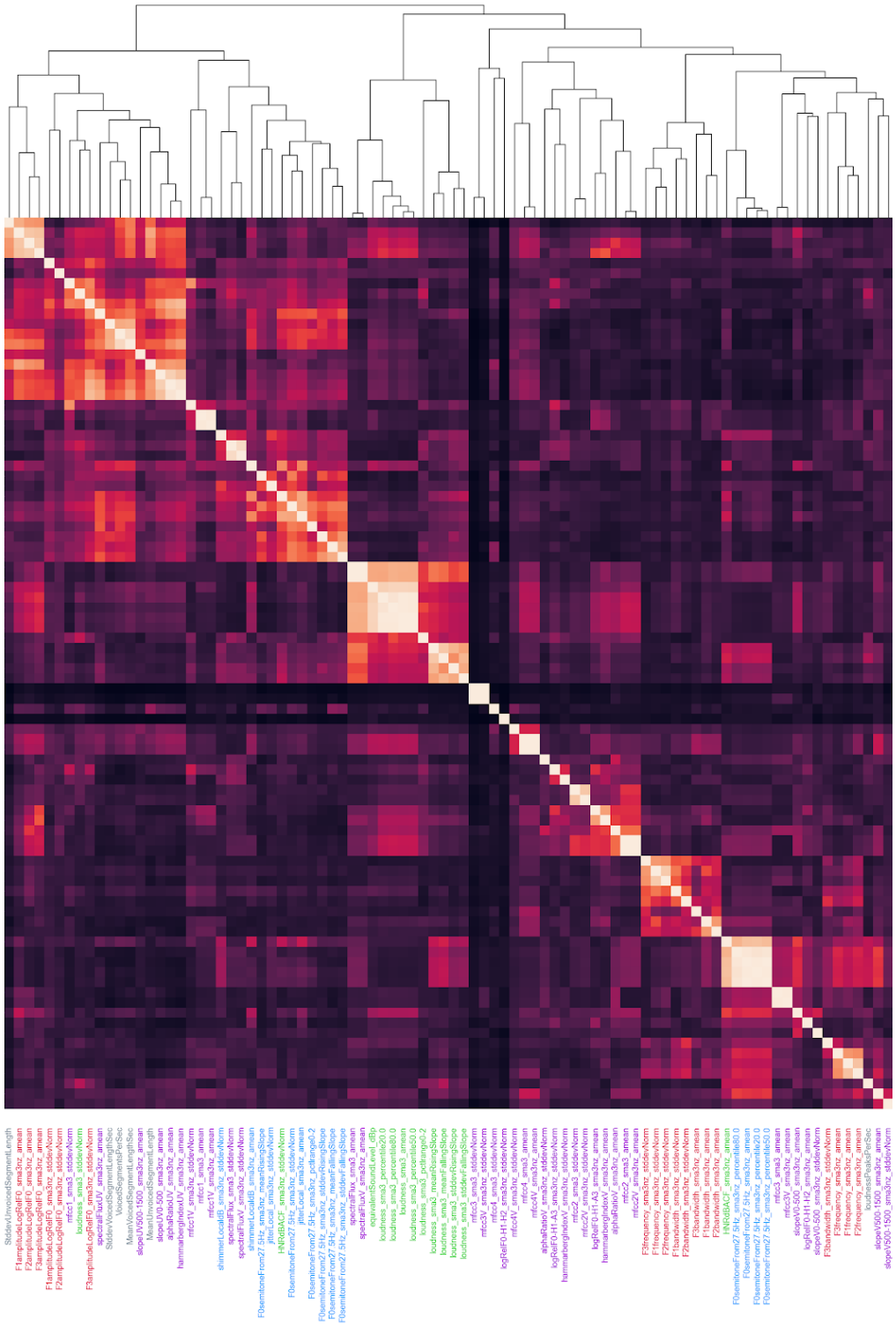

Figure S8. Controls, vowel task. Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features extracted. Squares are clusters of redundant features.

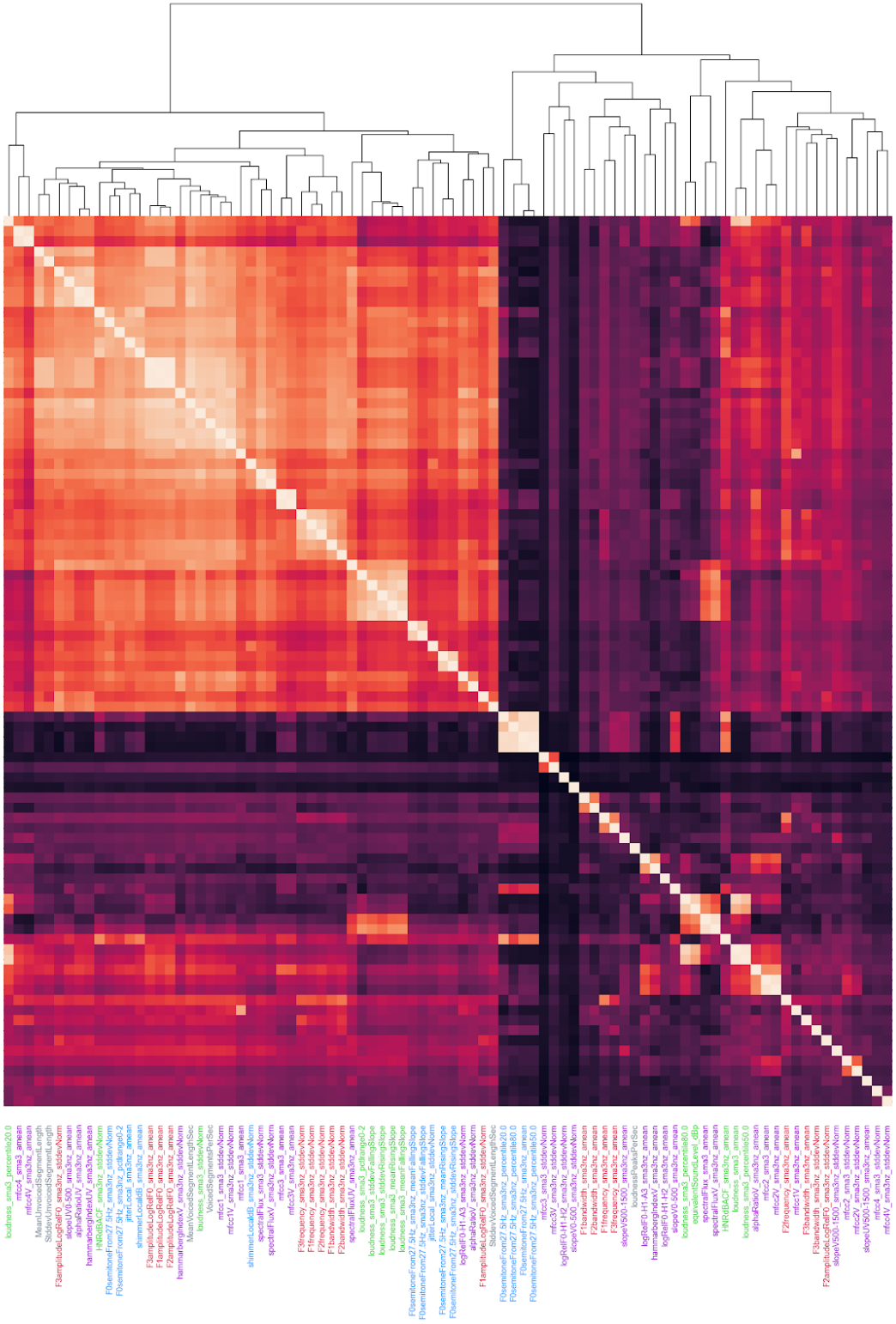

Figure S9. Controls, reading+vowel tasks. Visualization of features with shared information using pairwise distance correlation across the 88 eGeMAPs features. Squares are clusters of redundant features.

Feature Selection

To make sure information from the test sets is not having a strong influence on feature selection, we tested feature selection on 50 random train sets (80% of samples to match how models were trained) to make sure similar features were selected through this nested approach. If feature selection is relatively consistent across samples, removing features on the entire dataset should not be overfitting and is preferred for the explainability analysis to compare the same features. As seen in Table S1, all or most of the features used by selecting on the data set were also the most common across 50 splits and were selected in 91%, 83% and 76% of splits for reading, vowel and reading+vowel, respectively. Therefore, similar features are selected using both methods, but selecting on the entire dataset is preferred for explainability purposes (i.e., to rank the same features by their importance across all bootstrapping splits).

Selection using		**Reading**	**Vowel**	**Reading+Vowel**
Entire dataset	Optimal threshold and selected features	0.5	0.3	0.4
	Selected features	39	13	19
50 bootstrap train sets	Selected features (mean [95% CI])	35.8 [34–38]	12.3 [11–14]	17.9 [16–20]
	Match between both methods (entire dataset / most common across 50 train sets)	39/39	12/13	16/19
	Selected in percentage of runs	91%	83%	76%

Table S1. 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.

Performance removing participants that used other recording system

Given 24 patients were recorded using an iPad, we trained models without their samples to make sure these differences in recordings were not driving performance. 66, 72, and 138 samples were removed from the reading, vowel, and reading+vowel datasets, respectively. Performance did not drop considerably (see Supplementary Table S2).

Task	Features	LogisticRegression	MLP	RandomForest	SGD
Reading	88	.82 (.71–.87; .50)	.82 (.73–.88; .51)	.80 (.72–.88; .53)	.79 (.66–.87; .50)
Vowel	88	.78 (.71–.89; .50)	.79 (.68–.90; .54)	.81 (.73–.90; .52)	.74 (.60–.85; .45)
Reading+Vowel	88	.79 (.70–.87; .50)	.81 (.74–.88; .52)	.81 (.73–.88; .52)	.77 (.67–.84; .50)

Table S2. Performance of models without 24 patients recorded on iPad. Median ROC AUC score from 50 bootstrapping splits (95% confidence interval; median score of null model). The majority class (i.e., controls) represents 60% of the training samples of each dataset. Given the observed performance drop can also be due to removing training samples, the drop is not large enough to suspect that differences in recording are driving performance when using the full datasets. MLP: Multi-Layer Perceptron; SGD: Stochastic Gradient Descent Classifier.