THE Laryngoscope FOUNDED IN 1896

# Decoding Phonation with Artificial Intelligence (DEP AI): Proof of Concept

| | |
|---|---|
| Journal: | *The Laryngoscope* |
| Manuscript ID | Draft |
| Wiley - Manuscript type: | Original Reports |
| Date Submitted by the Author: | n/a |
| Complete List of Authors: | Powell, Maria; Vanderbilt University Medical Center, Vanderbilt Bill Wilkerson Center for Otolaryngology<br>Rodriguez Cancio, Marcelino; Vanderbilt University, Department of Information Technology<br>Young, David; Vanderbilt University Medical Center, Vanderbilt Bill Wilkerson Center for Otolaryngology<br>Nock, William; Vanderbilt University, Electrical Engineering and Computer Science<br>Abdelmessih, Beshoy; Vanderbilt University Medical Center, Vanderbilt Bill Wilkerson Center for Otolaryngology<br>Zeller, Amy; Vanderbilt University Medical Center, Vanderbilt Bill Wilkerson Center for Otolaryngology<br>Pérez Morales, Irvin; Central University Marta Abreu of Las Villas, Center of Research in Computational and Numerical Methods in Engineering; University of Brasilia, Infralab<br>Zhang, Peng; Vanderbilt University, Electrical Engineering and Computer Science<br>Garrett, C.; Vanderbilt University Medical Center, Vanderbilt Bill Wilkerson Center for Otolaryngology<br>Schmidt, Douglas; Vanderbilt University, Electrical Engineering and Computer Science<br>White, Jules; Vanderbilt University, Electrical Engineering and Computer Science<br>Gelbard, Alexander; Vanderbilt University Medical Center, Vanderbilt Bill Wilkerson Center for Otolaryngology |
| Keywords - Combo: | Voice/dysphonia < Laryngology, Speech language pathology < Laryngology, Laryngology |
| | |

# Decoding Phonation with Artificial Intelligence (DEP AI): Proof of Concept

Maria E. Powell, Ph.D*; Marcelino Rodriguez Cancio, Ph.D‡; David Young, M.D*; William Nock§; Beshoy Abdelmessih*; Amy Zeller, M.S*; Irvin Perez Morales, Ph.D†◊; Peng Zhang§; C Gaelyn Garrett, MD*; Douglas Schmidt, Ph.D§; Jules White, Ph.D§; Alexander Gelbard, MD*

* Vanderbilt Bill Wilkerson Center for Otolaryngology, Vanderbilt University Medical Center;
‡ Department of Information Technology, Vanderbilt University;
§ Department of Electrical Engineering and Computer Science, Vanderbilt University;
† Center of Research in Computational and Numerical Methods in Engineering, Central University Marta Abreu of Las Villas
◊ Infralab, University of Brasilia

# Abstract

**Purpose:** Acoustic analysis of voice has the potential to expedite detection and diagnosis of voice disorders. Applying an image-based, neural-network approach to analyzing the acoustic signal may be an effective means for detecting and differentially diagnosing voice disorders. The purpose of this study is to provide a proof-of-concept that embedded data within human phonation can be accurately and efficiently decoded with deep learning neural network analysis to differentiate between normal and disordered voices.

**Methods:** Acoustic recordings from 10 vocally-healthy speakers, as well as 70 patients with one of seven voice disorders (n=10 per diagnosis), were acquired from a clinical database. Acoustic signals were converted into spectrograms and used to train a convolutional neural network developed with the Keras library. The network architecture was trained separately for each of the seven diagnostic categories. Binary classification tasks (i.e., to classify normal vs disordered) were performed for each of the seven diagnostic categories. All models were validated using the 10-fold cross validation technique.

**Results:** Binary classification averaged accuracies ranged from 60%-80%. Models were most accurate in their classification of adductor spasmodic dysphonia, vocal fold polyp, polypoid corditis, and recurrent respiratory papillomatosis. Despite a small sample size, these findings are consistent with previously published data utilizing deep neural networks for classification of voice disorders.

**Conclusion:** Promising preliminary results support further study of deep neural networks for clinical detection and diagnosis of human voice disorders. Current models should be optimized with a larger sample size.


**Key Words:** Voice disorders, detection, acoustic analysis, convolutional neural network, classification

**Levels of Evidence:** Level III

# 1 Introduction

The clinical diagnosis of voice disorders relies on both the physical examination of laryngeal function and perceptual assessment of the acoustic output. While physical examination via endoscopy is the current gold standard for diagnosis, laryngoscopy and/or stroboscopy requires clinical expertise, and limited access to these clinical specialists may delay diagnosis. Perceptual assessment based on sound encoded within the speech signal is non-invasive and easily acquired, and has the potential to accelerate diagnosis; however, perceptual assessment of voice quality is subjective, and inter- and intra-rater reliability is highly influenced by clinician background, training, and experience.[1]

Acoustic analysis was initially introduced in the early 1990s as an objective, quantitative means to measure deviations from normal voice production.[2] Despite its widespread use for screening and progress monitoring, intrinsic limitations have prevented its effective application for automated detection and diagnosis of voice disorders.[3] Acoustic analysis has traditionally relied on the characterization of limited numbers of acoustic parameters.[4] The mechanism of human speech production is highly complex, however, and any given pathology affects multiple acoustic parameters simultaneously. Although a highly trained expert human brain can integrate and interpret these multiple deviations, a parameter-by-parameter approach to acoustic analysis has not replicated this functionality.

Artificial intelligence using deep neural networks may provide an alternative to the single or multi-dimensional parameter approach to acoustic analysis. Neural network learning has been extensively developed for automatic speech recognition applications since the 1980s.[4] Despite extensive research developing deep learning architectures to decode the speech signal for linguistic content, few studies have applied this technology to analysis of the disordered voice. Findings from these studies, which report classification accuracies for dysphonic voices between 40% and 96%, support the value of using deep neural networks for detection and

differential diagnosis of voice disorders.[5-7] These studies, however, not only require analysis of multiple acoustic parameters, which slows processing times, but they also rely heavily on the Mel frequency cepstral coefficient (MFCC), which filters frequency data to maintain the long-term temporal aspects of the frequency spectrum needed to extract critical linguistic elements of the speech signal. While this discarded frequency data may not be salient for speech recognition, it may be vital for the detection and distinction between certain voice disorders.

Recent studies have recommended moving away from the MFCC in favor of spectrograms.[8-9] Not only do spectrograms maintain the full frequency resolution of the acoustic signal, but they also have the unique characteristic of being data-rich *images* that can be analyzed via image analysis techniques. Since the early 2010s, a revolution in the field of image analysis has occurred. Tasks like image classification have been solved with near-human levels of accuracy.[5] Within medicine, image analysis with a neural network approach has made inroads into clinical diagnostics in both radiology[10] and dermatology[11], ultimately expediting accurate diagnoses using non-invasive techniques.

We hypothesize that applying an image-based neural network approach to classify voice disorders may result in similar advancements in laryngology. The purpose of this study is to provide a proof-of-concept that embedded data within human phonation can be accurately and efficiently decoded with deep learning neural network analysis to differentiate between normal and disordered voices.

## 2 Materials and methods

This study was performed in accordance with the Declaration of Helsinki, Good Clinical Practice, and was approved by the Institutional Review Board at Vanderbilt University Medical Center (IRB#: 181191). The study utilized previously-collected acoustic recordings from patients with voice disorders, as well as vocally-healthy individuals. As part of standard of care, individuals seen at the Vanderbilt Voice Center with a voice complaint are asked to provide a

standardized voice sample. These voice samples are captured at the time of evaluation and stored on a secure server (ImageStream, Image Stream Medical, Littleton, MA) as part of the patient's electronic medical record. Voice samples from vocally-healthy individuals are also included in this electronic database as reference.

## 2.1 Data Collection

### 2.1.1 Participants

Ten vocally-healthy adults and 70 adults with voice disorders were included in this study. Participants were identified by querying the electronic database by either diagnosis or normal voice status. The following diagnoses were included in the study (the sample size is n=10 in each diagnostic group): adductor spasmodic dysphonia (ADSD), essential tremor of voice (ETV), muscle tension dysphonia (MTD), polypoid corditis or Reinke's edema (PCord), unilateral vocal fold paralysis (UVFP), vocal fold polyp (Polyp), and recurrent respiratory papillomatosis (RRP). Diagnoses were confirmed by two independent, board-certified laryngologists at the Vanderbilt Voice Center. Table 1 shows the demographic information for each diagnostic group.

### 2.1.2 Acoustic Recordings

All participants were recorded reading the first three sentences of the phonetically-balanced Rainbow Passage.[12] Recordings were obtained in a quiet clinic room using an omnidirectional lapel microphone with a 44 KHz sampling rate (Olympus Visera Elite OTV-S19; Olympus Medical, Center Valley, PA) and stored on the clinical server as .mp4 files with audio compressed at 186kbps. Using the open-source audio editor Audacity® (Audacity® v2.2.1, Dec 2017), the acoustic signals were extracted from the video files, edited to include only the Rainbow Passage, and saved as uncompressed .wav files in a password protected, REDCap database.

### 2.1.3 Data Processing

To augment the limited amount of data, as well as make the model's predictions more robust, the raw .wav files were segmented into 3-second 'chunks'.[13] For the last chunk of each recording (which would not be a full 3 seconds), the chunk's frequencies were repeated with a small amount of noise (<5%) until the 3-second window was filled. This technique standardized the input to the neural network (despite the raw samples varying in duration) without losing information from the audio recordings (see Table 2 for the total number of spectrograms representing 3-second acoustic samples for each diagnostic group). Following segmentation, all .wav files were transformed into .png wide-band spectrogram images (standardized resolution: 256x256 pixels) using the short-time Fourier transform, as shown in Figure 1. While the source audio captures frequencies up to 22.5kHz, the frequency ceiling for the images was set at 5kHz in order to provide better resolution of relevant voice data and remove any compression artifact.

## 2.2 Data Exploration

### 2.2.1 Model Development

An open-source, deep-learning library named Keras was used for this study.[14] The Keras library is written in Python and was selected due to its ease of use and flexible interface that allows a combination of different types of layers in non-sequential architectures, with heterogeneous inputs and outputs.[15] Figure 2 shows the architecture of the convolutional neural network deep learning model used for the binary classification tasks, built in Keras.[14] The dimensions and number of parameters of each layer are shown, with a total of 6,795,457 parameters. The network is a Convolutional Neural Network (CNN) with a dropout used to reduce overfitting (when the model learns the training data too well, resulting in poor generalization) by means of regularization. CNNs are a special kind of neural network inspired by how the human brain perceives and classifies objects. The network works by taking an image and reducing it to simpler features that the computer can work with (such as edges and

color spots) through a series of convolutions and pooling operations (Figure 2, Conv2D and MaxPooling2D). The spatial information from the original image is preserved during these convolutions so that in the final layers, these features are combined together to produce a feature map. The network assigns a probability that the image belongs to a certain class based on the data it has previously been trained on.

### 2.2.2 Model Validation

Large data sets (preferably made up of thousands of images with balanced classes) are necessary to provide sufficient material to train deep learning models. Given the small size of the data set consisting of only 451 images (Table 2), we chose to perform a 10-fold cross validation to evaluate our models. All images corresponding to an individual subject belonged to the same fold to ensure independence between folds, preventing leakage of information from the training set to the validation set. In other words, for the classification problem corresponding to each disease, each fold contained all the spectrograms corresponding to one subject having the disease, and all the spectrograms corresponding to a normal subject. For example, the frames used in the $5^{th}$ validation fold include all spectrograms from the $5^{th}$ normal subject and the $5^{th}$ ADSD patient. Figure 3 shows the spectrograms from all normal subjects (left) and all patients with ADSD (right); the spectrograms included in the $5^{th}$ validation fold are surrounded by a dashed line. This neural network was trained on all the other frames in Figure 3, and was then used to perform the binary classification task on the frames surrounded by black lines. For each iteration of the model, sequential folds were withheld from the training set for validation. Each fold was therefore incorporated into the training set 9 times and served as the validation set once. This technique was employed to ensure some statistical significance in the results, which is even more necessary in this case of little data available.

### 2.2.3 Binary Classification Tasks

Seven binary classification tasks were conducted to categorize normal and disordered voice

samples. These classification tasks mimicked a clinical screening task to discriminate between normal and disordered for each of the 7 diagnostic groups. The network architecture of the deep learning model (shown in Figure 2) was trained separately for each fold within each of the seven diagnostic groups, for a total of 70 models. Training for each model required 10 full presentations of the data (epochs), which were iteratively optimized using gradient descent with 100 backpropagation steps and a learning rate equal to 10-4.

The primary metric for assessing our training was accuracy, defined as the fraction of all correctly classified instances with respect to the total number of instances. Baseline accuracy (the minimally acceptable level of accuracy) for each disorder was determined by a naïve algorithm that always predicted the disordered class ($Baseline\ accuracy = \frac{Spect_D}{Spect_N + Spect_D}$, where $Spect_D$ is the total number of disordered frames, and $Spect_N$ is the total number of normal frames). Table 2 lists the baseline accuracies for each diagnostic group. The accuracy of the model in the validation set provides an estimate of the model's performance with new data. In the ideal case, the accuracies of the training and validation sets should be similar.

Another important metric for training deep learning models is the loss function, which measures the difference between model predictions and the real values obtained from the binary classification task. Generally, high accuracy values should correspond to low loss values. To compare results between models, the presented values of the loss function have been normalized to values between 0 and 1.

# 3 Results

In some folds of some disorders, an almost perfect accuracy was obtained, such as the case of the 5$^{th}$ ADSD fold, which classifies all spectrograms from the 5$^{th}$ normal participant and the 5$^{th}$ ADSD patient (Figure 4). The classification accuracy from this task was 100% for 7 of the 10 epochs, stabilizing in this value after the 7$^{th}$ epoch. The absence of overfitting in this

favorable case is demonstrated by the similarity of the accuracy and loss values of the training set to the validation set.

Although similarly promising results were obtained on other individual folds, the accuracy of the models obtained by *averaging* the results of all folds within a diagnostic category is lower. For example, the highest validation accuracy of the averaged ADSD model is 76% in the third epoch (Figure 5a), compared to 100% in the same epoch for the 5$^{th}$ fold only, as shown in Figure 4. Despite the decrease in accuracy from the averaged ADSD data, the model still performed substantially better than the baseline accuracy of the naïve algorithm (55%). Similar results for averaged data were obtained for PCord, Polyp and RRP, with highest validation accuracies equal to 78%, 77% and 80% respectively, as shown in Figure 5b-d. The difference between the accuracy and loss values from the training data (dashed line) and the validation data (solid line) indicates that overfitting was prominent in all four models.

The averaged results for ETV, MTD and UFVP conditions were comparable to the naïve algorithm's baseline accuracy. Maximum validation accuracies for ETV, MTD, and UVFP were 64%, 60%, and 63% respectively. Despite these lower accuracies from the averaged data, the model performed much better than the naïve algorithm for classifying spectrograms from select individual speakers within these diagnostic groups, as shown in Figure 6.

# 4 Discussion

## 4.1 Current Findings

In this proof-of-concept study, we investigated the utility of employing image analysis with deep learning to differentiate between normal and disordered voices using spectrograms. The averaged models achieved substantially higher accuracy in the validation set compared to the naïve algorithm for classifying normal vs adductor spasmodic dysphonia (76%), polypoid corditis (78%), vocal fold polyp (77%), and recurrent respiratory papillomatosis (80%) voice samples, as shown in Figure 5a-d. While the average models for muscle tension dysphonia, unilateral vocal

fold paralysis, and essential tremor of voice were comparably less robust, results coincided with other studies that employ artificial intelligence models.[5] For all diagnostic groups, moreover, spectrograms from individual speakers (i.e., specific folds) were classified with accuracies up to 100%, as shown in Figure 4. We hypothesize that the variability in results from individual folds stem from individual patients' symptom severity, and subsequently, individuals with a more severe presentation of the voice disorder may be classified more accurately. These results are promising (despite the small dataset used for this study) and the accuracy of these models should improve with additional training data.

The cross-validation models for all seven classification tasks demonstrated overfitting, despite the dropout layers added for regularization. Overfitting occurs when the model adapts too closely to the idiosyncrasies of the training set and is unable to generalize to new data (i.e., the validation set). This type of modeling error is common in highly-complex models and is exacerbated by small sample sizes.

Implementation of an image augmentation technique commonly used to address overfitting did not improve our results, so that data is not reported. This augmentation technique boosts model performance by introducing random rotations and scalings of the original images.[13] We hypothesize that these findings are related to the inherent symmetry of the sound spectrogram images. The spectrogram is a two-dimensional visual representation of the frequency and intensity spectrums; all the spikes are parallel to the frame borders and are therefore relevant for this specific classification task. As such, the interpretation of the image is highly dependent on the orientation of the image and thus the orientation cannot be varied. Increasing the sample size would reasonably reduce overfitting and improve the generalizability of the model without the use of any image augmentation.

### 4.2 Challenges and Future Directions

The primary limiting factor in our proof-of-concept study is a lack of sufficient data. The current results are based on data from 80 individuals with a total sample size of 451

spectrograms. Each classification task, however, included data from the normal group and only one disordered group. The mean sample size for each classification task was therefore only 103, 3-second spectrograms (Table 2).

While these initial results are promising, a robust dataset that represents the full range of severities for each diagnostic group, as well as the wide variability among vocally-healthy speakers, is critical to improve the models. Current efforts are underway to gather thousands of new and existing voice samples from patients and vocally-healthy participants. However, the need for big data necessitates data collection protocols that minimize salient variabilities in recording conditions, and similarly requires models that are robust against these inconsistencies. Recordings must also be actively curated to maintain the fidelity of the training set, which is time-consuming and expensive.

Although these challenges are non-trivial, the potential clinical import of a robust, artificial intelligence-driven, acoustic analysis tool is worth the effort. Such a tool has the potential to improve diagnostic accuracy and reliability and provide a standardized metric for interpretation within and between clinical institutions.

# 5 Conclusion

In this paper we applied image classification techniques with deep learning to classify spectrograms into normal vs disordered voices. Despite the small size of the available dataset, satisfactory results were obtained for the adductor spasmodic dysphonia, polypoid corditis, vocal fold polyp, and recurrent respiratory papillomatosis diagnostic groups, with accuracy in the validation set substantially higher than the baseline accuracy of the naïve algorithm. These preliminary results support further study of deep neural networks for clinical detection and diagnosis of human voice disorders.

# References


1. Kreiman J, Gerratt BR, Precoda K. Listener experience and perception of voice quality. *J Speech, Lang, Hear Res*. 1990;33(1):103-115.
2. Lieberman P. Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. *J Acoust Soc Am*. 1963;35:344–353.
3. Saenz-Lechon N, Godino-Llorente JI, Osma-Ruiz, V, Gómez-Vilda P. Methodological issues in the development of automatic systems for voice pathology detection. *Biomed Signal Process Control*. 2006;1(2):120-128.
4. Hadjitodorov S, Mitev P. A computer system for acoustic analysis of pathologic voices and laryngeal diseases screening. *Med Eng Phys*. 2002;24:419-429.
5. Schönweiler R, Hess M, Wübbelt P, Ptok M. Novel approach to acoustical voice analysis using artificial neural networks. *J Assoc Res Otolaryngol*. 2000;1(4):270-82.
6. Linder R, Albers AE, Hess M, Pöppl SJ, Schönweiler R. Artificial neural network-based classification to screen for dysphonia using psychoacoustic scaling of acoustic voice features. *J Voice*. 2008;22(2):155-63.
7. 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):380-4.
8. Lee H, Pham P, Largman Y, Ng AY. Unsupervised feature learning for audio classification using convolutional deep belief networks. In: *Advances in neural information processing systems.* 2009:1096-1104.
9. Deng L, Li J, Huang JT, Yao K, Yu D, Seide F, Seltzer ML, Zweig G, He X, Williams JD, Gong Y. Recent advances in deep learning for speech research at Microsoft. In: *ICASSP*. 2013; 26:64-69.
10. Dheeba J, Singh NA, Selvi ST. Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. *J Biomed Inform*. 2014;49:45-52.
11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542:115-118.
12. Fairbanks G. *Voice and articulation drillbook.* 2nd ed. New York: Harper & Row; 1960:124-139.
13. Sprengle E, Jaggi M, Kilcher Y, Hofmann T. Audio based bird species identification using deep learning techniques. In: *LifeCLEF2016* (No. EPFL-CONF-229232). 2016:547-559.
14. Chollet F. *Deep learning with Python*. Manning Publications, 2017.
15. Sejdić E, Djurović I, Jiang J. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digit Signal Process*. 2009;19(1):153-183.


Figure 1: To standardize input into the neural network, acoustic signals were segmented into 3-second chunks (left) and transformed into spectrograms using the Fourier transform (middle). Spectrograms displayed frequency over time, with intensity coded in color (right).

190x43mm (150 x 150 DPI)

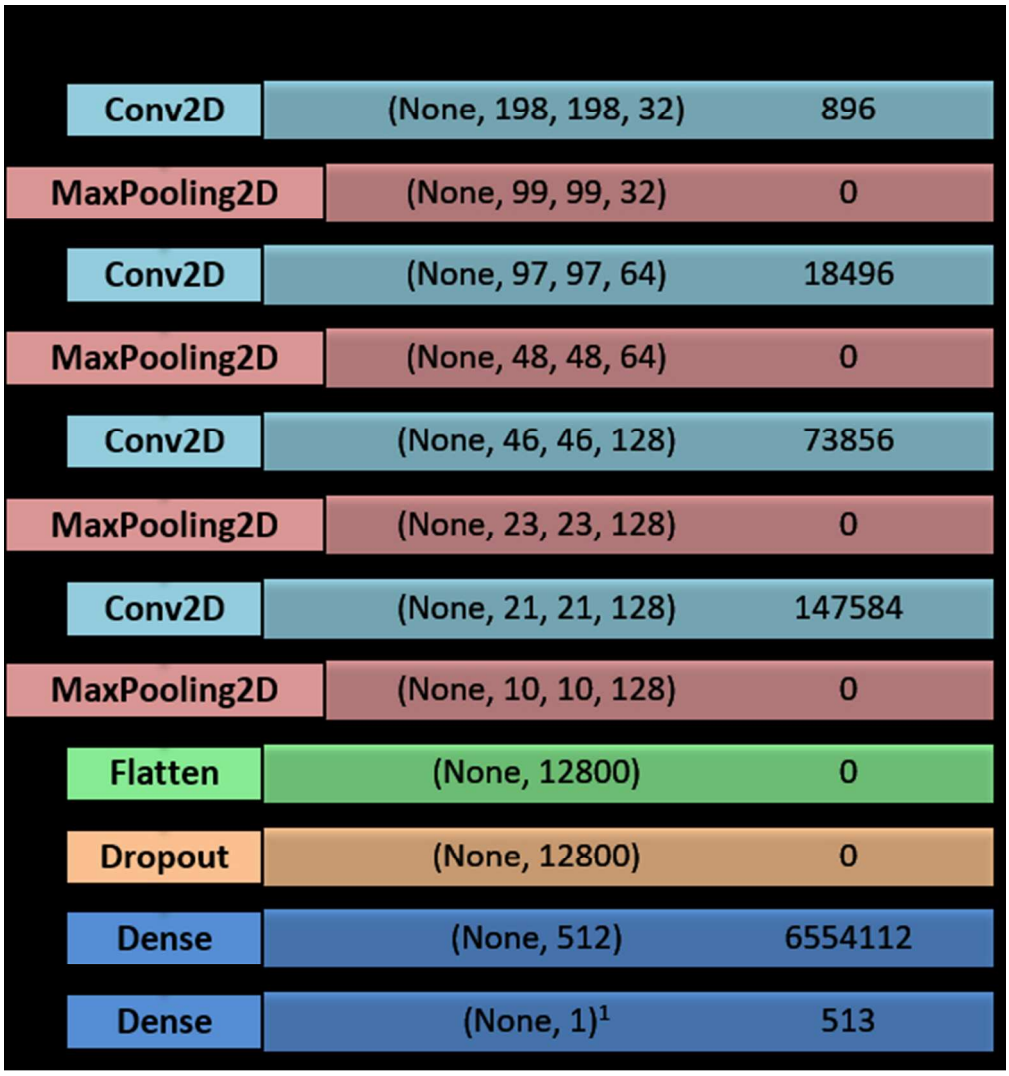


Figure 2: Summary of the Keras convolutional neural network models trained for the seven binary classification tasks. Conv2D = 2D convolutional layer; MaxPooling2D = 2D max-pooling layer.

103x110mm (150 x 150 DPI)

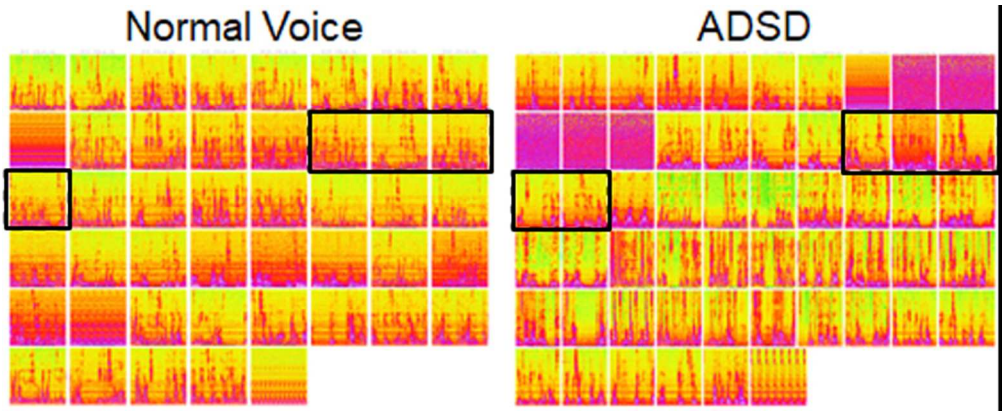


Figure 3: Spectrograms of all audio files from vocally-healthy individuals (left) and patients with adductor spasmodic dysphonia (right). The 5th validation fold classified all spectrograms from normal subject and ADSD patient 5. Frames used in the binary classification task are surrounded by lines. All other frames were used to train the model.

124x50mm (150 x 150 DPI)

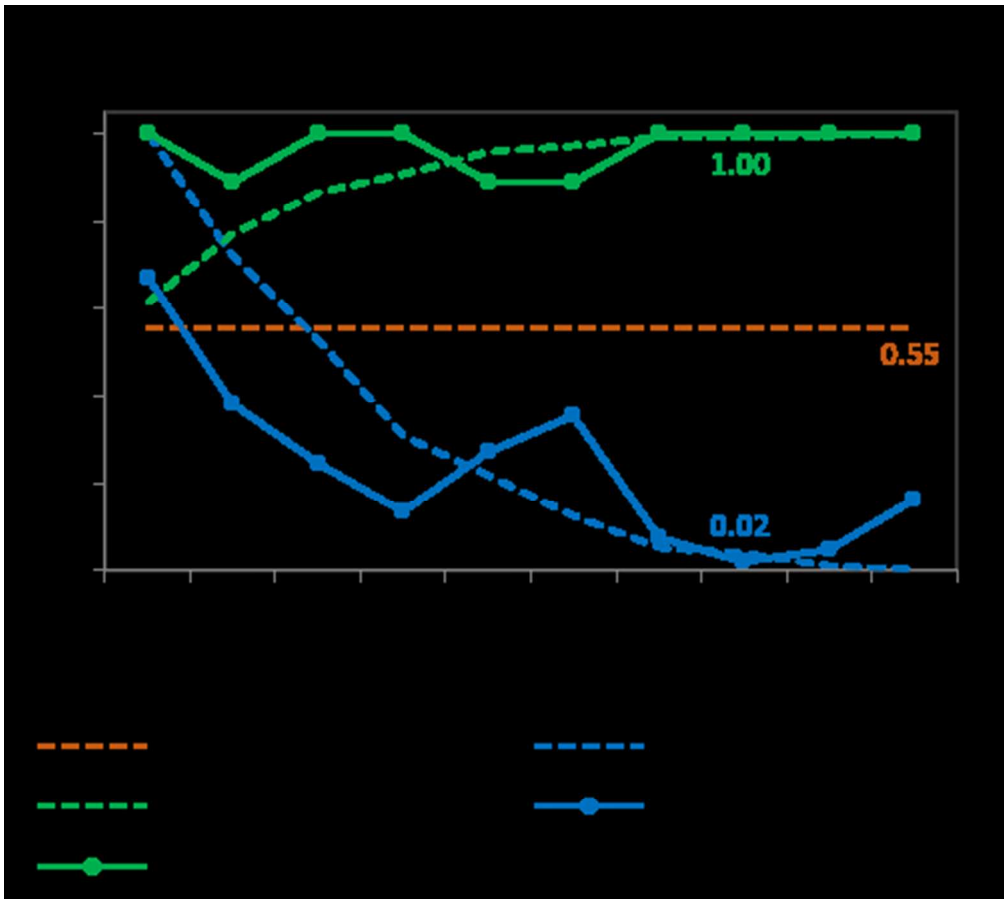


Figure 4: Accuracy and loss results for the 5th fold (best case) from the ADSD diagnostic category. Baseline accuracy, as well as the accuracy and loss results from the highest performing epoch (epoch 8) are labeled.

84x75mm (150 x 150 DPI)

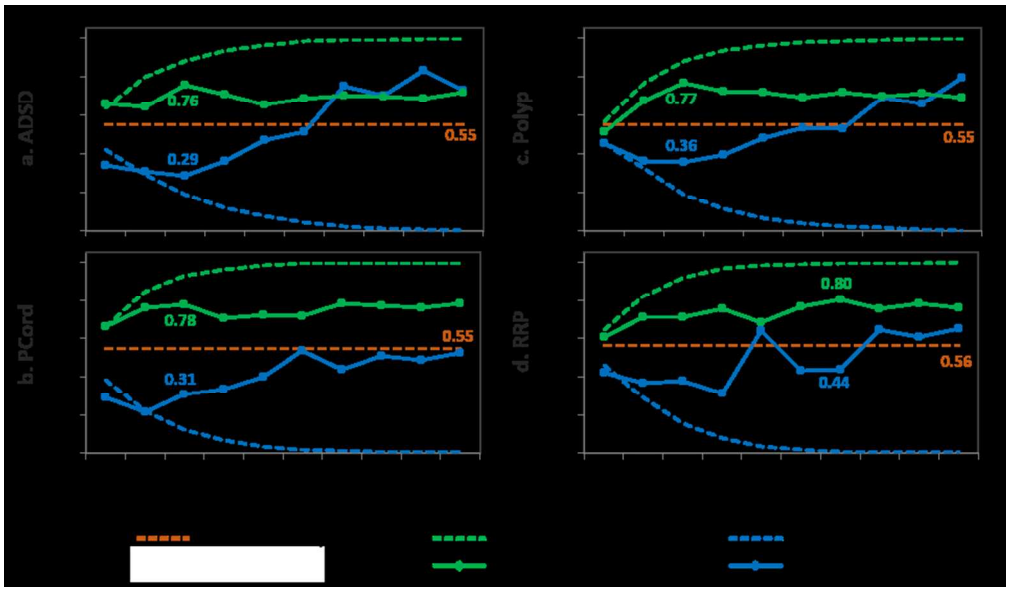


Figure 5: Average results of all folds obtained from 10-fold cross validation of (a) adductor spasmodic dysphonia, (b) polypoid corditis or Reinke's edema, (c) vocal fold polyp, (d) and recurrent respiratory papillomatosis. Baseline accuracies, as well as the accuracy and loss results from the highest performing epochs are labeled for each model.

162x94mm (150 x 150 DPI)

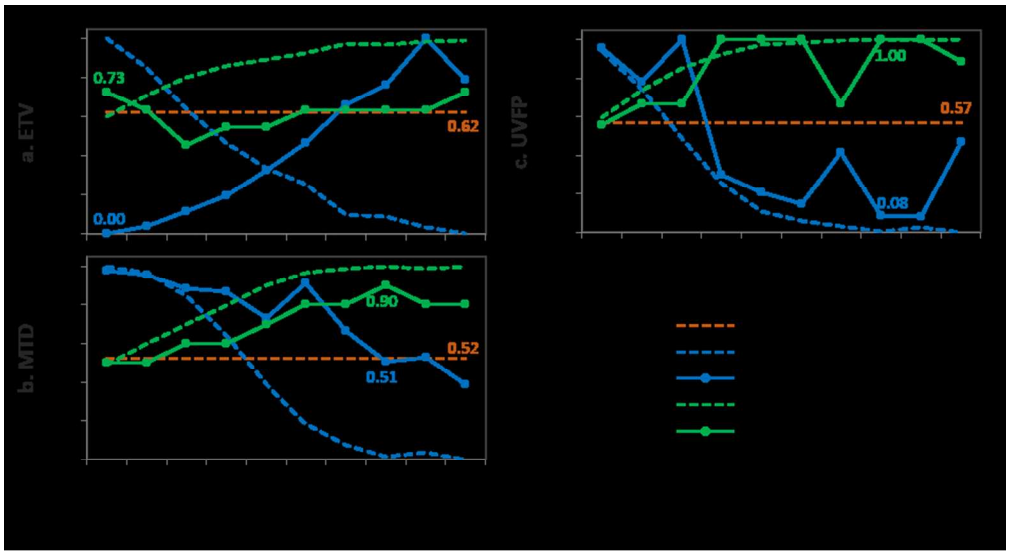


Figure 6: Accuracy and loss results for the best fold for (a) essential tremor of voice, (b) muscle tension dysphonia, and (c) unilateral vocal fold paralysis. Baseline accuracies, as well as the accuracy and loss results from the highest performing epochs are labeled for each model.

160x87mm (150 x 150 DPI)

**Table 1:** Demographic information for each diagnostic group.

| Diagnostic Group | Normal (n=10) | ADSD (n=10) | ETV (n=10) | MTD (n=10) | PCord (n=10) | UVFP (n=10) | Polyp (n=10) | RRP (n=10) | Total (n=80) |
|---|---|---|---|---|---|---|---|---|---|
| **Gender** F (M) | 8 (2) | 5 (5) | 10 (0) | 8 (2) | 8 (2) | 8 (2) | 5 (5) | 3 (7) | 55 (25) |
| **Age** Mean (Stdv) | 34 (10) | 56 (10) | 79 (4) | 47 (16) | 55 (10) | 53 (17) | 46 (11) | 53 (18) | 53 (17) |

Normal=vocally healthy; AdSD=adductor spasmodic dysphonia; ETV=essential tremor of voice; MTD=muscle tension dysphonia; PCord=polypoid corditis or Reinke's edema; UVFP=unilateral vocal fold paralysis; Polyp=vocal fold polyp; RRP=recurrent respiratory papillomatosis.

**Table 2:** Total sample size for each group and the derived baseline accuracy for each classification task.

| Diagnostic Group | Normal | ADSD | ETV | MTD | PCord | UVFP | Polyp | RRP | Total |
|---|---|---|---|---|---|---|---|---|---|
| **Total Spectrograms** | 45 | 56 | 74 | 49 | 54 | 59 | 56 | 58 | 451 |
| **Baseline Accuracy** | | | | | | | | | |
| Naïve algorithm (%) | -- | 56/101 (55%) | 74/119 (62%) | 49/94 (52%) | 54/99 (55%) | 59/104 (57%) | 56/101 (55%) | 58/103 (56%) | -- |

Normal=vocally healthy; AdSD=adductor spasmodic dysphonia; ETV=essential tremor of voice; MTD=muscle tension dysphonia; PCord=polypoid corditis or Reinke's edema; UVFP=unilateral vocal fold paralysis; Polyp=vocal fold polyp; RRP=recurrent respiratory papillomatosis.