End-to-end Deep Learning Framework for Vocal Pathology Classification (*VocalPathNet*)

Jordan Matthew Hodges
Stanford University
jmhodges@stanford.edu

Jingzhi Kevin Yu
Stanford University
Jyk306@stanford.edu

***Abstract*— In the field of otolaryngology, diseases of the throat can be difficult to diagnose due to limited access to trained professionals and procedures. However, with many laryngeal diseases having an accompanying vocal dysphonia, evaluation of auditory recordings has become a topic of interest in recent years. In this project, we aimed to create a novel end-to-end deep learning framework *VocalPathNet* for classifying specific vocal fold pathology that leveraged recordings of multiple vowels in order to harness vocal pathologies having different expression patterns during different auditory tasks. The described model utilizes a 1-dimensional convolutional neural network (1-D CNN) architecture that takes as input stacked audio data of an individual saying the vowels 'a', 'i' and 'u' in neutral, low and high pitch. The model achieved a final accuracy of 0.779 with an F1 score of 0.765 and an AUC of 0.865, surpassing models that take in only individual vowel recordings.**

I. Introduction

Changes in voice quality, or dysphonia, is a common symptom of many laryngeal diseases. While sometimes such a change in voice represents a short-term nuisance, in other situations it may be indicative of a more serious condition which may lead to chronic voice difficulties. Currently, diagnosis of many laryngeal diseases relies on expert opinion and often assessments that vary in their levels of invasiveness. Additionally, a laryngoscopy exam (ideally with stroboscopy) must be performed by a trained otolaryngologist, and access to an otolaryngologist may be limited. Additionally, acoustic and/or auditory-perceptual evaluation of the voice may not always be possible in many clinics. These requirements make detection and diagnosis costly and inaccessible in some cases [1]. Thus, reliable prediction models of specific vocal fold pathologies have become a topic of interest in recent years [2].

While there has been prior research on classifying both healthy and pathologic vocal recordings, limited work has been done in classifying the underlying categories of vocal fold pathology. Predicting the exact type of vocal fold pathology can provide additional insight for physicians diagnosing vocal fold pathologies in patients when compared with previous models that only distinguish between healthy and pathologic voices.

Machine learning based diagnosis of vocal fold pathology generally relies on participant recordings of either a standardized sentence or sustained vowel at various pitches [3]. In the case of sustained vowels, many models are trained on recordings of a single vowel in a single pitch. The rationale is that there are variations across the different vowels that do not generalize to each other. However, we believe that deep learning algorithms may identify latent relationships between these vowel recordings which can help classify between different vocal fold pathologies.

In this project, we aimed to create a novel end-to-end deep learning framework VocalPathNet for classifying specific vocal fold pathology that leveraged recordings of multiple sustained vowels. For our baseline, we created a model that took individual vowel recordings as inputs. Our novel framework took stacked raw audio files of sustained vowels produced by healthy and pathological as input, and sought to identify the specific vocal pathology, in this case being hyperkinetic dysphonia, hypokinetic dysphonia, and laryngitis.

II. Related Work

Classification models built with audio samples have been developed utilizing various strategies, including both deep learning and more traditional machine learning approaches. Many current models utilize support vector machines, multi-layer perceptrons, and random forests, with deep learning models gaining in popularity in recent years [2]. Almost always, feature extraction occurs as a distinct step before training and prediction. Mel-frequency cepstral coefficients (MFCCs) remain the most common features that vocal fold predictors rely on [2,3,4,5,6,7]. Alternatively, the analysis of spectrograms has also been explored in analyzing auditory data for similar tasks [8].

The popularity of the "2 step approach" leaves open the question of the efficacy of a fully end to end approach. Research conducted by Quan et al. recently reported accuracy as high as 92% on a similar task of identifying vocal patterns associated with the onset of Parkinson's disease[6]. There may be potential features that are overlooked by extracting MFCCs that are useful for classifying vocal fold pathology. In addition, the use of MFCCs is also dependent on the window size applied in the transformation process [1]. In this project, we opted to use raw audio samples instead of MFCCs to train a deep neural network that would predict categories of vocal fold pathology in a fully end-to-end approach.

III. Dataset

Voice data used in the project has been compiled from the Saarbruecken Voice Database (SVD) hosted by Saarland University. These voice samples have served as an ideal source as one of the largest databases of labeled healthy and pathological audio samples consisting of both sustained vowels (i.e., /a/ "ah", /i/ "ee", and /u/ "ooh") and sentences. The database includes recordings from over 2,000 individuals with over 70 class labels. For this project, we utilized all nine kinds of vowel sounds and a selection of 3 pathologies, including hyperkinetic dysphonia, hypokinetic dysphonia, and laryngitis.

Voice clips of the vowels 'a', 'i' and 'u' produced at a normal pitch were selected into our baseline model dataset. Our dataset consists of 674 healthy participants, 199 hyperkinetic dysphonia patients, 13 hypokinetic dysphonia patients and 115 laryngitis patients, each with 9 vowel recordings (Table 1). These classes were chosen based on their presence in related works [4].

Class	**Male**	**Female**	**Total**
Healthy	399	248	647
Dysphonia	158	41	199
Hypokinetic Dysphonia	5	8	13
Laryngitis	42	73	115
Total Pathological	205	122	654

Table 1**.** Counts of each recording type by condition and sex.

	Age Brackets									
	1-10	**11-20**	**21-30**	**31-40**	**41-50**	**51-60**	**61-70**	**71-80**	**81-90**	**Total**
M	76	28	120	75	47	21	17	0	0	384
F	1	190	214	68	71	55	28	10	1	638

Table 2. Counts of the age bin distribution across the participant dataset by sex. M: Male, F: Female

Data Preprocessing

The raw audio data was extracted with a sampling rate of 44100 Hz using the *librosa* package in Python. Due to the variability in the recording lengths, this sampling rate was chosen to provide a relatively large number of data points that could be included even after trimming. We also examined the presence of silence in the raw audio data and did not find any significant period of silence. All of the samples were trimmed from each end down to the same length of 21,000 values to avoid nonspecific variations at the starts and ends of phonated vowels.

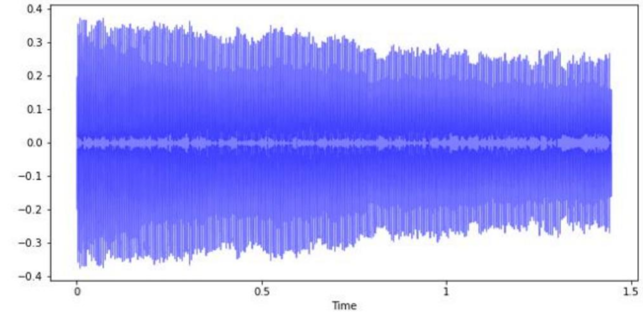

Figure 1. Audio waveform of a sustained 'a' in a neutral tone, extracted with a sampling rate of 44,100. The audio is from speaker 665 and is classified as having hyperfunctional dysphonia.

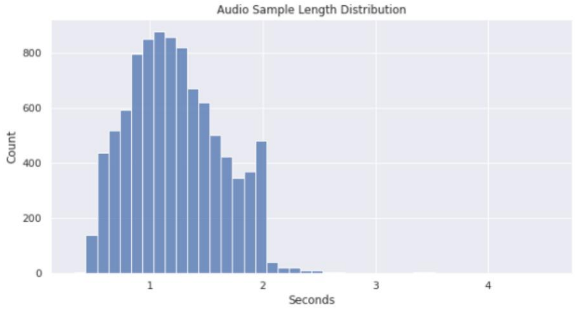

Figure 2. Distribution of the length of audio samples

IV. Methods

To explore the efficacy of our proposed novel deep learning framework, we created three different models: 1) a baseline model that takes in individual vowel recordings as inputs; 2) an intermediate model that takes in 3 vowels ('a', 'i', 'u') recordings stacked as inputs; 3) a full model that takes in the above 3 vowels in 3 different pitches ('low', 'neutral', 'high') stacked as inputs. To demonstrate proof of concept, we first used the models as binary classifiers for healthy versus pathologic voices. After reaching comparable performance with existing binary classifiers, we modified the models to classifying the specific class of vocal fold pathology.

For the baseline model, 3117 audio samples of 'a', 'i', 'u' were split into training, validation, and testing datasets. For the stacked vowel and stacked pitch models, audio files were organized into 3 channel stacks after trimming. For the stacked vowel input, the neutral pitch vowel recordings were standardized to the order [a, i, u] prior to being stacked into 3 channels per patient. The final sample count after stacking was 1039. For the stacked pitch input data, the recordings were standardized to the order [low, neutral and high] by vowel and by participant, and then were stacked into 3 channels with a final sample count of 3117. We split each dataset into training, validation and testing at a 6:2:2 ratio, stratified by their outcome.

For our deep learning framework, we utilized 1-dimensional convolutional neural networks (1-D CNN). This model type utilizes filters to extract features across a set length of the input data in one dimension, which in this case would be the time axis. The sequential nature of the audio data as well as the presence of semi-repetitive qualities, such as jitter and shimmer, should make the 1-D CNN effective at identifying hidden patterns within the recordings. A normalization layer was added after the input layer, which subtracted the input values from the mean and divided by the standard deviation of the inputs. Batch normalization of the activation values of each layer was also attempted, but we found both forms of normalization reduced the model performance and was therefore removed from the model. The models were implemented using *keras* [14].

$$L = \frac{1}{N} \Sigma_{\mathrm{N}}^{\mathrm{i=1}} - (\mathrm{y_i} * \log(p_i) + (1 - y_i) * \log(1 - p_i))$$

Equation 1. Binary cross-entropy loss equation

In the binary classification models, the binary cross-entropy loss function was used. This loss function is effective for binary classification applications as it effectively captures the difference between the true class and the model's prediction. Only one of the logarithmic terms is ever non-zero as $\hat{y}$ is always 0 or 1, and thus the function summarizes whether the predicted value was in fact close to the ground truth label. The negative sign is used to invert the log values such that the function will be minimized when the prediction is accurate.

The 3-channel stacked vowel and 3-channel stacked pitch input data models were trained on a similar 1-D CNN architecture utilizing the binary cross-entropy loss function. The 3-channel stacked vowel input model was used as an iterative step to verify that the stacking could indeed encode additional information. We then developed the 3-channel stacked pitch input data to test whether different pitches of the same vowel held more latent relationships. The binary classification task was used as an iterative step to validate the effectiveness of the stacked input models.

In the multiclass classification model, the loss function we used was multiclass cross-entropy loss.

$$L = \frac{1}{N}\ \Sigma_N^i\ \Sigma_M^j\ y_{ij} \log(\mathrm{p_{ij}})$$

Equation 2. Categorical cross-entropy loss equation. M represents the number of classes.

This loss function is calculated for each class and aggregated across the classes. Similar to the binary cross-entropy loss, only when the class is non-zero, will the loss function calculate the log value of $\hat{y}$, the predicted probability. As the probability values range between 0 and 1, the negative sign will inverse the log values so that the loss function minimizes as predictions become more accurate. In the multi-class classification models, the final output layer is replaced by a 4-unit *softmax* activation layer, which provides predicted probability of the four classes that sum to 1.

For the model training, we used the *Adam* optimization algorithm. The function is listed below:

$$m_t = \beta m_{t-1} + (1 - \beta_1)\left[\frac{\delta L}{\delta w_t}\right] v_t = \beta_2 v_{t-1} + (1 - \beta_2)\left[\frac{\delta L}{\delta w_t}\right]^2$$

Equation 3. Adaptive moment estimation, or *Adam*, optimization algorithm.

Adam provides updates based on the supplied learning rate and exponentially weighted moving averages of the gradient and squared gradient for each learnable parameter [13]. Each model's hyperparameters were tuned via the random search feature in the *keras_tuner* package that optimized the validation loss. The hyperparameters we tuned included the number of filters, filter size, size of the pooling layer, batch size and the learning rate of the model.

V. Results

Through hyperparameter tuning, we obtained the hyperparameters for our final model. The model specifications for our final model with stacked vowels at different pitches are in figure 3 below.

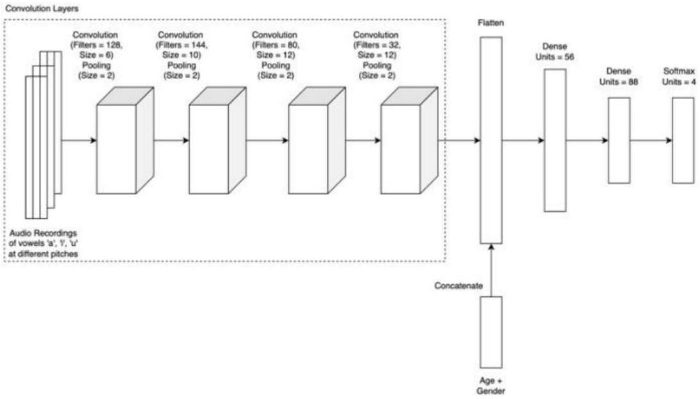
Convolution Layers
Convolution (Filters = 128, Size = 6) Pooling (Size = 2)
Convolution (Filters = 144, Size = 10) Pooling (Size = 2)
Convolution (Filters = 80, Size = 12) Pooling (Size = 2)
Convolution (Filters = 32, Size = 12) Pooling (Size = 2)
Audio Recordings of vowels 'a', 'i', 'u' at different pitches
Flatten
Dense Units = 56
Dense Units = 88
Softmax Units = 4
Concatenate
Age + Gender

Figure 3. Model specifications for *VocalPathNet* (Stacked vowels at different pitches)

To evaluate the performance of our models, we focused on three different metrics, including accuracy, area under the receiver operating curve (AUROC) and the F1 score. The accuracy reflects the percentage of predictions the model makes correctly. The AUROC reflects the model's ability to discriminate between the different vocal fold pathologies, specifically when the predicted probability of the specific type of vocal fold pathology is higher for the correct class than the others. In the multi-class classification task, there is a class imbalance issue, where we have significantly higher numbers of some vocal fold pathologies than others (e.g. 199 cases of hyperkinetic dysphonia versus 16 cases of hypokinetic dysphonia). This prompted us to weight the AUROC by the number in each type of vocal fold pathologies. The last metric is the F1 score, which is the harmonic mean of precision and recall. F1 score considers class imbalance and distinguishes between false positives and false negatives.

In the binary classification task, while the F1 score was similar (0.842 vs 0.847) between the two models, the accuracy (0.796 vs. 0.841) and AUROC (0.819 vs 0.890) of the stacked vowels model were higher than the baseline model (Table 3). Encouraged by the massive improvement in discrimination, we modified the models for the multi-class classification task.

The multi-class classification models with the stacked vowels performed significantly better than the baseline model on the testing dataset (Table 3). The accuracy of the baseline model was 0.702, and the 0.817 in the stacked vowel model. The AUROC of the baseline model was 0.812 compared with 0.846 in the stacked vowel model. The F1 score was significantly higher in the stacked vowel model with 0.805 compared with 0.650 in the baseline model. Optimal human accuracy on related vocal pathologies classification tasks were around 0.6 [7].

Model	**Accuracy**	**AUROC**	**F1 Score**
Baseline - Binary	0.796	0.819	0.842
Stacked Vowel - Binary	0.846	0.901	0.852
Stacked Vowel with Pitch - Binary	0.821	0.849	0.818
Baseline - Multiclass	0.702	0.812	0.650
Stacked Vowel - Multiclass	0.798	0.860	0.797
Stacked Vowel with Pitch - Multiclass	0.779	0.865	0.765

Table 3. Comparison of model performance

We examined the misclassification rates for each type of vocal fold pathologies (Table 4). In the baseline model, 8.0% of healthy voice samples were misclassified, and 53.6% of hyperkinetic dysphonia samples were misclassified. All of the hypokinetic dysphonia and laryngitis samples were misclassified. Whereas in the stacked vowel model, the number of healthy samples that were misclassified was slightly higher at 11.6%, but the misclassified percentage of hyperkinetic dysphonia patients and laryngitis patients were much lower (35.7% and 30.7% respectively). The hypokinetic dysphonia patients were all misclassified, most likely due to the low sample number in our dataset.

Model	**Healthy**	**Hyperkinetic Dysphonia**	**Hypokinetic Dysphonia**	**Laryngitis**
Baseline	8.0%	53.6%	100%	100%
Stacked Vowel	11.6%	35.7%	100%	30.7%
Stacked Vowel with Pitch	7.7%	54.4%	100%	38.9%

Table 4. Comparison of classification errors in the multiclass models

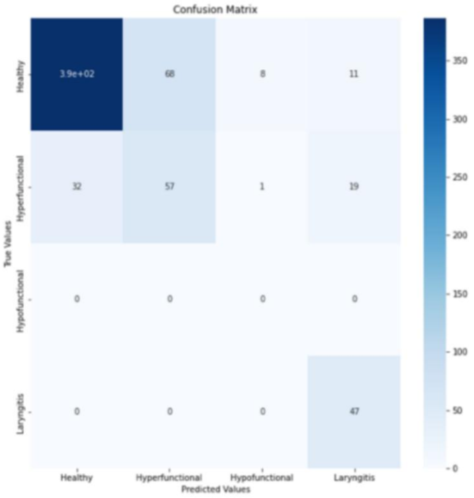

Figure 2. Confusion matrix of output classifications from stacked pitch model trained on training and development data subsets.

In addition to performance metrics, we examined the potential of overfitting in our models. In the stacked vowel model, the validation loss was slightly higher than the training loss, indicating the model is slightly overfit (Figure 2). In future iterations of the model, we intend to apply dropout to the convolution layers to see if it narrows the difference between the training and validation loss. Dropout is a form of regularization which shuts off certain nodes during the training process, helping with eliminating some effects of noise in the data.

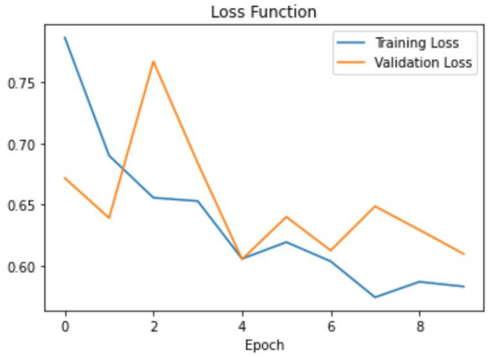

Figure 3. Training and validation loss of the stacked vowel model.

VI. Conclusion & Future Work

In this study, we demonstrated the efficacy of an end-to-end deep learning framework, *VocalPathNet*, that takes in multiple stacked raw vowels recordings to classify specific classes of vocal fold pathologies. This type of input data architecture was previously unexplored in deep learning classification models for classifying voice pathologies. We also expand on the research in the less explored areas of classifying specific vocal fold pathologies as well as using an end-to-end deep learning approach with raw audio recordings from patients. This proof-of-concept highlights the potential latent relationships between vowel recordings that can be used to better diagnose vocal fold pathologies.

For future work, we intend to improve the model through several approaches. First, we aim to address the problem of class imbalance, specifically for the hypokinetic dysphonia patients. We will augment the data through up-sampling on the recordings from hypokinetic dysphonia patients. Second, we wish to address overfitting through adding dropout to our convolution layers. Last, we would like to explore the specific regions of the recordings that the model focuses on for prediction. To achieve this, we will implement an attention mechanism within the convolution layers and observe the attention vector post model training.

CONTRIBUTIONS

Jordan Hodges designed the stacked vowels model, and implemented the data preprocessing pipeline and model. He also contributed to the writing of the final report and video recording.

Jingzhi Yu implemented the first iteration of the data preprocessing pipeline and the baseline model. He also contributed to the writing of the final report.

We acknowledge the expertise provided by Dr. George Liu and Dr. Philip Doyle for this project

REFERENCES

[1] Lili Chen, Junjiang Chen. Deep Neural Network for Automatic Classification of Pathological Voice Signals. Journal of Voice, Volume 36, Issue 2,2022. Pages 288.e15-288.e24. ISSN 0892-1997,https://doi.org/10.1016/j.jvoice.2020.05.029.

[2] Sáenz-Lechón, Nicolás, Juan I. Godino-Llorente, Víctor Osma-Ruiz, and Pedro Gómez-Vilda. "Methodological Issues in the Development of Automatic Systems for Voice Pathology Detection." Voice Models and Analysis for Biomedical Applications 1, no. 2 (April 1, 2006): 120–28. https://doi.org/10.1016/j.bspc.2006.06.003.

[3] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," in IEEE Transactions on Biomedical Engineering, vol. 51, no. 2, pp. 380-384, Feb. 2004, doi: 10.1109/TBME.2003.820386.

[4] Saska Tirronen, Sudarsana Reddy Kadiri, Paavo Alku. The Effect of the MFCC Frame Length in Automatic Voice Pathology Detection, Journal of Voice, 2022, ISSN 0892-1997, https://doi.org/10.1016/j.jvoice.2022.03.021.

[5] F. T. Al-Dhief et al., "Voice Pathology Detection and Classification by Adopting Online Sequential Extreme Learning Machine," in IEEE Access, vol. 9, pp. 77293-77306, 2021, doi: 10.1109/ACCESS.2021.3082565.

[6] Syed, Sidra Abid, Munaf Rashid, Samreen Hussain, and Hira Zahid. "Comparative Analysis of CNN and RNN for Voice Pathology Detection." Edited by Wen Si. BioMed Research International 2021 (April 15, 2021): 6635964. https://doi.org/10.1155/2021/6635964.

[7] Hu, H. C., Chang, S. Y., Wang, C. H., Li, K. J., Cho, H. Y., Chen, Y. T., Lu, C. J., Tsai, T. P., & Lee, O. K. (2021). Deep Learning Application for Vocal Fold Disease Prediction Through Voice Recognition: Preliminary Development Study. Journal of medical Internet research, 23(6), e25247. https://doi.org/10.2196/25247

[8] Powell ME, Rodriguez Cancio M, Young D, Nock W, Abdelmessih B, Zeller A, Perez Morales I, Zhang P, Garrett CG, Schmidt D, White J, Gelbard A. Decoding phonation with artificial intelligence (DeP AI): Proof of concept. Laryngoscope Investig Otolaryngol. 2019 Mar 25;4(3):328-334. doi: 10.1002/lio2.259. PMID: 31236467; PMCID: PMC6580072.

[9] Mohammed MA, Abdulkareem KH, Mostafa SA, Khanapi Abd Ghani M, Maashi MS, Garcia-Zapirain B, Oleagordia I, Alhakami H, AL-Dhief FT. Voice Pathology Detection and Classification Using Convolutional Neural Network Model. Applied Sciences. 2020; 10(11):3723. https://doi.org/10.3390/app10113723

[10] Changqin Quan, Kang Ren, Zhiwei Luo, Zhonglue Chen, Yun Ling. End-to-end deep learning approach for Parkinson's disease detection from speech signals. Biocybernetics and Biomedical Engineering, Volume 42, Issue 2, 2022, Pages 556-574.ISSN 0208-5216, https://doi.org/10.1016/j.bbe.2022.04.002.

[11] Harar, Pavol, Jesus B. Alonso-Hernandez, Jiri Mekyska, Zoltan Galaz, Radim Burget, and Zdenek Smekal. "Voice Pathology Detection Using Deep Learning: A Preliminary Study." In 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), 1–4. Funchal, Portugal: IEEE, 2017. https://doi.org/10.1109/IWOBI.2017.7985525.

[12] Schlegel, P., Kniesburges, S., Dürr, S. et al. Machine learning based identification of relevant parameters for functional voice disorders derived from endoscopic high-speed recordings. Sci Rep 10, 10517 (2020). https://doi.org/10.1038/s41598-020-66405-y

[13] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv; 2017. doi:10.48550/arXiv.1412.6980

[14] Chollet, F., & others. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras