Journal of Innovations in Computer Science and Engineering
Shahid Beheshti Unversity

January 2025, Special Issue on AI 4 All- 2

Efficient DL Models for Voice Pathology Detection in Healthcare Applications using Sustained Vowels

Sahar Farazi✉, , Code ORCID: 0000-0002-5244-4776
Faculty of Computer Science and Engineering , Shahid Beheshti University , Tehran, Iran , farazisahar75@gmail.com
Yasser Shekofteh, Code ORCID: 0000-0002-6733-3702
Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran,

Abstract— Voice Pathology Detection (VPD) aims to identify voice impairments through the analysis of speech signals, providing a foundation for developing diagnostic tools in advanced healthcare services to the public. This paper contributes to the development of efficient and accurate models based on deep learning (DL) for automatic VPD using sustained vowels of speech data. Therefore, this study explores the comparative efficacy of Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding (LPC) as acoustic features extracted from vowels /i/, /a/, and /u/. Using the AVFAD database, we utilized and optimized a Convolutional Neural Network (CNN) as a DL model to classify healthy and pathological voices, prioritizing both accuracy and computational efficiency for real-time applications. Our findings reveal that 20 MFCC features extracted from vowel /i/ achieve the highest accuracy, with the optimal model reaching approximately 88% on test data.

Keywords— Voice Pathology Detection, Sustained Vowel, Feature extraction, MFCC, LPC, CNN.

I. Introduction

Voice disorders, such as vocal fold nodules, polyps, and laryngitis, affect millions of people worldwide, often leading to challenges in communication and reduced quality of life. Despite the prevalence of these conditions, many individuals lack access to specialized medical professionals and diagnostic facilities, particularly in remote or underserved areas. Early detection of voice pathologies using an independent healthcare application, when the condition is still in its initial stages, is crucial for effective treatment and better outcomes. [1-2]

Recent advancements in artificial intelligence, particularly deep learning (DL), have opened new opportunities for leveraging speech signal analysis in voice pathology detection (VPD). There are many studies that have demonstrated the potential of using data-driven approaches and speech-based features to develop reliable diagnostic tools for healthcare purposes.

In [3], a VPD system was developed within a mobile healthcare framework. Smart devices were utilized to capture and process voice signals, leveraging transfer learning with CNN models such as *VGG-16* and *CaffeNet*. Using the *Saarbrücken Voice Disorder* (SVD) database, the system achieved an accuracy of 97.5%, emphasizing the potential of mobile platforms in improving voice pathology diagnostics. In [4], a VPD system was proposed within a smart healthcare framework, utilizing IoT devices like microphones and electroglottography (EGG) to capture voice and physiological signals. Spectrograms from these inputs were processed using a pretrained CNN for feature extraction and a bi-directional LSTM for classification. The system achieved 95.65% accuracy on the SVD database, demonstrating the benefits of bimodal input for enhanced diagnostic performance. In [5], a cloud-based smart healthcare monitoring framework was proposed to facilitate the interaction of smart devices and environments for providing accessible and affordable healthcare within smart cities. The framework included a VPD system that used voice and EGG signals, which were transmitted to the cloud for processing. Local and cepstral features were extracted from the signals, and a Gaussian Mixture Model (GMM) was applied for classification. The system achieved over 93% accuracy, demonstrating the potential of cloud-oriented solutions for improving healthcare delivery in smart cities.

In this study, we aim to contribute to this growing field by developing an optimized CNN model for VPD. Using the *Advanced Voice Function Assessment Database* (AVFAD), which includes sustained vowels /a/, /i/, and /u/, we compare the performance of two widely acoustic speech features: Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding (LPC). Our goal is to identify the vowel and feature extraction technique that yields the best accuracy. After determining the optimal feature and vowel, we further refine the CNN model by optimizing its parameters to reduce memory usage and enhance processing speed during testing, making it more suitable for real-time applications. In [6], it is mentioned that among Deep Neural Networks (DNN), Bidirectional Long Short-Term Memory (Bi-LSTM), and CNN, CNN achieved the highest accuracy in classifying healthy samples from those with voice disorders in spontaneous speech of AVFAD. Additionally, numerous studies have highlighted CNN as a highly effective model for VPD [7]. As a result, we selected the model proposed in [6] as the base CNN model for our study.

Submit Date: 2024-12-01
Accept Date: 2025-05-05
✉ Corresponding author

A key novelty of this work lies in comparing MFCC and LPC features across the three sustained vowels in the AVFAD dataset—an important and publicly available database used in VPD studies [8]. To the best of our knowledge, no prior research has conducted such a comparative analysis on this dataset. Additionally, we address the practical need for efficient and accurate models by focusing on optimizing CNN architectures. Our findings aim to advance the development of functional diagnostic tools, contributing to the accessibility and accuracy of VPD technologies for healthcare framework.

In the following sections of this paper, we will first review related works in the field of VPD. Then, we will describe the dataset used for our study and explain the extracted features. Following that, we will provide an overview of the utilized CNN-based model that we implemented and detailing its architecture. In Section 4, we will present the evaluation results and discuss the performance of our smaller models. Finally, we will conclude by reflecting on the findings from our study and propose potential directions for future research in VPD.

II. Related works

Various studies in the domain of automatic VPD have leveraged DL and machine learning (ML) models to identify patterns of voice disorders from speech signals. These investigations primarily aimed to differentiate healthy individuals from those with voice pathologies or to classify specific voice disorders among pathological cases [9]. In many approaches, features representing acoustic characteristics were extracted from speech signals. However, some studies have also explored the use of raw speech data for analysis [10]. Notably, most of these studies concentrated on sustained vowel sounds as a key data type.

Sustained vowel sounds refer to recordings where speakers produce and hold a specific vowel sound for a certain duration, enabling detailed acoustic analysis. These vowels are typically chosen from a common set that includes sounds such as /a/, /e/, /i/, /o/, and /u/. Several datasets, including SVD, AVPD, MEEI, and AVFAD, offer sustained vowel samples [11, 12]. The collection of this type of data plays a crucial role in evaluating acoustic features, as changes in these features can provide significant insights into potential voice disorders. Such information makes sustained vowels as a valuable resource for diagnosing voice pathologies. For instance, in [13], researchers utilized the SVD dataset, focusing specifically on the sustained vowel /a/. They used the *ResNet34* model as a pre-trained CNN. Spectrograms are their model's input. These spectrograms were configured with octave-based band filters, with at least 20 bands. Using this approach, the system achieved an impressive accuracy of approximately 96% in detecting voice pathologies.

In another study [14], researchers proposed a novel automatic VPD system rooted in voice production theory. This system emphasized features related to the vocal tract, particularly the glottal region, which is closely associated with voice pathologies. Their database consisted of sustained vowel /AH/ recordings from the MEEI dataset. The study extracted vocal tract irregularity features and employed an ML-based classifier to perform the detection, achieving a remarkable accuracy of 99.02% ± 0.01 on the clean MEEI dataset.

Study [15] investigated the effect of combining multiple vowel sounds (/a/, /i/, and /u/) on wavelet coefficient extraction using the AVFAD and SVD datasets. By applying a Random Forest classifier, it was found that merging vowel sounds improved the separability of wavelet coefficients, enhancing the system's accuracy in distinguishing between healthy and pathological sounds. This approach highlighted the advantage of using multiple vowels to improve feature discrimination. For combined gender models applied to sustained vowels in the AVFAD dataset, the highest accuracy achieved was approximately 79% for vowel /i/ and 78% for vowel /a/.

In [16], the SVD and AVFAD datasets were utilized to analyze sustained vowel sounds (/a/) and read speech, with participants reading pre-written sentences. The study employed a support vector machine (SVM) classifier to differentiate between pathological and healthy samples, using traditional acoustic features like MFCC for classification. The highest accuracy achieved was approximately 83% for read speech in the SVD dataset, while sustained vowel sounds in the AVFAD dataset reached a maximum accuracy of about 86%.

III. material and methods

A. AVFAD Database

In this study, we utilized the AVFAD dataset, created by the University of Aveiro in Portugal. This dataset features audio recordings from individuals with various speech disorders as well as those with healthy voices. For each participant, the dataset includes three distinct types of recordings. The first type comprises sustained vowel sounds, specifically the vowels /a/, /u/, and /i/, with each audio file containing three repetitions of the respective vowel. The second type consists of recordings of participants reading a specific text and six predefined sentences, while the third type includes samples of spontaneous speech. All audio recordings in the dataset were captured at a sampling rate of 48 kHz.

The AVFAD dataset contains recordings from 709 individuals in total, with 346 diagnosed with vocal pathologies and 363 classified as healthy controls. However, due to issues with the audio files of three participants, we excluded them from the analysis and proceeded with data from 706 individuals. To conduct our analysis, we partitioned the data into three subsets: training, testing, and validation sets. These subsets were allocated as 65% for training, 20% for testing, and 15% for validation. We tried to have a balanced distribution across these subsets, considering both the participants' gender and their health status. For splitting the data, we used the methodology outlined in [6]. Table 1 shows the gender and health distribution of the dataset. We opted for a custom split to ensure a balanced distribution of healthy and pathological samples as well as gender representation. To the best of our knowledge, there is no officially established data-splitting protocol for this dataset. Furthermore, apart from [6], no study has explicitly detailed their methodology for a custom split of the AVFAD dataset based on pathology (healthy vs. pathological) and gender balance across the three categories of training, validation, and test sets. In [6], the authors provided a well-documented methodology and achieved fair results, which is why we chose their approach.

Since the duration of each sample is different, we extracted the maximum, minimum, mean, and mean + std duration for all samples for the vowels /a/, /i/, and /u/. These values are presented in Table 2.

Table 1. Data Splitting Methodology in This Study for the AVFAD Dataset

Data	Train		Test		Validation	
Gender	***Male***	***Female***	***Male***	***Female***	***Male***	***Female***
Normal	73	162	22	50	18	37
Pathologic	64	161	20	49	13	37
Total (Gender)	137	323	42	99	31	74
Total (All)	460		141		105	

Table 2. Duration (sec) statistics for the vowels /a/, /i/, and /u/.

Vowels	Min	Max	Mean	Mean + STD
/a/	3.81	110.92	14.61	21.81
/i/	3.81	121.32	14.81	22.34
/u/	3.63	344.51	14.55	29.09

B. Feature Extraction

1) Mel Frequency Cepstral Coefficient (MFCC)

The Mel Frequency Cepstral Coefficients (MFCCs) method is commonly used as a foundational technique for feature extraction in speech processing, speech recognition, and audio pattern recognition [17]. They represent the short-term power spectrum of sound. They are computed by transforming the audio signal into the *Mel* scale, which approximates human ear perception of frequencies. The MFCCs are calculated by first applying a Fourier transform to the signal to extract its spectral components, followed by a *Mel* filter bank to map the frequencies to the *Mel* scale. Then, the logarithm of the filter bank energies is taken, and a discrete cosine transform (DCT) is applied to obtain the MFCCs. The formula for MFCC computation is:

$$c_i = \sum_{n=1}^{N_f} S_n \cos\left(i(n-0.5)\frac{\pi}{N_f}\right) \qquad (1)$$

where C_i is the i-th MFCC, N_f is the total number of triangular filters in the Mel filter bank, and S_n is the logarithmic energy of the n-th filter.

2) Linear Predictive Coding (LPC)

LPC features are extracted as time-domain based acoustic features by modeling the human vocal tract as an all-pole filter. Using a linear prediction (LP) method, they tried to estimate the speech signal as a weighted sum of its past samples. In the LP analysis, the speech signal *s[n]* is modeled as a linear combination of its past samples. The LPC coefficients a_k are determined by minimizing the squared prediction error *e[n]* over a windowed frame, where:

$$e[n] = s[n] - \sum_{k=1}^{p} a_k \, s[n-k] \qquad (2)$$

Here, p is the order of the prediction, which determines the number of coefficients used. Using the Autocorrelation Method or the Covariance Method, a set of linear equations (known as Yule-Walker equations) can be formed and then solved to obtain the LPC coefficients a_k. In this study, we extracted 20 LPC coefficients for each speech frame.

C. Based CNN Model

We trained our model using the CNN architecture proposed in [6]. Conventional CNNs are highly effective for processing two-dimensional (2D) data, such as the MFCC and LPC coefficient frames in our study [18]. The baseline CNN model employed in this study comprises three convolutional layers, three max-pooling layers, one global average pooling layer [19], two dropout layers, one dense layer, and one classification layer. The CNN model parameters and details are shown in Table 3.

Table 3: The CNN model parameters and details.

Input Layer	Input Shape (Number of Frames, Feature Size, 1)
Convolutional Layer 1	Kernel Size: 3 x 3, Filters: 16, Activation: RELU, Padding: Same
Max pooling Layer 1	Pool Size: 2 x 2
Convolutional Layer 2	Kernel Size: 3 x 3, Filters: 32, Activation: RELU, Padding: Same
Max pooling Layer 2	Pool Size: 2 x 2
Convolutional Layer 3	Kernel Size: 3 x 3, Filters: 64, Activation: RELU, Padding: Same
Max pooling Layer 3	Pool Size: 2 x 2
Global Average Pooling	2 Dimensional
Dropout Layer 1	Rate: 0.5
Dense Layer 1	Units: 128, Activation: RELU
Dropout Layer 2	Rate: 0.3
Dense Layer 2	Units: 1, Activation: Sigmoid

Additionally, the optimizer used in the model is Adam, with a learning rate of 0.001 and the binary cross-entropy loss function.

IV. evaluation and results

In this section, we present the experiments conducted in this study. Our objective was to develop a simple yet effective DL model capable of accurately classifying healthy and unhealthy samples in the AVFAD database. To achieve this, we evaluated our base model on three types of data preprocessing approaches:

1. **Using Mean + STD Duration (~2045 frames for vowel/a/, 2095 frames for vowel /i/, and 2725 frames for vowel /u/)**: For this approach, we calculated the mean + standard deviation (STD) duration for all the data. For audio files with fewer frames than this value, we padded the data by adding zeros to replace the missing frame coefficients.
2. **Using Mean + STD without Silence (~1800 frames for vowel/a/, 1825 frames for vowel /i/, and 1825 frames for vowel /u/)**: In the second approach, we used the same mean + STD duration but first removed all silence parts from all audios before applying the padding strategy.
3. **Using Fixed 15-Second Duration (~1430 frames for vowel/a/, 1430 frames for vowel /i/, and 1430 frames for vowel /u/)**: For the third approach, we standardized all audio files to the first 15 seconds. For files shorter than 15 seconds, zeros were added to replace the missing frame coefficients.

These pre-processing parts, using the values of Table 2, were performed for all three vowels (/a/, /i/, and /u/). We used the frame length of approximately 42 milliseconds (msec) with an overlap time of 32 msec to extract the frames. Subsequently, LPC and MFCC features are extracted. In the first part of our study to reach the best feature and model, we used 13 MFCCs and 20 LPC Coefficients. However, in the second part, we extracted 20 MFCCs to improve the learning process of the model. Also, features were standardized using z-score normalization. The results of the experiments by the CNN model and LPC features are summarized in Table 4 and by MFCC features are shown in Table 5. For all the experiments in this study, we ran our models twice and reported the best accuracy from the two runs for each experiment.

Table 4: LPC-Based CNN Model - Validation & Test Accuracies for Different Vowels

Vowel type	Frames (Mean+STD) with silence	Frames (Mean+STD) without Silence	First 15 Seconds (with silence)
Vowel /i/	Valid:0.8952 **Test: 0.8591**	Valid:0.8571 **Test: 0.8098**	Valid:0.8857 Test: 0.8380
Vowel /a/	Valid:0.8666 Test: 0.8239	Valid:0.8476 Test: 0.7676	Valid:0.9142 **Test: 0.8450**
Vowel /u/	Valid:0.7619 Test: 0.8028	Valid:0.7333 Test: 0.7183	Valid:0.7809 Test: 0.7816

Table 5: MFCC-Based CNN Model - Validation & Test Accuracies for Different Vowels

Vowel type	Frames (Mean+STD) with silence	Frames (Mean+STD) Without Silence	First 15 Seconds (with silence)
Vowel /i/	Valid:0.8761 **Test:0.8661**	Valid:0.8476 **Test:0.8239**	Valid:0.8666 **Test:0.8521**
Vowel /a/	Valid:0.8857 Test:0.8309	Valid:0.8761 Test:0.7535	Valid:0.8761 Test:0.8380
Vowel /u/	Valid:0.9238 Test:0.8591	Valid:0.9047 Test:0.8098	Valid:0.9047 Test:0.8309

Based on the results presented in Tables 4 and 5, we observe that the highest test accuracy is achieved when using the mean + STD duration of the audio files combined with MFCC features. Among the vowels /a/, /i/, and /u/, the best result is obtained for the vowel /i/. Therefore, we decided to focus our study on this vowel and evaluate smaller CNN architectures to determine how much we can reduce the model size while maintaining accuracy close to the 86% achieved by the base model. In addition, to the best of our knowledge, no study has specifically aimed to compare and analyze the impact of different vowel types, the presence or absence of silence, and different sustained vowel file durations in the AVFAD dataset, as well as to determine whether LPC or MFCC features yield the best results for voice pathology detection and the classification of healthy and pathological samples. As a result, direct comparisons between our results and those of previous studies may not be entirely accurate.
To create smaller CNN models, we systematically reduced the number of filters in the convolutional layers, eliminated certain layers, and adjusted the number of neurons in the dense layer. The detailed configurations and results are as follows:

- **Base CNN Model 1**
  - **Architecture**: Three convolutional layers with 16, 32, and 64 filters, respectively, followed by max-pooling layers. First Dense layer neurons: 128.
  - **Parameters**: 31,745 (~124 KB)
  - **Validation Accuracy**: 0.8761
  - **Test Accuracy**: 0.8661
- **Small CNN Model 2**
  - **Architecture**: Two convolutional layers with 8 and 16 filters, followed by max-pooling layers. The third convolutional layer and its max-pooling layer were discarded. First Dense layer neurons: 64.
  - **Parameters**: 1,721 (~6.72 KB)
  - **Validation Accuracy**: 0.8380
  - **Test Accuracy**: 0.7464
- **Small CNN Model 3**
  - **Architecture**: Two convolutional layers with 16 and 32 filters, followed by max-pooling layers. The third convolutional layer and its max-pooling layer were discarded. First Dense layer neurons: 64.
  - **Parameters**: 6,977 (~27.25 KB)
  - **Validation Accuracy**: 0.8571
  - **Test Accuracy**: 0.8309
- **Small CNN Model 4**
  - **Architecture**: Three convolutional layers with 8, 16, and 32 filters, each followed by max-pooling layers. First Dense layer neurons: 64.
  - **Parameters**: 8,065 (~31.5 KB)
  - **Validation Accuracy**: 0.8666
  - **Test Accuracy**: 0.8521

The optimal model, with three convolutional layers (8, 16, and 32 filters) and a dense layer of 64 neurons, achieves a test accuracy of 0.8521 while significantly reducing the number of parameters compared to the base model.
Results show that the model with 8,065 parameters (Model 4) strikes an optimal balance between size and performance, achieving a test accuracy of 0.8521 while significantly reducing the number of parameters compared to the base model. Additionally, the inference time of testing the base model on test data is about 1 second and half (1.54 sec) while this time for the optimal model is less than one second (0.77 sec).

To explore whether our optimal CNN model could achieve higher accuracy, we extracted 20 MFCCs instead of 13 coefficients to evaluate its performance. We obtained an accuracy of 0.8857 on the validation data and 0.8802 on the test data, which is an improvement compared to using 13 MFCCs for the vowel /i/. This demonstrates that using 20 MFCCs, combined with our optimal CNN model, resulted in better accuracy.
Figure 1 illustrates the validation accuracy curve during the training of the optimal model using 20 MFCCs for the vowel /i/. Additionally, Table 6 presents the precision, recall, and F1-score for healthy samples (0) and unhealthy samples (1).

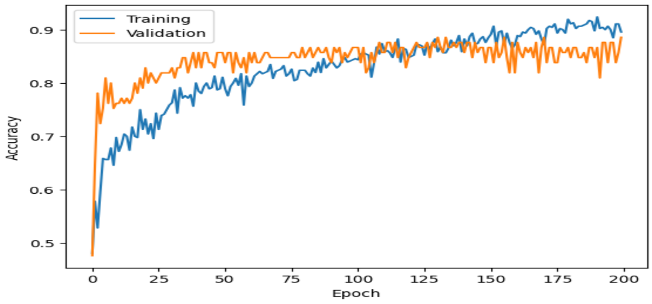

Figure 1: Validation Accuracy Curve for Optimal CNN Model on 20 MFCCs (Vowel /i/)

Table 6: Precision, Recall, and F1-Score for Healthy (0) and Unhealthy (1) Samples of 20 MFCCs for Vowel /i/ Using the Optimal CNN Model

Label	Precision	Recall	F1-Score
0	0.95	0.81	0.87
1	0.83	0.96	0.89

The model demonstrates high performance for healthy samples (0), with a precision of 0.95 and an F1-score of 0.87, indicating that it effectively identifies healthy samples. For unhealthy samples (1), the model achieves a precision of 0.83 and a recall of 0.96, reflecting its ability to accurately detect most unhealthy samples, though with a slightly higher false-positive rate that the model mistakenly classifies healthy samples as unhealthy. Overall, the F1-scores of 0.87 for healthy and 0.89 for unhealthy samples suggest a well-balanced performance.

V. Future works

For future work, exploring LPC features further could provide valuable insights. Additionally, experimenting with other acoustic features (PLP, PLPC, recurrence plots, and multivariate AR) and varying coefficient numbers may enhance classification. Combining multiple vowel sounds and leveraging pre-trained models could further improve accuracy and efficiency in voice pathology detection [20, 21, 22]. Although we selected our CNN-based model from those proposed in [6] and [23], and in [23] this model achieved approximately 84% accuracy using MFCC features and the sustained vowel /a/, future studies could explore modifications to improve performance. Specifically, it would be worthwhile to experiment with different kernel sizes better suited for speech signals and to investigate the use of GELU instead of ReLU as the activation function.

References

[1] Abdulmajeed, N.Q., et. al., *A review on voice pathology: Taxonomy, diagnosis, medical procedures and detection techniques, open challenges, limitations, and recommendations for future directions*, Journal of Intelligent Systems, 2022. **31**(1): p. 855-875.

[2] Latif, S., et al., *Speech technology for healthcare: Opportunities, challenges, and state of the art*, IEEE Reviews in Biomedical Engineering, 2020. **14**: p. 342-356.

[3] Alhussein, M. and G. Muhammad, *Voice pathology detection using deep learning on mobile healthcare framework.* IEEE Access, 2018. **6**: p. 41034-41041.

[4] Muhammad, G. and M. Alhussein, *Convergence of artificial intelligence and internet of things in smart healthcare: a case study of voice pathology detection.* IEEE Access, 2021. **9**: p. 89198-89209.

[5] Hossain, M.S., G. Muhammad, and A. Alamri, *Smart healthcare monitoring: a voice pathology detection paradigm for smart cities.* Multimedia Systems, 2019. **25**(5): p. 565-575.

[6] Farazi, S., and Shekofteh, Y., *Voice pathology detection on spontaneous speech data using deep learning models*, International Journal of Speech Technology, 2024. **27**(3), p. 739–751.

[7] Syed, S.A., et. al., *Comparative Analysis of CNN and RNN for Voice Pathology Detection,* BioMed Research International, 2021. **2021**(1), p. 1–8.

[8] Jesus, L.M., et al., *The advanced voice function assessment databases (AVFAD): Tools for voice clinicians and speech research*, in *Advances in Speech-language Pathology*. 2017, IntechOpen.

[9] Xie, X., et. al., *A Voice Disease Detection Method Based on MFCCs and Shallow CNN,* Journal of Voice, 2023. In press.

[10] Harar, P., et al. Voice pathology detection using deep learning: a preliminary study. in 2017 international conference and workshop on bioinspired intelligence (IWOBI). 2017. IEEE.

[11] Pützer, M. and W. Barry, Saarbrücken Voice Database, Institute of Phonetics, Saarland University. 2009.

[12] Mesallam, T.A., et al., *Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms,* Journal of healthcare engineering, 2017. **2017**(1): p. 8783751.

[13] Mohammed M.A., *et al.*, "Voice pathology detection and classification using convolutional neural network model," *Appl. Sci.*, vol. 10, no. 11, 2020, doi: 10.3390/app10113723.

[14] Muhammad, G. *Voice pathology detection using vocal tract area*. in *2013 European Modelling Symposium*. 2013. IEEE.

[15] Oliveira, B. F., et. al., Combined sustained vowels improve the performance of the Haar wavelet for pathological voice characterization. In 2020 IWSSIP, pp: 381-386, IEEE.

[16] Ribas, D., et. al., On the Problem of Data Availability in Automatic Voice Disorder Detection, In HEALTHINF, 2023. pp. 330–337.

[17] Sidhu, M.S., et. al., *MFCC in audio signal processing for voice disorder: a review.* Multimedia Tools and Applications, 2024, In press, p. 1-21.

[18] Li, Z., et al., *A survey of convolutional neural networks: analysis, applications, and prospects.* IEEE transactions on neural networks and learning systems, 2021. **33**(12): p. 6999-7019.

[19] Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting.* The journal of machine learning research, 2014. **15**(1): p. 1929-1958.

[20] Firooz, S., et al., *Improvement of automatic speech recognition systems utilizing 2D adaptive wavelet transformation applied to recurrence plot of speech trajectories*. Signal, Image and Video Processing, 2024. **18**(2): p. 1959-1967.

[21] Shekofteh, Y. *What can phone attractors in RPS tell us? A study of dynamic information in speech signals for phone classification purposes*, Applied Acoustics, 2023. **211**: p. 109534.

[22] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv.org*, Apr. 10, 2015. https://arxiv.org/abs/1409.1556.

[23] S. Farazi and Y. Shekofteh, “Evaluation of phone posterior probabilities for pathology detection in speech data using deep learning models,” *International Journal of Speech Technology*, Jan. 2025, doi: 10.1007/s10772-024-10166-w.