Utilizing a Multimodal Deep Learning Model to Identify Parkinson's Disease from Voice Samples

Andrew Oh[1]

[1]Crescenta Valley High School, Glendale, CA, USA

Abstract

Dysarthria, a motor speech disorder affecting control of speech muscles, is a common early symptom of Parkinson's Disease (PD), making voice analysis a promising tool for early detection. Acoustic biomarkers derived from sustained vowel phonations have shown potential for PD detection. This study develops a multimodal transformer model that integrates engineered acoustic features with log-Mel spectrogram embeddings to classify PD from sustained /a/ phonations. Eight acoustic features—mean fundamental frequency (F0), local jitter, local shimmer, detrended fluctuation analysis (DFA) exponent, pitch period entropy (PPE), recurrence period density entropy (RPDE), pitch variability, and harmonics-to-noise ratio (HNR)—were extracted using Librosa and Parselmouth and processed through a numerical branch. In parallel, log-Mel spectrograms were encoded with pretrained CNN backbones (ResNet18, EfficientNet_B0, MobileNet_V3_Large) in an image branch. A transformer layer integrated both modalities, with a final classifier predicting PD vs. healthy controls. Two datasets were combined: D1 with 81 recordings (41 HC, 40 PD; average ages 47.7 ± 14.3 and 67.0 ± 9.0 years) and D2 with 99 recordings (44 HC, 55 PD; average ages 67.1 ± 5.2 and 67.2 ± 8.7 years), yielding 180 recordings (85 HC, 95 PD). Data were split 80% training, 10% validation, and 10% testing. Models were trained with a learning rate of 1e-4, batch size 32, and 10 epochs across 5 independent runs. The best model achieved an accuracy of 0.93 ± 0.07, precision of 0.96 ± 0.05, recall of 0.91 ± 0.08, and F1-score of 0.94 ± 0.06, with stable training and validation loss convergence. These findings suggest that lightweight multimodal fusion of engineered acoustic and spectrographic features outperforms single-modality baselines and holds promise for scalable, noninvasive voice-based PD screening.

Keywords: *Parkinson's Disease, voice analysis, multimodal deep learning, acoustic features, log-mel spectrograms, transformer models*

1. Introduction

Parkinson’s Disease (PD) is a progressive neurodegenerative disorder that impairs motor function and cognition, manifesting in tremors, rigidity, bradykinesia, and postural instability. As medical technology advances and life expectancy increases, the prevalence of PD and other chronic conditions has risen correspondingly. Despite this, PD remains challenging to diagnose, particularly in its early stages. Even expert clinical assessments misclassify up to 15% of cases, with higher error rates reported among non-experts. Most PD symptoms remain undetected for long periods, with diagnosis often delayed by up to 10 years [1]. Earlier and more efficient means of detection of PD may be a key factor in providing proper treatment and halting disease progression, and a supplementary tool that can help identify the disease could play a major role in saving lives.

The clinical diagnosis of PD involves the identification of four key motor symptoms. Postural instability can lead to changes in posture and a loss in balance which can increase the risk of dangerous falls. Tremors commonly develop in the hands or in the jaw, leading to a rhythmic back-and-forth motion in the affected area. Rigidity can also affect the passive movement of limbs or the jaw and cause pain. Bradykinesia, or the slowness of movement, is the most characteristic of PD. Planning, initiation, and execution of movements are significantly slowed, leading to decreased reaction times and movement speeds. Bradykinesia can also manifest as hypophonic dysarthria, characterized by a monotone and breathy voice that is reduced in volume [2-4].

Tremors, rigidity, and bradykinesia can be identified as the physiological basis of speech impairment in PD. The lips, tongue, jaw, and vocal tract that serve as the producers and controllers of speech can be affected by tremors and rigidity, leading to decreased velocity and amplitude of movement [5]. This phenomenon can explain the reduced volume and imprecise articulation of speech in PD patients.

Previous studies have leveraged deep learning frameworks to identify PD through a variety of different modalities including speech, handwriting, and gait [6-10]. However, relatively few have combined engineered numerical features with spectrogram representations in a multimodal framework.

The multimodal approach processes multiple types of data during the learning process to improve decision making and classification. Multimodal deep learning models have previously displayed improved feature learning performance, as displayed in the study shown in [11] using both audio and video data for improved speech recognition. Multimodal approaches are also often common in medical research, particularly image segmentation to process multiple forms of medical imaging such as MRI, CT, PET, and others [12].

In this study, we leverage multimodal learning to analyze sustained vowel phonations from PD patients. Specifically, we propose a transformer-based architecture that integrates engineered acoustic features with spectrogram embeddings for binary classification of PD versus healthy controls. By incorporating both numerical and visual representations of speech, the proposed model has the potential to serve as a supplementary clinical tool, improving early detection of PD.

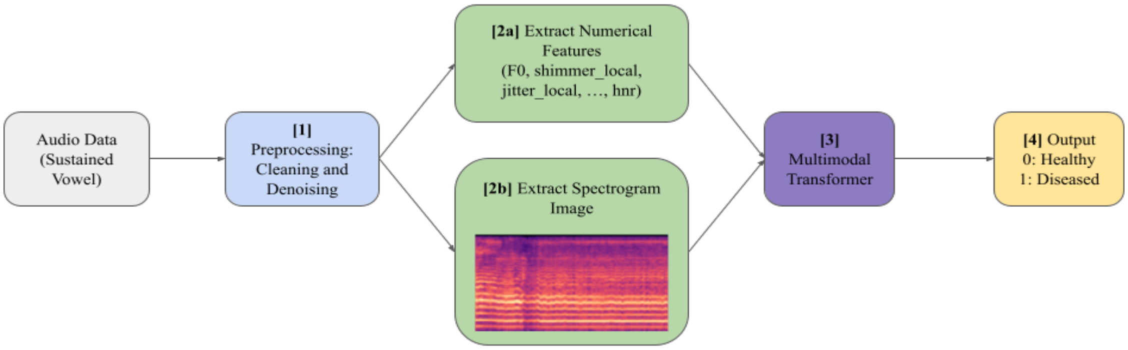
Audio Data (Sustained Vowel)
[1] Preprocessing: Cleaning and Denoising
[2a] Extract Numerical Features (F0, shimmer_local, jitter_local, …, hnr)
[2b] Extract Spectrogram Image
[3] Multimodal Transformer
[4] Output 0: Healthy 1: Diseased

Figure 1. Flowchart of the proposed PD deep learning detection procedure.

2. Methods

2.1 Dataset

To improve robustness and generalizability, we combined two distinct speech datasets (D1 and D2), each containing sustained vowel recordings from both healthy and diseased participants. This integration helped increase the total sample size, enabling more effective model training and preventing overfitting.

All recordings capture a sustained /a/ phonation. D1 [13] consists of 81 total recordings (41 HC, 40 PD) from 41 English-speaking HC participants (25 female, 16 male, average age 47.7±14.3 years) and 40 English-speaking PD participants (19 female, 21 male, average age 67.0±9.0 years). D2 [14] consists of 99 total recordings (44 HC, 55 PD) from 22 Italian-speaking HC participants (12 female, 10 male, average age 67.1±5.2 years) and 28 Italian-speaking PD participants (9 female, 19 male, average age 67.2±8.7 years. The combined dataset contains 180 total recordings (85 HC, 95 PD) allowing for more effective training and reducing the overfitting risk.

2.2 Features

We extracted a total of eight numerical features commonly used in PD voice analysis [15] from each audio file for the numerical branch. We also generated log-Mel spectrograms for the image branch. Numerical features were extracted using the librosa and parselmouth packages. Spectrograms were created using the librosa package [16, 17].

2.2.1 Mean Fundamental Frequency (F0)

F0 (Hz) is the average rate at which the vocal folds vibrate across the voiced segment. Dysarthria often shifts average pitch, and F0 statistics are consistently included in voice feature sets in PD speech pathology studies.

2.2.2 Local Jitter (Jitter)

Jitter (%) is the small, random variations in pitch period, or the duration of a cycle of a vocal fold vibration. A higher jitter indicates higher vocal fold instability and impaired laryngeal control.

2.2.3 Local Shimmer (Shimmer)

Shimmer (dB) is the variation in amplitude in irregular vocal fold vibration, indicating variation in volume. A higher shimmer indicates less control over vocal amplitude.

2.2.4 Detrended Fluctuation Analysis (DFA) Exponent

DFA quantifies long-range correlations in the F0 signal. It is a nonlinear measurement of the turbulent noise in a speech signal caused by unstable air flow in the vocal tract shown to be indicative of PD. The DFA correlates to dysphonias caused by incomplete vocal fold closure.

2.2.5 Pitch Period Entropy (PPE)

PPE is a nonlinear measurement of the unpredictability in pitch variation correlated to deteriorating muscle coordination in the vocal folds.

2.2.6 Recurrence Period Density Entropy (RPDE)

RPDE is a nonlinear measurement of the predictability of recurring voice structures and patterns, correlating to the ability of the vocal folds to sustain simple vibration.

2.2.7 Pitch Variability

Pitch variability is the spread of fundamental frequency across the phonation, corresponding to a more monotonic voice characteristic of PD.

2.2.8 Average Harmonics-to-Noise Ratio (HNR)

HNR is a measurement of the clarity of voice, measuring the ratio between harmonic and noise components of the signal. Lower HNR is indicative of increased breathiness and noise caused by dysphonia in PD patients.

2.2.9 Spectrogram Images

To capture time-frequency patterns of the sustained /a/ phonations, we created a log-Mel spectrogram for the middle 1.5 second window of each recording. The Mel scale provides a perceptual scale where equal “mel” increments correspond to equal perceived pitch differences [18]. Instead of linear frequency scales, we convert the frequency axis to the mel scale before applying a logarithmic compression. This mimics human perception of both pitch differences and loudness differences, making log-Mel spectrograms the mode of choice for image representations of speech pathology in deep learning models [19, 20]. We then normalized the spectrograms to a 0 to 1 scale to create more stable training and cross-speaker comparisons.

2.3 Model Design

The proposed multimodal transformer model jointly analyzes spectrogram images and the eight engineered acoustic features. We constructed the model using PyTorch [21]. The image branch is based on a lightweight convolutional neural network (CNN). To identify the most effective architecture, we evaluated three pretrained backbones: ResNet18, EfficientNet_B0, and MobileNet_V3_Large. For each, the final classification layers were removed, and the penultimate feature embeddings were retained for fusion.

The numerical branch is implemented as a feed-forward neural network with two fully connected layers. Input features are first projected to 64 dimensions, followed by a second projection to 128 dimensions. Each layer is followed by ReLU activation, batch normalization, and dropout (p=0.2) to mitigate overfitting.

Outputs from the two branches are concatenated to form a unified multimodal feature representation. This combined vector is processed by a classification head consisting of a fully connected layer (128 dimensions), ReLU activation, batch normalization, and dropout (p=0.4). A final linear layer outputs the binary prediction (PD vs. control).

2.4 Training Procedure

The model was trained on an 80% training, 10% testing, and 10% validation split. For each CNN backbone, we ran five independent trials with different random seeds and identical hyperparameters. We report mean ± standard deviation across the five runs.

2.4.1 Hyperparameters

We used the Adam optimizer, a learning rate of 1e-4, the CrossEntropyLoss function, a batch size of 32, and 10 epochs for training.

Table 1. List of hyperparameters used to train the proposed multimodal transformer model

Optimizer	Adam
Learning Rate	1e-4
Loss Function	CrossEntropyLoss
Batch Size	32
Epochs	10

2.4.2 Evaluation Metrics

We collected four evaluation metrics: accuracy, precision, recall, and F1-score.

2.4.3 Accuracy, Precision, Recall, and F1-Score

Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy indicates the proportion of correctly classified samples.

Precision

$$\frac{TP}{TP + FP}$$

Precision indicates the proportion of predicted positive samples that were true positives.

Recall

$$\frac{TP}{TP + FN}$$

Recall indicates the proportion of true positive samples that were correctly identified.

F1-Score

$$2 \cdot \frac{precision \cdot recall}{precision + recall}$$

F1-score indicates the harmonic mean of precision and recall.

3. Results

Table 2. Performance measures of the multimodal transformer model using different pretrained CNN backbones (mean ± standard deviation).

	ResNet18	EfficientNet_B0	MobileNet_V3_Large
Accuracy	0.93 ± 0.07	0.73 ± 0.09	0.74 ± 0.07
Precision	0.96 ± 0.05	0.77 ± 0.17	0.75 ± 0.09
Recall	0.91 ± 0.08	0.75 ± 0.13	0.85 ± 0.13
F1-Score	0.94 ± 0.06	0.74 ± 0.11	0.79 ± 0.09

The best results came from the model with the ResNet18 CNN architecture as the model achieved an average accuracy of 0.93 ± 0.07, average precision of 0.96 ± 0.05, average recall of 0.91 ± 0.08, and average F1-score of 0.94 ± 0.06 across the five runs. Overall, the model demonstrated strong and consistent performance, with stable training and validation loss curves across all runs, outperforming existing approaches.

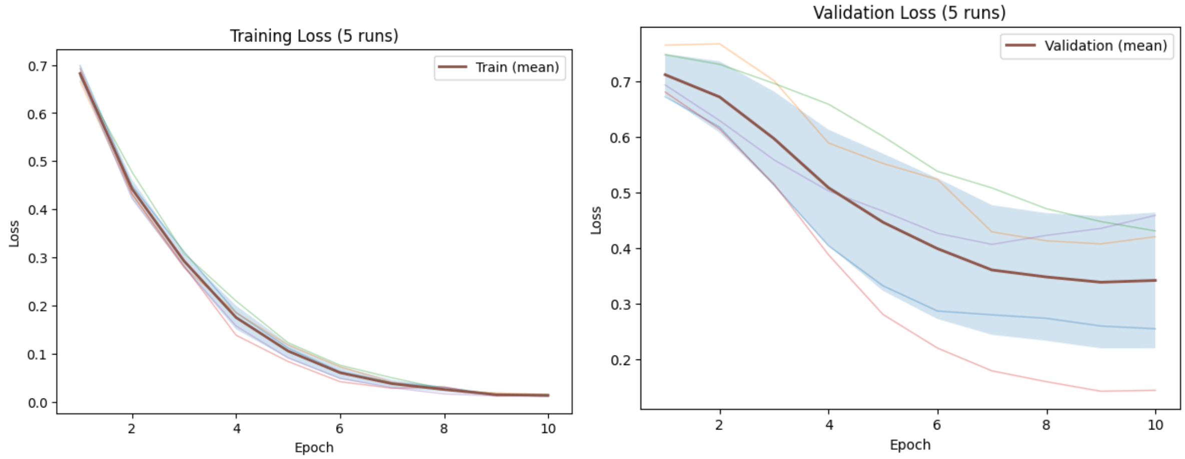

Figure 2. Left: Training loss curves across 5 independent runs using the pretrained ResNet18 CNN as the backbone for the image branch of the multimodal transformer.
Right: Validation loss curves across 5 independent runs using the pretrained ResNet18 CNN as the backbone for the image branch of the multimodal transformer.

4. Discussion

This study demonstrates the feasibility of using a multimodal deep learning framework that integrates spectrogram and acoustic numerical representations to detect PD from sustained vowel phonations. Our high results reinforce the hypothesis that combining complementary modalities can improve PD classification. While spectrograms only capture temporal and frequency-based information, numerical features can provide well-established clinical markers of PD through representations of the vocal tract or signs of dysphonia. Integrating these representations through a multimodal network allows the model to learn relationships between frequency-time and quantitative speech patterns.

Previous efforts using multimodal approaches have reported promising results but run high risks of overfitting, largely due to limited training data. Several multimodal studies report high performance but rely on smaller cohorts and aggressive model capacity. Studies performed in [6], [7], and [9] use relatively small datasets and do not show training or validation loss curves, leaving very little information to do with training-stability. [8] uses a dataset larger than the others (N=584), but only displays the history of the risk function on the testing data for just one cross-validation data set.

By combining multiple datasets, we mitigated the overfitting risk considerably and demonstrated stable training, as evidenced by relatively consistent validation loss curves across five independent runs. The stability underscores the robustness of the proposed architecture.

From a clinical perspective the results highlight the potential of voice-based biomarker analysis as a non-invasive supplementary tool for PD screening. The ability to collect sustained phonation recordings and analyze them automatically makes this approach particularly promising for cheap and accessible early detection. Incorporating such tools into clinical workflows could increase diagnostic accuracy with minimal effort and reduce delays in identification of PD onset, facilitating earlier intervention and improved patient outcomes.

4.1 Limitations and Future Work

Despite the promising results, several limitations must be acknowledged. First, although the dataset used in this study is larger than those in previous studies, it remains relatively small for deep learning. Though validation loss curves are relatively stable, they plateaued near epoch 7, indicating that the risk of overfitting, while reduced, has not been fully eliminated. Second, the dataset was constructed by combining recordings from multiple sources. Although this approach increased both sample size and diversity, it may also have produced unintended consequences affecting model learning. Combining datasets may have introduced confounding variability such as differences in microphone quality, ambient noise, and recording protocols. Third, the recordings were restricted to sustained /a/ vowel phonations. Though studies have shown that sustained vowel phonations are more successful in indicating the presence of PD, simple phonations may limit the model's ability to capture more nuanced and dynamic features of natural speech.

Future work should focus on expanding the dataset both in size and variety. Consolidating larger collections of voice recordings could help reduce overfitting and improve generalizability. Using techniques such as data augmentation and random transformations could help increase dataset size. To improve clinical relevance, continuous speech and additional phonation types could be included to capture subtler patterns of dysphonia and bradykinesia. Finally, external validation on independent cohorts will be essential for assessing robustness and ensuring translational potential in real-world clinical settings. With these improvements and development of workflow integration, deep learning models such as the one presented in this study shows promise in incorporation into such real-world clinical settings.

References

[1] Bloem, B. R., Okun, M. S., & Klein, C. (2021). *Parkinson's disease. The Lancet, 397*(10291), 2284-2303. https://doi.org/10.1016/S0140-6736(21)00218-X

[2] Ibarra, E. J., Arias-Londoño, J. D., Zañartu, M., & Godino-Llorente, J. I. (2023). Towards a Corpus (and Language)-Independent Screening of Parkinson's Disease from Voice and Speech Through Domain Adaptation. *Bioengineering, 10*(11), Article 1316. https://doi.org/10.3390/bioengineering10111316

[3] Marsden, C. D. (1994, June). *Parkinson's disease. Journal of Neurology, Neurosurgery & Psychiatry, 57*(6), 672-681. https://doi.org/10.1136/jnnp.57.6.672

[4] Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A. E., Halliday, G., Goetz, C. G., Gasser, T., Dubois, B., Chan, P., Bloem, B. R., Adler, C. H., & Deuschl, G. (2015, October). *MDS clinical diagnostic criteria for Parkinson's disease. Movement Disorders, 30*(12), 1591-1601. https://doi.org/10.1002/mds.26424

[5] Skodda, S., Grönheit, W., & Schlegel, U. (2012, February 28). Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease. *PLOS ONE, 7*(2), e32132. https://doi.org/10.1371/journal.pone.0032132

[6] Vásquez-Correa, J. C., Arias-Vergara, T., Orozco-Arroyave, J. R., Eskofier, B., Klucken, J., & Nöth, E. (2018). *Multimodal assessment of Parkinson's disease: A deep learning approach.* IEEE. https://doi.org/10.1109/ISBI.2018.8363564

[7] Barukab, O., Abuzaid, M., Al-Sharif, A., Ali, N., Alsharif, M., & Aslam, N. (2022). Analysis of Parkinson's Disease Using an Imbalanced Voice Dataset. *Diagnostics, 12*, Article 3000. https://doi.org/10.3390/diagnostics12103000

[8] Wang, W., Lee, J., Harrou, F., & Sun, Y. (2020, August 21). *Early detection of Parkinson's disease using deep learning and machine learning. IEEE Access.* https://doi.org/10.1109/ACCESS.2020.3016062

[9] Grover, S., Bhartia, S., Akshama, Yadav, A., & Seeja, K. R. (2018). *Predicting severity of Parkinson's disease using deep learning. Procedia Computer Science, 132*, 1788-1794. https://doi.org/10.1016/j.procs.2018.05.154

[10] Iyer, A., Kemp, A., Rahmatallah, Y., Pillai, L., Glover, A., Prior, F., Larson-Prior, L., & Virmani, T. (2023). A machine learning method to process voice samples for identification of Parkinson's disease. *Scientific Reports, 13*(1), 20615. https://doi.org/10.1038/s41598-023-47568-w

[11] Guo, Z., Li, X., Huang, H., Guo, N., & Li, Q. (2019, March). Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences, 3*(2), 162–169. https://doi.org/10.1109/TRPMS.2018.2890359

[12] Zhou, T., Ruan, S., & Canu, S. (2019). A review: Deep learning for medical image segmentation using multi-modality fusion. *Array, 3–4*, 100004. https://doi.org/10.1016/j.array.2019.100004

[13] "Voice Samples for Patients with Parkinson's Disease and Healthy Controls. (2023). Figshare. https://figshare.com/articles/dataset/Voice_Samples_for_Patients_with_Parkinson_s_Disease_and_Healthy_Controls/23849127

[14] Italian Parkinson's Voice and Speech. (2020). IEEE DataPort. https://ieee-dataport.org/open-access/italian-parkinsons-voice-and-speech

[15] Tsanas, A., Little, M., McSharry, P., & Ramig, L. (2009, October 29). Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *Nature Precedings*. https://doi.org/10.1038/npre.2009.3920.1

[16] McFee, B., et al. *Librosa: Audio and music signal processing in Python* [Software documentation]. Retrieved from https://librosa.org/doc/latest/index.html

[17] Jadoul, Y., Thompson, B., & de Boer, B. *Parselmouth documentation* [Software documentation]. Retrieved from https://parselmouth.readthedocs.io/en/stable/

[18] Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *The American Journal of Psychology, 53*(3), 329-353. https://doi.org/10.2307/1417526

[19] Tesfai, S. (2024, May 24). *Multimodal ensemble models for Parkinson's disease diagnosis using log-Mel spectrograms and acoustic features.* IEEE. https://doi.org/10.1109/URTC60662.2023.10534982

[20] Suhas, B. N., Mallela, J., Illa, A., Yamini, B. K., Atchayaram, N., & Yadav, R. (2020, August 28). *Speech task based automatic classification of ALS and Parkinson's disease and their severity using log Mel spectrograms.* IEEE. https://doi.org/10.1109/SPCOM50965.2020.9179503

[21] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. *PyTorch documentation*. https://docs.pytorch.org/docs/stable/index.html