DOI: 10.61782/fa.2025.0303

DISCRIMINATING PARKINSON'S DISEASE FROM NORMOPHONIC VOICES USING A SINCNET MODEL

J.A. Gómez-García[1]* **H. Fandiño-Toro**[2] **D. Torricelli**[1]
[1] Spanish National Research Council, Arganda del Rey, Spain
[2] Instituto Tecnológico Metropolitano, Medellín, Colombia

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative disorder that often manifests vocal symptoms, making voice and speech analysis a valuable tool for noninvasive monitoring and diagnosis. This paper investigates the use of a deep learning model, comprising a SincNet front-end coupled with an EfficientNetV2-L backbone, to discriminate between pathological voices of individuals with PD and normophonic voices in Spanish-speaking individuals in the Neurovoz database. Using an 11-fold stratified group cross-validation methodology, our model achieved a mean accuracy of 76. 08% to discriminate between PD patients and healthy controls (HC). The results demonstrate the capabilities of the Sinc network for the characterization of voice pathologies using custom filterbanks.

Keywords: Parkinson's disease, pathological speech, SincNet, learnable frontend

1. INTRODUCTION

Parkinson's disease (PD) is a progressive neurological disorder characterized by motor and nonmotor symptoms, including characteristic speech impairments known as hypokinetic dysarthria (e.g., reduced volume, monotonic pitch, imprecise articulation) [1]. The analysis of sustained vowels provides a reliable and non-invasive means to objectively assess these vocal changes and monitor the progression of PD [2].

Detecting pathological voices has traditionally involved extracting hand-crafted acoustic features, such as jitter, shimmer, harmonic-to-noise ratio (HNR), or the widespread Mel-frequency cepstral coefficients (MFCC) [3, 4]. While MFCC efficiently captures spectral data through fixed filterbanks grounded in human auditory models, this fixed structure might limit its ability to detect subtle or atypical acoustic patterns associated with particular pathologies. In contrast, learnable frontend architectures allow filterbank characteristics to be optimized directly from data during model training. This data-driven adaptability enables filters to specialize for the specific acoustic discrimination task, potentially enhancing the detection of pathological voice characteristics, such as those present in PD.

SincNet is a prominent example of learnable filterbanks. This architecture, described in [5], is designed to process raw audio. By using parameterized sinc functions in its initial convolutional layer, SincNet is designed to optimize and learn task-specific bandpass filters. This improves interpretability and has demonstrated promising results in multiple speech-related tasks. For example, its application in pathology detection, as reported in [6], showed an increase in accuracy of approximately 7% to classify various voice disorders (such as neoplasm, functional dysphonia, vocal palsy and phonotrauma) compared to controls, using sustained phonation recordings coming from the Far Eastern Memorial Hospital dataset.

With these antecedents in mind, this paper evaluates the effectiveness of SincNet to discriminate between patients affected by PD and healthy controls (HC), using sustained vowel /a/ recordings from the NeuroVoz database [7]. The proposed architecture includes features extracted by a learnable SincNet front-end and by a Convolutional Neural Network (CNN) backbone, which are then for clas-

****Corresponding author***: jorge.gomez.garcia@csic.es.*

sification PD vs HC by a subsequent multilayer perceptron (MLP) head.

2. MATERIALS AND METHODS

2.1 Dataset: NeuroVoz

The NeuroVoz dataset is used in this paper. It contains voice and speech signals from 108 native Spanish-Castilian speakers (55 HC and 53 individuals with PD). PD participants were recorded while in their ”ON” medicated state, having taken their standard medication between 2 to 4 hours before the recording sessions. The complete data set includes recordings of different speech tasks, such as sustained vowel phonations, diadochokinetic tests, 16 listen-and-repeat phrases, and a monologue. However, this paper focuses exclusively on the multiple sustained phonations of the vowel /a/ contained in the dataset. As a result, 111 recordings from HC speakers (45.5%) and 133 recordings from PD participants (54.5%) are employed.

2.2 Methodology

2.2.1 Preprocessing

All audio recordings were first resampled at a sample rate of 16 kHz. Then, each recording was normalized in amplitude to the range [-1, 1]. To handle variable recording lengths and ensure consistent input dimensionality, normalized audio signals were segmented into fixed length chunks. Non-overlapping segments of 5 seconds duration were extracted from each recording using a stride equal to the segment length. A subject-independent 11-fold stratified group cross-validation strategy was used for model evaluation. In this way, the signal chunks were partitioned so that all segments belonging to a single subject were contained entirely within a single fold, preventing data leakage between the training and validation sets within each fold iteration.

2.2.2 Model Architecture

An end-to-end deep learning model was developed, processing raw audio waveforms directly. The architecture comprises three main components: a learnable frontend, a pre-trained backbone, and a classifier head. A summary graphic of this architecture is presented in Figure 1.

- **SincNet Frontend:** The initial layer employed SincNet, a convolutional layer with 128 filters, a kernel size of 251 samples, and a stride of 160 samples. The minimum low-frequency cut-off and the minimum filter bandwidth were set to 10 Hz. The SincNet layers implementations in Speechbrain were used in this paper [8].

- **Backbone Network:** The spectral representation resulting from the convolution of the input waveforms and filterbanks trained by the SincNet layer was processed by a CNN convolutional backbone based on an EfficientNetV2-L [9] architecture through the timm library [10]. The backbone was used in the feature extraction mode, where the final classification layer of the original EfficientNetV2-L was removed. Global average pooling was applied to the backbone output to generate a fixed-size feature vector for each input segment.

- **Classifier Head:** A Multi-Layer Perceptron (MLP) served as the classification head. It consisted of one hidden layer with 256 units and a ReLU activation function, followed by a dropout layer with a rate of 0.3 for regularization. The final output layer produced logits for the two target classes (HC and PD).

The model was trained using the Cross-Entropy loss function. An AdamW optimizer [11] was used with a weight decay of 0.001. A component-specific learning rate strategy was adopted: the SincNet frontend used an initial learning rate of 0.01, the EfficientNetV2-L backbone used 0.0005 (to enable fine-tuning), and the classifier head used 0.001. These learning rates were dynamically adjusted during training using a Cosine Annealing schedule with warm restarts, annealing to a minimum learning rate of $1e-6$ for all components. Gradient clipping with a maximum norm of 1.0 was applied to stabilize the training. The models were trained for a maximum of

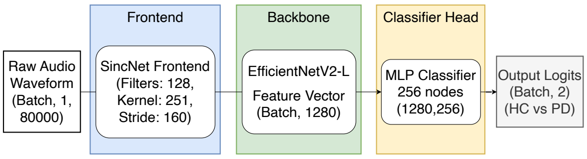

Figure 1. Deep learning strategy followed in this paper. Three layers are used for classifying between PD and HC spakers: a SincNet frontend layer, a backbone using a Efficientnet pretrained model and a MLP classifier.

150 epochs using a batch size of 16. Early stopping was implemented with a patience of 20 epochs based on the validation accuracy: training stopped if no improvement in validation accuracy was observed for 20 consecutive epochs. The model achieving the highest validation accuracy during the training within each fold was saved as the best model for that fold. All experiments were implemented using PyTorch and tracked using WandB [12]. Reproducibility was ensured by setting a fixed random seed (420) for all relevant libraries (Python random, numpy, Pytorch, Speechbrain).

3. RESULTS

The proposed audio classification model, which combines a SincNet front-end with an EfficientNetV2-L backbone, was evaluated on the Neurovoz dataset for discriminating between PD vs HC. The crossvalidated results with the average over the 11-folds are included in Table 1. Likewise, the consolidated confusion matrix of the 11-folds cross-validation is presented in Figure 2.

Table 1. Crossvalidated Performance Metrics (Mean ± SD over 11 Folds)

Metric	Value (Mean ± SD).
Accuracy (%)	76.08 ± 8.88
ROC AUC	0.66 ± 0.19
F1 Score	0.72 ± 0.19
Sensitivity	0.73 ± 0.26
Specificity	0.62 ± 0.34

4. DISCUSSION

This study investigated the efficacy of a deep learning model, comprising a SincNet frontend coupled with a pre-trained EfficientNetV2-L backbone, for the detection of Parkinson's Disease (PD) using sustained phonations of the vowel /a/ recordings from the NeuroVoz database [7]. By using an 11-fold stratified group cross-validation methodology, our model achieved a mean accuracy of 76.08% to discriminate between PD patients and Healthy Controls (HC). The performance level seem promising compared to the baseline results shown in the original NeuroVoz dataset paper [7]. In this paper, traditional machine learning methods, using the AVCA-ByO feature set [4], and employing Random Forest and Logistic Regression classifiers, reported balanced accuracies of 64% and 65%. Likewise, a deep learning baseline using Mel-spectrogram inputs with a fine-tuned ResNet-18 reached a balanced accuracy of 69%. While our main metric was overall accuracy instead of balanced accuracy, and our cross-validation strategies vary slightly, our model's performance indicates an improvement on these baselines for the specific task of sustained vowels in the NeuroVoz dataset. The potential advantage of our approach, particularly compared to the ResNet-18 baseline, may come from the SincNet frontend. As discussed in [5], SincNet's ability to learn filterbank parameters directly from the raw audio waveform allows for the optimization of filters specific to the task at hand. This contrasts with the fixed, psychoacoustically motivated Mel filterbanks used for the creation of Mel-spectrograms. It is plausible that these learnable filters capture subtle acoustic markers indicative of Parkinsonian voice quality more effectively than standard feature representations.

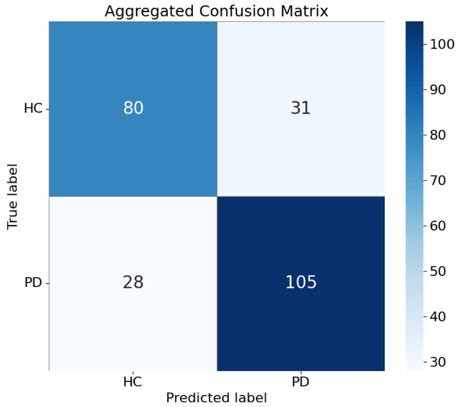

Figure 2. Crossvalidated confusion matrix aggregating individual results by each one of the folds.

Although the results are promising, several limitations must be addressed. Our analysis focused only on the sustained vowel /a/ task in NeuroVoz and did not include other speech tasks like sentence reading or monologue, which might yield different performance results in continuous speech contexts. Furthermore, the NeuroVoz dataset is limited in size and restricted to Castilian Spanish, potentially impacting the generalization of the findings. Furthermore, our chosen architecture is effective but not exhaustive; we did not investigate alternative learnable frontends such as Leaf [13] or other convolutional backbone architectures.

These limitations define different directions for future research. For instance, the current methodology could be evaluated on continuous speech from NeuroVoz to assess its robustness and performance in more complex acoustic contexts. Addressing generalizability requires validating the approach on larger, more diverse, and multilingual datasets. Furthermore, systematic exploration of alternative deep learning architectures, including different combinations of frontends and backbones, could lead to performance enhancements. Investigating the acoustic characteristics captured by the learned SincNet filters may also provide valuable information on the specific markers identified for the detection of PD. Finally, exploring the fusion of acoustic features with other relevant modalities could offer further diagnostic improvements.

In conclusion, while the presented methodology demonstrates potential, further investigation is necessary to address the limitations mentioned above, to refine its performance, confirm its robustness, and establish its suitability in voice pathology detection.

5. ACKNOWLEDGMENTS

This work is partially funded by the grant AI4HealthyAging with reference MIA.2021.M02.0007 funded by the ”Spanish ministry of Economy and Digital transformation” and by the “European Union NextGenerationEU/PRTR” and by the NEUROMARK project (PID2020-120491RA-I00) funded by MCIN/AEI/10.13039/501100011033.

6. REFERENCES

[1] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2019.

[2] L. Moro-Velazquez, J. A. Gomez-Garcia, J. D. Arias-Londoño, N. Dehak, and J. I. Godino-Llorente, “Advances in Parkinson’s disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects,” *Biomedical Signal Processing and Control*, vol. 66, p. 102418, 2021.

[3] J. A. Gómez-García, L. Moro-Velázquez, J. D. Arias-Londoño, and J. I. Godino-Llorente, “On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art,” *Biomedical Signal Processing and Control*, vol. 51, pp. 181–199, 2021.

[4] J. A. Gómez-García, L. Moro-Velázquez, J. D. Arias-Londoño, and J. I. Godino-Llorente, “On the design of automatic voice condition analysis systems. Part III: Review of acoustic modelling strategies,” *Biomedical Signal Processing and Control*, vol. 66, p. 102049, 2021.

[5] M. Ravanelli and Y. Bengio, “Speaker recognition from raw waveform with SincNet,” *arXiv preprint arXiv:1808.00158*, 2018.

[6] C. H. Hung, S. S. Wang, C. T. Wang, and S. H. Fang, “Using SincNet for learning pathological voice disorders,” *Sensors*, vol. 22, no. 17, p. 6634, 2022.

[7] J. Mendes-Laureano, J. A. Gómez-García, A. Guerrero-López, E. Luque-Buzo, J. D. Arias-Londoño, F. J. Grandas-Pérez, and J. I. Godino-Llorente, “Neurovoz: a castillian spanish corpus of parkinsonian speech,” *Scientific Data*, vol. 11, no. 1, p. 1367, 2024.

[8] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, *et al.*, “Speechbrain: A general-purpose speech toolkit,” *arXiv preprint arXiv:2106.04624*, 2021.

[9] M. Tan and Q. V. Le, “EfficientNetV2: Smaller models and faster training,” in *International Conference on Machine Learning*, pp. 10096–10106, PMLR, 2021. Also available as arXiv:2104.00298.

[10] R. Wightman, “PyTorch Image Models.” `https://github.com/rwightman/pytorch-image-models`, 2019. GitHub repository.

[11] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.

[12] L. Biewald, “Experiment tracking with Weights & Biases.” `https://www.wandb.com/`, 2020. Software available from wandb.com.

[13] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, “Leaf: A learnable frontend for audio classification,” *ICLR*, 2021.