Integrating Synchronized Dysarthria and Hypomimia Deep Patterns to Quantify Parkinson Disease

WSSFN 2025 Interim Meeting. Abstract 0168

William Omar Contreras López,[1] Brayan Valenzuela, [2] John Archila,[2] John Arevalo,[3] Fabio Martíinez.[2]
[1] International Neuromodulation Center NEMOD. Colombia.
[2] Biomedical Imaging, Vision and Learning Laboratory (BIVL2AB). Universidad Industrial de Santader. Colombia.
[3] Researcher with Machine Learning Analysis and Computer Vision (MLACV). Colombia.

Corresponding author: William Omar Contreras López email: wcontreras127@unab.edu.co

How to Cite: Contreras López WO, Valenzuela B, Archila J, Arévalo J, Martíinez F. Integrating Synchronized Dysarthria and Hypomimia Deep Patterns to Quantify Parkinson Disease: WSSFN 2025 Interim Meeting. Abstract 0168. NeuroTarget. 2025;19(2):148-9.

Abstract

Introduction: Parkinson's disease (PD) is a progressive neurological disorder that often leads to impairments in speech (dysarthria) and facial expressiveness (hypomimia). These signs are clinically relevant but often evaluated subjectively. Current methods treat these impairments separately. We propose a multimodal deep learning framework that jointly analyzes audio and video data to detect patterns associated with dysarthria and hypomimia.

Method: Facial movement and voice signals were encoded using deep neural representations to identify PD-related patterns. A retrospective study involved 14 native Colombian Spanish speakers: 7 PD patients (mean age 65 ± 4, 4M/3F) and 7 controls (mean age 61 ± 3, 2M/5F). Diagnoses were confirmed by a neurologist and partially labeled via the Hoehn and Yahr scale; 2 patients were off medication and 5 under Levodopa. Data were acquired in controlled conditions using a Nikon D3500 camera (1080p, 60 fps) with monaural audio (48 kHz). Participants performed vowel, phoneme, and word tasks to elicit PD symptoms. This is the first Spanish-speaking synchronized audio-video corpus for PD orofacial analysis. Voice recordings were transformed into Mel spectrograms. Three 2D CNN architectures (ResNet-50, VGG-16, and a custom model) were trained to classify PD vs. controls. Facial videos were processed using a compact 3D-CNN and an inflated 3D (I3D) model to extract spatiotemporal features linked to hypomimia. A late fusion strategy combined audio and video predictions via weighted linear integration. Models were evaluated using leave-one-out cross-validation across pronunciation tasks, reporting precision, recall, F1-score, accuracy, and AUC. All networks were trained with binary cross-entropy, Adam optimizer, early stopping, and a learning rate of 1 \times 10^{-5} over 25 epochs.

Results: Audio-based classification using Mel spectrograms showed that VGG-16 outperformed ResNet-50 and a baseline CNN, achieving an average AUC of 69.72%, with peak performance (73.55%) during vowel articulation. For video analysis, a custom 3D-CNN surpassed both a 2D ResNet-50 and an inflated 3D (I3D) model, effectively capturing spatiotemporal facial patterns linked to hypomimia. Multimodal fusion of VGG-16 (audio) and 3D-CNN (video) improved diagnostic accuracy across tasks, with the highest AUC (85.39%) observed for phonemes at fusion weight $\lambda = 0.5$. This approach demonstrates the complementary value of combining speech and facial cues.

Discussion: The multimodal approach effectively integrates facial and vocal biomarkers. Phoneme-based facial cues via 3D-CNN yielded superior performance (AUC = 82.59%), likely due to expressive dynamics from plosive articulations. Vowel-based audio features processed via VGG-16 achieved an AUC of 73%, benefiting from stable acoustic phases. Fusion enhanced classification across all tasks, supporting the hypothesis that synchronized motor and speech signals offer complementary diagnostic value. Complementary studies have focused on audio-only representations, achieving high accuracy but overlooking facial-vocal interplay. Our synchronized dataset contributes a valuable resource for integrated modeling. Correlations between jaw motion and respiratory control support multimodal analysis. Compared to prior fusion methods, our approach captures richer temporal dynamics and articulatory complexity.

Conclusions:The proposed framework improves PD classification by combining deep audio and visual embeddings, with phoneme articulation as the most informative task. Convex fusion outperformed unimodal models, with highest gains for phonemes (AUC = 85.39 at $\lambda = 0.5$), reinforcing the clinical relevance of synchronized motor-speech impairments. Future work will explore end-to-end models capa-

ble of learning fusion weights and temporal dependencies directly from raw inputs. Expanding the dataset to include spontaneous and emotionally expressive speech, as well as broader phonetic coverage, will enhance ecological validity. Longitudinal studies will track disease progression and therapy response, enabling personalized diagnostic tools and early intervention.

References

1. Poewe W, Seppi K, Tanner CM, Halliday GM, Brundin P, Volkmann J, et al. Parkinson disease. Nat Rev Dis Primers. 2017;3:17013. Available from: https://pubmed.ncbi.nlm.nih.gov/28332488.
2. Ricciardi L, De Angelis A, Marsili L, Faiman I, Pradhan P, Pereira EA, et al. Hypomimia in Parkinson's disease: an axial sign responsive to levodopa. Eur J Neurol. 2020;27(12):2422–9. Available from: http://dx.doi.org/10.1111/ene.14452.
3. Valenzuela B, Arevalo J, Contreras W, Martinez F. A spatio-temporal hypomimic deep descriptor to discriminate Parkinsonian patients. Annu Int Conf IEEE Eng Med Biol Soc. 2022:4192–5. Available from: https://pubmed.ncbi.nlm.nih.gov/36085867.
4. Vasquez-Correa JC, Arias-Vergara T, Orozco-Arroyave JR, Eskofier B, Klucken J, Noth E. Multimodal assessment of Parkinson's disease: A deep learning approach. IEEE J Biomed Health Inform. 2019;23(4):1618–30. Available from: https://pubmed.ncbi.nlm.nih.gov/30137018.
5. Reshma S, Chennakesavulu M, Patil SS, Lamani MR. Efficient feature fusion model withmodified bidirectional LSTM for automatic Parkinson's disease classification. Int J Inf Technol. 2024;16(6):3963–71. Available from: http://dx.doi.org/10.1007/s41870-024-01886-y