A Mixed audio-video SPD network for online classification of Parkinsonian speech patterns

John Archila[1], Antoine Manzanera[2], and Fabio Martínez[1]

[1] Biomedical Imaging, Vision and Learning Laboratory (BivL^{2}ab). Universidad Industrial de Santander (UIS), Colombia famarcar@saber.uis.edu.co

[2] U2IS/Robotics & Autonomous Systems, ENSTA Paris, Institut Polytechnique de Paris, France antoine.manzanera@ensta-paris.fr

Abstract. Parkinson's disease (PD) is a neurodegenerative disease that produces progressive motor impairments. Dysarthria (speech disorders) and hypomimia (face rigidity) are two major Parkinsonism patterns observed even at the early stages of the disease. Nonetheless, the clinical diagnosis is mainly observational and dependent on the specialists' expertise. Besides, the categorization of each of these patterns is isolated, which may lead to delayed diagnosis and misplanning of treatments. This work introduces a non-invasive multimodal strategy that integrates video and audio modalities into the online characterization of speech exercises. Subjects were invited to pronounce sustained vowels while video and audio were recorded. Then, a temporal window is run along the sequence to build online covariance matrices of synchronized face landmarks position and characteristic voice frequencies. From these temporal covariance matrices are learned Riemannian descriptors that allow to discriminate between Parkinson's and control subjects. From a study with 14 subjects, the proposed approach achieved a mean accuracy of 70% in sustained vowel pronunciation. Considering online predictions, the proposed approach evidenced a consistent accuracy of 0.77 during pronunciation of close vowels.

Keywords: Mixed audio-video SPD networks · online Parkinson's Disease prediction.

1 INTRODUCTION

Parkinson's disease (PD) is a chronic neurodegenerative disease with no cure, characterized by progressive degeneration of nerve cells, decreasing the production of dopamine, resulting in serious impairments regarding the control of movement and coordination [4]. Early motor impairments are usually manifested as dysarthria (speech affectation associated with rigidity of muscles) and hypomimia (facial expression affectation associated with movement slowness and rigidity) [16, 19]. Patients with such symptoms may experience difficulties in articulating words or changing the tone of their voice, resulting in difficult and monotonous speech. These symptoms are manifested between 7 and 11 years before the definitive diagnosis of Parkinson [5, 13]. Nowadays, these patterns are characterized only by observational tests, highly dependent on the specialist's expertise [2, 6]. Additionally, they have low sensitivity in early stages, and researchers need to spend a significant amount of time developing the skills for an adequate evaluation [17].

The main contribution of this work is a geometrical online learning method to support Parkinson classification considering multimodal sources (audio and video). Thus, characterizing dysarthria and hypomimia, the proposed approach use a set of video landmarks that, together with fundamental frequencies, form a compact covariance descriptor. From this second-order representation, geometrical learning is herein implemented to learn covariative patterns associated to the disease at different temporal intervals. The paper is structured as follows: Section 2 provides an overview of the literature on Parkinson's disease focusing on methodologies to support hypomimia and dysarthria. Section 3 describes the proposed approach integrating audio and video modalities. Section 4 presents the classification results. Section 5 discusses the advantages and limitations of the proposed online geometrical representation.

2 Related works

Communication is a fundamental daily life task, involving the coordination of multiple muscular, respiratory, and facial functions [10]. The facial expression during communication is based on the gesticulation of words, producing mouth movements and the coordination of the zygomatic muscles and the orbicular muscle. For patients affected by PD at early stages, there exist evidences of gesture limitations, which causing slowness and rigidity, known as mask face or hypomimia [19, 16]. These persons may experience difficulties in articulating words or changing the voice tone, resulting also in speech difficulties known as dysarthria. Today, there are no significant advances on the characterization of such pattern and even worst, in the combination of dysarthria and hypomimia patterns, from multimodal approaches.

The quantification of hypomimia has been previously estimated using strategies to classify single images [7, 15] or videos [12, 21, 24]. Approaches based on single-image classification consider the identification of facial landmarks whose spatial characteristics allow classification through classical machine learning methods [15] or statistical analysis [7]. Other proposed approaches have attempted to temporally characterize the most significant expressions during classification from activation maps of a 3D convolutional networks [21]. However, the retrieved activation maps only coarsely distinguish regions, so that differentiating patients and control subjects remains challenging. Recurrent networks have also been used to extract the temporal embedding to classify PD [24]. Alternatively, landmarks have been located in face to associate emotions expressions with Parkinson patterns and carry out the classification [12].

Regarding dysarthria, the frequency analysis of voice has been used as descriptor to classify patients with PD, in particular harmonic analysis and signal-to-noise parameters [1, 11]. Other works have incorporated deep learning stages using a CNN [8, 23] or recurrent architectures [14], where these architectures learned new representations based on the frequency characteristics of emotional expression [8] and vowel pronunciation [14, 23]. These computational approaches have evidenced remarked scores to classify Parkinson's disease, but their application is yet limited to operate in clinical scenarios, without complex setups of recording. Besides, to the best of our knowledge, there exists limited information about how to fuse hypomimia and dysarthria information to enhance Parkinson's representation.

3 PROPOSED APPROACH

This work introduces an online multimodal approach that fuses orofacial patterns, following an early fusion method based on covariance patterns. The covariance descriptor encodes both face landmarks trajectories and fundamental frequencies of the audio speech, aligned in intervals of time. Then, a geometrical representation is learned on the Riemannian manifold, to classify Parkinsonian patterns. The general pipeline of the proposed approach is illustrated in Figure 1.

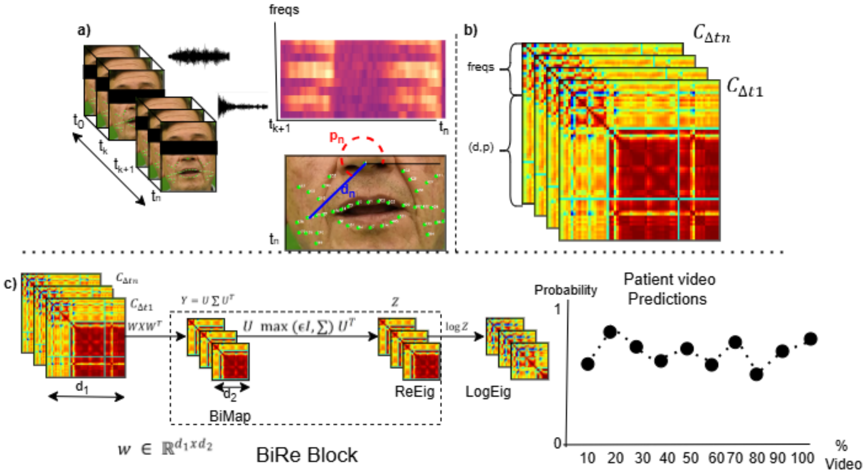

Fig. 1. Multimodal Architecture: a) The position of each key-point in polar coordinates $d(t,k)$ and $p(t,k)$ where d is the distance between the nose to the landmark and p is the angle, is combined with short time spectrogram $\sigma(t,f)$ through b) covariance matrices in time intervals $C_{\Delta t}$. c) Then, the model learns new representations more compact for quantification of PD, with the capacity to output a prediction for each video slice. Thus, this approach characterizes the patient's pronunciation temporally, by predicting the probability of PD during the vocalization (bottom right plot).

3.1 Facial and Audio low-level features

In this work, we first computed low-level features, at each sequence time, to encode dysarthria and hypomimia disorders. For dysarthria, we computed short-time spectrograms $\sigma(t, f)$ as fundamental representations (Fig. 1(a), top right), capturing the essential frequency dynamics for frequencies f over sliding window at time t. Consequently, an audio sequence is represented by a spectrogram map with dimensions $N_f \times N_t$ where N_f is the number of frequencies and N_t is the number of time samples.

Regarding hypomimia, we computed the displacement of face key points in regions around the mouth because of the association with facial muscles involved in lip expression. The MediaPipe architecture was used to compute facial landmarks using only video information [9]. We selected 44 landmarks near the mouth and muscles involved in jaw movement during pronunciation. These landmarks allow summarizing the dynamics of the subject's face during various expressions and movements. Specifically, at each time synchronised with the audio spectrogram samples, we encode the position of each keypoint in polar coordinates, using as centre the tip of the nose (Fig. 1(a), bottom right), resulting in a sequence $\{d(t, k), p(t, k)\}$ of dimensions $2N_k \times N_t$, where d is the distance between the nose to the landmark and p is the angle. N_k is the number of keypoints and N_t the number of time samples. Using the nose as centre of coordinates allow to eliminate head movements and to focus on the motion of the mouth.

3.2 Temporal Covariance Computation

Now, for each time interval Δt, made of consecutive N_t time samples, we calculate the covariance matrix of the synchronised features $\Phi(t, i)$ composed of concatenated spectrogram frequencies $\sigma(t, f)$ and face keypoints $\{d(t, k), p(t, k)\}$:

$$C_{\Delta t}(i, j) = \mathbb{E}_{\Delta t} \left(\Phi(t, i) \Phi(t, j) \right) - \mathbb{E}_{\Delta t} \Phi(t, i) \mathbb{E}_{\Delta t} \Phi(t, j)$$

where $\mathbb{E}_{\Delta t}$ refers to the expectancy calculated over the N_t samples $t \in \Delta t$. This temporal covariance matrix, with dimension $(N_f + 2N_k)^2$ (Fig. 1(b)), encodes the dynamic relationships among integrated facial and speech features, providing a comprehensive description of their temporal dependencies. This representation helps with classification performance but also results self-explainable to support recognition of coordination patterns, which is crucial for unraveling the intricate temporal interplay between facial and voice features.

3.3 Covariance-based learning for temporal video predictions.

Covariance matrices are Symmetric Positive Definite (SPD) matrices that lie in Riemannian manifolds with particular geometry, and need to be processed in a dedicated framework. For each temporal covariance $C_{\Delta t}$, we then learn a geometrical representation, capturing the inherent temporal dependencies between the different modalities. To do so, we first code a BiMap Layer following a bilinear mapping in each layer l, as: $\mathcal{C}_l = W_l\mathcal{C}_{l-1}W_l^T$, with $C_{l-1} \in \mathbb{R}_*^{d_{l-1}\times d_{l-1}}$ being the SPD matrix output of the layer $l-1$ and $W_l \in \mathbb{R}_*^{d_l\times d_{l-1}}$ the weight matrix transformation [3]. Hence, to ensure SPD property, an eigenvalue rectification layer is carried out, as: $\mathcal{C}_l = U_{l-1}\max(\varepsilon I, \varSigma_{l-1})U_{l-1}^T$ where U_{l-1} and $\varSigma_{l-1}$ are defined by the diagonal decomposition $\mathcal{C}_{l-1} = U_{l-1}\varSigma_{l-1}U_{l-1}^T$. Here, $\varepsilon > 0$ is a rectification threshold value, I is the identity matrix and $\varSigma_{l-1}$ the diagonal matrix of the eigenvalues of $\mathcal{C}_{l-1}$. This operation adjusts the eigenvalues, avoiding negative values and improving discriminative performance. This specialized block facilitates the extraction of relevant information from the input data, contributing to the computation of effective covariation patterns.

Finally, to carry out the classification task, the learned matrix is projected onto a tangent plane (i.e. back to a Euclidean space), following a logarithm map $\mathbf{log}(\mathcal{C}) = U\log(\varSigma)U^T$. Then, classical dense layers are implemented to achieve the classification of the multimodal pronunciation exercise input.

3.4 Dataset description

This study involved 14 participants, consisting of 7 patients diagnosed with Parkinson's disease (PD) and 7 control patients. The PD group had an average age of 65 ± 4, while the control group had an average age of 61 ± 3. All PD patients were on (Levodopa) medication during data acquisition. Informed consent was obtained from each participant, and the study was approved by the ethics committee of the Universidad Industrial de Santander. The dataset captured synchronized audio and video modalities, with participants performing sustained vowel pronunciation used in the clinical routine. All recordings were conducted in the same environment using a Nikon D3500 digital camera with an integrated monaural microphone. Video was recorded at 1080p resolution and 60 fps, focusing on the face region, while audio was captured at a sampling rate of 48 kHz. Phonation patterns included the pronunciation of five vowels, each vowel being repeated three times, providing a comprehensive dataset for phonation and articulatory analysis. In the study, participants are asked to sustain the pronunciation of vowels for about 5 seconds. This exercise is incorporated into clinical routines to detect voice abnormalities and to observe the facial expressions of the individuals.

4 Evaluation and Results

The proposed approach was validated with the oral task of sustained vowels, which allows the identification of voice impairments such as dysarthria during PD diagnosis, but also to peculiar conditions such as strengthening vocal muscles and motor coordination during rehabilitation therapies. For validation was followed leave-one-patient-out cross-validation, where at each iteration, one patient is left out for testing and the remaining ones (13 subjects in our experiment) are used for training. To evaluate the performance of the multimodal prediction, the implemented model configurations were assessed for the sensitivity, specificity, accuracy, precision, and F1-score per video. A video was considered correctly predicted by majority vote of its temporal predictions. Specifically, table metrics were quantified by considering either 5, or 10 or 15 predictions during each video.

A first validation was carried out to establish the best video representation to classify PD according to hypomimia-encoded patterns. In this experiment, the temporal covariance matrices were built from landmarks information using only phase (dimension of 44×44), only distance (dimension of 44×44), and integrating both variables (dimension of 88×88). These experiments were also evaluated in different temporal intervals, by evenly dividing the video in five, ten, and fifteen slices respectively. Table 1 summarizes the achieved results, reporting the best performance with the covariance descriptor using only phase information. These results highlight a high sensitivity of 78%, with an accuracy of 65%, evidencing a capability to capture motor coordination changes, especially with 10 slices per video.

Table 1. Hypomimia video classification (facial expression alone) with different number of video slices and polar coordinates of landmarks.

Facial Features	**Predictions per video**	**Ac**	**Pr**	**Sen**	**Spec**	**F1-s**
Phase	5	0.5	0.5	0.69	0.3	0.58
	10	**0.65**	**0.62**	**0.78**	**0.52**	**0.69**
	15	0.4	0.4	0.41	0.39	0.41
Distance	5	0.59	0.58	0.64	0.54	0.61
	10	0.56	0.55	0.66	0.46	0.6
	15	0.55	0.55	0.56	0.53	0.56
Phase and Distance	5	0.58	0.58	0.56	0.6	0.57
	10	0.57	0.57	0.58	0.56	0.57
	15	0.41	0.4	0.39	0.43	0.4

In a second evaluation the audio branch was assessed concerning its capability to classify dysarthria patterns from temporal covariance matrices of spectrograms only, with 20 and 50 frequency bands. Each configuration was also evaluated with five, ten, and fifteen slices per video. Table 2 summarizes the achieved results, reporting a better score with the configuration of 20 frequencies and ten slices (sensitivity of 64%). The improvement in results with 20 frequencies in sustained vowel pronunciation could be attributed to a higher generalization capacity or efficiency in representing relevant features for detecting individuals with Parkinson. It is possible that the learning covariance model can extract more discriminative information with fewer dimensions, facilitating the identification of distinctive patterns in the case of 20 frequencies.

Then, in a third experiment, the proposed approach was evaluated by fusing vocal spectrogram frequencies with facial landmark phases and distances. In such cases, it was considered 20 frequency bands for audio, and whole facial configura-

Table 2. Dysarthria Audio classification with different frequencies and different number of video slices

Freqs	**Predictions per video**	**Ac**	**Pr**	**Sen**	**Spec**	**f1-s**
20	5	0.52	0.52	0.6	0.45	0.55
	10	**0.62**	**0.61**	**0.64**	**0.6**	**0.62**
	15	0.57	0.57	0.58	0.56	0.57
50	5	0.54	0.54	0.5	0.57	0.52
	10	0.55	0.55	0.51	0.58	0.53
	15	0.53	0.54	0.5	0.56	0.52

Table 3. Multimodal (audi-video) classification with 20 speech frequencies, phase and distance facial features

Fusion Features	**Predictions per video**	**Ac**	**Pr**	**Sen**	**Spec**	**f1-s**
20 freqs, phase	5	0.44	0.44	0.44	0.45	0.44
	10	0.66	0.65	0.65	0.65	0.66
	15	0.65	0.64	0.64	0.62	0.67
20 freqs, Distance	5	0.6	0.59	0.69	0.56	0.61
	10	0.58	0.61	0.61	0.58	0.59
	15	0.64	0.63	0.63	0.6	0.65
20 freqs, Distance, Phase	5	0.58	0.58	0.58	0.54	0.6
	10	**0.70**	**0.69**	**0.73**	**0.68**	**0.71**
	15	0.62	0.62	0.62	0.64	0.61

tions. Table 3 summarizes the achieved results with multimodal configurations, being the best performance achieved in the third experiment, where vocal frequencies were fused with both facial landmark phase and distance, improving accuracy to 70% (10 intervals). These results highlight the complementarity and synergy of features extracted from both modalities. Also, the temporal interval of ten frames shows an appropriate trade-off to capture pronunciation dynamics and avoiding excessive fragmentation of the task.

5 Discussion and conclusive remarks

This work introduced an online multimodal approximation to classify Parkinson disease from facial expression (hypomimia) and voice patterns (dysarthria). In the literature there exist evidences that dysarthria, through the pronunciation of sustained vowels, can identify speech difficulties, associated with early Parkinson's disease [18]. Additionally, treatments have been proposed that use vowel pronunciation in the attempt to improve these impairments [22]. Considering that, this work reported a multimodal approach that integrates visual and audio information to recover hypomimia and dysarthria-associated patterns. For doing so, the proposed approach captured face landmarks in video, and coded spectrograms from audio, which are integrated into temporal covariance descriptors, allowing to obtain a representation of bimodal vocalization. Then, this temporal covariance embedding is projected to a geometrical deep architecture to obtain a refined second-order representation with the ability to distinguish Parkinson patterns from control signals. Thanks to the sliding nature of time covariance descriptors, the geometrical net can bring a prediction at each time interval, allowing to detect abnormal patterns associated to PD, during the exercise, in clinical routine.

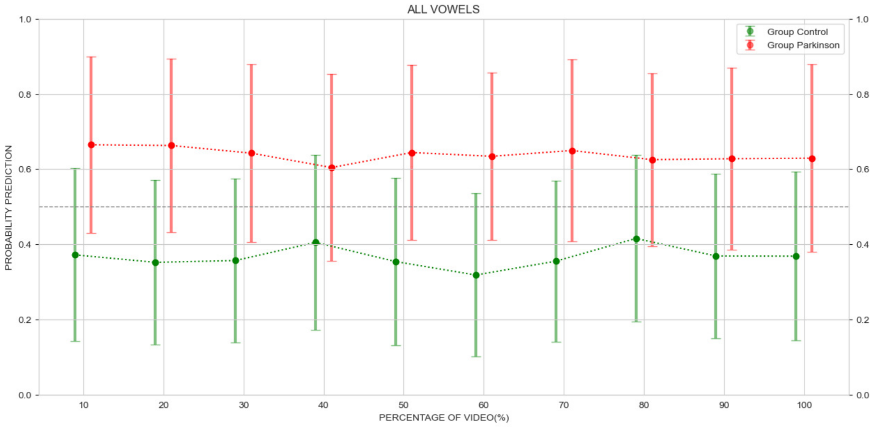

Fig. 2. Probability prediction per interval of video (red line and green line), for all vowels

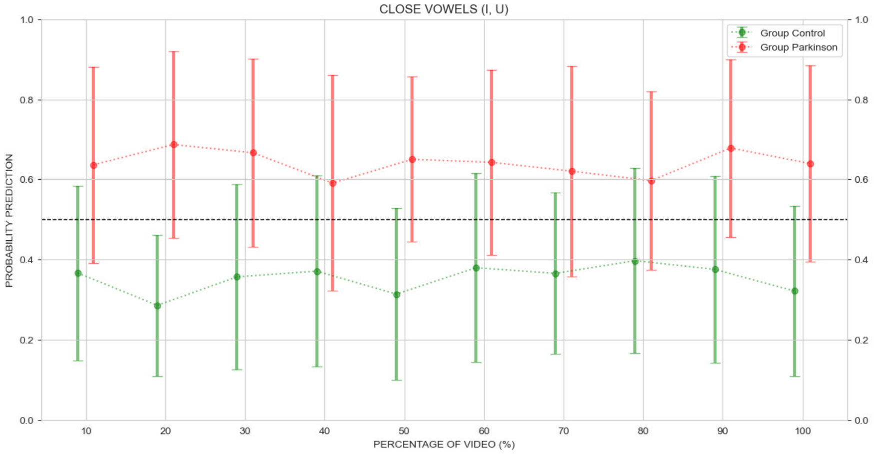

Fig. 3. Probability per interval of video (red line and green line), for close vowels

The proposed geometrical representation was validated with respect to isolated video and audio patterns, and also with the integration of both modalities. Using only videos, the proposed approach encodes temporal covariance matrices using only the correlation among face landmarks. In such case, the proposed approach achieved 65% of accuracy, a f1-score of 69%, and a total of 4 Parkinson and 5 Control subjects were correctly classified. The mistakes in classification may be partially associated to instability of landmarks and recording conditions,

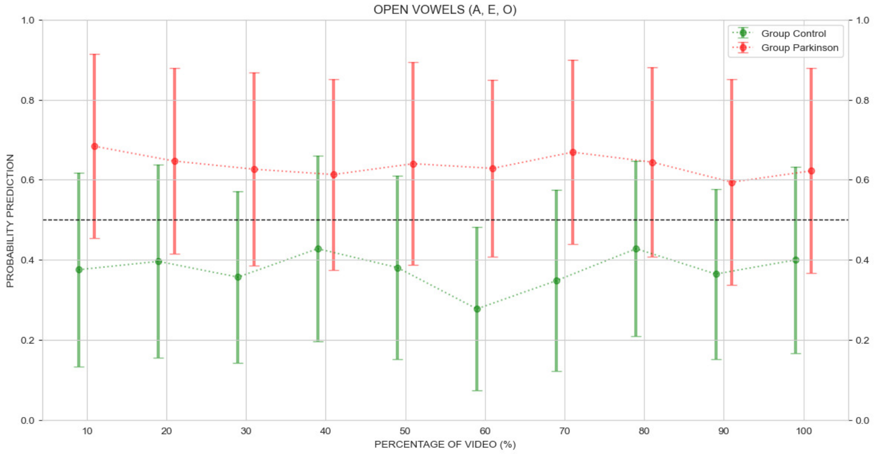

Fig. 4. Probability per interval of video (red line and green line), for open vowels

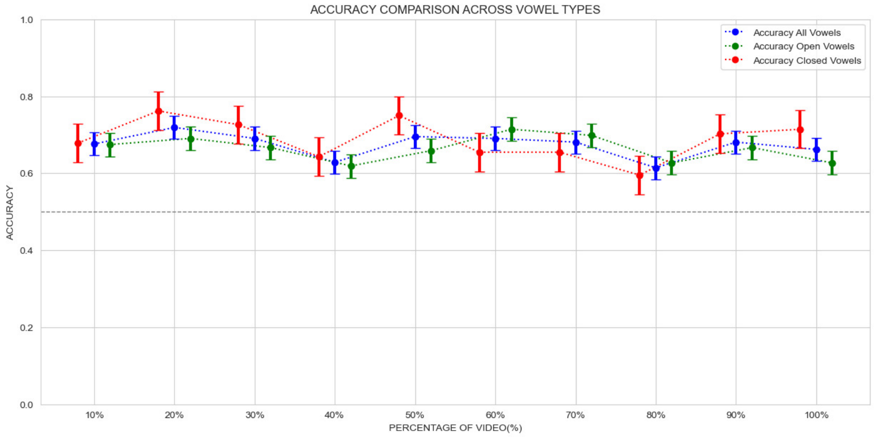

Fig. 5. Accuracy per interval of video for close vowels (red line), open vowels (green line) and all vowels (blue line).

but also to the limitation of visual information alone to determine Parkinsonian patterns.

Regarding, an audio geometrical net, trained using only spectrogram voice information, was obtained an accuracy of 62% and a f1-score of 62%. These scores were achieved from a configuration of 20 frequencies an 10 intervals per video. Then, we conducted multimodal experiments using a geometrical net, learning from covariance matrices encoding the two modalities. In such case, the multimodal approximation has a gain of 5% and 8% in accuracy, and a gain of 2% and 9% in f1-score.

Besides, the probability for the multimodal approach was calculated for patients and control subjects, for each video percentage during the sustained vowel pronunciation (see Figure 2). The performance remains stable for both Parkinson's and control groups, suggesting that all vocalization phases can yield similar predictions. Figure 4 (resp. Figure 3) shows the probability predictions and accuracy for the pronunciation of open vowels, in Spanish: A, E and O (resp. closed vowels: I and U), for Control and Parkinson groups at each interval per video. Interestingly, this categorization is related to movement: Closed vowels are produced with minimal mouth cavity amplitude, while open vowels involve greater mouth cavity expansion with the tongue positioned low. The Figure 3 of the closed vowels shows greater consistency in the control groups, maintaining the average probability and its stable variability. As for the Parkinson group, higher and more variable results were observed during the initial pronunciation of closed vowels. Similarly, in Figure 4 of the open vowels, the Parkinson group presents greater variability in the initial intervals. But in contrast to the group of closed vowels for the Control group, the best-predicted values (closer to zero) are found in intermediate pronunciation stages. These results show different dynamics for each vowel group in patients and control subjects. The pronunciation is divided into three phases: initial, stabilization, and decay [20]. Figure 5 indicate in the initial phase (predictions at 10%, 20%, and 30%), there is significant effort, with pronounced facial muscle movements. The most discriminative predictions in this phase are 20% considering all vowels (blue line) with a mean accuracy of 72%. The stabilization phase (predictions from 40% to 70%) represents the maximum vocal production stability, with constant acoustic characteristics and minimal facial movement. The most discriminative intervals here are at 50% of videos with a mean accuracy of 70% (blue line). Finally, the decay phase (predictions at 80%, 90%, and 100%) shows a decline in vocal production and increased facial movement until the mouth closes. The most discriminative intervals in this phase are at 90% of video with a mean accuracy of 68% (blue line).

The red and green line indicate that accuracy trends for both open and closed vowels remain relatively stable across video percentages, suggesting that prediction variability does not significantly change, indicating robustness in results. For control subjects, the most discriminative percentages are 20% for closed vowels (red line) with a mean accuracy of 76% (initial stage) and 60% for open vowels (green line) with a mean accuracy of 72% (stabilization stage). Future works will include the analysis of enriched representations with other input modalities, as well as an investigation toward an end-to-end processing of the complete information, since vowels are versatile and can combine with a variety of consonants to create a wide range of sounds and words. Also, this study will be extended to other voice instructions to explore the capabilities of the proposed approach.

ACKNOWLEDGMENT

To the Ministry of Science, Technology and Innovation of Colombia, through the project: *Caracterización de movimientos anormales del Parkinson desde patrones oculomotores, de marcha y enfoques multimodales basados en visión computacional* with code 92694.

References

1. Ahmed, I., Aljahdali, S., Khan, M.S., Kaddoura, S.: Classification of parkinson disease based on patient's voice signal using machine learning. Intelligent Automation and Soft Computing **32**(2), 705 (2022)
2. Alegre-Ayala, J., et al.: The impact of parkinson's disease severity on performance of activities of daily living: an observational study. Revista de Neurología **76**(8), 249 (2023)
3. Bronstein, M.M., et al.: Geometric deep learning: going beyond Euclidean data. IEEE Signal Processing Magazine **34**(4), 18–42 (2017)
4. Feigin, V.L., et al.: Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the global burden of disease study 2016. The Lancet Neurology **18**(5), 459–480 (2019)
5. Fereshtehnejad, S.M., et al.: Evolution of prodromal parkinson's disease and dementia with lewy bodies: a prospective study. Brain **142**(7), 2051–2067 (2019)
6. Friedman, J.H.: Misperceptions and parkinson's disease. Journal of the Neurological Sciences **374**, 42–46 (2017)
7. Grammatikopoulou, A., Grammalidis, N., Bostantjopoulou, S., Katsarou, Z.: Detecting hypomimia symptoms by selfie photo analysis: for early parkinson disease detection. In: Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments. pp. 517–522 (2019)
8. Khan, H., Ullah, M., Al-Machot, F., Cheikh, F.A., Sajjad, M.: Deep learning based speech emotion recognition for parkinson patient. Electronic Imaging **35**, 298–1 (2023)
9. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
10. Miller, N., Noble, E., Jones, D., Burn, D.: Life with communication changes in parkinson's disease. Age and ageing **35**(3), 235–239 (2006)
11. Nayak, S.S., Darji, A.D., Shah, P.K.: Identification of parkinson's disease from speech signal using machine learning approach. International Journal of Speech Technology **26**(4), 981–990 (2023)
12. Pegolo, E., et al.: Quantitative evaluation of hypomimia in parkinson's disease: A face tracking approach. Sensors **22**(4), 1358 (2022)

13. Postuma, R.B., et al.: How does parkinsonism start? prodromal parkinsonism motor changes in idiopathic rem sleep behaviour disorder. Brain **135**(6), 1860–1870 (2012)
14. Quan, C., Ren, K., Luo, Z.: A deep learning based method for parkinson's disease detection using dynamic features of speech. IEEE Access **9**, 10239–10252 (2021)
15. Rajnoha, M., Mekyska, J., Burget, R., Eliasova, I., Kostalova, M., Rektorova, I.: Towards identification of hypomimia in parkinson's disease based on face recognition methods. In: 2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). pp. 1–4. IEEE (2018)
16. Ricciardi, L., De Angelis, A., et al.: Hypomimia in parkinson's disease: an axial sign responsive to levodopa. European Journal of Neurology **27**(12), 2422–2429 (2020)
17. Rissardo, J.P., et al.: Parkinson's disease rating scales: A literature review. Annals of Movement Disorders **3**(1), 3–22 (2020)
18. Roland, V., Huet, K., Harmegnies, B., Piccaluga, M., Verhaegen, C., Delvaux, V.: Vowel production: a potential speech biomarker for early detection of dysarthria in parkinson's disease. Frontiers in Psychology **14**, 1129830 (2023)
19. Rusz, J., et al.: Distinct patterns of speech disorder in early-onset and late-onset de-novo parkinson's disease. npj Parkinson's Disease **7**(1), 98 (2021)
20. Tripathi, K., Rao, K.S.: Robust vowel region detection method for multimode speech. Multimedia Tools and Applications **80**(9), 13615–13637 (2021)
21. Valenzuela, B., et al.: A spatio-temporal hypomimic deep descriptor to discriminate parkinsonian patients. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 4192–4195. IEEE (2022)
22. Wight, S., Miller, N.: Lee silverman voice treatment for people with parkinson's: audit of outcomes in a routine clinic. International journal of language & communication disorders **50**(2), 215–225 (2015)
23. Wodzinski, M., et al.: Deep learning approach to parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). pp. 717–720. IEEE (2019)
24. Xu, Z., Lv, D., Li, H., Li, H., Gao, H.: Application of reslstm in hypomimia video detection for parkinson's disease. In: 2023 International Conference on New Trends in Computational Intelligence (NTCI). vol. 1, pp. 243–247. IEEE (2023)