CLASIFICACIÓN DE PATRONES PARKINSONIANOS INTEGRANDO INFORMACIÓN GESTO-AUDITIVA POR MEDIO DE UNA ESTRATEGIA MULTIMODAL

JOSÉ DANIEL VALERA SÁNCHEZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2025

CLASIFICACIÓN DE PATRONES PARKINSONIANOS INTEGRANDO INFORMACIÓN GESTO-AUDITIVA POR MEDIO DE UNA ESTRATEGIA MULTIMODAL

JOSÉ DANIEL VALERA SÁNCHEZ

Trabajo de Grado presentado en cumplimiento de los requisitos para optar al título de:
Ingeniero de Sistemas

Director:
Fabio Martínez Carrillo
Doctor en Ingeniería de Sistemas y Computación

Codirectora:
Alejandra Moreno Tarazona
Magíster en Ingeniería de Sistemas e Informática

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2025

AGRADECIMIENTOS

Primeramente, agradezco a Dios por guiar mi camino y brindarme la fuerza necesaria para superar cada reto y continuar adelante.

Quiero expresar mi profundo agradecimiento a mi director, el profesor Fabio Martínez, por su constante guía, el conocimiento compartido y, sobre todo, el tiempo dedicado. Son cosas de las cuales siempre estaré agradecido. Gracias también por haberme abierto las puertas al grupo de investigación BIVL^{2}ab, donde tuve la oportunidad de aprender y rodearme de personas excepcionales, cuyos recuerdos llevaré conmigo por siempre. Agradezco de manera muy especial a mi codirectora, Alejandra Moreno, por su apoyo constante, sus consejos y su oportuna retroalimentación, que fueron fundamentales para el desarrollo de este trabajo. Siempre le tendré un gran cariño, pero principalmente una admiración enorme. Muchas gracias por todo, Aleja.

A mis compañeros y amigos del grupo, les agradezco por todo el cariño mostrado y por hacerme sentir en familia. Aprendí muchísimo de ustedes y espero haberles dejado un grato recuerdo.

A mis padres, a quienes siempre dedicaré todos mis logros, todo lo hago por ustedes. Las palabras no me alcanzan para agradecer todo el esfuerzo que han puesto para mi formación y bienestar. Mi mayor satisfacción es verlos felices. Agradezco su sabiduría y el haberme apoyado en este camino. Siempre han sido y serán mi ejemplo a seguir. De igual manera, agradezco a toda mi familia por siempre demostrarme su amor y su gran apoyo. Gracias tías, tíos, hermano, hermanas, primas, primos, abuela, sobrinos, sobrinas, cuñada y cuñado. Y a mi querido perro Cotri, quien vino a este mundo con un propósito y fue llenar nuestro hogar de alegría, por siempre te recordaré.

A mi novia, Laura Méndez, quien me acompañó en todo este camino, de principio a fin. Has sido mi confidente y mi apoyo incondicional. Gracias por tu amor y por nunca dejarme caer, lo logramos. También agradezco a tu familia por brindarme tanto cariño y hacerme sentir como uno más.

A mis amigos, todos tienen un lugar especial en mi corazón, no importa la distancia ni el lugar donde nos encontremos, sé que siempre podré contar con ustedes.

Finalmente, a la Universidad Industrial de Santander y a la Escuela de Ingeniería de Sistemas por brindarme la formación académica y profesional que me ha permitido alcanzar esta meta.

CONTENIDO

pág.

LISTA DE FIGURAS

pág.

LISTA DE TABLAS

pág.

RESUMEN

TÍTULO: Clasificación de patrones parkinsonianos integrando información gesto-auditiva por medio de una estrategia multimodal [*]

AUTOR: José Daniel Valera Sánchez [**]

PALABRAS CLAVE: Enfermedad de Parkinson, clasificación, metodologías convolucionales, mecanismo de auto-atención, representaciones discriminativas, información audiovisual.

DESCRIPCIÓN: La enfermedad de Parkinson se caracteriza por una degeneración progresiva del sistema nervioso, afectando las neuronas dopaminérgicas. Es la segunda enfermedad neurodegenerativa más prevalente a nivel mundial. A nivel global, se estiman entre 5 y 35 nuevos casos por cada $100,000$ individuos, con una prevalencia del 3% en personas mayores de 80 años. En Colombia, en 2016, se estimaron $26,000$ casos y se reportaron 800 muertes asociadas a esta enfermedad. Entre los síntomas más característicos se encuentran desórdenes del habla, como disminución del volumen de voz, mala articulación y falta de inflexión tonal, así como la hipomimia facial, afectando la calidad de vida del paciente. Debido al carácter multisintomático de esta enfermedad, es fundamental desarrollar esquemas de diagnóstico multimodales que integren patrones motores y de habla para mejorar su detección y tratamiento. Este trabajo propuso un enfoque basado en una red de auto-atención multimodal para analizar datos audiovisuales de pacientes con Parkinson y sujetos control. Las señales de audio y video fueron procesadas mediante arquitecturas profundas diseñadas para extraer características relevantes de cada modalidad. Posteriormente, estas representaciones fueron integradas mediante un mecanismo de auto-atención para capturar relaciones internas entre modalidades. La red fue ajustada durante una tarea de clasificación binaria (Control vs. Parkinson) utilizando fonemas, vocales sostenidas y palabras como ejercicios evaluativos. Los resultados obtenidos fueron competitivos, alcanzando una precisión de 74.19%, *recall* de 73.02% y un *AUC* de 75.26% para fonemas. Para vocales sostenidas, el modelo alcanzó una precisión de 65.19%, *recall* de 83.81% y un *AUC* de 70.78%, demostrando la efectividad del método en la discriminación de patrones relacionados con la enfermedad de Parkinson.

---

[*] Trabajo de investigación

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez, PhD en ingeniería de sistemas y computación, análisis de imágenes y análisis de vídeo. Codirectora: Alejandra Moreno Tarazona, Magíster en Ingeniería de Sistemas e Informática

ABSTRACT

TITLE: Parkinsonian patterns classification integrating gest-auditive information using a multimodal strategy [*]

AUTHOR: José Daniel Valera Sánchez [**]

KEYWORDS: Parkinson's disease, classification, convolutional neural networks, self-attention mechanism, discriminative representations, audiovisual information.

DESCRIPTION: Parkinson's disease is characterized by a progressive degeneration of the nervous system that primarily affects the neurons responsible for producing dopamine. It is currently recognized as the second most prevalent neurodegenerative condition worldwide. Globally, an estimated 5 to 35 new cases per $100,000$ individuals occur each year, with a prevalence of 3% in individuals over the age of 80. In Colombia, by 2016, there were $26,000$ reported cases, with 800 deaths attributed to this disease. Among its most characteristic symptoms are speech disorders, including reduced voice volume, poor articulation, and lack of tonal inflection, as well as facial hypomimia, which significantly impacts patients' quality of life. Due to the multisymptomatic nature of this disease, it is essential to develop multimodal diagnostic schemes that integrate motor and speech patterns to enhance detection and treatment strategies. This work proposed a multimodal self-attention-based approach to analyze audiovisual data from Parkinson's patients and control subjects. Audio and video signals were represented using deep learning architectures specifically designed to extract relevant features from each modality. Subsequently, these representations were integrated using a self-attention mechanism to capture internal relationships and relevant patterns between modalities. The network was trained for a binary classification task (Control vs. Parkinson) using phonemes, sustained vowels, and words as evaluative exercises. The obtained results were competitive, achieving an accuracy of 74.19%, a recall value of 73.02%, and an AUC of 75.26% for phonemes. For sustained vowels, the model reached an accuracy of 65.19%, a recall value of 83.81%, and an AUC of 70.78%, demonstrating the effectiveness of the proposed method in discriminating patterns related to Parkinson's disease.

---

[*] Research work

[**] Faculty of Physics-Mechanics Engineering. School of Systems Engineering and Informatics. Advisor: Fabio Martínez Carrillo, PhD. Computer and systems engineering, medical image analysis and video analysis. Co-advisor: Alejandra Moreno Tarazona, Master in Systems and Informatics Engineering.

INTRODUCCIÓN

El Parkinson es conocido como la segunda enfermedad neurodegenerativa con mayor prevalencia después del Alzheimer, con una incidencia significativa en personas mayores a los 65 años[1]. Entre 1990 y 2019, la prevalencia mundial se duplicó, alcanzando los $8,5$ millones de casos[1] y se espera un aumento para el año 2040, estimando que aproximadamente $14,2$ millones de personas sufrirán de Parkinson[2]. En Colombia, estas cifras no son atípicas, dado que durante el período 2016 y 2020 se atendieron un total de 148.224 personas con diagnóstico de Parkinson, de los cuales 33.687 fueron atendidas en el año 2020 [3].

La enfermedad del Parkinson está asociada con la pérdida de células en el cerebro que producen dopamina, la cual es un neurotransmisor encargado de controlar las funciones motrices y el movimiento, provocando síntomas como temblores, movimiento pausado y rígido, y pérdida del equilibrio[4]. Dichos síntomas empeoran a medida que el tiempo avanza dado el carácter progresivo de la enfermedad. Aún más importante, actualmente, no existe cura y las causas del Parkinson permanecen desconocidas[4].

La hipomimia es uno de los síntomas más representativos del Parkinson, estando presente en el 70% de los pacientes. La hipomimia consiste en una marcada disminución de los

---

[1] Zejin OU et al. “Global trends in the incidence, prevalence, and years lived with disability of Parkinson’s disease in 204 countries/territories from 1990 to 2019”. In: *Frontiers in public health* 9 (2021), p. 776847.

[2] Michelle Hyczy S TOSIN et al. “Nursing and Parkinson’s disease: a scoping review of worldwide studies”. In: *Clinical Nursing Research* 31.2 (2022), pp. 230–238.

[3] MINISTERIO DE SALUD DE COLOMBIA. *Día Mundial del Parkinson: Colombia se destaca en atención*. Último acceso: 17 de febrero de 2025. 2020. URL: https://www.minsalud.gov.co/Paginas/Dia-Mundial-del-Parkinson-Colombia-se-destaca-en-atencion.aspx.

[4] Suzy L WONG; Heather Lynne GILMOUR, and Pamela L RAMAGE-MORIN. *Parkinson’s disease: Prevalence, diagnosis and impact*. 2014.

gestos expresivos del rostro [5]. Sumado a esto, entre un 85% y 90% de los pacientes que padecen Parkinson desarrollan desórdenes en la comunicación, caracterizados por cambios en el volumen de la voz (hipofonía), disminución de la articulación (disartria), y pérdida de inflexión en el tono (aprosodia) [6]. Las áreas neurológicas donde se originan la hipomimia y la disartria están correlacionadas significativamente y se activan durante la comunicación verbal [65]. Estos trastornos de expresión facial y habla ocasionan diversas consecuencias, afectando de manera considerable la calidad de vida del paciente [7]. De hecho, ambos síntomas forman parte de los primeros estadíos de la enfermedad, por lo cual su detección temprana podría potencialmente mejorar el resultado del tratamiento[8]. No obstante, los protocolos clínicos para la evaluación del Parkinson no incluyen la caracterización de estos síntomas y son empleados en etapas avanzadas de la enfermedad, cuando los síntomas motores empiezan a ser bastante evidentes[9]. Dichos protocolos suelen ser inaccesibles para los pacientes con problemas de movilidad, barreras geográficas, o costos asociados. Sumado a esto, se sigue dependiendo en gran medida de la opinión de los expertos, diagnosticando erróneamente a 2 de cada 10 pacientes[10]. Además, dichos protocolos se basan en un análisis no correlacionado de los síntomas, analizándolos de

---

[5] Teresa MAYCAS-CEPEDA et al. "Hypomimia in Parkinson's disease: what is it telling us?" In: *Frontiers in Neurology* 11 (2021), p. 603582.

[6] Michael D MCCLEAN and Stephen M TASKO. "Association of orofacial with laryngeal and respiratory motor output during speech". In: *Experimental brain research* 146 (2002), pp. 481–489.

[7] Margaret PRENGER et al. "Social symptoms of Parkinson's disease". In: *Parkinson's Disease* 2020 (2020).

[8] Sigurlaug SVEINBJORNSDOTTIR. "The clinical symptoms of Parkinson's disease". In: *Journal of neurochemistry* 139 (2016), pp. 318–324.

[9] Fundación de PARKINSON. *Sitio web de la Fundación de Parkinson*. Fecha de acceso: 1 de agosto de 2023. Sin fecha. URL: https://www.parkinson.org/espanol.

[10] Giovanni RIZZO et al. "Accuracy of clinical diagnosis of Parkinson disease: a systematic review and meta-analysis". In: *Neurology* 86.6 (2016), pp. 566–576.

manera individual y causando pérdida de información fundamental para diagnosticar la enfermedad[11].

Diversas metodologías computacionales han emergido en el estado del arte para caracterizar estos patrones parkinsonianos logrando discernir entre características de pacientes control. Específicamente, para los desórdenes del habla existen enfoques que a partir de grabaciones de audio permiten analizar y clasificar características como el tono, la frecuencia, la intensidad, la fluctuación[12] [13]. Otros enfoques usan el *Relevance Vector Machine (RVM)* para realizar dicha clasificación, mediante el análisis de la amplitud y periodicidad del discurso[14]. De igual manera, existen otras metodologías que logran caracterizar la disminución de gestos expresivos en el rostro permitiendo encontrar patrones parkinsonianos, usando representaciones convolucionales *(CNN)* [15] [16] [17]. Sin embargo, el análisis individual de estos síntomas desaprovecha la correlación de la información

---

[11] Movement Disorder Society Task Force on RATING SCALES FOR PARKINSON'S DISEASE. "The unified Parkinson's disease rating scale (UPDRS): status and recommendations". In: *Movement Disorders* 18.7 (2003), pp. 738–750.

[12] Timothy J WROGE et al. "Parkinson's disease diagnosis using machine learning and voice". In: *2018 IEEE signal processing in medicine and biology symposium (SPMB)*. IEEE. 2018, pp. 1–7.

[13] John M TRACY et al. "Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease". In: *Journal of biomedical informatics* 104 (2020), p. 103362.

[14] María GOÑI et al. "Smartphone-based digital biomarkers for Parkinson's disease in a remotely-administered setting". In: *IEEE access* 10 (2022), pp. 28361–28384.

[15] Jacek JAKUBOWSKI et al. "A study on the possible diagnosis of Parkinson's disease on the basis of facial image analysis". In: *Electronics* 10.22 (2021), p. 2832.

[16] Bo JIN et al. "Diagnosing Parkinson disease through facial expression recognition: video analysis". In: *Journal of medical Internet research* 22.7 (2020), e18697.

[17] Bhakti SONAWANE and Priyanka SHARMA. "Review of automated emotion-based quantification of facial expression in Parkinson's patients". In: *The Visual Computer* 37 (2021), pp. 1151–1167.

audiovisual, para realizar una clasificación de la enfermedad[18] [19].

En este trabajo se introduce una arquitectura multimodal que aprovecha secuencias sincronizadas donde se registró tanto en audio como en video a los pacientes mientras realizaban ejercicios de pronunciación, entonación y articulación. El método propuesto incluye una representación profunda que integra un mecanismo de atención para aprender patrones conjuntos no lineales y no locales, que permiten la discriminación entre pacientes diagnosticados con Parkinson y sujetos control.

---

[18] Wee Shin LIM et al. “An integrated biometric voice and facial features for early detection of Parkinson’s disease”. In: *npj Parkinson’s Disease* 8.1 (2022), p. 145.

[19] Justyna SKIBIŃSKA and Jiri HOSEK. “Computerized analysis of hypomimia and hypokinetic dysarthria for improved diagnosis of Parkinson’s disease”. In: *Heliyon* 9.11 (2023).

1. FUNDAMENTOS Y TRABAJOS PREVIOS

1.1. Desórdenes orofaciales y de comunicación en el Parkinson.

El Parkinson es una enfermedad neurodegenerativa, producida por la deficiencia de dopamina, conduciendo a un desorden del movimiento caracterizado por diversos síntomas, tanto motores (rigidez, lentitud, temblores, entre otros), como no motores (dificultad para dormir, depresión, entre otros)[20]. A pesar de que los desórdenes motores son típicamente asociados al Parkinson, estas características suelen manifestarse de manera notoria entre 5 y 10 años después del diagnóstico. En cambio, existe evidencia científica que indica que los desórdenes faciales y de comunicación se presentan en las primeras etapas de la enfermedad (aproximadamente entre 1 y 5 años)[9].

De hecho, la literatura ha reportado una notable correlación entre las áreas neurológicas donde se producen los desórdenes faciales y de comunicación durante la coordinación motora del habla. Los anteriores hallazgos experimentales han sido soportados al evidenciar un vínculo neuronal de los músculos orofaciales (presentes en labios, lengua y mandíbula), laríngeos (ubicados en la parte superior de la tráquea, específicamente en la laringe) y respiratorios (principalmente presentes en la región del tórax y el abdomen)[6]. A continuación, se detallarán los desórdenes del habla y del movimiento facial.

Desórdenes orofaciales. El sistema orofacial agrupa diversos órganos responsables de las funciones fisiológicas como la respiración, la deglución, la succión , el habla, la fonación y la expresión facial [21]. En este sistema se desarrollan numerosos síntomas del

---

[20] Ronald F PFEIFFER. “Non-motor symptoms in Parkinson’s disease”. In: *Parkinsonism & related disorders* 22 (2016), S119–S122.

[21] Braedan RJ PRETE and Aviv OUANOUNOU. “Medical management, orofacial findings, and dental care for the patient with Parkinson’s disease”. In: *J Can Dent Assoc* 87.l10 (2021), pp. 1488–2159.

Parkinson, los cuales están presentes en los primeros estadios y se hacen más evidentes con el progreso de la enfermedad[22]. Entre estos síntomas está la hipomimia, comúnmente conocida como “máscara facial”, la cual consiste en una reducción o pérdida de los gestos faciales espontáneos, incluyendo el movimiento de las cejas que acompaña al discurso y la disminución de expresiones faciales[23]. Respecto a la parte superior del rostro, la hipomimia se manifiesta en la reducción del porcentaje de parpadeo y ausencia de contracción en los músculos involucrados en la producción del habla[23]. Para la parte inferior de la cara, se presentan problemas para sonreír de forma espontánea y apertura involuntaria de la boca, entre otros[24]. Adicionalmente, se manifiesta comúnmente en un solo lado de la cara[23].

Desórdenes de comunicación. Los trastornos del habla son una manifestación común y significativa que impacta la calidad de vida de quienes viven con esta condición neurodegenerativa[9]. Estos desórdenes hacen parte de los primeros indicadores de la enfermedad y consisten en una reducción del volumen del habla, fluctuación del tono, ritmos de habla inconsistentes y articulación imprecisa[25]. De hecho, estas características son colectivamente conocidas como ”disartria hipocinética”[25]. Particularmente, las deficiencias en la entonación y articulación son de las alteraciones de comunicación más observadas en los pacientes[25]. Irregularidad en la prosodia, reducción en la variabilidad de la frecuencia del discurso y la incapacidad de expresar estrés léxico (énfasis puesto en ciertas palabras para

---

22 Arthur H FRIEDLANDER et al. “Parkinson disease: systemic and orofacial manifestations, medical and dental management”. In: *The Journal of the American Dental Association* 140.6 (2009), pp. 658–669.

23 Nomi VINOKUROV et al. “Quantifying hypomimia in parkinson patients using a depth camera”. In: *International Symposium on Pervasive Computing Paradigms for Mental Health*. Springer. 2015, pp. 63–71.

24 Matteo BOLOGNA et al. “Facial bradykinesia”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 84.6 (2013), pp. 681–685.

25 Khashayar DASHTIPOUR et al. “Speech disorders in Parkinson’s disease: pathophysiology, medical management and surgical approaches”. In: *Neurodegenerative disease management* 8.5 (2018), pp. 337–348.

transmitir significado) son también características importantes que impactan la fluidez y comprensión del lenguaje[25]. Estos síntomas son usualmente analizados con el objetivo de identificar patrones que no son distinguibles desde un punto de vista clínico, ayudando así a frenar la progresión de la enfermedad por medio de tratamientos y disminuir el impacto en las actividades diarias del paciente[26].

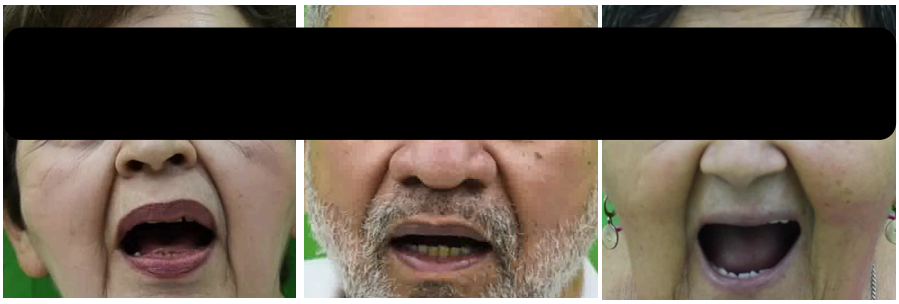

Figura 1. Ilustración de los desórdenes orofaciales presentes en pacientes con la enfermedad de Parkinson.

1.2. Metodologías computacionales.

1.2.1. Representaciones convolucionales. Últimamente, se han propuesto diversas metodologías relacionadas al reconocimiento de patrones; incluyendo el procesamiento de imágenes y señales de audio que permiten realizar clasificación de pacientes con la enfermedad del Parkinson[15] [14]. Las *CNN* son arquitecturas altamente conocidas en el aprendizaje profundo y utilizadas en gran medida para tareas de clasificación. Las capas convolucionales están compuestas de distintos *kernels*, los cuales aprenden representaciones locales de los elementos de entrada, generando diferentes mapas de características. Dichos mapas relacionan elementos de entrada dependiendo de la dimensionalidad del *kernel*, esta puede ser espacial (2D) o volumétrica (3D). En las

[26] Anna FAVARO et al. “Multilingual evaluation of interpretable biomarkers to represent language and speech patterns in Parkinson’s disease”. In: *Frontiers in Neurology* 14 (2023), p. 1142642.

convoluciones 2D (Figura 2-A), el *kernel* se desliza sobre la entrada $X \in \mathbb{R}^{W \times H}$ a lo largo de las dimensiones de ancho (W) y alto (H), permitiendo la caracterización de patrones gestuales asociados con el Parkinson. En el caso de los desórdenes de comunicación, las convoluciones 2D (Figura 2-A) son capaces de extraer patrones locales relevantes sobre representaciones de sonoridad en espectrogramas de Mel, donde se realzan las variaciones de frecuencia en la voz humana[27]. En las convoluciones 3D (Figura 2-B) el *kernel* se desplaza sobre una imagen $X \in \mathbb{R}^{W \times H \times S'}$, con dimensión espacial (W, H) y S' *slices*. Esto permite capturar no solo características espaciales sino información temporal, siendo ideal para modelar patrones de movimiento alterados que maximizan las diferencias entre pacientes control y Parkinson[28].

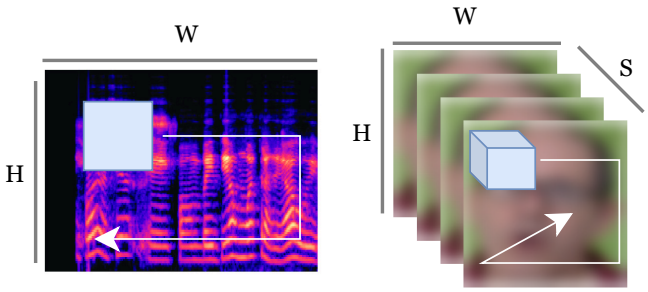

Figura 2. Convoluciones según el desplazamiento del *kernel*. A) Desplazamiento del *kernel* en dos dimensiones para capturar variaciones de frecuencia en la voz sobre espectrogramas de Mel. B) Desplazamiento del *kernel* en tres dimensiones logrando detectar patrones anormales de movimiento en secuencias de videos faciales.

---

[27] Zhijing XU et al. "Voiceprint recognition of Parkinson patients based on deep learning". In: *arXiv preprint arXiv:1812.06613* (2018).

[28] Du TRAN et al. "Learning spatiotemporal features with 3d convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4489–4497.

1.2.2. Aprendizaje multimodal. El principal interés de este trabajo es la integración de información del habla y gestos faciales en una estrategia multimodal, considerando la naturaleza multifactorial del Parkinson (Figura 3). Particularmente, en la literatura, los enfoques multimodales se refieren a estrategias que integran datos provenientes de diversas fuentes, como imagen y sonido; con el fin de realizar tareas como el reconocimiento de patrones. Entonces, el objetivo del aprendizaje multimodal es desarrollar modelos capaces de manejar y relacionar datos de diversos dominios, identificando las relaciones entre modalidades y proyectándolas a un espacio de representación común en donde se logre evidenciar una separabilidad y discriminación entre los pacientes que padecen la enfermedad de Parkinson y los pacientes control[29].

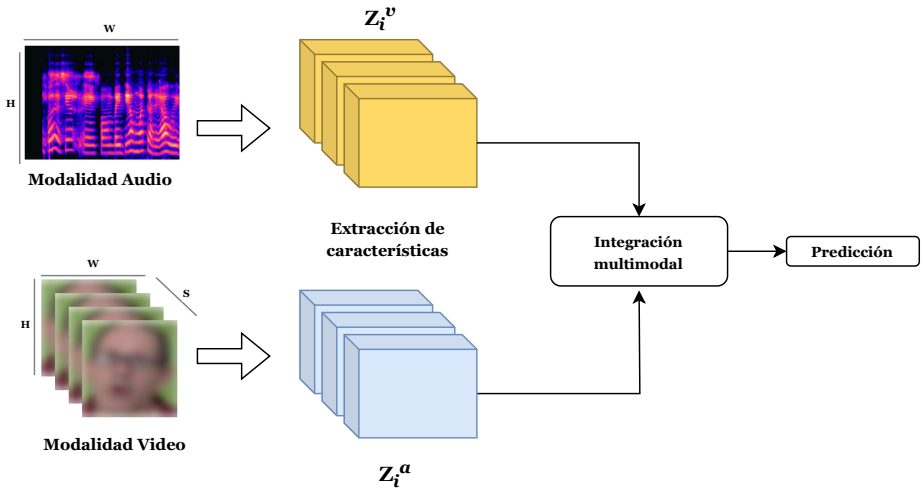

Figura 3. Ilustración de aprendizaje multimodal para relacionar información audiovisual y realizar una clasificación entre pacientes control y Parkinson.

En general, estos enfoques multimodales se pueden expresar desde un conjunto N de datos pareados de al menos dos modalidades, siendo $\{\mathbf{X}_i^a\}_{i=1}^{\mathbf{N}}$ y $\{\mathbf{X}_j^v\}_{j=1}^{\mathbf{N}}$. Desde

[29] Tadas BALTRUŠAITIS; Chaitanya AHUJA, and Louis-Philippe MORENCY. "Multimodal machine learning: A survey and taxonomy". In: *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2018), pp. 423–443.

estos datos, se puede hacer una proyección o generación de descriptores o vectores de características $\mathbf{Z}_i^a$ y $\mathbf{Z}_j^v$. A continuación, se detallan estas estrategias de fusión.

Fusión temprana. Esta estrategia de fusión considera aprender representaciones conjuntas, desde el inicio de la proyección. Es decir, los vectores de proyección en cada modalidad $(\mathbf{Z}_i^a \oplus \mathbf{Z}_j^v)$ son rápidamente fusionados en la arquitectura, para lograr relaciones de bajo nivel entre las dos fuentes de información. Así, la fusión temprana consiste en la integración de las características de bajo nivel obtenidas por cada modalidad antes de la predicción, usando técnicas como la correlación o concatenación.

Esta técnica podría facilitar la captura de las interacciones entre movimientos faciales y características vocales que son distintivas en pacientes con Parkinson (ver Figura 4). Esto se logra mediante el ajuste de parámetros para optimizar la representación integrada de las modalidades. Esto permite al modelo aprender cómo se manifiestan conjuntamente estos síntomas, potencialmente mejorando la precisión en la detección y caracterización de la enfermedad.

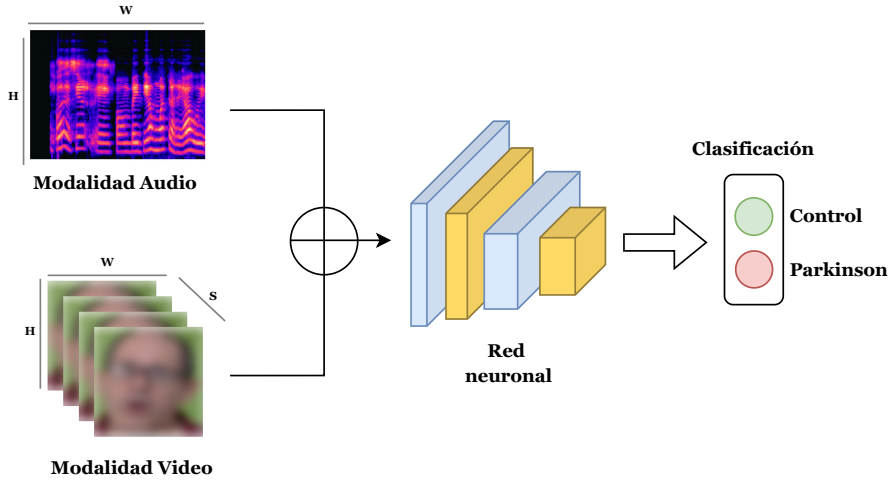

Figura 4. Ilustración de la técnica de fusión temprana entre modalidades de audio y video para realizar una clasificación entre paciente Parkinson y control.

Fusión tardía. En esta estrategia, las fuentes de información se procesan y modelan de forma independiente, generando vectores de características que capturan las relaciones de cada modalidad por separado $(\mathbf{Z}^a i \oplus \mathbf{Z}^v j)$. Cada modalidad realiza una clasificación individual basada en estas representaciones, lo que permite aprovechar al máximo las fortalezas específicas de cada fuente de información. Posteriormente, los resultados de clasificación se combinan en una etapa de fusión tardía, integrando representaciones comunes y capturando relaciones discriminativas entre ambas modalidades[30].

Esta técnica permite que cada modelo se especialice en capturar los detalles más relevantes de su respectiva modalidad antes de combinar los resultados para una clasificación final, como se ilustra en la figura 5. Esto es particularmente útil en el contexto del Parkinson, donde ciertos síntomas pueden ser más evidentes en una modalidad específica, logrando así una integración más efectiva de las características discriminativas entre ambas modalidades.

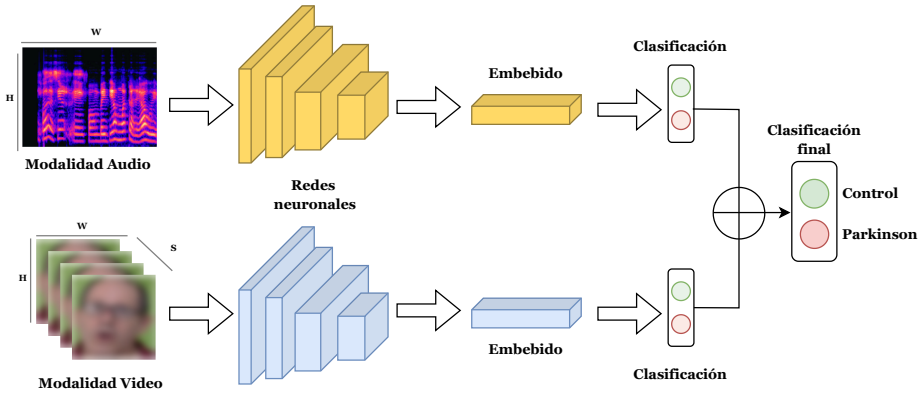

Figura 5. Ilustración de la técnica de fusión tardía entre modalidades de audio y video para realizar una clasificación entre paciente Parkinson y control.

---

[30] Khaled BAYOUDH et al. "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets". In: *The Visual Computer* 38.8 (2022), pp. 2939–2970.

Fusión intermedia. Técnicas alternativas que involucran un procesamiento parcial de las modalidades individuales y luego, estos vectores o representaciones parciales son fusionados, aprovechando lo mejor de las dos estrategias de fusión. Así, permitiendo una integración continua y complementaria de las fuentes de información.

La fusión intermedia combina características de ambas modalidades en múltiples capas intermedias del modelo, como se ilustra en la Figura 6. Este enfoque permite capturar relaciones complejas y no lineales entre las modalidades de video y audio en diferentes niveles de abstracción. En el contexto del Parkinson, esta técnica puede mejorar la capacidad del modelo para detectar y caracterizar síntomas que se manifiestan de manera conjunta en ambas modalidades.

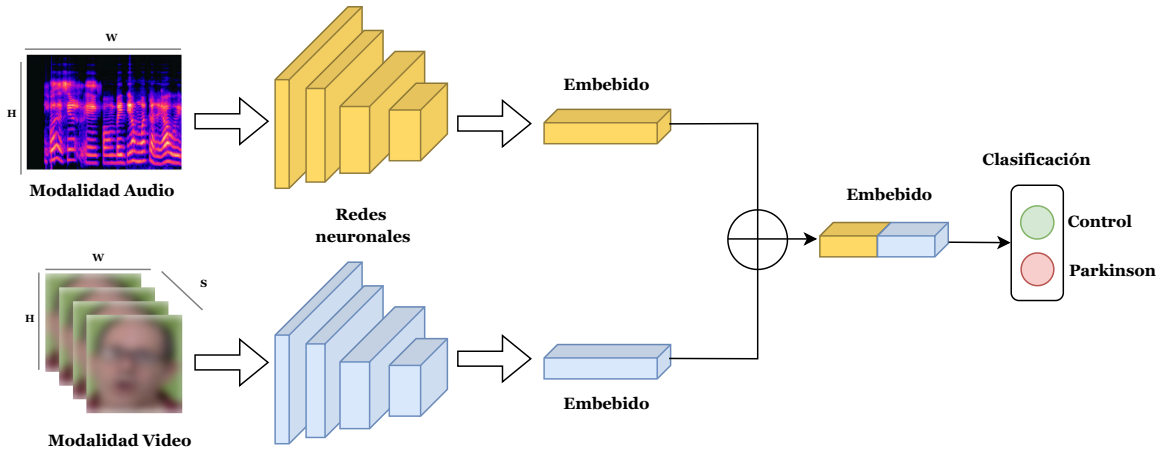

Figura 6. Ilustración de la técnica de fusión intermedia entre modalidades de audio y video para realizar una clasificación entre paciente Parkinson y control.

Fusión con métodos de atención Los métodos de atención han emergido en el estado del arte permitiendo extraer patrones de las características, capturar relaciones no locales y obtener resultados competitivos en aplicaciones como visión por computador y extracción

de características audiovisuales[31] [32] [33]. En este contexto, las arquitecturas basadas en atención pueden ser clasificadas en diferentes tipos, según la forma en que se relacionan las entradas. La *self-attention* (auto-atención) es un mecanismo que permite capturar relaciones internas dentro de una única entrada, resaltando interacciones clave entre los diferentes elementos que conforman la entrada. Por otro lado, la *cross-attention* (atención cruzada) se utiliza para modelar relaciones entre múltiples modalidades o fuentes de datos. En general, el cálculo de atención se basa en la proyección de las entradas en tres representaciones principales: consulta $\mathbf{Q}$, clave $\mathbf{K}$ y valor $\mathbf{V}$. El producto punto es la operación central de los métodos de atención, y consiste en la multiplicación entre la matriz de consulta $\mathbf{Q}$ y la matriz transpuesta de clave $\mathbf{K^T}$, es decir $\mathbf{QK^T}$. Dicha operación representa el grado de similitud o alineación entre ambas matrices; cuanto mayor sea el valor del producto punto, mayor será la relación entre las matrices. Las puntuaciones resultantes suelen ser normalizadas al dividir entre la raíz cuadrada de la dimensión de la clave $\sqrt{d_K}$, contribuyendo a la estabilidad de los gradientes durante el aprendizaje. Dicho proceso resulta en la matriz de similaridad, a la cual se aplica la función de activación *softmax* para obtener una distribución de probabilidad que representa la importancia relativa o la proporción de atención de cada clave para la consulta dada. Esta distribución de probabilidad se usa para ponderar los elementos de la matriz de valor $\mathbf{V}$, generando una representación refinada de las entradas. El enfoque general de atención puede ser descrito matemáticamente como:

---

[31] Baozhou ZHU et al. “An attention module for convolutional neural networks”. In: *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part I 30*. Springer. 2021, pp. 167–178.

[32] Xinjie FAN et al. “Bayesian attention modules”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 16362–16376.

[33] Jan K CHOROWSKI et al. “Attention-based models for speech recognition”. In: *Advances in neural information processing systems* 28 (2015).

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)\mathbf{V} \quad (1)$$

Estos métodos son ampliamente utilizados en el diseño de modelos para tareas multimodales, donde se requiere integrar información de distintas fuentes, como el audio y el video. La flexibilidad de la atención para capturar relaciones tanto locales como globales hace que sea una herramienta clave en aplicaciones relacionadas con el Parkinson, permitiendo explorar la interacción entre síntomas motores y de comunicación.

1.3. Estrategias para la clasificación del Parkinson.

En cuanto a la caracterización de patrones de hipomimia, diversas metodologías computacionales han basado su análisis en el estudio de imágenes o secuencias de fotogramas de pacientes realizando ejercicios faciales, logrando caracterizar patrones como la apertura involuntaria de la boca o la reducción de expresiones. Por ejemplo, Jakubowski *et al.* llevaron a cabo un estudio basado en el análisis de imágenes estáticas capturadas de pacientes con Parkinson y sujetos control mientras realizaban expresiones faciales como sonrisas y gestos neutros. El objetivo fue caracterizar la magnitud de la expresión y el temblor de los músculos faciales a partir de estas imágenes. Cabe destacar que en este trabajo se incluyeron también pacientes bajo los efectos de Levodopa, esto para reducir la severidad de los síntomas y simular las condiciones presentes en los primeros estadíos, ayudando a la capacidad de caracterizar la enfermedad del paciente[15]. Por otro lado, Huan et al. propusieron una metodología para la clasificación de la enfermedad utilizando imágenes sintetizadas desde muestras de pacientes con Parkinson, generadas mediante una *StarGAN*. Esta red generativa fue capaz de producir expresiones faciales de ira, disgusto, miedo, alegría, tristeza y sorpresa, permitiendo concatenar estas representaciones para generar un descriptor denso y, además, a través de metodologías preentrenadas con un *backbone*, clasificar a los pacientes como Parkinson o control. De esta manera, se abordó la limitación de disponibilidad de datos. Pese a ello, se presenta una limitación en el uso

de imágenes sintetizadas al existir la posibilidad de no reproducir fielmente características sutiles, como cambios en la posición o el movimiento de músculos faciales que indican diferentes emociones[34]. En un estudio posterior, Gómez *et al.* incluyeron un conjunto de datos de imágenes individuales y fotogramas secuenciales de expresiones faciales anotadas de forma artificial o por expertos para la detección de Unidades de Acción Facial *(FAU)*, un término utilizado para describir movimientos específicos de los músculos faciales. Para ello, se realizó la adaptación de modelos preentrenados como *VGG* y *ResNet* para la detección de *(FAUs)*[35].

Desde la caracterización de desórdenes de habla asociados al Parkinson, Faragó *et al.* lograron recolectar grabaciones de pacientes leyendo textos que permitían evidenciar características como entonación, volumen, fonación, prosodia y articulación [36]. Dichas grabaciones fueron representadas por medio de espectrogramas de Mel, reconocidos por su capacidad de proporcionar una representación visual de la percepción auditiva humana, explorando la variación y el cambio fonético. Además, también utilizaron espectrogramas de habla y espectrogramas de energía de la voz para lograr caracterizar la potencia y energía de la voz a lo largo del tiempo. No obstante, el enfoque en habla continua limitó la evaluación de variaciones estándar en la fonación, especialmente en pacientes con Parkinson, debido a la corta duración de los segmentos vocales analizados. También Govindu y Palwe emplearon audios de pacientes de distintas regiones, obtenidos a través

---

[34] Wei HUANG et al. “Auto Diagnosis of Parkinson’s Disease Via a Deep Learning Model Based on Mixed Emotional Facial Expressions”. In: *IEEE Journal of Biomedical and Health Informatics* 28.5 (2024), pp. 2547–2557. DOI: 10.1109/JBHI.2023.3239780.

[35] L. F. GOMEZ et al. “Exploring Facial Expressions and Action Unit Domains for Parkinson Detection”. In: *PLOS ONE* (2023).

[36] Paul FARAGÓ et al. “CNN-Based Identification of Parkinson’s Disease from Continuous Speech in Noisy Environments”. In: *Bioengineering* 10.5 (2023), p. 531.

del estudio colaborativo PPMI[37], que incluían fluctuaciones, vibraciones y parámetros acústicos de las fonaciones de vocales. Específicamente se analizaron indicadores como la relación ruido-tono armónico *(NHR)* y tono armónico a ruido *(HNR)*, mostrando que el ruido o las distorsiones en el habla aumentan con la progresión de la enfermedad, estando correlacionado con una limitada calidad de voz[38]. Más adelante, Di Cesare *et al.* utilizaron datos no solo de pacientes leyendo textos sino también el audio obtenido de diálogos abiertos, los cuales fueron representados en espectrogramas de Mel y espectrogramas *Gammatone*, cuya diferencia radica en el uso de filtros diseñados para simular de manera más estrecha la respuesta del oído humano a las ondas sonoras[39]. No obstante, la no inclusión de información visual sincronizada de los pacientes puede impactar la representación de los enfoques propuestos.

También en la literatura se han propuesto alternativas multimodales para abarcar el carácter multifactorial del Parkinson, integrando información de audio y video. Por ejemplo, Lim *et al.* realizaron un análisis audiovisual de pacientes que realizaban un ejercicio de lectura. Se usaron puntos de referencia faciales que evaluaban características como el porcentaje de parpadeo, variación en la amplitud de la boca, variaciones de distancia entre la boca y los ojos, entre otros. Para el caso del audio, se evaluaron características como variaciones en el volumen y tono. Se realizó una integración de las mejores características de audio y video usando los clasificadores de *Random Forest (RF)* y Regresión logística[18]. De la misma manera, Skibinska *et. al.* realizaron grabaciones tanto en audio como en video a pacientes,

---

[37] *Parkinson's Progression Markers Initiative*. Fecha de acceso: 18 de agosto de 2024. Sin fecha. URL: https://www.ppmi-info.org/access-data-specimens/download-data.

[38] Aditi GOVINDU and Sushila PALWE. "Early detection of Parkinson's disease using machine learning". In: *Procedia Computer Science* 218 (2023). International Conference on Machine Learning and Data Engineering, pp. 249–261.

[39] Michele Giuseppe DI CESARE et al. "Machine Learning-Assisted Speech Analysis for Early Detection of Parkinson's Disease: A Study on Speaker Diarization and Classification Techniques". In: *Sensors* 24.5 (2024).

mientras realizaban ejercicios como la pronunciación de vocales, sentencias, palabras, trabalenguas, poemas, entre otros. Los puntos de referencia faciales también fueron utilizados en este trabajo para identificar características faciales asociadas al Parkinson y mediante el algoritmo *XGBoost* se realizó una clasificación, integrando características de audio como insuficiencia de flujo de aire y fluctuación irregular del tono. Los ejercicios de pronunciación de trabalenguas resultaron ser los más informativos, evidenciando una fuerte correlación entre las variaciones en las distancias y ángulos faciales con la presencia de la enfermedad[19]. Ambos trabajos concluyeron que la integración de características biométricas asociadas a la voz y las expresiones faciales podría ayudar a identificar a pacientes con Parkinson en etapas tempranas de la enfermedad. Una limitación de estos enfoques es el uso de puntos de referencia faciales, los cuales pueden ser sensibles a movimientos rápidos y factores ambientales como iluminación.

2. PROBLEMA DE INVESTIGACIÓN

El Parkinson es el trastorno neurodegenerativo de más rápido crecimiento en el mundo. Particularmente, la prevalencia mundial de la enfermedad del Parkinson se ha duplicado en los últimos 25 años, alcanzando los $8,5$ millones de casos en el 2019. En donde se reportaron 329.000 muertes causadas por el Parkinson, significando un incremento del 100% con respecto al año 2000[40]. Esta enfermedad causa limitaciones en la coordinación motora y rigidez en los músculos, típicamente caracterizadas por la reducción de la expresividad y lentitud de los músculos faciales (hipomimia), así como dificultades para articular palabras a la hora de hablar (disartria e hipofonía), siendo síntomas característicos que impactan significativamente la calidad de vida de los pacientes, afectando su interacción social y bienestar emocional[7]. Los protocolos para el diagnóstico de la enfermedad del Parkinson se basan principalmente en un análisis individual de los síntomas, perdiendo la complementariedad de la información que puede dar un mayor soporte a la caracterización de la enfermedad[11].

Existen alternativas computacionales que permanecen limitadas a un estudio individual de los síntomas, desaprovechando las ventajas de un análisis multifactorial. Algunos esfuerzos recientes de metodologías multimodales podrían no capturar las interacciones y dependencias específicas entre ambas modalidades debido a la selección previa de características, su integración tardía y reglas lineales de fusión.

¿Cómo implementar una arquitectura de aprendizaje profundo multimodal que integre relaciones intermedias de información gestual y oral para clasificar patrones parkinsonianos?

---

[40] World Health ORGANIZATION. *Launch of WHO's Parkinson disease technical brief*. Fecha de acceso: 27 de mayo de 2024. 2022. URL: https://www.who.int/news/item/14-06-2022-launch-of-who-s-parkinson-disease-technical-brief.

3. OBJETIVOS

3.1. Objetivo general.

Desarrollar una estrategia de aprendizaje profundo multimodal integrando información audiovisual para clasificar patrones parkinsonianos.

3.2. Objetivos específicos.

- Seleccionar un conjunto de videos con información gesto-auditiva, de pacientes con Parkinson y sujetos control.
- Implementar una estrategia de aprendizaje profundo para extraer características gestuales del conjunto de videos.
- Implementar una estrategia de aprendizaje profundo capaz de obtener características auditivas del audio presente en el conjunto de videos.
- Elaborar una estrategia multimodal que integre representaciones orales y gestuales de un observador.
- Evaluar la capacidad de la representación multimodal para clasificar pacientes Parkinson y control.

4. MÉTODO PROPUESTO

En este trabajo se implementó una estrategia de aprendizaje profundo multimodal que integra información audiovisual para detectar patrones característicos de la enfermedad de Parkinson, relacionados tanto con las deficiencias del habla como con los patrones de hipomimia facial. Como se puede ver en la Figura 7, el método está compuesto por una red multimodal que extrae características orales y gestuales de un conjunto de videos. Estas características se integran en un embebido combinado, que sirve como entrada a un modelo de auto-atención, mejorando la captura de relaciones complejas entre modalidades y destacando la naturaleza multisintomática de la enfermedad. Posteriormente, se lleva a cabo el proceso de clasificación. El entrenamiento de esta arquitectura fue *end-to-end*, logrando un aprendizaje de patrones multimodales a lo largo de la estrategia. A continuación, se detallan los componentes clave del enfoque desarrollado.

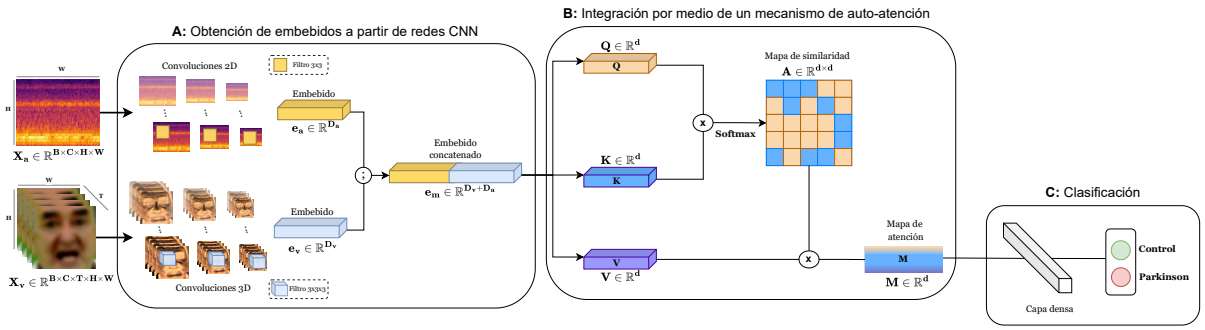

Figura 7. Esquema del enfoque propuesto. Se obtienen embebidos a partir de las redes convolucionales de audio y video, los cuales son concatenados (A) y sirven como entrada al modelo de auto-atención (B) para luego realizar la clasificación entre control y Parkinson (C).

4.1. Representación de video.

La extracción de patrones espacio-temporales asociados al movimiento facial es clave para caracterizar la hipomimia, ya que permite realizar la captura de movimientos anormales o

incluso con diferentes velocidades, mejorando la capacidad de clasificación de síntomas parkinsonianos. En este contexto, se utilizó una red neuronal convolucional tridimensional (*CNN 3D, por sus siglas en inglés*), la cual está diseñada para capturar tanto la información espacial como la información temporal. De esta forma, la *CNN 3D* caracteriza los movimientos faciales a lo largo del tiempo de manera local, permitiendo detectar patrones alterados.

En este trabajo, la red recibe como entrada videos de pacientes mientras realizan algunos ejercicios lingüísticos, como la pronunciación de fonemas, vocales y palabras. Esta entrada se denota como $\mathbf{X}_{\mathrm{v}} \in \mathbb{R}^{B\times C\times T\times H\times W}$, donde: B es el tamaño del *batch*, C es el número de canales, T es el número de *frames* en la secuencia, y $(H \times W)$ son las dimensiones espaciales de cada frame. Específicamente, dada una video secuencia $\mathbf{X}_{\mathrm{v}}$ y asumiendo una arquitectura con L capas, en cada capa $\ell \in \{1, \ldots, L\}$ se calculan transformaciones no lineales y secuenciales, usando un banco de filtros K_ℓ. Estos filtros convolucionales son aprendidos y tienen una composición 3D, definidos como: Ψ_k^ℓ con un valor de sesgo escalar b_k^ℓ, posteriormente se aplica una activación no lineal $\mathbf{a}_k^\ell$ y una operación de *pooling* Π_ℓ. Como en una red neuronal típica, esta proyección se realiza como una composición de funciones, siguiendo un proceso secuencial, definido como:

$$\Phi_k^\ell = \Pi_\ell \left(\mathbf{a}_k^\ell \left(\mathbf{\Phi}^{\ell-1} * \Psi_k^\ell + b_k^\ell \right) \right) \tag{2}$$

Considerando en este caso: $\mathbf{\Phi}^\ell = \left\{ \Phi_k^\ell \right\}_{k=1}^{K_\ell}$, $\mathbf{\Phi}^0 = \mathbf{X}_{\mathrm{v}}$. De esta representación resulta un banco de filtros volumétricos $\mathbf{\Phi^L} = \left\{ \boldsymbol{F}_v^L \right\}_{i=1}^{N}$ que aprenden no solo los componentes cinemáticos locales de los gestos faciales, sino la estructura espacial de los *frames*. Posteriormente, se aplana la salida y luego ingresa a una capa totalmente conectada para generar un embebido $\mathbf{e}_{\mathrm{v}} \in \mathbb{R}^{D_{\mathrm{v}}}$, el cual es una representación compacta y de alta dimensionalidad que encapsula la información más relevante sobre los patrones visuales y temporales.

4.2. Representación de audio.

Un aspecto crucial en la caracterización de la enfermedad de Parkinson es el análisis vocal de los pacientes durante la realización de ejercicios de habla. Patrones como la disartria son anomalías frecuentes que presentan una alta correlación con esta enfermedad. Por esta razón, el estudio de la gesticulación debe complementarse con información proveniente del audio, lo que enriquece el análisis y mejora la discriminación frente a pacientes control. En este trabajo, nos enfocamos en el aprendizaje multimodal procesando y analizando el audio de los pacientes mientras realizan ejercicios de fonemas, vocales y palabras.

Para el tratamiento computacional del audio, los espectrogramas de Mel han representado una de las principales herramientas que permiten una representación espacial sobre la percepción auditiva humana, codificada en un mapa de frecuencia. Específicamente, las señales de audio suelen adquirirse como vectores unidimensionales denotados como $\mathbf{x(t)}$, donde t representa la duración del audio. Estas señales se transforman al espectro de frecuencia, a través de la Transformada de Fourier de Tiempo Reducido (*STFT, por sus siglas en inglés*), obteniendo una nueva representación $\hat{\mathbf{x}}(\mathbf{k}, \mathbf{j})$, donde k es el índice de frecuencia y j es el índice de la ventana temporal. Con el fin de eliminar redundancias y mantener las características más importantes, la representación espectral obtenida se proyecta en la escala Mel mediante un banco de filtros, donde la frecuencia en la escala Mel m está relacionada con la frecuencia lineal $\hat{x}$ (en Hz) según: $m = 2595 \cdot \log_{10}\left(1 + \frac{\hat{x}}{700}\right)$. Por tanto, esto genera el espectrograma de Mel $S_{\text{Mel}}(t, m)$, que utiliza bandas perceptualmente distribuidas para representar la energía en cada ventana temporal.

Posteriormente, el espectrograma de Mel en potencia se transforma a una escala logarítmica para aumentar la sensibilidad a las frecuencias bajas, donde suelen encontrarse alteraciones características del habla, como la frecuencia fundamental (F_0, típicamente entre 80-300 Hz), donde los pacientes con disartria hipocinética tienden a mostrar una

variabilidad reducida en el tono[41]. Esta transformación logarítmica se define como: $S_{\text{dB}}(t, m) = 10 \cdot \log_{10}(S_{\text{Mel}}(t, m) + \epsilon)$, donde ϵ es un valor de corrección que evita problemas relacionados con valores nulos en el cálculo logarítmico.

Una vez procesada la información de audio, los espectrogramas de Mel entran a una red convolucional para representar patrones de audio anormales en pacientes con Parkinson, siguiendo tareas de articulación y fonación. Para ello, se implementó una red neuronal convolucional bidimensional (*CNN 2D, por sus siglas en inglés*) permitiendo capturar patrones espaciales que representan características relevantes de las señales acústicas, aprovechando las propiedades perceptuales de la escala Mel. Esta red fue inicialmente entrenada de forma independiente, para la clasificación entre Parkinson y control.

Así, la *CNN 2D* podría capturar patrones acústicos anormales asociados con la enfermedad de Parkinson, como alteraciones en la prosodia y en la articulación del habla. Específicamente, la red recibe como entrada espectrogramas de Mel denotados como $\mathbf{X}_{\text{a}} \in \mathbb{R}^{B \times C \times H \times W}$, donde B es el tamaño del lote, C es el número de canales de la imagen, ya que el espectrograma de Mel es monocromático, y (H, W) las dimensiones espaciales del espectrograma, que representan la frecuencia y el tiempo, respectivamente. La red aplica filtros bidimensionales en sus capas convolucionales generando mapas de activación $\mathbf{F}_{\mathbf{a}}^{\mathbf{L}}$, que aprenden características locales relacionadas con la variación de energía en diferentes bandas de frecuencia y a lo largo del tiempo. Posteriormente, la salida convolucional se aplana y se pasa a una capa completamente conectada que genera un embebido $\mathbf{e}_{\text{a}} \in \mathbb{R}^{D_{\text{a}}}$. Este embebido encapsula la información más relevante de los patrones acústicos presentes en el espectrograma de Mel, proporcionando una representación compacta de alta dimensionalidad adecuada para su integración en la estrategia multimodal.

---

[41] Francisco MARTÍNEZ-SÁNCHEZ. “Trastornos del habla y la voz en la enfermedad de Parkinson”. In: *Revista de neurologia* 51 (Jan. 2010), pp. 542–550.

4.3. Integración multimodal desde un mecanismo de auto-atención.

En el enfoque propuesto, los embebidos generados por las modalidades de audio ($\mathbf{e}_{\text{a}}$) y video ($\mathbf{e}_{\text{v}}$) son combinados mediante un mecanismo de concatenación para integrar las representaciones acústicas y visuales ($\mathbf{e}_{\text{m}} = [\mathbf{e}_{\mathbf{a}}; \mathbf{e}_{\mathbf{v}}]$). Este paso es crucial en el contexto multimodal, ya que permite integrar la información proveniente de diferentes dominios, lo que enriquece la representación conjunta de los patrones multisensoriales característicos de la enfermedad de Parkinson. Sin embargo, al procesar este vector concatenado mediante capas convolucionales, únicamente se capturan patrones locales, sin modelar explícitamente las interacciones entre ambas modalidades. Además, una proyección densa posterior podría asignar relevancia a componentes dominantes, ignorando relaciones sutiles entre audio y video, afectando la discriminación multimodal.

Considerando lo anterior, y para realizar una captura no-local entre las modalidades, en este trabajo se implementó un mecanismo de auto-atención, generado a partir de las modalidades de audio y video. Este enfoque permite resaltar patrones relevantes relacionados con los síntomas del Parkinson, como las alteraciones en el habla y la hipomimia facial, que pueden manifestarse de manera compleja y distribuida en las diferentes modalidades. De esta manera, teniendo el embebido multimodal concatenado, se proyectan tres matrices de representaciones: la matriz de consulta $\mathbf{Q}$, la matriz de clave $\mathbf{K}$ y la matriz de valor $\mathbf{V}$. Estas se definen mediante transformaciones lineales aprendibles:

$$\mathbf{Q} = \mathbf{e}_{\text{m}}\mathbf{W}^{Q}, \quad \mathbf{K} = \mathbf{e}_{\text{m}}\mathbf{W}^{K}, \quad \mathbf{V} = \mathbf{e}_{\text{m}}\mathbf{W}^{V}, \tag{3}$$

donde $\mathbf{W}^{Q}, \mathbf{W}^{K}, \mathbf{W}^{V} \in \mathbb{R}^{D\times d}$ son matrices de pesos que transforman el embebido de un espacio $D = D_{\text{a}} + D_{\text{v}}$ a un espacio latente de dimensión d y $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{B\times d}$ son matrices proyectadas por medio de capas lineales que capturan representaciones específicas.

El mecanismo de atención utiliza el producto matricial entre $\mathbf{Q}^{\top}$ y $\mathbf{K}$, permitiendo extraer representaciones de largo alcance, expresándose formalmente como $\mathbf{S} = \frac{\mathbf{Q}^{\top}\mathbf{K}}{\sqrt{d}}$, donde

$\sqrt{d}$ es un término de escalamiento que estabiliza los gradientes durante el aprendizaje. La matriz de similitud $\mathbf{S} \in \mathbb{R}^{d \times d}$ captura qué tan relevantes son las distintas partes del embebido en relación con otras.

Esta capacidad es crucial para identificar conexiones no triviales entre patrones de voz y gestos faciales, como cambios sutiles en la prosodia (ritmo y tono del habla) que podrían correlacionarse con expresiones faciales atenuadas. Para destacar las relaciones más importantes, la matriz de similitud $\mathbf{S}$ se normaliza utilizando la función *softmax*, produciendo un mapa de similaridad $\mathbf{A} = \text{softmax}(\mathbf{S})$ donde cada elemento de $\mathbf{A}$ representa la importancia relativa de cada componente del embebido para una consulta dada. Este paso es fundamental en el análisis multimodal, ya que permite asignar mayor peso a características específicas del audio o video que reflejen directamente los síntomas del Parkinson. Logrando así identificar patrones relevantes en la entonación mientras ignora detalles redundantes en las expresiones faciales. El mapa de similaridad $\mathbf{A}$ pondera la matriz de valores $\mathbf{V}^\top$ para generar el vector $\mathbf{M} = \mathbf{A}\mathbf{V}^\top$. Este vector contiene una representación refinada del embebido multimodal, destacando interacciones internas relevantes. De hecho, este vector embebido puede contener relaciones semánticas que integran información no-local, como la correlación entre pausas prolongadas en el habla y expresiones faciales alteradas, que son difíciles de capturar mediante enfoques tradicionales. De esta forma, el mecanismo de auto-atención busca no solo mejorar la capacidad del método para identificar patrones distribuidos en las dos modalidades, sino que también busca priorizar las características más discriminativas, contribuyendo al análisis multimodal del Parkinson, donde los síntomas multimodales suelen estar interrelacionados.

Finalmente, el vector refinado $\mathbf{M}$ es introducido en una capa totalmente conectada, cuya salida permite realizar la clasificación entre sujetos control y pacientes con Parkinson.

5. DISEÑO EXPERIMENTAL

5.1. Datos.

El conjunto de datos propios utilizado presenta grabaciones sincronizadas de audio y video de 14 participantes: 7 pacientes diagnosticados con Parkinson (edad promedio de 65 ± 4) y 7 sujetos control (edad promedio de 61 ± 3). La distribución de género fue de 4 hombres y 3 mujeres en la población con Parkinson, y 2 hombres y 5 mujeres en la población de control. Dos de los pacientes con Parkinson no estaban medicados al momento de la adquisición de los datos, mientras que los 5 restantes recibían tratamiento con Levodopa. A cada participante se le grabó en una secuencia de video con la señal de audio sincronizada, mientras realizaban ejercicios orales como la pronunciación sostenida de vocales y la articulación fonemas y palabras. Las grabaciones se realizaron bajo condiciones controladas para ambas poblaciones, con el objetivo de mantener consistentes los niveles de ruido de fondo y las variables de iluminación. Todos los pacientes fueron diagnosticados por un neurólogo experto y etiquetados parcialmente según la escala Hoehn and Yahr (H & Y)[42].

Ejercicios orofaciales. Este estudio incluye diferentes tareas de fonación y articulación, que se resumen a continuación:

- Para registrar los patrones asociados a la fonación, se realizaron tres repeticiones de las 5 vocales.
- La capacidad de articulación se evaluó mediante la pronunciación de fonemas que fuerzan el movimiento de los músculos afectados por el Parkinson. Los fonemas

[42] Roongroj BHIDAYASIRI et al. “Parkinson’s disease: Hoehn and Yahr scale”. In: *Movement disorders: a video atlas: a video atlas* (2012), pp. 4–5.

incluidos en el estudio son: *(pa, pe, ta, ka): pa-ta-ka, pa-ka-ta, pe-ta-ka*.

- También se incluyó la repetición de tres grupos de palabras para enriquecer la fonación y articulación. El primer grupo incluye palabras como: *petaca, bodega, pato, apto, campana, presa y plato*. El segundo grupo incluye verbos motores: *acariciar, aplaudir, agarrar y dibujar*. El tercer grupo consiste en sustantivos de objetos bien conocidos: *barco, bosque, ciudad, establo, hospital, luna y montaña*.

5.2. Implementación de la arquitectura propuesta

Configuración de la arquitectura de video. Para recuperar los patrones de hipomimia a partir de grabaciones de video, se seleccionó una *CNN 3D* compacta que consiste en 2 bloques convolucionales, cada uno con una capa convolucional 3D con un *kernel* de tamaño = $(3, 3, 3)$, filtros $(32$ y $64)$, y activación *Relu*. Luego de ello, se realizó una operación de *3D max pooling* con un tamaño de filtro de $2 \times 2 \times 2$. Seguido de esto, se aplica una capa densa de 128 neuronas, de donde se obtiene el embebido.

Configuración de la arquitectura de audio. Con la finalidad de realizar el reconocimiento de audio, cada grabación fue transformada al dominio de los espectrogramas de Mel, utilizando una transformada de Fourier de Tiempo Reducido (*STFT*). En este proceso, a cada señal de audio se le realizó un preprocesamiento inicial donde, en caso de grabaciones estéreo (multi-canal), se seleccionó un único canal (izquierdo) con la finalidad de realizar el cálculo. A continuación, se aplicó la *STFT* con una ventana de $n_{\text{fft}} = 1024$ puntos y un salto entre ventanas (*hop length*) de $n_{\text{fft}}/2 = 512$, generando una superposición del 50% entre segmentos consecutivos, permitiendo capturar transiciones suaves y garantizar una representación robusta de las variaciones temporales y frecuenciales en la señal de audio. Posteriormente, se proyectó la energía espectral en $n_{\text{mels}} = 64$ bandas utilizando un banco de filtros ajustado con la escala Mel para así realizar una transformación a una escala logarítmica, permitiendo obtener espectrogramas menos ruidosos y robustos para

caracterizar la enfermedad del Parkinson. Estos espectrogramas fueron normalizados para garantizar la uniformidad en los datos, permitiendo ser la entrada de una *CNN 2D* diseñada específicamente para discriminar entre la población control y Parkinson. La red constaba de 2 bloques compuestos por una capa convolucional 2D, tamaño de *kernel* = $(3, 3)$, filtros $(32$ y $64)$ respectivamente, y activación *Relu*, seguida de una operación de *2D max pooling* con un tamaño de filtro de 3×3. El embebido resultante fue obtenido de la salida de una capa densa de 128 neuronas.

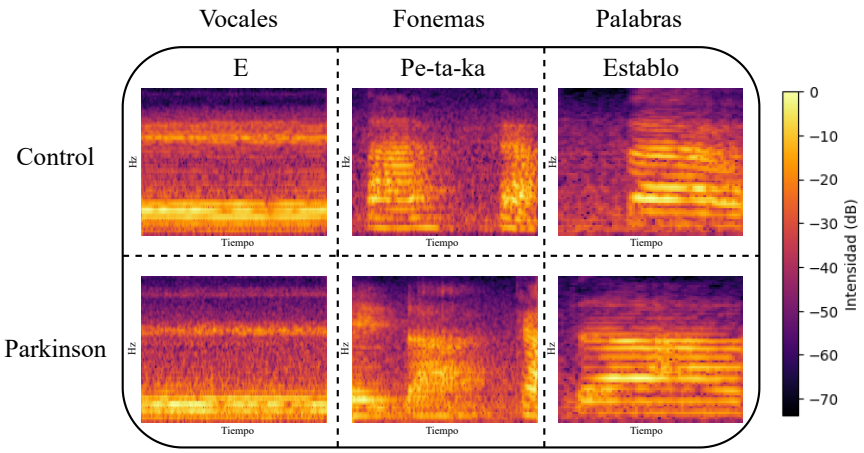

Figura 8. Espectrogramas de Mel correspondientes a pacientes control y Parkinson durante la pronunciación de los distintos ejercicios de habla. La primera columna muestra la vocal sostenida *E*, la segunda columna el fonema *pe-ta-ka* y la tercera columna la articulación de la palabra *establo*. Se observa que los pacientes con Parkinson presentan una menor definición en las frecuencias y mayor irregularidad en la estructura temporal de los espectrogramas, lo que sugiere alteraciones en la fonación y coordinación articulatoria.

Configuración del mecanismo de auto-atención. El mecanismo de auto-atención empleado en este trabajo procesó un embebido concatenado de audio y video e_m con una dimensión de entrada de $[5, 256]$, donde 256 corresponde a la suma de las dimensiones de los embebidos de audio $(D_{\text{a}} = 128)$ y video $(D_{\text{v}} = 128)$, y 5 al número de *batch* por cada iteración. Cada uno de los componentes de la atención, consulta $(\mathbf{Q})$, clave

$(\mathbf{K})$ y valor $(\mathbf{V})$, fue proyectado a un espacio latente de dimensión $d = 128$ mediante capas lineales. El mapa de atención obtenido fue normalizado mediante *LayerNorm* y se aplicó una función de activación *ReLU*, logrando obtener representaciones no-lineales mejorando la caracterización de los embebidos. Luego, se calculó una transformación lineal (capa completamente conectada) intermedia de dimensión 64 para así obtener la capa de clasificación binaria (2 *logits*). El mecanismo de auto-atención se entrenó utilizando una regla binaria de entropía cruzada, optimizador Adam, 70 épocas con una parada temprana de 25 épocas y una tasa de aprendizaje de 1×10^{-5}.

5.3. Validación

La validación del modelo propuesto se realizó siguiendo un esquema de *Leave-One-Out Cross-Validation (LOOCV)*. En este esquema, por cada iteración, el conjunto de validación pertenecía a los datos de audio y video correspondientes a un único paciente. Por el contrario, el entrenamiento fue definido por los demás pacientes. Este procedimiento se repitió iterativamente hasta que cada paciente fue utilizado como conjunto de validación. Cada experimento se dividió según los ejercicios realizados por los pacientes: vocales, fonemas y palabras. Los resultados de estos experimentos se analizaron de manera independiente para cada tipo de ejercicio, evaluando así el desempeño del modelo en diferentes contextos de datos audiovisuales. Las métricas empleadas para evaluar el desempeño del modelo fueron: exactitud, precisión, sensibilidad, *F1-score*, y el área bajo la curva *ROC* (*AUC*). Estas métricas permitieron evaluar la capacidad del modelo multimodal para clasificar patrones asociados a la enfermedad de Parkinson a partir de las representaciones audiovisuales generadas.

6. EVALUACIÓN Y RESULTADOS

El enfoque desarrollado permite analizar descriptores provenientes de datos gesto-auditivos para discriminar patrones asociados a la enfermedad de Parkinson. Para ello, se ajustaron representaciones profundas dedicadas que codifican tanto las señales acústicas como las gestuales de manera independiente, optimizando las configuraciones de cada arquitectura según la modalidad. Una vez identificadas las configuraciones más efectivas, se implementó una estrategia de integración multimodal basada en atención, con el objetivo de combinar y potenciar las capacidades discriminativas de ambas fuentes de información y robustecer su poder de clasificación. A continuación, se relacionan los resultados obtenidos.

Clasificación de Parkinson desde el audio.

El análisis de las señales de audio se centró en evaluar la capacidad discriminativa de dos arquitecturas convolucionales profundas: *CNN 2D* y *VGG-16*. La arquitectura *CNN 2D* tiene un diseño propio de dos capas convolucionales y sus pesos no han sido inicializados desde otras representaciones. Esta arquitectura es compacta para poder ser ajustada desde un conjunto limitado de datos. En cuanto a la arquitectura *VGG-16*, esta es una arquitectura clásica, reconocida en la literatura por su efectiva extensión en otros dominios. En este trabajo, esta red se adoptó para validar el aprovechamiento de un ajuste desde representaciones pre-entrenadas con grandes volúmenes de datos. Estas redes fueron ajustadas para procesar espectrogramas de frecuencia Mel derivados de los audios de los pacientes. En la Tabla 1 se presentan las métricas de desempeño para cada arquitectura, considerando ejercicios basados en fonemas, vocales y palabras. Dada la tarea de clasificación binaria, se consideraron como métricas de desempeño la precisión, sensibilidad, medida F1 (*F1 score*), exactitud y área bajo la curva (*AUC* por sus siglas en

inglés).

Red	Ejercicio	Precisión (%)	Sensibilidad (%)	F1 (%)	Exactitud (%)	AUC (%)
VGG-16	Fonemas	*58.49*	*49.21*	*53.45*	*57.14*	*63.57 ± 0.051*
	Vocales	*46.34*	*36.19*	*40.64*	*47.14*	*47.31 ± 0.039*
	Palabras	*62.20*	*61.38*	*61.78*	*62.04*	*63.25 ± 0.020*
Método Propuesto Audio CNN 2D	Fonemas	***71.79***	**64.44**	**54.90**	**63.49**	**80.75 ± 0.038**
	Vocales	**68.64**	**77.14**	**72.65**	**70.95**	**78.84 ± 0.031**
	Palabras	**72.64**	**56.88**	**63.80**	**67.72**	**77.78 ± 0.016**

Tabla 1. Resultados de clasificación entre Parkinson y control a partir de patrones de audio utilizando una *CNN 2D* y *VGG-16*.

Como se puede apreciar, la red *CNN 2D* demostró un desempeño superior en todas las métricas evaluadas, especialmente en términos de *AUC*, alcanzando valores de hasta 80.75% en ejercicios de fonemas. Este comportamiento podría atribuirse a la estructura menos compleja de la red, que permite capturar patrones acústicos esenciales sin sobreajustarse a los datos de entrenamiento. Por otro lado, la *VGG-16*, aunque muestra resultados consistentes, tiende a obtener desempeños más bajos en precisión y sensibilidad. Esto podría deberse a su diseño, que incluye una mayor cantidad de parámetros y capas profundas. Además, los datos de preentrenamiento de la arquitectura se realizan sobre imágenes naturales, lo cual puede ser distante de los espectrogramas de Mel. Estas características, si bien son útiles para tareas espaciales complejas, hacen que la red sea más propensa al sobreajuste cuando se trabaja con conjuntos de datos limitados.

El análisis por tipo de ejercicio muestra que la clasificación basada en las vocales presenta un mejor desempeño general. Esto puede explicarse por la naturaleza prolongada y repetitiva de los audios, facilitando la extracción de características frecuenciales esenciales a partir de los espectrogramas. En contraste, los ejercicios de fonemas y palabras, que incluyen mayor complejidad estructural, variabilidad en la pronunciación y transiciones rápidas con patrones acústicos menos uniformes, presentan una mayor dificultad para ambas redes. Considerando lo anterior, se selecciona la *CNN 2D* como modelo base para la representación de patrones acústicos en la integración multimodal.

Clasificación de Parkinson desde el video.

El análisis de las señales de video se centró en evaluar patrones espacio-temporales relacionados con la enfermedad de Parkinson durante la ejecución de ejercicios lingüísticos, ya que los movimientos faciales y gestuales son indicadores clave en la caracterización de la enfermedad. Para ello, se implementaron dos arquitecturas: una *CNN 3D*, diseñada para capturar dinámicas temporales y gestuales en secuencias completas de video, y una *CNN 2D*, enfocada en la extracción de características espaciales a partir de fotogramas individuales. Particularmente para esta arquitectura bidimensional, desde cada video, se seleccionó un único fotograma, el cual fue procesado de manera aislada para la extracción de patrones gestuales estáticos. La Tabla 2 reporta los resultados obtenidos para las dos arquitecturas consideradas, evaluando ejercicios basados en fonemas, vocales y palabras.

Red	Ejercicio	Precisión (%)	Sensibilidad (%)	F1 (%)	Exactitud (%)	AUC (%)
CNN 2D	Fonemas	**87.50**	*44.44*	*58.95*	**69.05**	*74.25 ± 8.23*
	Vocales	*57.73*	*53.33*	*55.45*	*57.14*	*63.82 ± 7.17*
	Palabras	*54.55*	*36.51*	*43.74*	*53.04*	*57.42 ± 4.01*
Método Propuesto Video CNN 3D	Fonemas	*73.08*	**60.32**	**66.09**	**69.05**	**76.04 ± 0.043**
	Vocales	**64.65**	**60.95**	**62.75**	**63.81**	**69.53 ± 0.035**
	Palabras	**68.88**	**52.15**	**59.34**	**64.20**	**75.86 ± 0.018**

Tabla 2. Resultados de clasificación entre Parkinson y control a partir de patrones de video utilizando una *CNN 3D* y *CNN 2D*.

La CNN 3D demostró un desempeño superior en el análisis de ejercicios de fonemas, donde alcanzó un *AUC* de 76.04%. Este resultado destaca la capacidad de esta arquitectura para capturar tanto patrones gestuales como dinámicas temporales, que son esenciales para discriminar entre pacientes con Parkinson y controles. Por otro lado, la *CNN 2D* evidenció un desempeño destacable en ejercicios de fonemas y vocales, donde la repetitividad y la relativa estabilidad de los gestos facilitaron la extracción de características discriminativas (*AUC* de 74.24% en fonemas). Sin embargo, esta arquitectura mostró una caída significativa en tareas que demandan un análisis más profundo de patrones temporales, como en las palabras, alcanzando un *AUC* de solo 57.42%. Además, su dependencia de fotogramas aislados puede provocar la pérdida de información contextual

importante inherente a la dinámica de los gestos en el video.
Los resultados obtenidos resaltan la importancia de las arquitecturas 3D para capturar de manera efectiva patrones temporales y gestuales en ejercicios más complejos. Esto sugiere que la *CNN 3D* es la opción más robusta para la representación de información visual, lo que la posiciona como el modelo base en los experimentos de integración multimodal.

6.1. Integración de las modalidades de audio y video.

La integración multimodal permite aprovechar las representaciones conjuntas de cada modalidad (señales acústicas y gestuales) logrando capturar patrones representativos y robustos de manera multi-factorial y mejorando la clasificación de la enfermedad de Parkinson. Es por ello que se realizó una estrategia de fusión intermedia (*intermediate fusion*) por medio de un mecanismo de auto-atención (*self-attention*), diseñado para procesar de manera simultánea ambas modalidades y capturar relaciones complementarias entre ellas. Como se puede ver, la Tabla 3 evidencia que la integración realizada basada en auto-atención supera significativamente a las técnicas alternativas de integración multimodal, como la concatenación simple de embebidos y la atención cruzada, en métricas clave como precisión (74.19% en fonemas, 68.15% en vocales y 59.43% en palabras), *F1 score* (73.60% en fonemas y 76.67% en vocales) y *AUC* (75.26% en fonemas y 77.01% en vocales). En comparación, la concatenación simple mostró resultados más limitados, con un *F1 score* máximo de 54.15% y un *AUC* de 55.76%, mientras que la atención cruzada obtuvo mejores resultados en sensibilidad para palabras (89.42%) pero no logró un balance consistente en las demás métricas.
El mejor desempeño del modelo basado en auto-atención puede atribuirse a su capacidad para identificar y resaltar características complementarias relevantes entre las modalidades de audio y video. A diferencia de la concatenación simple, que trata las características de ambas modalidades como independientes, y de la atención cruzada, que depende en gran

medida de las representaciones iniciales para generar conexiones entre las relaciones, el mecanismo de auto-atención analiza todas las interacciones posibles entre las características de ambas modalidades, permitiendo una integración más completa y efectiva. Esto se traduce en un mejor balance entre todas las métricas presentes, demostrando que el modelo propuesto es más robusto para tareas de clasificación multimodal en este contexto.

Red	Ejercicio	Precisión (%)	Sensibilidad (%)	F1 (%)	Exactitud (%)	AUC (%)
Concatenación de Embebidos	Fonemas	41.25	*52.38*	46.15	38.89	30.69 ± 0.044
	Vocales	50.00	*59.05*	54.15	50.00	56.79 ± 0.020
	Palabras	54.12	55.56	54.83	54.23	53.54 ± 0.015
Atención Cruzada	Fonemas	*49.12*	44.44	*46.67*	*49.21*	*51.10 ± 0.085*
	Vocales	*61.80*	52.38	*56.70*	*60.00*	*62.07 ± 0.018*
	Palabras	*57.19*	**89.42**	**69.76**	**61.24**	**69.76 ± 0.079**
Integración Multimodal Auto-atención	Fonemas	**74.19**	**73.02**	**73.60**	**73.81**	**75.26 ± 0.044**
	Vocales	**68.15**	**87.62**	**76.67**	**73.33**	**77.01 ± 0.032**
	Palabras	**59.43**	*66.67*	*62.84*	*60.58*	*62.44 ± 0.021*

Tabla 3. Comparación de los resultados obtenidos con el modelo multimodal de auto-atención frente a otras técnicas de integración como la concatenación simple de embebidos de audio y video, y la aplicación de un mecanismo de atención cruzada, donde el *query* se proyectó a partir del embebido de audio, mientras que el *key* y *value* del embebido de video.

Por otro lado, se decidió hacer una comparación entre el enfoque multimodal con respecto a las modalidades individuales. Como se puede visualizar, la tabla 4 indica que, de forma general, aunque el enfoque multimodal logra superar a las modalidades individuales en métricas como precisión (74.19% en fonemas), sensibilidad (73.02%, 87.62%, 66.67% para fonemas, vocales y palabras, respectivamente) y *F1 score* (76.67% en vocales), su desempeño en términos de *AUC* no supera el mejor resultado en ninguna de las categorías evaluadas con respecto a la modalidad de video. Este comportamiento refleja que, si bien el mecanismo de auto-atención puede identificar interacciones útiles entre las modalidades, la integración puede enfrentar limitaciones para combinar de manera óptima ambas fuentes de información en ciertas condiciones.

Específicamente, en los ejercicios de fonemas, el modelo multimodal obtiene una mejor clasificación en todas las métricas, centrándose especialmente en métricas como *F1 score*

(73.60%) y sensibilidad (73.02%) en comparación con las arquitecturas individuales de audio y video (*CNN 2D* y *CNN 3D*), lo que evidencia su capacidad para aprender patrones relevantes provenientes de ambas modalidades e incluso, proveer una representación mejorada en la clasificación. Esto, porque a pesar de tener un *AUC* (75.26%) inferior al de la *CNN 2D* (80.75%), su separación provee una confiabilidad más alta.

Por otro lado, en los ejercicios de vocales, el modelo destaca con una sensibilidad del 87.62%, y un *F1 score* de 76.67% superando ampliamente a las arquitecturas individuales (*CNN 2D*: 77.14%, 72.65%; *CNN 3D*: 60.95%, 62.75% respectivamente). Este resultado muestra que el modelo multimodal es efectivo para capturar patrones consistentes en ejercicios repetitivos, donde ambas modalidades aportan información complementaria. Sin embargo, el *AUC* del modelo multimodal (77.01%) sigue siendo ligeramente inferior al de la *CNN 2D* (78.84%). Esto sugiere que, aunque el mecanismo de auto-atención integra información tanto del audio como del video, no siempre logra equilibrar de manera óptima las fortalezas de ambas modalidades. En particular, las características frecuenciales y temporales del audio, que son especialmente relevantes en ejercicios repetitivos como las vocales, podrían estar siendo parcialmente diluidas al combinarse con las representaciones gestuales del video.

En el caso de las palabras, estas presentan mayor complejidad estructural y variabilidad teniendo en cuenta su naturaleza pausada, el modelo multimodal alcanza a superar en términos de sensibilidad por al menos un 10%, mostrando que la tasa de verdaderos positivos es considerablemente alta. Por otro lado, muestra un *F1 score* (62.84%) similar al de la *CNN 2D* (63.80%) y superior al de la *CNN 3D* (59.34%). Sin embargo, la caída en *AUC* (62.44%) sugiere que el modelo no logra integrar de manera óptima las características acústicas y gestuales en estas condiciones más desafiantes.

Para poder visualizar bien cada una de las predicciones realizadas, la Figura 9 muestra la distribución de probabilidades predichas para los ejercicios de vocales en las modalidades de audio, video y la integración multimodal. Aunque la integración multimodal ofrece

Red	Ejercicio	Precisión (%)	Sensibilidad (%)	F1 (%)	Exactitud (%)	AUC (%)
Método Propuesto Audio CNN 2D	Fonemas	71.79	*64.44*	54.90	63.49	**80.75 ± 0.038**
	Vocales	**68.64**	*77.14*	*72.65*	*70.95*	**78.84 ± 0.030**
	Palabras	**72.64**	*56.88*	**63.80**	**67.72**	**77.78 ± 0.016**
Método Propuesto Video CNN 3D	Fonemas	*73.08*	60.32	*66.09*	*69.05*	76.04 ± 0.043
	Vocales	64.65	60.95	62.75	63.81	69.53 ± 0.035
	Palabras	*68.88*	52.15	59.34	*64.20*	*75.86 ± 0.018*
Integración Multimodal Auto-atención	Fonemas	**74.19**	**73.02**	**73.60**	**73.81**	*75.26 ± 0.044*
	Vocales	*68.15*	**87.62**	**76.67**	**73.33**	*77.01 ± 0.032*
	Palabras	59.43	**66.67**	*62.84*	60.58	62.44 ± 0.021

Tabla 4. Resultados de los tres experimentos considerados: modalidad de audio utilizando una red *CNN 2D*, modalidad de video utilizando una red *CNN 3D*, y el método propuesto de integración utilizando un mecanismo de auto-atención. Para cada uno de los experimentos, los ejercicios de vocales, fonemas, y palabras fueron evaluados individualmente.

una mayor consistencia para pacientes como C0 y P5, persisten limitaciones notables en casos como P7 y C1, donde el modelo no logra una mejora significativa respecto a las modalidades individuales. Este comportamiento podría estar relacionado con la dificultad de integrar patrones acústicos y gestuales de manera complementaria en pacientes con características más complejas o atípicas. Además, cabe destacar que factores externos, como el estado de medicación, podrían influir en la variabilidad de los resultados. Por ejemplo, en este conjunto de datos, todos los pacientes, excepto P0 y P4, se encontraban bajo medicación durante la captura de las señales, lo que podría explicar los patrones inconsistentes observados en ciertos casos.

Comparación con estado del arte La Tabla 5 compara los resultados del método propuesto con un enfoque previamente desarrollado por Archila *et al.*[43]. Este trabajo implementó una estrategia de fusión temprana basada en descriptores de covarianza para integrar información de video y audio. Específicamente, se calcularon matrices de covarianza temporales que combinan características faciales, obtenidas a partir de 44

[43] John ARCHILA; Antoine MANZANERA, and Fabio MARTINEZ CARRILLO. "A Mixed audio-video SPD network for online classification of Parkinsonian speech patterns". In: *IBERAMIA 2024: 18th Ibero-American Conference on Artificial Intelligence*. hal-04879377. Montevideo, Uruguay, Nov. 2024.

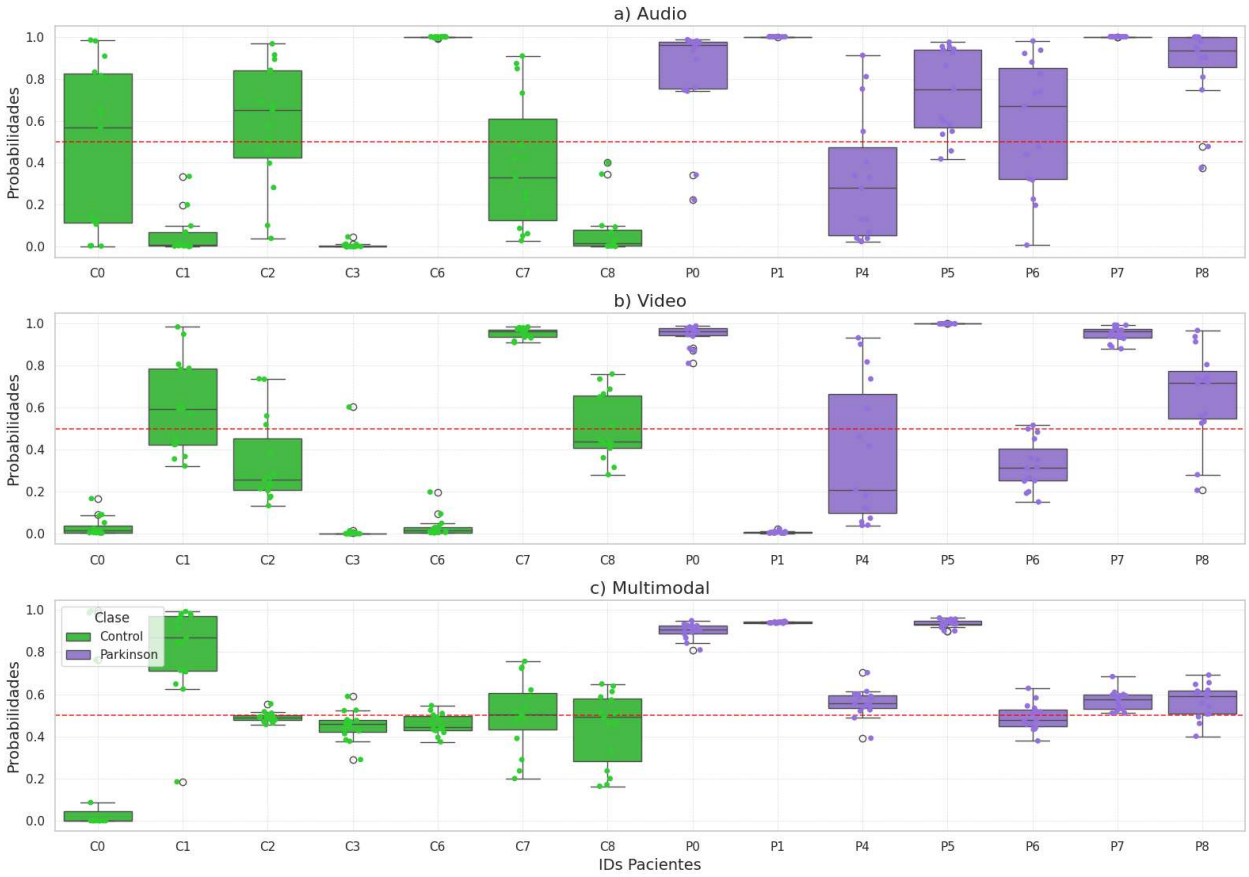

Figura 9. Distribución de probabilidades para las modalidades de audio, video y su integración por cada paciente incluido en el estudio, utilizando *box plot*, *scatter plots* y los ejercicios de vocales. Figura a) muestra la distribución de probabilidad estimada para la modalidad de audio, la figura b) muestra la distribución de probabilidad estimada para la modalidad de video. Finalmente, la figura c) muestra la distribución de probabilidad estimada para la modalidad multimodal. Una linea roja horizontal es usada como valor de referencia (0.5) para determinar si una muestra fue clasificada como paciente Parkinson o control.

puntos clave cercanos a la boca, y frecuencias fundamentales extraídas de espectrogramas de audio generados mediante la Transformada de Fourier de Corto Tiempo (*STFT*). Estas matrices fueron procesadas en un espacio Riemanniano para aprender representaciones que capturan las relaciones dinámicas entre ambas modalidades, las cuales luego fueron clasificadas para identificar patrones parkinsonianos.

Sin embargo, es importante señalar que este trabajo evaluó su metodología únicamente en

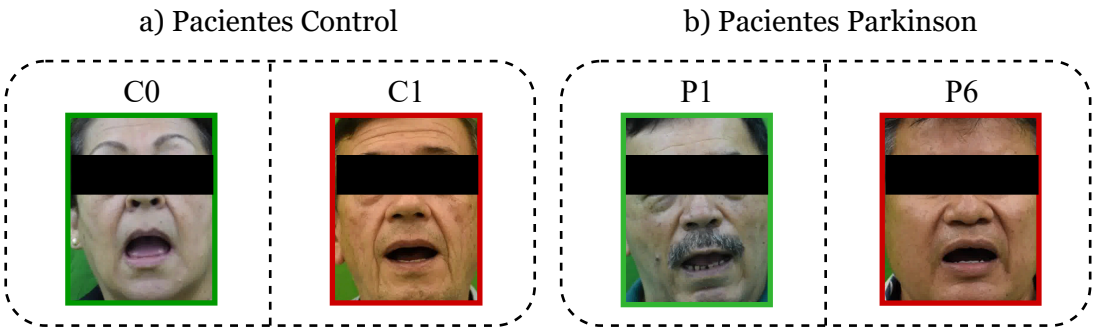

Figura 10. Ejemplos de muestras de pacientes control a) y Parkinson b). En verde se marca el contraste de los pacientes correctamente identificados en su clase, mientras que en rojo se identifican aquellos clasificados erróneamente en el ejercicio de vocales utilizando la modalidad multimodal.

ejercicios de vocales, lo que limita la comparación directa con nuestro método propuesto, el cual abarca una evaluación más amplia en distintos ejercicios. En términos de resultados, Archila et al. reportaron una precisión del 69.00%, sensibilidad del 73.00% y un *F1-score* de 71.00% en la clasificación de pacientes con Parkinson frente a sujetos control durante la pronunciación de vocales, alcanzando una exactitud global del 70.00%. Si bien estos valores reflejan un desempeño sólido en la integración multimodal, nuestro método basado en atención propia presenta fortalezas en sensibilidad (87.62%), *F1-score* (76.67%) y exactitud (73.33%), lo que sugiere una mayor capacidad para capturar patrones relevantes en señales acústicas y gestuales, favoreciendo la caracterización de la enfermedad.

Red	Ejercicio	Precisión (%)	Sensibilidad (%)	F1 (%)	Exactitud (%)	AUC (%)
Archila, J[44]	Vocales	**69.00**	*73.00*	*71.00*	*70.00*	-
Integración Multimodal Auto-atención	Vocales	68.15	**87.62**	**76.67**	**73.33**	77.01 ± 0.032

Tabla 5. Comparación entre los resultados obtenidos por el método propuesto de integración utilizando un mecanismo de auto-atención y la metodología propuesta por Archila *et al.* basada en descriptores de covarianza para integrar información de audio y video, obtenida de los ejercicios de vocales.

7. CONCLUSIONES Y TRABAJO FUTURO

En este trabajo se desarrolló una estrategia de aprendizaje profundo multimodal que permite integrar información audiovisual y clasificar patrones parkinsonianos. A través de la implementación de modelos basados en redes convolucionales 2D y 3D, se extrajeron características gestuales y acústicas de videos de pacientes con Parkinson y sujetos control, generando representaciones que capturan dinámicas espaciales, temporales y frecuenciales. Posteriormente, dichas representaciones fueron integradas mediante un mecanismo de auto-atención, lo que permitió identificar interacciones complementarias entre ambas modalidades. La estrategia multimodal demostró mejoras específicas en métricas como precisión (74.19% en fonemas), sensibilidad (87.62% en vocales) y *F1-Score* (76.67% en vocales). En términos generales, la integración multimodal permitió capturar patrones complementarios, reportando un desempeño en *AUC* (77.01% en vocales, 75.26% en fonemas, 62.44% en palabras).

Para llevar a cabo este trabajo, en una etapa inicial se seleccionó un conjunto de datos compuesto por 1092 videos con información gesto-auditiva, distribuidos en 9 videos para fonemas, 15 para vocales y 54 para palabras por paciente, administrados por el grupo de investigación BIVL^{2}ab. Este conjunto incluyó 7 pacientes con Parkinson y 7 sujetos control, asegurando una representación balanceada de ambas poblaciones. A partir de estos datos, se implementaron estrategias basadas en redes convolucionales para extraer representaciones profundas de cada modalidad: una *CNN 2D* para modelar la progresión temporal de la fonación y resaltar características frecuenciales clave del audio, mientras que una *CNN 3D* permitió capturar información espacio-temporal relevante de los movimientos orofaciales, los cuales están fuertemente asociados con signos motores del Parkinson. Posteriormente, la integración multimodal mediante un mecanismo de auto-atención permitió capturar interacciones complementarias entre ambas modalidades, combinando las representaciones de audio y video para mejorar la discriminación entre

pacientes con Parkinson y sujetos control. Esta estrategia evidenció mejoras en ejercicios como vocales y fonemas, con incrementos en sensibilidad y *F1-score* en comparación con los modelos unimodales. En vocales, la sensibilidad aumentó en aproximadamente 10.48% respecto a la *CNN 2D* y 26.67% en relación con la *CNN 3D*, mientras que en fonemas el *F1-score* presentó una ganancia de 7.15% y 7.45%, respectivamente. Sin embargo, a pesar de estas mejoras, el desempeño global en términos de *AUC* no mostró diferencias estadísticamente significativas con respecto a las modalidades individuales ($p > 0.05$, prueba de *ANOVA*), lo que sugiere que la integración multimodal aún enfrenta desafíos en la combinación óptima de ambas fuentes de información, especialmente en ejercicios con mayor variabilidad estructural, como las palabras.

En la literatura, diversos estudios han abordado la caracterización de la enfermedad de Parkinson, típicamente, a partir de representaciones acústicas, destacando el uso de espectrogramas y modelos de aprendizaje profundo. Algunos trabajos han explorado la variación fonética en el habla de pacientes mediante espectrogramas de Mel y espectrogramas de energía de la voz, alcanzando una precisión de hasta 96% (estudio realizado con 27 pacientes, divididos en 16 pacientes Parkinson y 11 pacientes Control) en la clasificación de la enfermedad[36]. Otros estudios han evaluado la discriminación entre pacientes Parkinson y sujetos control a partir de espectrogramas de Mel y *Gammatone*, con una precisión máxima de 92.3% (el estudio incluyó un total de 37 grabaciones de audio, 16 de las cuales pertenecían a pacientes Parkinson y 21 de pacientes Control), evidenciando la efectividad de los modelos basados en características frecuenciales[39]. Sin embargo, la ausencia de información gestual en estos enfoques podría limitar su capacidad de capturar la complejidad motora de la enfermedad, especialmente en manifestaciones combinadas de disartria e hipomimia. Por otra parte, en cuanto al análisis de señales gestuales, se han propuesto estrategias basadas en modelos volumétricos y descriptores de expresiones faciales. Algunos estudios han evaluado la variabilidad de las expresiones en pacientes con Parkinson bajo tratamiento con Levodopa, empleando imágenes estáticas de sonrisas

y expresiones neutras para evaluar la rigidez facial, alcanzando un *F1-score* de 94.1% (un total de 48 personas hicieron parte del estudio, en donde 24 pertenecían a pacientes Parkinson de los cuales 21 estaban bajo los efectos de Levodopa) en la clasificación de la enfermedad[15]. Otros enfoques han utilizado arquitecturas de aprendizaje profundo para la detección de unidades de acción facial (*FAUs*), logrando un *AUC* superior al 80% (estudio realizado con 24 pacientes Control y 30 pacientes Parkinson, cuyos estadíos se encontraban en el rango 2-3 de la escala *Hoehn & Yahr*) en la identificación de movimientos faciales específicos asociados a la enfermedad[35]. Aunque estos métodos han demostrado ser efectivos en la caracterización de hipomimia, el análisis de imágenes estáticas puede omitir información relevante sobre la progresión dinámica de los gestos y la sincronización con patrones vocales, limitando su aplicabilidad en escenarios más complejos. Además, existen diversos fenotipos de la enfermedad que requieren evaluación complementaria para distinguir el inicio particular de los síntomas de cada paciente. Más allá de los enfoques unimodales, algunos estudios han explorado estrategias de integración multimodal para mejorar la discriminación de la enfermedad a partir de información audiovisual. Algunos trabajos han realizado un análisis de pacientes pronunciando distintas unidades lingüísticas, integrando características faciales y acústicas mediante algoritmos de clasificación, alcanzando una sensibilidad del 88%, precisión balanceada de 83% en ejercicios de prosodia[19]. En comparación con estos enfoques, la estrategia basada en auto-atención utilizada en este trabajo permitió alcanzar en una población de 14 pacientes una precisión de 74.19, y sensibilidad de 87.62, evidenciando un mejor comportamiento al integrar la información de las dos fuentes de información. A diferencia de estos enfoques, la estrategia basada en auto-atención utilizada en este trabajo permite una integración más flexible y adaptativa, capturando relaciones complementarias entre señales acústicas y gestuales sin depender de representaciones explícitas de puntos clave faciales o características de audio preseleccionadas.

Los resultados obtenidos en este trabajo evidencian el potencial de la estrategia multimodal

basada en auto-atención para capturar interacciones complementarias entre señales gestuales y acústicas, logrando mejoras en sensibilidad y *F1-score* en ejercicios como fonemas y vocales. Sin embargo, en ejercicios de mayor complejidad como las palabras, donde la coordinación de los movimientos del habla y la variabilidad fonética juegan un papel fundamental, la integración multimodal puede diluir patrones críticos que son mejor representados en cada modalidad por separado. Esta dificultad sugiere que, aunque la fusión de información mejora la complementariedad entre ambas fuentes, el modelo podría beneficiarse de estrategias que preserven los aspectos más discriminativos de cada señal antes de la integración final. Como línea de trabajo futuro, se plantea la exploración de arquitecturas multimodales que implementen mecanismos de atención en etapas más avanzadas del procesamiento, considerando que la representación de los embebidos puede perder información relevante al combinar características dominantes del audio y el video. Asimismo, la incorporación de modelos auto-supervisados, capaces de aprender representaciones a partir de grandes volúmenes de datos sin necesidad de anotaciones manuales, podría fortalecer la capacidad del modelo para capturar patrones robustos en la fonación y la gestualidad, favoreciendo su adaptación a la variabilidad clínica y mejorando su aplicabilidad en entornos reales. Adicionalmente, es fundamental ampliar la representación de pacientes y ejercicios para mejorar la generalización del modelo y evaluar su desempeño en condiciones más diversas. Finalmente, la variabilidad intrínseca en las señales gestuales y auditivas podría haber afectado el rendimiento general del modelo, dificultando su capacidad para capturar de manera consistente los patrones asociados a la enfermedad, lo que resalta la necesidad de estrategias que mitiguen este efecto en futuros desarrollos.

BIBLIOGRAFÍA

ARCHILA, John; MANZANERA, Antoine, and MARTINEZ CARRILLO, Fabio. “A Mixed audio-video SPD network for online classification of Parkinsonian speech patterns”. In: *IBERAMIA 2024: 18th Ibero-American Conference on Artificial Intelligence*. hal-04879377. Montevideo, Uruguay, Nov. 2024 (cit. on pp. 45, 47).

BALTRUŠAITIS, Tadas; AHUJA, Chaitanya, and MORENCY, Louis-Philippe. “Multimodal machine learning: A survey and taxonomy”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2018), pp. 423–443 (cit. on p. 18).

BAYOUDH, Khaled et al. “A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets”. In: *The Visual Computer* 38.8 (2022), pp. 2939–2970 (cit. on p. 20).

BHIDAYASIRI, Roongroj et al. “Parkinson’s disease: Hoehn and Yahr scale”. In: *Movement disorders: a video atlas: a video atlas* (2012), pp. 4–5 (cit. on p. 35).

BOLOGNA, Matteo et al. “Facial bradykinesia”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 84.6 (2013), pp. 681–685 (cit. on p. 15).

CHOROWSKI, Jan K et al. “Attention-based models for speech recognition”. In: *Advances in neural information processing systems* 28 (2015) (cit. on p. 22).

DASHTIPOUR, Khashayar et al. “Speech disorders in Parkinson’s disease: pathophysiology, medical management and surgical approaches”. In: *Neurodegenerative disease management* 8.5 (2018), pp. 337–348 (cit. on pp. 15, 16).

DI CESARE, Michele Giuseppe et al. “Machine Learning-Assisted Speech Analysis for Early Detection of Parkinson’s Disease: A Study on Speaker Diarization and Classification Techniques”. In: *Sensors* 24.5 (2024) (cit. on pp. 25, 49).

FAN, Xinjie et al. “Bayesian attention modules”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 16362–16376 (cit. on p. 22).

FARAGÓ, Paul et al. “CNN-Based Identification of Parkinson’s Disease from Continuous Speech in Noisy Environments”. In: *Bioengineering* 10.5 (2023), p. 531 (cit. on pp. 24, 49).

FAVARO, Anna et al. “Multilingual evaluation of interpretable biomarkers to represent language and speech patterns in Parkinson’s disease”. In: *Frontiers in Neurology* 14 (2023), p. 1142642 (cit. on p. 16).

FRIEDLANDER, Arthur H et al. “Parkinson disease: systemic and orofacial manifestations, medical and dental management”. In: *The Journal of the American Dental Association* 140.6 (2009), pp. 658–669 (cit. on p. 15).

GOMEZ, L. F. et al. “Exploring Facial Expressions and Action Unit Domains for Parkinson Detection”. In: *PLOS ONE* (2023) (cit. on pp. 24, 50).

GOÑI, María et al. “Smartphone-based digital biomarkers for Parkinson’s disease in a remotely-administered setting”. In: *IEEE access* 10 (2022), pp. 28361–28384 (cit. on pp. 12, 16).

GOVINDU, Aditi and PALWE, Sushila. “Early detection of Parkinson’s disease using machine learning”. In: *Procedia Computer Science* 218 (2023). International Conference on Machine Learning and Data Engineering, pp. 249–261 (cit. on p. 25).

HUANG, Wei et al. “Auto Diagnosis of Parkinson’s Disease Via a Deep Learning Model Based on Mixed Emotional Facial Expressions”. In: *IEEE Journal of Biomedical and Health Informatics* 28.5 (2024), pp. 2547–2557. DOI: 10.1109/JBHI.2023.3239780 (cit. on p. 24).

JAKUBOWSKI, Jacek et al. “A study on the possible diagnosis of Parkinson’s disease on the basis of facial image analysis”. In: *Electronics* 10.22 (2021), p. 2832 (cit. on pp. 12, 16, 23, 50).

JIN, Bo et al. “Diagnosing Parkinson disease through facial expression recognition: video analysis”. In: *Journal of medical Internet research* 22.7 (2020), e18697 (cit. on p. 12).

LIM, Wee Shin et al. “An integrated biometric voice and facial features for early detection of Parkinson’s disease”. In: *npj Parkinson’s Disease* 8.1 (2022), p. 145 (cit. on pp. 13, 25).

MARTÍNEZ-SÁNCHEZ, Francisco. “Trastornos del habla y la voz en la enfermedad de Parkinson”. In: *Revista de neurologia* 51 (Jan. 2010), pp. 542–550 (cit. on p. 32).

MAYCAS-CEPEDA, Teresa et al. “Hypomimia in Parkinson’s disease: what is it telling us?” In: *Frontiers in Neurology* 11 (2021), p. 603582 (cit. on p. 11).

MCCLEAN, Michael D and TASKO, Stephen M. “Association of orofacial with laryngeal and respiratory motor output during speech”. In: *Experimental brain research* 146 (2002), pp. 481–489 (cit. on pp. 11, 14).

MINISTERIO DE SALUD DE COLOMBIA. *Día Mundial del Parkinson: Colombia se destaca en atención*. Último acceso: 17 de febrero de 2025. 2020. URL: https://www.minsalud.gov.co/Paginas/Dia-Mundial-del-Parkinson-Colombia-se-destaca-en-atencion.aspx (cit. on p. 10).

ORGANIZATION, World Health. *Launch of WHO's Parkinson disease technical brief*. Fecha de acceso: 27 de mayo de 2024. 2022. URL: https://www.who.int/news/item/14-06-2022-launch-of-who-s-parkinson-disease-technical-brief (cit. on p. 27).

OU, Zejin et al. "Global trends in the incidence, prevalence, and years lived with disability of Parkinson's disease in 204 countries/territories from 1990 to 2019". In: *Frontiers in public health* 9 (2021), p. 776847 (cit. on p. 10).

PARKINSON, Fundación de. *Sitio web de la Fundación de Parkinson*. Fecha de acceso: 1 de agosto de 2023. Sin fecha. URL: https://www.parkinson.org/espanol (cit. on pp. 11, 14, 15).

Parkinson's Progression Markers Initiative. Fecha de acceso: 18 de agosto de 2024. Sin fecha. URL: https://www.ppmi-info.org/access-data-specimens/download-data (cit. on p. 25).

PFEIFFER, Ronald F. "Non-motor symptoms in Parkinson's disease". In: *Parkinsonism & related disorders* 22 (2016), S119–S122 (cit. on p. 14).

PRENGER, Margaret et al. "Social symptoms of Parkinson's disease". In: *Parkinson's Disease* 2020 (2020) (cit. on pp. 11, 27).

PRETE, Braedan RJ and OUANOUNOU, Aviv. "Medical management, orofacial findings, and dental care for the patient with Parkinson's disease". In: *J Can Dent Assoc* 87.l10 (2021), pp. 1488–2159 (cit. on p. 14).

RATING SCALES FOR PARKINSON'S DISEASE, Movement Disorder Society Task Force on. "The unified Parkinson's disease rating scale (UPDRS): status and recommendations". In: *Movement Disorders* 18.7 (2003), pp. 738–750 (cit. on pp. 12, 27).

RIZZO, Giovanni et al. “Accuracy of clinical diagnosis of Parkinson disease: a systematic review and meta-analysis”. In: *Neurology* 86.6 (2016), pp. 566–576 (cit. on p. 11).

SKIBIŃSKA, Justyna and HOSEK, Jiri. “Computerized analysis of hypomimia and hypokinetic dysarthria for improved diagnosis of Parkinson’s disease”. In: *Heliyon* 9.11 (2023) (cit. on pp. 13, 26, 50).

SONAWANE, Bhakti and SHARMA, Priyanka. “Review of automated emotion-based quantification of facial expression in Parkinson’s patients”. In: *The Visual Computer* 37 (2021), pp. 1151–1167 (cit. on p. 12).

SVEINBJORNSDOTTIR, Sigurlaug. “The clinical symptoms of Parkinson’s disease”. In: *Journal of neurochemistry* 139 (2016), pp. 318–324 (cit. on p. 11).

TOSIN, Michelle Hyczy S et al. “Nursing and Parkinson’s disease: a scoping review of worldwide studies”. In: *Clinical Nursing Research* 31.2 (2022), pp. 230–238 (cit. on p. 10).

TRACY, John M et al. “Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson’s disease”. In: *Journal of biomedical informatics* 104 (2020), p. 103362 (cit. on p. 12).

TRAN, Du et al. “Learning spatiotemporal features with 3d convolutional networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4489–4497 (cit. on p. 17).

VINOKUROV, Nomi et al. “Quantifying hypomimia in parkinson patients using a depth camera”. In: *International Symposium on Pervasive Computing Paradigms for Mental Health*. Springer. 2015, pp. 63–71 (cit. on p. 15).

WONG, Suzy L; GILMOUR, Heather Lynne, and RAMAGE-MORIN, Pamela L. *Parkinson's disease: Prevalence, diagnosis and impact*. 2014 (cit. on p. 10).

WROGE, Timothy J et al. "Parkinson's disease diagnosis using machine learning and voice". In: *2018 IEEE signal processing in medicine and biology symposium (SPMB)*. IEEE. 2018, pp. 1–7 (cit. on p. 12).

XU, Zhijing et al. "Voiceprint recognition of Parkinson patients based on deep learning". In: *arXiv preprint arXiv:1812.06613* (2018) (cit. on p. 17).

ZHU, Baozhou et al. "An attention module for convolutional neural networks". In: *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part I 30*. Springer. 2021, pp. 167–178 (cit. on p. 22).