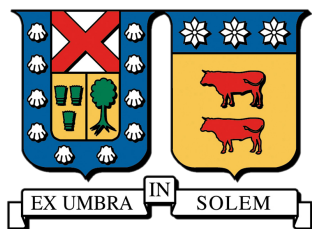


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA  
DEPARTAMENTO DE INFORMÁTICA  
VALPARAÍSO - CHILE



“Robustez y adaptabilidad de los descriptores acústicos  
en la detección de alteraciones fonológicas mediante  
un enfoque plurilingüe y multi-patológico:  
caso Refracted Speech”

Joaquín Ignacio García Ramírez

Memoria para optar al título de Ingeniero Civil en Informática

Profesor Guía: Dr. Raquel Pezoa, UTFSM

Junio - 2025



## CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

### 1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

**Tipo de monografía (marcar una opción):** ☒ Memoria o trabajo de título; ☐ Tesis de Postgrado;

**Título del trabajo:** Robustez y adaptabilidad de los descriptores acústicos en la detección de alteraciones fonológicas mediante un enfoque plurilingüe y multi-patológico: caso Refracted Speech

**Nombre del candidato(a):** Joaquín Ignacio García Ramírez

**Carrera / Grado:** Ingeniería Civil Informática

**Campus:** Casa Central Valparaíso; **Departamento:** Informática

### 2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Raquel Pezoa Rivera, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución

### 3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo **NO contiene información que amerite confidencialidad** y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (embargo) por:

☐ 6 meses; ☐ 12 meses; ☐ 2 años; ☐ 3 años; ☐ 5 años; ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

### 4.- FIRMAS

**Profesor(a) guía o director(a) de memoria o tesis:**

**Fecha:** 01/08/2025

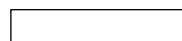
**; Firma:**

**Estudiante o Candidato(a):**

**Fecha:** 01/08/2025

**; Firma:**

*Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.*



## Dedicatoria

Dedicado a mi madre, que a pesar de los miles de kilómetros de distancia, no la he sentido lejos ni un minuto. A mis hermanos, a mi tío Rober, a mi nina y a mi weli. Estas son las personas que me forjaron, me alojaron y me enseñaron a vivir. Dedicado también a Clifford y Maya.

*“I wake up in the morning thinking I’m not gonna make it. Then I laugh inside,  
remembering all the times I felt that way.”*

— Charles Bukowski

## Agradecimientos

Agradezco enormemente a mi familia, quienes no se han dejado hundir por todas las dificultades que nos han tocado. Quienes, a pesar de la enorme distancia física, han sabido mantenerse presentes. Gracias por reconocer todas las cosas que he hecho, de las cuales ni yo mismo me he logrado percatar. Gracias también por el mero hecho de existir, de estar ahí, de dejarme saber que no estoy a la deriva en este extraño mundo.

Gracias también a mis amigos. Un gracias general y no tan general, pues aquellos a quienes más agradezco son un grupo muy pequeño y particular. Son aquellos quienes, de alguna forma u otra, han compartido fragmentos importantes de su vida conmigo. Muchas gracias, en especial, por traer momentos súbitos de claridad a este fabricado tablero de frenesís incalculables.

Gracias a mi profesora guía Raquel Pezoa, por permitirme trabajar con ella en este arduo proyecto, y proveerme de dirección cuando no entendía mucho hacia dónde iba.

Gracias a las experiencias; a las cotidianas, y a las que con violencia rompen la serenidad de la rutina.

Gracias a los recuerdos, o a mi capacidad de recordar, o a los pensamientos, o a mi capacidad de pensar.

Hace tiempo no agradecía a tantas cosas. Quizá incluso sea la primera vez que lo hago. Así que de paso, me agradezco en parte también a mí mismo. Sé que podría haber hecho más. Sé que tal vez lo podría haber hecho mejor. Aún así, lo hice.

Se agradece también al proyecto ANID PIA/APOYO AFB230003.

## Resumen

Las alteraciones fonológicas asociadas a trastornos neurológicos como la disartria o la enfermedad de Parkinson afectan gravemente la comunicación oral, y su diagnóstico automático sigue siendo un desafío en entornos multilingües. Esta tesis propone una evaluación sistemática de distintos descriptores acústicos con el objetivo de identificar aquellos más robustos, interpretables y generalizables para la detección automática de patologías del habla. A partir de un marco experimental basado en Refracted Speech—una aplicación originalmente apuntada al diagnóstico automático de disartria en pacientes angloparlantes—, se implementó un pipeline de extracción de características, selección mediante correlación y RFECV, entrenamiento y evaluación de modelos de aprendizaje automático (Random Forest, SVM, MLP), y análisis explicativo con SHAP sobre múltiples idiomas y patologías. Los resultados revelan que un subconjunto reducido de features—entre ellos tVSA, VAI, Pause Percentage y Autocorr Energy—logra altos niveles de precisión diagnóstica y consistencia transversal. Asimismo, se validó el aporte de características basadas en correlación de señales acústicas, proponiendo alternativas más efectivas que descriptores clásicos como jitter o shimmer. Esta investigación aporta herramientas concretas para el desarrollo de sistemas diagnósticos automáticos, adaptables a contextos clínicos reales y multilingües.

## Abstract

Phonological impairments associated with neurological disorders such as dysarthria and Parkinson’s disease severely impact oral communication, and their automatic diagnosis remains challenging, particularly in multilingual contexts. This thesis proposes a systematic evaluation of acoustic descriptors to identify the most robust, interpretable, and generalizable features for the automatic detection of speech pathologies. Starting from an experimental framework based on Refracted Speech—an application originally aimed at the automatic diagnosis of dysarthria in English-speaking patients—a feature extraction pipeline was implemented, incorporating correlation-based selection methods, Recursive Feature Elimination with Cross-Validation (RFECV), training and evaluation of machine learning models (Random Forest, SVM, MLP), and explanatory SHAP analysis, applied across multiple languages and pathologies. Results indicate that a small subset of acoustic descriptors—including tVSA, VAI, Pause Percentage, and Autocorr Energy—achieve high diagnostic accuracy and cross-contextual consistency. Moreover, the study validates novel features derived from acoustic signal correlation, proposing more effective alternatives compared to traditional descriptors such as jitter or shimmer. Overall, this research provides concrete tools for the development of automatic diagnostic systems, readily adaptable to real-world clinical and multilingual contexts.

## Glosario

**MFCC**: Mel-Frequency Cepstral Coefficients  
**ML**: Machine Learning  
**MLP**: Multi-Layer Perceptron  
**PCA**: Principal Component Analysis  
**RBF**: Radial Basis Function  
**RF**: Random Forest  
**RFECV**: Recursive Feature Elimination with Cross-Validation  
**SHAP**: SHapley Additive exPlanations  
**SVC**: Support Vector Classifier  
**SVM**: Support Vector Machine  
**t-SNE**: t-distributed Stochastic Neighbor Embedding  
**tVSA**: Total Vowel Space Area  
**VAI**: Vowel Articulation Index  
**ACV**: Accidente cerebrovascular  
**ALMA**: Atacama Large Millimeter/submillimeter Array  
**APA**: Evaluación Perceptual Auditiva  
**CNN**: Convolutional Neural Network  
**DFT**: Transformada Discreta de Fourier  
**ELA**: Esclerosis Lateral Amiotrófica  
**EMG**: Electromiografía  
**EP**: Enfermedad de Parkinson  
**F0**: Frecuencia fundamental  
**FS**: Feature Selection  
**GCI**s: Glottal Closure Instants  
**GMM**: Gaussian Mixture Models  
**HC**: Healthy Control  
**HD**: Huntington's Disease  
**LEASgram**: Logarithmic Envelope Analysis Spectrum

# Índice de contenidos

|                                                                                                     |           |
|-----------------------------------------------------------------------------------------------------|-----------|
| Introducción . . . . .                                                                              | 1         |
| <b>1. Definición del Problema</b>                                                                   | <b>3</b>  |
| Definición del Problema . . . . .                                                                   | 3         |
| 1.1. Problema clínico y social . . . . .                                                            | 3         |
| 1.2. Limitaciones de los métodos diagnósticos tradicionales . . . . .                               | 4         |
| 1.3. Limitaciones identificadas en <i>Refracted Speech</i> . . . . .                                | 5         |
| 1.4. Desafío crítico: escasez de datos y generalización de modelos . . . . .                        | 6         |
| 1.5. Necesidad de generalización multi-patológica y multilingüe . . . . .                           | 7         |
| 1.6. Desafío crítico: Interpretabilidad clínica . . . . .                                           | 8         |
| <b>2. Marco conceptual</b>                                                                          | <b>10</b> |
| 2.1. Definición clínica y neurológica de la disartria . . . . .                                     | 10        |
| 2.1.1. Tipos, causas y manifestaciones de la disartria . . . . .                                    | 11        |
| 2.1.2. Disartria hipocinética y su vínculo con la Enfermedad de Parkinson . . . . .                 | 12        |
| 2.2. Mel-Frequency Cepstral Coefficients (MFCCs) . . . . .                                          | 13        |
| 2.2.1. Metodología de Cálculo . . . . .                                                             | 13        |
| 2.2.2. Representación y Agregación . . . . .                                                        | 15        |
| 2.2.3. Justificación de Uso . . . . .                                                               | 15        |
| 2.3. Features extraídos manualmente . . . . .                                                       | 15        |
| 2.3.1. Acoustic Features . . . . .                                                                  | 15        |
| 2.3.2. Acoustic Landmarks . . . . .                                                                 | 16        |
| 2.3.3. Características de Articulación Vocal . . . . .                                              | 17        |
| 2.3.4. Jitter y Shimmer . . . . .                                                                   | 18        |
| 2.3.5. Descriptores Basados en Correlación: Propuesta de Nuevas Características Acústicas . . . . . | 20        |
| 2.3.6. Extracción y análisis de características lingüísticas . . . . .                              | 23        |
| 2.4. Modelos de Clasificación en Machine Learning . . . . .                                         | 24        |
| 2.5. Estado del Arte y Enfoque del Proyecto . . . . .                                               | 28        |
| <b>3. Propuesta de solución</b>                                                                     | <b>30</b> |
| 3.1. Descripción General del Pipeline . . . . .                                                     | 30        |
| 3.2. Conjunto de Datos Utilizados . . . . .                                                         | 31        |
| 3.3. Extracción y preprocesamiento de características . . . . .                                     | 32        |
| 3.3.1. Descriptores basados en autocorrelación y correlación cruzada . . . . .                      | 37        |
| 3.3.2. Selección de características . . . . .                                                       | 39        |

|           |                                                                                                     |           |
|-----------|-----------------------------------------------------------------------------------------------------|-----------|
| 3.4.      | Modelos de clasificación y estrategias de entrenamiento . . . . .                                   | 43        |
| 3.4.1.    | Modelos de clasificación utilizados . . . . .                                                       | 44        |
| 3.4.2.    | Clasificación multiclase directa . . . . .                                                          | 44        |
| 3.4.3.    | Clasificación en dos etapas (two-tiered) . . . . .                                                  | 44        |
| 3.5.      | Estrategias de evaluación . . . . .                                                                 | 45        |
| 3.5.1.    | Métricas de evaluación . . . . .                                                                    | 46        |
| 3.5.2.    | Criterios comparativos entre modelos . . . . .                                                      | 46        |
| 3.6.      | Interpretabilidad del modelo . . . . .                                                              | 47        |
| 3.6.1.    | Cálculo de valores SHAP . . . . .                                                                   | 47        |
| <b>4.</b> | <b>Resultados Experimentales</b>                                                                    | <b>49</b> |
| 4.1.      | Justificación del Enfoque Experimental sin MFCC . . . . .                                           | 49        |
| 4.2.      | Comparación de métricas entre modelos . . . . .                                                     | 49        |
| 4.2.1.    | Descripción General . . . . .                                                                       | 49        |
| 4.2.2.    | Análisis de la clasificación multiclase directa . . . . .                                           | 50        |
| 4.2.3.    | Análisis de la clasificación mediante pipeline escalonada (two-tiered) .                            | 53        |
| 4.3.      | Impacto del Umbral de Correlación en la Arquitectura Two-Tiered . . . . .                           | 56        |
| 4.3.1.    | Resultados en la Fase Binaria . . . . .                                                             | 56        |
| 4.3.2.    | Resultados en la Fase Multiclase . . . . .                                                          | 60        |
| 4.3.3.    | Síntesis global . . . . .                                                                           | 63        |
| 4.4.      | Selección de Características en la Two-Tiered Pipeline . . . . .                                    | 64        |
| 4.4.1.    | Sin MFCC . . . . .                                                                                  | 64        |
| 4.4.2.    | Con MFCC . . . . .                                                                                  | 66        |
| 4.5.      | Interpretabilidad de los Modelos: Análisis SHAP de la Selección de Características . . . . .        | 68        |
| 4.5.1.    | Análisis SHAP para Clasificación Multiclase Directa (sin MFCC, umbral 0.95) . . . . .               | 68        |
| 4.5.2.    | Análisis SHAP para Clasificación mediante arquitectura two-tiered (sin MFCC), umbral 0.95 . . . . . | 74        |
| 4.5.3.    | Clasificación de universalidad multilingüe y multipatológica de cada feature . . . . .              | 83        |
| <b>5.</b> | <b>Conclusiones</b>                                                                                 | <b>86</b> |

# Índice de figuras

|                                                                                                                                         |    |
|-----------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1. Prevalencia de disartria en distintos contextos neurológicos según la literatura (González & Bevilacqua, 2012). . . . .            | 4  |
| 1.2. Espectrograma en escala Mel generado a partir de una muestra de voz perteneciente a un hablante sano. . . . .                      | 9  |
| 2.1. Ejemplo de detección automática de <i>acoustic landmarks</i> utilizando la herramienta SpeechMark. . . . .                         | 17 |
| 2.2. Visualización conceptual de la correlación cruzada entre dos señales. . . . .                                                      | 21 |
| 2.3. Implementación de un filtro adaptado (matched filter) usando correlación cruzada. . . . .                                          | 22 |
| 2.4. Visualización en 3D de la frontera de decisión generada por un modelo SVM binario lineal . . . . .                                 | 25 |
| 2.5. Distribución de las muestras tras proyección con t-SNE sobre dos dimensiones. . . . .                                              | 26 |
| 2.6. Ejemplo generado de red neuronal multicapa (MLP) con dos capas ocultas. . . . .                                                    | 27 |
| 3.1. Matriz de correlación entre características acústicas prosódicas. . . . .                                                          | 33 |
| 3.2. Matriz de correlación entre características acústicas derivadas de <i>landmarks</i> . . . . .                                      | 34 |
| 3.3. Matriz de correlación entre características de articulación vocal. . . . .                                                         | 35 |
| 3.4. Matriz de correlación entre descriptores basados en autocorrelación, correlación cruzada, <i>jitter</i> y <i>shimmer</i> . . . . . | 36 |
| 3.5. Proyección t-SNE del subconjunto de características acústicas de tipo landmark. . . . .                                            | 39 |
| 3.6. Proyección PCA del subconjunto de características acústicas de tipo landmark. . . . .                                              | 40 |
| 3.7. Curva de validación cruzada del algoritmo RFECV utilizando LinearSVC como modelo base. . . . .                                     | 42 |
| 3.8. Curva de validación cruzada del algoritmo RFECV utilizando Random Forest como modelo base . . . . .                                | 42 |
| 3.9. Comparación directa de la precisión de validación cruzada entre LinearSVC y Random Forest . . . . .                                | 43 |
| 3.10. Esquema general del pipeline de clasificación en dos etapas . . . . .                                                             | 45 |
| 3.11. Ejemplo ilustrativo del funcionamiento de SHAP en modelos de árboles. . . . .                                                     | 48 |
| 4.1. Heatmap de métricas para clasificación multiclase directa con (arriba) y sin MFCC (abajo). Umbral Pearson 0.95. . . . .            | 50 |
| 4.2. Heatmap de métricas para clasificación multiclase directa con (arriba) y sin MFCC (abajo). Umbral Pearson 0.70. . . . .            | 51 |

|                                                                                                                                                                    |    |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 4.3. Heatmap de métricas para clasificación multiclase directa con (arriba) y sin MFCC (abajo). Umbral Pearson 0.40. . . . .                                       | 52 |
| 4.4. Heatmap de métricas para clasificación multiclase directa utilizando únicamente MFCC. Umbral Pearson 0.95. . . . .                                            | 53 |
| 4.5. Heatmap de métricas para pipeline two-tiered. Arriba: con MFCC, abajo: sin MFCC. Umbral Pearson 0.95. . . . .                                                 | 54 |
| 4.6. Heatmap de métricas para pipeline two-tiered. Arriba: con MFCC, abajo: sin MFCC. Umbral Pearson 0.70. . . . .                                                 | 55 |
| 4.7. Heatmap de métricas para pipeline two-tiered. Arriba: con MFCC, abajo: sin MFCC. Umbral Pearson 0.40. . . . .                                                 | 56 |
| 4.8. Evolución de la <b>precisión</b> por clase y modelo a través de distintos umbrales de correlación de Pearson (fase binaria, two-tiered sin MFCC). . . . .     | 57 |
| 4.9. Evolución del <b>recall</b> por clase y modelo según el umbral de correlación de Pearson (fase binaria, two-tiered sin MFCC). . . . .                         | 58 |
| 4.10. Evolución del <b>f1-score</b> por clase y modelo en función del umbral de correlación de Pearson (fase binaria, two-tiered sin MFCC). . . . .                | 58 |
| 4.11. Accuracy binaria global para cada modelo en función del umbral de Pearson (Two-Tiered, sin MFCC). . . . .                                                    | 59 |
| 4.12. Evolución de la <b>precisión</b> por clase y modelo a través de distintos umbrales de correlación de Pearson (fase multiclase, two-tiered sin MFCC). . . . . | 61 |
| 4.13. Evolución del <b>recall</b> por clase y modelo según el umbral de correlación de Pearson (fase multiclase, two-tiered sin MFCC). . . . .                     | 61 |
| 4.14. Evolución del <b>f1-score</b> por clase y modelo en función del umbral de correlación de Pearson (fase multiclase, two-tiered sin MFCC). . . . .             | 62 |
| 4.15. Accuracy multiclase global para cada modelo en función del umbral de Pearson (Two-Tiered, sin MFCC). . . . .                                                 | 62 |
| 4.16. Comparación RFECV en la fase binaria sin MFCC. . . . .                                                                                                       | 64 |
| 4.17. Comparación RFECV en la fase multiclase sin MFCC. . . . .                                                                                                    | 65 |
| 4.18. Comparación RFECV en la fase binaria con MFCC. . . . .                                                                                                       | 66 |
| 4.19. Comparación RFECV en la fase multiclase con MFCC. . . . .                                                                                                    | 67 |
| 4.20. Comparación de importancia SHAP para la clase Dysarthria-English. . . . .                                                                                    | 69 |
| 4.21. Comparación de importancia SHAP para la clase Parkinson-Italian. . . . .                                                                                     | 70 |
| 4.22. Comparación de importancia SHAP para la clase Healthy-Spanish. . . . .                                                                                       | 71 |
| 4.23. Beeswarm plots para Healthy-English: dispersión y perfil de impacto de los descriptores. . . . .                                                             | 72 |
| 4.24. Beeswarm plots para Parkinson-Spanish: dispersión y perfil de impacto de los descriptores. . . . .                                                           | 73 |
| 4.25. Comparación de <i>dependence plots</i> SHAP para la característica Loudness Variability en modelos Random Forest y Linear SVC. . . . .                       | 74 |
| 4.26. Barplots de valores medios absolutos de SHAP para la fase binaria, comparando LinearSVC y Random Forest. . . . .                                             | 75 |
| 4.27. Beeswarm plots de SHAP para la fase binaria, comparando LinearSVC y Random Forest. . . . .                                                                   | 76 |
| 4.28. Importancia de las características para la clase Dysarthria-English (fase multiclase, Random Forest). . . . .                                                | 78 |

|                                                                                                                     |    |
|---------------------------------------------------------------------------------------------------------------------|----|
| 4.29. Importancia de las características para la clase Parkinson-Italian (fase multi-clase, Random Forest). . . . . | 79 |
| 4.30. Importancia de las características para la clase Parkinson-Spanish (fase multi-clase, Random Forest). . . . . | 80 |
| 4.31. Importancia de las características para la clase Parkinson-English (fase multi-clase, Random Forest). . . . . | 81 |

# Índice de tablas

|                                                                                                                                                            |    |
|------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1. Distribución demográfica y clínica de los participantes del estudio de Vásquez-Correa et al. (2021). . . . .                                          | 8  |
| 3.1. Resumen de correlaciones altas entre características acústicas . . . . .                                                                              | 37 |
| 3.2. Métricas de evaluación utilizadas para la clasificación . . . . .                                                                                     | 46 |
| 4.1. Universalidad y utilidad binaria de los descriptores acústicos manuales según la importancia SHAP y la evidencia multilingüe/multipatológica. . . . . | 85 |

## Introducción

Los trastornos del habla constituyen un grupo heterogéneo de condiciones que afectan significativamente la capacidad comunicativa de una persona, repercutiendo de forma profunda en su calidad de vida. No estamos hablando solo de la dificultad para pronunciar ciertas palabras: cuando alguien pierde, en parte o totalmente, la capacidad de expresarse verbalmente, se afecta también su autoestima, su vida social, su desarrollo emocional y muchas veces hasta sus oportunidades académicas y laborales. La comunicación oral no es simplemente una herramienta para transmitir información: es un eje central en la forma en que construimos identidad, formamos vínculos y participamos activamente en la sociedad [34].

Frente a esto, la detección temprana de los trastornos del habla se vuelve una pieza clave. Una intervención a tiempo no solo permite frenar el deterioro comunicativo, sino que puede marcar la diferencia en la integración personal, emocional y social del paciente. Especialmente en poblaciones infantiles o en etapas de neuroplasticidad activa, un diagnóstico temprano puede cambiar radicalmente el pronóstico y la calidad de vida futura.

Dentro del amplio abanico de trastornos del habla, la disartria destaca por su frecuencia y por las consecuencias que trae consigo. Se trata de una alteración motora que afecta la articulación, la fonación, la prosodia y, en general, la inteligibilidad del habla. En Chile, representa aproximadamente el 54 % del total de casos diagnosticados de trastornos de la comunicación [11], con más de 30.000 personas afectadas. Esta cifra no solo habla de un problema clínico serio, sino también de un desafío social: ¿Cómo aseguramos acceso a un diagnóstico oportuno y certero para quienes lo necesitan? ¿Cómo reducimos la brecha que separa a quienes tienen acceso a especialistas de quienes no?

La disartria puede presentarse como consecuencia de una larga lista de condiciones neurológicas: Parkinson, esclerosis múltiple, ELA, accidentes cerebrovasculares, traumatismos craneoencefálicos, entre otros. Y si bien la causa varía, lo que tienen en común estas patologías es el impacto que generan en la motricidad del habla. Esto produce un espectro amplio de síntomas que pueden variar en intensidad, pero que terminan afectando siempre la claridad y fluidez del habla [11], [5].

El proyecto *Refracted Speech* [30] desarrollado inicialmente como parte de la Feria de Software 2024, nació con la intención de aportar una solución concreta a este problema. Su objetivo fue ofrecer una herramienta accesible, rápida y confiable para detectar automáticamente la disartria, especialmente pensada para personas que no cuentan con acceso fácil a un especialista. Más allá del diagnóstico, el sistema busca también apoyar el proceso terapéutico, ofreciendo ejercicios de vocalización adaptados a la severidad de los síntomas detectados. Sin embargo, en su fase inicial, el modelo fue entrenado exclusivamente con el dataset TORGO [5], lo que limitó su uso al idioma inglés y dejó en evidencia una debilidad importante: su falta de generalización a otros idiomas y a otras patologías del habla.

Esta limitación no es exclusiva de *Refracted Speech*. En general, el campo del reconocimiento automático de trastornos del habla se enfrenta a una carencia estructural: faltan

datasets grandes, diversos y clínicamente validados que permitan entrenar modelos robustos. La mayoría de los corpus existentes tienen restricciones importantes tanto en tamaño como en variedad lingüística, lo que dificulta seriamente la capacidad de generalización de los modelos actuales de IA y Machine Learning [35], [28].

En este contexto, la presente investigación propone abordar esta limitación desde un enfoque práctico y bien definido: la evaluación sistemática de un conjunto reducido —pero potente— de descriptores acústicos, diseñados para ser a la vez clínicamente interpretables y capaces de generalizar entre múltiples patologías e idiomas. La idea central es que ciertas alteraciones neuromotoras subyacentes al habla patológica presentan patrones acústicos comunes que pueden ser detectados más allá de las diferencias fonológicas superficiales entre lenguas [37], [16].

Por otra parte, el desarrollo de tecnologías basadas en ML ofrece una oportunidad real y concreta para automatizar y mejorar la detección de estos trastornos. Técnicas como Support Vector Machines (SVM), Random Forests (RF), Redes Neuronales Convolucionales (CNN) e incluso modelos más recientes como Transformers han demostrado un rendimiento destacable en tareas de clasificación clínica del habla [5]. Una de sus mayores ventajas es la capacidad para reducir significativamente el número de variables necesarias para hacer una predicción confiable. Esto no solo mejora el desempeño computacional, sino que también facilita la interpretación clínica de los resultados, permitiendo a los especialistas focalizar su atención en un conjunto acotado de indicadores acústicos claros y medibles [9], [12].

En este sentido, la tesis que aquí se presenta tiene como objetivo aportar al desarrollo de sistemas automáticos de diagnóstico que sean robustos, interpretables y aplicables en entornos clínicos reales. No se trata solo de alcanzar una buena precisión en condiciones controladas, sino de construir herramientas que realmente puedan ser utilizadas por profesionales de la salud —y eventualmente por los mismos pacientes— para facilitar el acceso temprano al diagnóstico. Esto es especialmente relevante en contextos donde los recursos son limitados y donde un sistema automático puede funcionar como un primer filtro, ayudando a identificar casos prioritarios y optimizando el uso de los recursos disponibles.

# Capítulo 1

## Definición del Problema

### 1.1. Problema clínico y social

Las patologías del habla constituyen un conjunto de trastornos comunicativos que afectan gravemente la calidad de vida, dificultando la integración social, educativa y laboral de quienes las padecen. Entre ellas, la disartria ocupa un lugar particularmente relevante, al ser un trastorno motor del habla provocado por alteraciones neurológicas como accidentes cerebrovasculares (ACV), enfermedad de Parkinson, parálisis cerebral o esclerosis lateral amiotrófica (ELA) (ver Figura 1.1). En Chile, su prevalencia es especialmente alta: representa aproximadamente el 54 % de los trastornos comunicativos diagnosticados [11]. Junto con esto, se estima que unas 30.000 personas conviven con disartria en el país [7], lo que resalta la urgencia de abordar esta problemática con herramientas eficaces.

Las dificultades comunicativas derivadas de estos trastornos no solo generan aislamiento social y una profunda merma en la calidad de vida, sino que también interfieren con el desempeño académico y laboral. Un diagnóstico temprano y preciso es clave para iniciar intervenciones terapéuticas que logren mitigar estos impactos. Sin embargo, el acceso a tales diagnósticos suele estar restringido, especialmente en zonas rurales o sectores donde escasean profesionales de la fonoaudiología. Incluso en regiones donde sí hay especialistas, las barreras económicas pueden dificultar el acceso a estos servicios.

Aunque estas limitaciones no siempre impidan completamente el proceso diagnóstico, sí lo entorpecen de forma significativa, lo que puede disuadir a muchas familias de iniciar o sostener tratamientos a largo plazo. De ahí surge la necesidad de diseñar herramientas que amplíen el acceso a evaluaciones o pre-evaluaciones, promoviendo así una mayor equidad en salud.

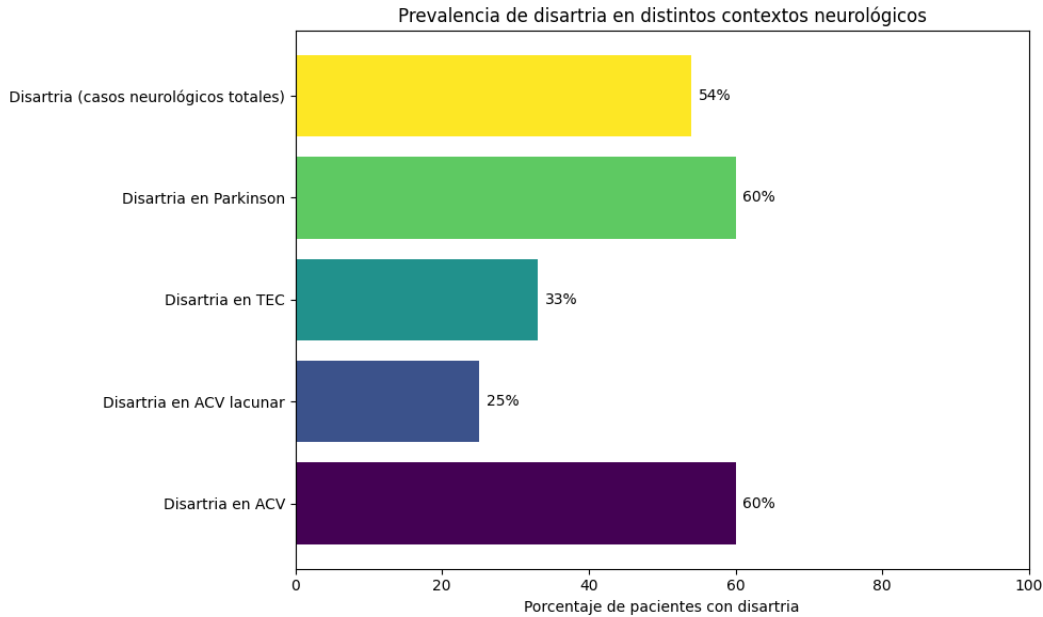


Figura 1.1: Prevalencia de disartria en distintos contextos neurológicos según la literatura (González & Bevilacqua, 2012). Se observa que la disartria aparece con alta frecuencia en enfermedades como el Parkinson y el ACV, lo que refuerza su relevancia clínica.

## 1.2. Limitaciones de los métodos diagnósticos tradicionales

El enfoque clínico dominante para diagnosticar trastornos del habla es la evaluación perceptual auditiva (APA), donde un profesional —generalmente un fonoaudiólogo— evalúa el desempeño del paciente mediante tareas como repetición de palabras, lectura en voz alta o ejercicios de articulación. Si bien esta metodología ha demostrado ser útil, presenta limitaciones críticas:

- **Costo y accesibilidad:** Requiere especialistas capacitados, muchas veces inaccesibles en zonas rurales o para comunidades de bajos recursos.
- **Subjetividad y variabilidad:** Depende del juicio perceptual del evaluador, lo que introduce una variabilidad significativa tanto entre evaluadores distintos como en un mismo evaluador en distintos momentos [35]. Aunque se trate de un profesional experimentado, los sentidos humanos no son inmunes a sesgos o inconsistencias.
- **Consumo de tiempo:** El proceso puede ser largo y agotador, lo que limita su aplicabilidad a gran escala y dificulta el diagnóstico temprano. Además, muchos pacientes —especialmente niños— no toleran sesiones prolongadas, lo que enlentece aún más el proceso diagnóstico y retrasa el inicio de terapias.

Estas limitaciones evidencian la necesidad de sistemas complementarios que sean más accesibles, objetivos, rápidos y reproducibles.

### 1.3. Limitaciones identificadas en *Refracted Speech*

El proyecto Refracted Speech surgió con el objetivo de construir una herramienta automatizada que facilitara el diagnóstico preliminar de disartria, utilizando técnicas de aprendizaje automático —específicamente, Support Vector Machines (SVM)— entrenadas exclusivamente con el dataset TORGO. Este corpus, uno de los pocos disponibles públicamente con muestras de voz etiquetadas de pacientes con disartria, está compuesto por grabaciones en inglés provenientes de hablantes nativos con y sin patologías del habla. La elección de este dataset respondió a una limitación concreta durante la etapa de desarrollo: la falta de corpora alternativos que fueran suficientemente grandes, balanceados y clínicamente validados como para permitir el entrenamiento efectivo del modelo.

A pesar de estas restricciones, el sistema logró una accuracy destacable —cercana al 89 %— en la detección automática de disartria dentro de ese conjunto de datos [5]. Sin embargo, el enfoque adoptado en esta primera versión presenta dos limitaciones fundamentales que comprometen su aplicabilidad clínica más allá del contexto de experimentación:

- **Restricción lingüística:** Al haber sido entrenado exclusivamente con datos en inglés, el modelo no generaliza adecuadamente a otros idiomas. Esto impide su utilización directa en entornos hispanohablantes como el chileno —donde fue originalmente concebido—, o en contextos multilingües donde se espera que el sistema funcione con distintos acentos, estructuras fonológicas y normas fonotácticas. La variabilidad lingüística no es solo una cuestión superficial: muchas de las características acústicas que un modelo aprende a reconocer están profundamente ancladas en la estructura del idioma con el que fue entrenado. Así, aplicar el modelo fuera de ese dominio introduce un riesgo importante de errores diagnósticos, lo que limita gravemente su escalabilidad y su implementación práctica en contextos clínicos reales.
- **Foco diagnóstico limitado:** El sistema fue diseñado para detectar únicamente la disartria general, sin incorporar distinciones clínicas entre diferentes tipos o manifestaciones específicas de la patología. Por ejemplo, no considera variaciones como la disartria hipocinética asociada a la enfermedad de Parkinson [7], ni explora diferencias con otras alteraciones del habla neurológicas como la apraxia o la afasia. Esta simplificación, si bien comprensible desde el punto de vista técnico en una primera etapa, no responde a las necesidades diagnósticas reales, donde distinguir entre distintas condiciones es crucial para definir tratamientos personalizados. La heterogeneidad clínica de los trastornos del habla exige herramientas que no solo detecten si hay una alteración, sino que puedan caracterizar su tipo, severidad y evolución probable.
- **Dependencia estructural del dataset:** A pesar de que Refracted Speech logró man-

tener un desempeño sólido en pruebas realizadas con hablantes externos y en condiciones con cierto nivel de ruido ambiental, no deja de ser relevante señalar que su entrenamiento se basó exclusivamente en el dataset TORGO. Esto implica que, aunque el modelo demostró cierta capacidad de generalización, sus límites aún no han sido explorados completamente frente a la diversidad de escenarios clínicos y lingüísticos que se pueden presentar en la práctica real. El entorno de grabación controlado del dataset original, los tipos de micrófono utilizados, las tareas fonológicas específicas y el perfil demográfico de los participantes siguen constituyendo un marco restringido. Por lo tanto, si bien el modelo mostró una resistencia inicial al cambio de condiciones, avanzar hacia una mayor robustez y validación cruzada con datos más diversos continúa siendo una necesidad fundamental para su evolución hacia una herramienta clínica confiable y ampliamente aplicable.

En conjunto, estas limitaciones subrayan la necesidad urgente de rediseñar el enfoque del sistema con miras a ampliar su cobertura lingüística y clínica, incorporar nuevas fuentes de datos más diversas, y fortalecer la capacidad del modelo para adaptarse a distintos contextos sin comprometer su precisión. Si bien Refracted Speech demostró que es posible automatizar parte del proceso diagnóstico con buenos resultados iniciales, su versión actual está lejos de constituir una herramienta lista para su implementación en ambientes reales de evaluación clínica. La siguiente etapa del proyecto debe, por tanto, abordar estos puntos de forma explícita, con énfasis en la generalización, robustez e interpretabilidad del sistema.

## **1.4. Desafío crítico: escasez de datos y generalización de modelos**

Una de las barreras más persistentes en el desarrollo de sistemas automáticos para el diagnóstico de patologías del habla es la escasez de datos clínicos etiquetados, particularmente en distintos idiomas y patologías. La mayoría de los datasets existentes son reducidos y están sesgados, lo que dificulta la generalización de los modelos. Frente a esta limitación, resulta clave identificar un conjunto reducido pero potente de características acústicas que no solo tengan alta capacidad discriminativa, sino que también sean clínicamente interpretables y transferibles entre contextos.

El problema se agudiza considerando la gran heterogeneidad de los trastornos del habla y la alta dimensionalidad del espacio de salida (output space) que generan [28]. Esto complica tanto el desarrollo como la validación de sistemas automáticos, ya que exige datasets grandes, diversos y bien etiquetados, los cuales rara vez están disponibles. Además, muchos de estos conjuntos de datos de voz son de uso restringido por razones éticas, dado que contienen datos sensibles, lo que dificulta aún más la replicación y comparación de resultados entre estudios.

Ante este escenario, emergen enfoques que buscan reducir la dependencia de grandes volúmenes de datos. Uno de los más prometedores es el uso de modelos fundacionales co-

mo Whisper, que pueden adaptarse a tareas clínicas mediante *fine-tuning* sobre corpus más pequeños pero bien curados. Este tipo de modelos aprovecha conocimiento previo para generalizar más allá de las muestras de entrenamiento. Un ejemplo notable es el estudio de Sanguedolce et al. [28], donde un modelo ajustado con datos de pacientes post-ACV logró extender su rendimiento a otras condiciones como demencia y disartria, demostrando un grado de transferencia patológica impensado con técnicas tradicionales.

## 1.5. Necesidad de generalización multi-patológica y multilingüe

Una de las hipótesis centrales de este proyecto es que un conjunto cuidadosamente seleccionado de descriptores acústicos puede permitir la distinción entre múltiples trastornos del habla, incluso en contextos multilingües. Por ejemplo, aunque el Parkinson puede causar disartria, esta condición tiene características acústicas específicas que permiten diferenciarla de otros cuadros disártricos [7].

Desarrollar modelos que funcionen en contextos multilingües y multi-patológicos implica enfrentarse a la alta variabilidad del habla humana, tanto en lo lingüístico como en lo clínico. Aún dentro de una misma lengua, existen variaciones fonológicas, dialectales y prosódicas significativas. Y cuando se agregan alteraciones patológicas —que afectan aspectos como articulación, ritmo, prosodia e inteligibilidad— el espacio de salida se vuelve aún más complejo.

Como señalan Irizarry-Pérez et al. [16], incluso en niños bilingües con trastornos fonológicos, la transferencia cruzada entre lenguas puede tanto facilitar como obstaculizar el diagnóstico. Esto sugiere que los modelos deben ser lo suficientemente flexibles como para capturar patrones profundos de la producción vocal, más allá de los detalles superficiales de cada lengua.

Por ello, la búsqueda de descriptores acústicos universales —o al menos compartidos entre lenguas y patologías— representa una línea de investigación estratégica. Estudios como el de Vásquez-Correa et al. [37] han demostrado que técnicas de aprendizaje por transferencia pueden mejorar significativamente la precisión diagnóstica en escenarios multilingües, siempre que los modelos base hayan sido entrenados sobre corpus suficientemente representativos. Lograr esto podría tener un enorme impacto en entornos clínicos donde el acceso a especialistas es limitado.

|                       |   | PD-Spanish  |             | PD-German   |             | PD-Czech    |             | HD-Czech    |             |
|-----------------------|---|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                       |   | PD          | HC          | PD          | HC          | PD          | HC          | HD          | HC          |
| Subjects              | M | 25          | 25          | 47          | 44          | 30          | 30          | 20          | 20          |
|                       | F | 25          | 25          | 41          | 44          | 30          | 30          | 20          | 20          |
| Range of age          | M | 33–81       | 31–86       | 44–82       | 26–83       | 43–82       | 41–77       | 23–67       | 26–70       |
|                       | F | 49–75       | 49–76       | 42–84       | 28–85       | 41–72       | 40–79       | 27–69       | 27–69       |
| Age                   | M | 61.3 (11.4) | 60.5 (11.6) | 66.7 (8.7)  | 63.8 (12.7) | 65.3 (9.6)  | 60.3 (11.5) | 47.7 (12.2) | 48.3 (12.3) |
|                       | F | 60.7 (7.3)  | 61.4 (7.0)  | 66.2 (9.7)  | 62.6 (15.2) | 60.1 (8.7)  | 63.5 (11.1) | 49.5 (14.1) | 50.1 (13.9) |
| MDS-(UPDRS-III/UHDRS) | M | 37.8 (22.1) | –           | 22.1 (9.9)  | –           | 21.4 (11.5) | –           | 26.8 (12.7) | –           |
|                       | F | 37.6 (14.1) | –           | 23.3 (12.0) | –           | 18.1 (9.7)  | –           | 27.1 (10.7) | –           |
| MDS speech item       | M | 1.4 (0.9)   | –           | 1.4 (0.6)   | –           | 0.9 (0.5)   | –           | 0.9 (0.3)   | –           |
|                       | F | 1.3 (0.8)   | –           | 1.2 (0.5)   | –           | 0.7 (0.6)   | –           | 0.7 (0.5)   | –           |

Cuadro 1.1: Distribución demográfica y clínica de los participantes del estudio de Vásquez-Correa et al. (2021).

## 1.6. Desafío crítico: Interpretabilidad clínica

Más allá de lograr buenos resultados en métricas de precisión, los sistemas de diagnóstico automático deben ser comprensibles para los profesionales de la salud. Este es un requisito central en entornos clínicos reales, donde las decisiones deben poder ser auditadas, explicadas y justificadas. Los MFCCs (Mel-Frequency Cepstral Coefficients), por ejemplo, son altamente eficaces desde el punto de vista computacional, pero su nivel de abstracción los hace poco útiles desde la perspectiva interpretativa. Son difíciles de traducir en términos clínicos concretos, lo que limita su utilidad en el trabajo cotidiano de un especialista. Esto se ve ejemplificado en la Figura 1.2.

Esto evidencia la necesidad de desarrollar características acústicas que, además de tener poder discriminativo, conserven trazabilidad semántica. Es decir, que permitan vincular los resultados del modelo con desviaciones fonológicas observables, como irregularidades en la fonación o distorsiones rítmicas.

En esta línea, la selección de características juega un rol fundamental. Técnicas como la Eliminación Recursiva de Características con Validación Cruzada (RFECV) permiten reducir la dimensionalidad sin comprometer el rendimiento. Además, combinaciones entre algoritmos como Random Forest y SVM han demostrado generar representaciones más compactas y explicables [9].

Guyon y Elisseeff destacan que una buena selección de *features* no solo mejora la precisión, sino que también permite entender mejor la estructura interna de los datos. Y en contextos médicos, esta comprensión es esencial para generar confianza en el sistema. No se trata solo de predecir con precisión, sino de construir modelos que hablen el mismo idioma que los profesionales clínicos [12].

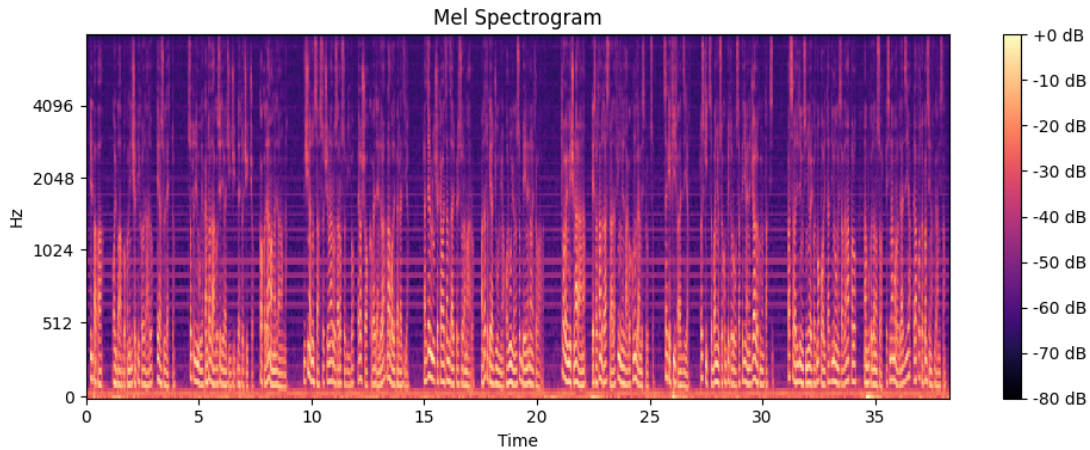


Figura 1.2: **Imagen propia.** Espectrograma en escala Mel generado a partir de una muestra de voz perteneciente a un hablante sano. Se visualiza la energía distribuida a lo largo del tiempo y el espectro de frecuencias en escala perceptual. Este tipo de representación es ampliamente utilizada en tareas de clasificación automática del habla debido a su capacidad de capturar componentes acústicos relevantes, aunque presenta limitaciones de interpretabilidad clínica directa.

# Capítulo 2

## Marco conceptual

### 2.1. Definición clínica y neurológica de la disartria

La disartria es, en términos simples pero clínicamente precisos, un trastorno motor del habla que tiene su origen en el sistema nervioso. Aparece cuando las lesiones en el sistema nervioso —ya sea central o periférico— comprometen la ejecución de los movimientos necesarios para una producción vocal normal. Es importante aclarar que no se trata de una alteración del lenguaje en sí mismo (como ocurre con la afasia), sino de una dificultad específica para realizar físicamente el habla. En otras palabras, el paciente sabe perfectamente lo que quiere decir, pero su sistema motor no logra ejecutar correctamente los movimientos necesarios para expresarlo verbalmente [11].

Desde la perspectiva clínica, las disartrias pueden manifestarse con diferentes grados de severidad —desde una afectación muy leve hasta la anartria, donde el habla es prácticamente inexistente— y tienen una diversidad considerable de causas: desde accidentes cerebrovasculares (ACV) y traumatismos craneoencefálicos (TEC), hasta enfermedades neurodegenerativas como la esclerosis lateral amiotrófica (ELA) o la enfermedad de Parkinson.

El compromiso funcional depende directamente de dónde se localice y qué tan extenso sea el daño, pudiendo afectar de manera diferente los procesos motores básicos del habla: respiración, fonación, articulación, resonancia y prosodia. Estos elementos —que en la comunicación diaria funcionan de manera automática— requieren una coordinación extremadamente fina y continua de múltiples estructuras neuromusculares. Cuando se alteran, el impacto va mucho más allá de la simple inteligibilidad del habla: afecta directamente la autoestima, la participación social y la calidad de vida de las personas [11].

Estos subsistemas del habla trabajan juntos de forma dinámica:

- **Respiración:** proporciona el soporte aéreo fundamental para la producción del habla. Cuando el control respiratorio se ve alterado, puede afectar tanto la intensidad como

la duración del habla.

- **Fonación:** es la producción de sonido a través de la vibración de las cuerdas vocales. Las disfunciones en este subsistema generan voces débiles, ásperas o monótonas.
- **Articulación:** se refiere a la coordinación precisa de labios, lengua, paladar y mandíbula para moldear los sonidos del habla. Su deterioro lleva a imprecisión consonántica y reduce considerablemente la inteligibilidad.
- **Resonancia:** corresponde al equilibrio acústico entre las cavidades orales y nasales. Las alteraciones aquí pueden dar lugar a hipernasalidad o hiponasalidad.
- **Prosodia:** implica el ritmo, la entonación y la acentuación del habla. La monotonía, la poca variación melódica y los patrones de habla robotizados suelen ser claros indicadores de afectación prosódica.

### 2.1.1. Tipos, causas y manifestaciones de la disartria

La clasificación de las disartrias suele organizarse según la fisiopatología predominante, considerando el tipo de alteración neuromuscular que genera el trastorno del habla:

- **Disartria flácida:** causada por lesión de la motoneurona inferior, se caracteriza por debilidad e hipotonía. La voz tiende a ser apagada, con hipernasalidad y disminución del volumen [7].
- **Disartria espástica:** resultado de lesiones bilaterales de la vía corticoespinal, se presenta con rigidez, habla forzada y distorsión consonántica [11].
- **Disartria atáxica:** asociada a daño cerebeloso. Produce habla entrecortada, con alteraciones en la prosodia y en la coordinación de los movimientos articulatorios [11].
- **Disartria hipercinética:** asociada a trastornos extrapiramidales como la enfermedad de Huntington. Se caracteriza por alteraciones prosódicas impredecibles, movimientos involuntarios que afectan la fonación, y fenómenos acústicos como *tremor*, *flutter* o variaciones rítmicas abruptas. Estos patrones han sido identificados en análisis espectrales de pacientes con patologías extrapiramidales, donde se observan inestabilidades vocales y afectaciones en el control respiratorio y articulatorio [19].
- **Disartria hipocinética:** fuertemente asociada a la enfermedad de Parkinson. Se caracteriza por hipofonía, monotonía, ritmo acelerado y articulación imprecisa. Es precisamente sobre esta forma específica de disartria donde nos enfocaremos más adelante, dada su relevancia directa con el objeto de esta investigación [7].
- **Disartrias mixtas:** resultado de daños en múltiples estructuras neuromusculares, como ocurre en la ELA. Sus características varían según la combinación de factores presentes [11].

Es importante tener en cuenta que, además de estas clasificaciones ampliamente aceptadas según el tipo de lesión neuromuscular, estas categorías no siempre se presentan de forma pura en la práctica clínica. De hecho, múltiples estudios han demostrado que incluso profesionales con amplia experiencia enfrentan dificultades para diferenciar con precisión entre los distintos tipos de disartria basándose únicamente en el análisis perceptual. Esto sugiere que estas clasificaciones pueden solaparse en ciertos casos clínicos y necesitan herramientas complementarias para una identificación confiable [40].

### **2.1.2. Disartria hipocinética y su vínculo con la Enfermedad de Parkinson**

En pacientes con enfermedad de Parkinson (EP), la disartria hipocinética aparece como una manifestación muy frecuente —presente en hasta el 90 % de los casos según algunas estimaciones— y se atribuye al daño en los circuitos extrapiramidales, particularmente aquellos que involucran a los ganglios basales [7].

Este tipo de disartria afecta todos los procesos motores del habla: la fonación se vuelve débil (hipofonía), la prosodia se aplanan (monotonía), la articulación se torna imprecisa y el flujo de habla puede acelerarse de forma irregular (lo que se conoce como “rushes of speech”). Además, suelen aparecer pausas inapropiadas y una inteligibilidad global reducida, lo que genera un impacto comunicativo severo para el paciente.

Sin embargo, aquí viene un punto que muchas veces pasa desapercibido: la disartria hipocinética no agota todas las manifestaciones lingüísticas del Parkinson. Es decir, no todos los trastornos del habla presentes en la EP corresponden exclusivamente a una disartria hipocinética. De hecho, algunos pacientes pueden presentar rasgos de disartria mixta, alteraciones pragmáticas, dificultades de evocación léxica o cambios en la prosodia emocional que no encajan estrictamente dentro del cuadro de disartria. Por esta razón, desde una perspectiva tanto clínica como computacional, tiene sentido tratar al Parkinson como una categoría diagnóstica independiente y no simplemente como un subgrupo de la disartria general.

La disartria hipocinética, asociada casi exclusivamente con la enfermedad de Parkinson, se caracteriza por alteraciones acústicas específicas como hipofonía, monotonía en tono y volumen, aceleración del ritmo del habla, articulación imprecisa, y una voz áspera o entrecortada [26]. Estos síntomas derivan de una disminución de la amplitud de movimiento en los órganos articulatorios, producto de la disfunción extrapiramidal causada por la degeneración nigroestriatal característica del Parkinson.

Además, como se ha señalado, el impacto del Parkinson sobre el habla no se limita únicamente a la disartria hipocinética. Los estudios funcionales de neuroimagen han mostrado patrones de sobreactivación cortical y subcortical durante tareas de habla, lo cual sugiere que las alteraciones del habla en la EP involucran redes motoras más amplias, incluyendo el área motora suplementaria y la corteza prefrontal dorsolateral [26].

Este hallazgo refuerza la idea de que la enfermedad de Parkinson, como entidad clínica, puede presentar múltiples patrones de afectación del habla que no se limitan exclusivamente a un subtipo de disartria. Por esta razón, desde una perspectiva computacional y diagnóstica, resulta válido —y necesario— tratar a la EP como una categoría diferenciada de análisis. Aunque la mayoría de los pacientes parkinsonianos presentan rasgos compatibles con una disartria hipocinética, también pueden manifestar alteraciones mixtas o presentar síntomas que no se encuadran en ninguna de las categorías tradicionales. Este fenómeno ha sido observado incluso en estudios de clasificación perceptual, donde el tipo hipocinético fue el único con tasas altas de identificación correcta, mientras que los otros subtipos eran comúnmente confundidos entre sí [40].

## 2.2. Mel-Frequency Cepstral Coefficients (MFCCs)

Los Coeficientes Cepstrales en la Escala Mel, más conocidos como MFCCs, constituyen uno de los descriptores acústicos más utilizados en procesamiento del habla, tanto en contextos de reconocimiento automático como en el análisis clínico de patologías. Su eficacia reside en que logran capturar de manera compacta y robusta la envolvente espectral del habla, es decir, la distribución de energía en distintas bandas de frecuencia a lo largo del tiempo, que es precisamente donde residen muchas de las alteraciones provocadas por los trastornos del habla.

Inspirados en el funcionamiento del sistema auditivo humano, los MFCCs incorporan un mapeo no lineal de las frecuencias reales a una escala perceptiva conocida como escala Mel, que refleja cómo percibimos los cambios en el tono: lineal para frecuencias bajas y logarítmica para las altas. Esta transformación se justifica neurofisiológicamente: el oído humano es más sensible a cambios de frecuencia en los rangos bajos (por ejemplo, distinguir 500 Hz de 700 Hz), pero necesita un cambio mayor para notar una diferencia en frecuencias altas (como entre 5000 Hz y 5200 Hz). La conversión se realiza típicamente mediante la siguiente fórmula:

$$M(f) = 1125 \ln \left( 1 + \frac{f}{700} \right) \quad (2.1)$$

y su inversa:

$$f(M) = 700 \left( e^{M/1125} - 1 \right) \quad (2.2)$$

### 2.2.1. Metodología de Cálculo

El proceso de obtención de MFCCs implica varias etapas bien definidas, que permiten transformar una señal de habla cruda en un vector numérico representativo de sus carac-

terísticas acústicas:

**Preénfasis:** Se aplica un filtro de alta frecuencia para compensar la atenuación natural de las frecuencias altas durante la producción del habla. Esto mejora la relación señal-ruido y resalta los formantes superiores relevantes para el reconocimiento de patrones.

**Enmarcado y ventana de Hamming:** La señal continua se divide en ventanas solapadas de 20 a 40 ms (25 ms es estándar), ya que el habla se considera cuasi-estacionaria en ese rango. Cada ventana se multiplica por una función de Hamming para reducir discontinuidades en los bordes y evitar distorsiones espectrales [29].

**Transformada de Fourier:** Se calcula la Transformada Rápida de Fourier (FFT) para obtener el espectro de potencia de cada frame. Esta operación transforma la señal al dominio de la frecuencia, permitiendo analizar su contenido espectral. La DFT del frame  $i$  se define como:

$$S_i(k) = \sum_{n=0}^{N-1} s_i(n) \cdot e^{-2\pi jkn/N} \quad (2.3)$$

El espectro de potencia se obtiene como:

$$P_i(k) = \frac{1}{N} |S_i(k)|^2 \quad (2.4)$$

**Filtro en la Escala Mel:** Se aplican entre 20 y 40 filtros triangulares espaciados en la escala Mel (26 es común). Cada filtro captura la energía en un rango específico del espectro, generando una representación perceptualmente relevante de la energía en distintas bandas.

**Logaritmo de la Energía:** Se calcula el logaritmo natural de cada energía filtrada. Esto simula la percepción logarítmica del oído humano y mejora la separabilidad estadística.

**Transformada Discreta del Coseno (DCT):** Se aplica la DCT para decorrelacionar los coeficientes obtenidos y obtener una representación compacta, conocida como cepstrum. Generalmente se conservan los primeros 12 o 13 coeficientes.

**Deltas y Delta-Deltas:** Se calcula la derivada temporal ( $\Delta$ ) y la segunda derivada ( $\Delta\Delta$ ) de cada MFCC para capturar la dinámica temporal del habla. Esto es crucial ya que muchas alteraciones patológicas no solo afectan la fonación estática, sino también sus variaciones temporales.

$$\Delta c_t = \frac{\sum_{n=1}^N n \cdot (c_{t+n} - c_{t-n})}{2 \sum_{n=1}^N n^2} \quad (2.5)$$

### 2.2.2. Representación y Agregación

En la práctica, para cada frame de audio se extraen 13 MFCCs, 13 deltas y 13 delta-deltas, lo que da un vector de 39 dimensiones por frame. Posteriormente, se agregan estadísticas como la media, desviación estándar, máximo, mínimo y mediana de cada coeficiente a lo largo de todo el archivo. Esto da lugar a un total de 195 características acústicas por archivo de audio, ampliamente utilizadas en modelos de clasificación.

### 2.2.3. Justificación de Uso

Los MFCCs son particularmente útiles porque permiten capturar de manera eficiente la forma de los formantes vocales, que son altamente sensibles a alteraciones en la articulación, la resonancia y la prosodia —tres de los subsistemas comúnmente afectados en las disartrias. Debido a su alta capacidad de discriminación y su larga trayectoria en tareas de reconocimiento de habla, han sido adoptados en numerosos trabajos sobre detección de patologías como disartria, Parkinson y trastornos del ritmo [18], [31].

En nuestro proyecto RefractedSpeech, los MFCCs constituyeron el núcleo de la representación acústica inicial. Su eficacia quedó demostrada por los altos niveles de precisión obtenidos al clasificar entre habla patológica y control, especialmente en entornos donde el ruido estaba relativamente controlado. Sin embargo, como se discutirá más adelante, su capacidad de generalización interlingüística y robustez ante interferencias acústicas es limitada, lo que ha motivado la exploración de descriptores alternativos.

Las definiciones matemáticas incluidas en esta sección se basan en textos clásicos de procesamiento de señales y habla como Oppenheim & Schaffer [24].

## 2.3. Features extraídos manualmente

### 2.3.1. Acoustic Features

Entre los descriptores acústicos más directamente relacionados con la fonación y la prosodia se encuentran el **promedio de pitch**, su **variabilidad**, la **intensidad promedio**, su **varianza**, y el **porcentaje de pausas**. Estos parámetros reflejan aspectos fisiológicos del habla que suelen verse alterados en cuadros neurológicos como la enfermedad de Parkinson.

El **pitch promedio** corresponde a la frecuencia fundamental ( $F0$ ) de la voz, y está estrechamente ligado a la tensión y longitud de las cuerdas vocales. Su **variabilidad** refleja el rango prosódico con el que un hablante modula su entonación: una menor variabilidad tiende a asociarse con habla monótona, un rasgo típico en pacientes con Parkinson. Por otro lado, la **intensidad vocal** (volumen) se relaciona con la amplitud de vibración de las cuerdas

vocales y la presión subglótica, y su disminución —conocida como *hipofonía*— es uno de los signos más comunes en etapas tempranas de la enfermedad.

En nuestro sistema, estas medidas se obtuvieron a través del análisis de señales de voz con herramientas como Parselmouth y Pydub [17], [27]. La **varianza de volumen** se calculó sobre los valores de energía de cada frame, mientras que el **porcentaje de pausas** se estimó como la proporción de frames silenciosos respecto al total de la grabación, usando umbrales dinámicos basados en la energía local. Las medidas de pitch se extrajeron con seguimiento de F0 frame a frame, utilizando algoritmos robustos para habla patológica.

La inclusión de estos parámetros se basa en estudios recientes que los han identificado como efectivos para discriminar entre pacientes con Parkinson y controles sanos, incluso en contextos multilingües. En particular, Lim et al. (2024) desarrollaron un modelo de detección cruzado entre coreano y taiwanés utilizando precisamente estos descriptores, obteniendo resultados significativos en cohortes independientes [20].

### 2.3.2. Acoustic Landmarks

Los *acoustic landmarks* son eventos fonéticamente significativos en la señal de voz, definidos por cambios abruptos en propiedades acústicas tales como sonoridad, fricación o estallidos consonánticos. Estos landmarks permiten una representación intermedia entre el nivel acústico y el fonológico, facilitando el análisis clínico del habla sin requerir alineación fonética detallada ni reconocimiento automático de palabras.

La herramienta más utilizada para su extracción es la **SpeechMark Toolbox** [4] (ver Figura 2.1), presentada en INTERSPEECH 2012, la cual implementa una serie de heurísticas sobre el espectrograma para identificar landmarks como:

- **+g / -g**: inicio y fin de la vibración glótica (voicing).
- **+b / -b**: estallidos explosivos de sonidos oclusivos.
- **+f / -f**: fricación no sonora.
- **+v / -v**: fricación sonora.
- **+s / -s**: transiciones asociadas a consonantes sonorantes.

Además de los landmarks individuales, SpeechMark permite agruparlos en *clusters silábicos* (SC) —agrupaciones temporales que reflejan la estructura rítmica del habla— y calcular *n-gramas de landmarks* (como bigramas y trigramas), los cuales capturan relaciones secuenciales entre eventos acústicos. También pueden calcularse razones entre tipos de landmarks (por ejemplo, bigramas/unigramas), así como estadísticas sobre su *fuerza acústica* (intensidad relativa del evento). Estas características han demostrado ser útiles para distinguir entre poblaciones con y sin trastornos del habla.

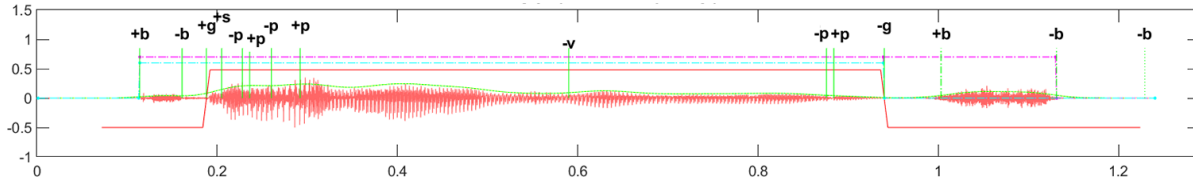


Figura 2.1: Ejemplo de detección automática de *acoustic landmarks* utilizando la herramienta SpeechMark. Se observan eventos como +g (inicio de sonoridad), +s (inicio de consonante sonorante), y +b (explosión consonántica), entre otros. **Fuente:** Suthar et al. (2022).

En esta tesis, se implementó una aproximación en Python que emula parcialmente la lógica de SpeechMark, utilizando análisis de energía, pitch y derivadas temporales. Esta aproximación incluye landmarks simulados como:

- **+v / -v:** encendido/apagado del pitch (seguimiento con Parselmouth).
- **+s / -s:** fricación sin voicing detectada por umbral de energía.
- **+b / -b:** picos abruptos en la derivada de la intensidad.
- **Clusters silábicos:** agrupaciones con ritmo compatible a sílabas ( 120–300 ms).
- **Unigramas y Bigramas:** frecuencia de ocurrencia de secuencias.
- **Strength:** energía local en torno al evento.

El valor de estas representaciones radica en su interpretabilidad y sensibilidad a alteraciones motoras sutiles. Tal como señalan Suthar et al. (2022), los patrones de landmarks y clusters silábicos permiten distinguir de forma efectiva entre niños con y sin trastornos fonológicos, con mejoras significativas en precisión cuando se incluyen bigramas, tasas, duraciones y razones normalizadas [35].

En nuestro caso, la incorporación de estos descriptores buscó complementar los MFCCs y otros features clásicos con una visión más estructurada del habla, enfocada en su organización articulatoria.

### 2.3.3. Características de Articulación Vocal

Las vocales constituyen uno de los elementos acústicos más informativos en el habla humana, ya que su producción depende de movimientos precisos de los articuladores orales —principalmente la lengua, los labios y la mandíbula— que modifican la forma del tracto vocal y generan cavidades de resonancia. Estas cavidades amplifican ciertas bandas de frecuencia del espectro de la voz, conocidas como *formantes*, que definen acústicamente a cada vocal.

En particular, los dos primeros formantes (F1 y F2) han demostrado ser altamente representativos de la configuración articulatoria de las vocales. El primero (F1) se asocia con la altura de la lengua, mientras que el segundo (F2) refleja su posición anterior/posterior en la cavidad oral. Por tanto, cualquier restricción en el rango de movimiento de los articuladores —como la observada en pacientes con enfermedad de Parkinson (EP)— afecta directamente la producción de formantes, dando lugar a una articulación de vocales menos precisa y más centralizada.

Uno de los métodos clásicos para medir esta afectación es el *Triangular Vowel Space Area* (tVSA), que se calcula a partir de los valores promedio de F1 y F2 para las vocales /a/, /i/ y /u/, trazando un triángulo en el espacio acústico. Una reducción en esta área se interpreta como una contracción del “espacio de trabajo articulatorio”, fenómeno consistente con la hipocinesia observada en la disartria parkinsoniana. Sin embargo, estudios recientes han señalado que el tVSA puede ser poco sensible en pacientes con disartria leve o moderada, especialmente en etapas tempranas de la enfermedad.

Frente a esta limitación, se ha propuesto el uso del *Vowel Articulation Index* (VAI), un índice acústico que combina los valores de F1 y F2 de las vocales mencionadas en una fórmula que busca capturar de forma más robusta la centralización vocálica. A diferencia del tVSA, el VAI ha mostrado ser sensible incluso en pacientes con disartria leve, manteniendo su capacidad discriminativa aun cuando la inteligibilidad del habla está preservada. En la literatura, se ha observado que el VAI tiende a estar reducido en hablantes con Parkinson, en ambos géneros, mientras que la reducción de tVSA parece ser más pronunciada en varones [32].

Ambos indicadores —tVSA y VAI— son calculados a partir de los valores promedio de F1 y F2 extraídos de las vocales contenidas en palabras seleccionadas de una tarea de lectura. En este trabajo se adoptaron ambos enfoques, utilizando herramientas de análisis espectral para estimar los formantes en segmentos temporales centrados en cada vocal, replicando la metodología empleada por Skodda y colaboradores en su análisis acústico de la disartria parkinsoniana.

### **2.3.4. Jitter y Shimmer**

Los descriptores acústicos *jitter* y *shimmer* son fundamentales para detectar irregularidades en la voz, al capturar variaciones sutiles en la frecuencia y la amplitud que reflejan alteraciones en el control neuromotor de las cuerdas vocales. Gracias a su sensibilidad para evidenciar inestabilidades fonatorias, estos parámetros se han utilizado extensamente tanto en investigación como en entornos clínicos para el diagnóstico y seguimiento de patologías vocales.

## Definición y Características Técnicas

El *jitter* mide la variabilidad ciclo a ciclo de la frecuencia fundamental, es decir, las fluctuaciones en la duración de cada periodo vocal. Se expresa comúnmente en milisegundos o como porcentaje relativo. Por su parte, el *shimmer* cuantifica las variaciones en la amplitud entre ciclos consecutivos, reflejando la regularidad de la intensidad vocal. Ambos parámetros están estrechamente relacionados con la estabilidad fonatoria y pueden ser indicadores de disfunción laríngea o alteraciones neurológicas, como las que ocurren en la disartria.

## Evidencia Científica y Aplicaciones Clínicas

El estudio de Sripriya et al. (2017) propuso una técnica no intrusiva para clasificar voces normales y patológicas usando medidas de *jitter* y *shimmer* extraídas a partir de los *glottal closure instants* (GCIs), lo cual mejoró la precisión respecto a técnicas tradicionales basadas en el periodo de *pitch*, alcanzando un 85 % de exactitud en la clasificación binaria [33].

Por otro lado, Hernández et al. (2020) incorporaron *jitter* y *shimmer* dentro de un conjunto de características de calidad vocal y prosodia para evaluar la severidad de la disartria en inglés y coreano. Su inclusión permitió mejorar sustancialmente la detección de casos leves, elevando el *recall* del 42.42 % al 83.34 % en inglés, lo que destaca su valor diagnóstico incluso en etapas tempranas del deterioro [15].

## Implementación en RefractedSpeech

En el sistema *RefractedSpeech*, estos descriptores se combinaron con coeficientes cepstrales en escala Mel (MFCCs) para formar un vector híbrido que integra información temporal y espectral. Esta integración permitió alcanzar una precisión del 89 % en la clasificación binaria de voz patológica, superando el rendimiento obtenido al usar estas características por separado.

La razón de esta mejora se atribuye a la complementariedad de los descriptores: mientras el *jitter* y *shimmer* capturan inestabilidades microtemporales asociadas a disfunciones fonatorias, los MFCCs representan aspectos espectrales globales de la señal.

## Relevancia para esta Investigación

En esta tesis, se reincorpora el análisis de *jitter* y *shimmer* con el objetivo de evaluar su capacidad de generalización en contextos multilingües. Se busca determinar si estos descriptores conservan su valor discriminativo cuando se aplican a voces en distintos idiomas y acentos, un criterio crucial para el desarrollo de sistemas diagnósticos de alcance global.

Además, se los contrastará con descriptores más complejos, lo cual permitirá valorar si el incremento en complejidad computacional se traduce en mejoras significativas en precisión diagnóstica, considerando también la eficiencia y la interpretabilidad clínica.

### 2.3.5. Descriptores Basados en Correlación: Propuesta de Nuevas Características Acústicas

La **autocorrelación** es la correlación de una señal consigo misma desplazada en el tiempo (lags). Matemáticamente, la autocorrelación  $R_{xx}$  de una señal discreta  $x[n]$  se define como:

$$R_{xx}[k] = \sum_{n=-\infty}^{\infty} x[n]x[n+k] \quad (2.6)$$

La **correlación cruzada** mide la similitud entre dos señales diferentes desplazadas temporalmente. Dados dos arreglos discretos  $x[n]$  y  $y[n]$ , la correlación cruzada  $R_{xy}$  se define como:

$$R_{xy}[k] = \sum_{n=-\infty}^{\infty} x[n]y[n+k] \quad (2.7)$$

La decisión de incluir este tipo de descriptores en el presente trabajo surgió a partir de la experiencia del autor como Ingeniero de Software en el observatorio astronómico ALMA (Atacama Large Millimeter/submillimeter Array), donde ha estado constantemente expuesto a conceptos como la *correlación cruzada de ondas electromagnéticas*, fundamentales para el procesamiento interferométrico de señales astronómicas. A partir del estudio de estos principios y su aplicación en un sistema de tan alta precisión, surgió la motivación por explorar ideas análogas en el ámbito del procesamiento de señales de voz patológica, proponiendo métricas basadas en correlación que permitan capturar patrones relevantes de alteración vocal.

En esta investigación, se implementaron los siguientes descriptores basados en correlación:

- **Autocorr Peak:** Máximo valor de la función de autocorrelación.
- **Autocorr Energy:** La energía de autocorrelación se calcula como:

$$E_{xx} = \sum_{k=-(N-1)}^{N-1} |R_{xx}[k]|^2, \quad \text{con} \quad R_{xx}[k] = \frac{1}{N} \sum_{n=0}^{N-1} x[n]x[n+k] \quad (2.8)$$

o equivalentemente en el dominio frecuencial (Teorema de Wiener-Khinchin):

$$E_{xx} = \frac{1}{N} \sum_{f=0}^{N-1} |X[f]|^4 \quad (2.9)$$

donde  $X[f]$  es la DFT no normalizada de  $x[n]$ .

- **Autocorr Zero Crossings:** Número de cruces por cero en la función de autocorrelación.
- **Crosscorr Peak:** Máximo valor de la correlación cruzada entre dos segmentos contiguos de la señal.
- **Crosscorr Entropy:** Entropía  $H$  calculada a partir de la distribución normalizada de valores absolutos de la correlación cruzada:

$$H = - \sum_i p_i \log(p_i), \quad \text{donde} \quad p_i = \frac{|R_{xy}[i]|}{\sum_j |R_{xy}[j]|} \quad (2.10)$$

- **Crosscorr Std:** Desviación estándar de la correlación cruzada.

Estos descriptores buscan capturar aspectos como periodicidad, regularidad, grado de orden o desorden y similitud entre segmentos vocales, ayudando a diferenciar emisiones patológicas de sanas.

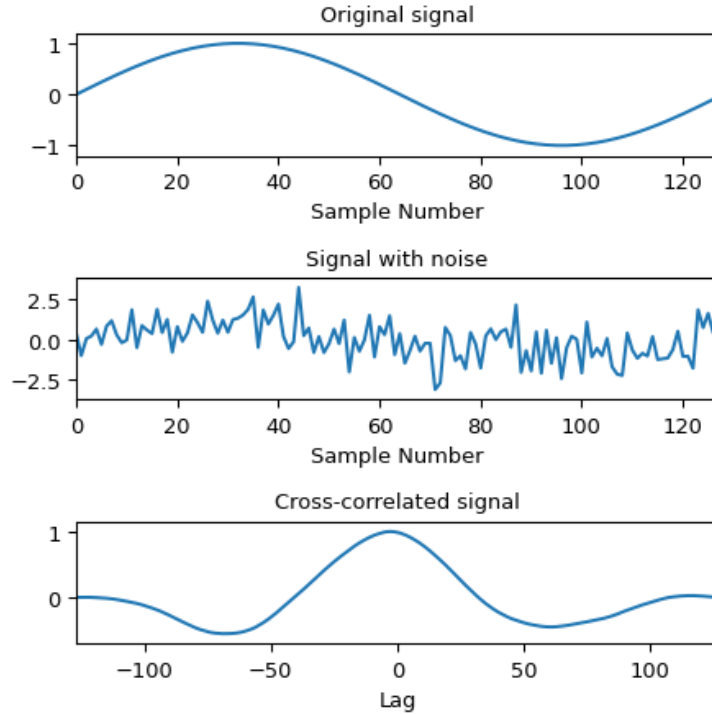


Figura 2.2: Visualización conceptual de la correlación cruzada entre dos señales. Fuente: `scipy.signal.correlate`

La Figura 2.2 ejemplifica la correlación cruzada para detectar similitudes entre señales desplazadas temporalmente, destacando su utilidad para extraer patrones estructurales ocultos en presencia de ruido.

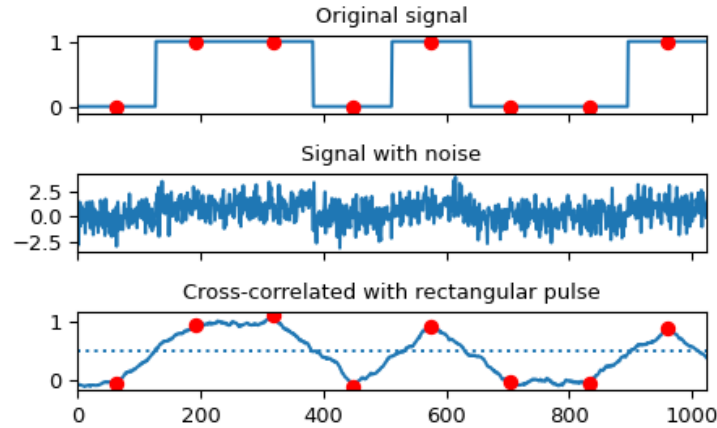


Figura 2.3: Implementación de un filtro adaptado (matched filter) usando correlación cruzada. Fuente: `scipy.signal.correlate`

La Figura 2.3 muestra cómo un filtro adaptado basado en correlación cruzada puede recuperar señales en entornos ruidosos, demostrando así la robustez de estos métodos.

Estas representaciones visuales y fórmulas permiten comprender intuitivamente los descriptores propuestos y justifican claramente su inclusión como herramientas efectivas en el análisis acústico de señales patológicas.

## Evidencia en la Literatura

Diversos estudios recientes respaldan el uso de la correlación como fuente de características discriminativas en distintas áreas de análisis de señales:

- Bose et al. (2016) propusieron un sistema de clasificación de enfermedades neuromusculares basado en la correlación cruzada de señales EMG. Al correlacionar señales de pacientes con miopatía y neuropatía contra una señal de referencia sana, extrajeron características estadísticas, Hjorth y autorregresivas adaptativas, que permitieron alcanzar una precisión de hasta un 100 % al combinarse con clasificadores como SVM y kNN [3].
- Pestana-Nobles et al. (2023) aplicaron análisis de correlación cruzada bidimensional y dimensión fractal para identificar especies de diatomeas en imágenes microscópicas. La correlación permitió comparar texturas estructurales complejas entre especies, evidenciando su valor como herramienta para tareas de clasificación visual [25].
- He et al. (2020) implementaron autocorrelación para detectar fallas en engranajes mediante el método LEASgram, el cual analiza el autoespectro de la envolvente logarítmica de vibraciones. Este enfoque permitió identificar impulsos periódicos indicativos de defectos mecánicos, incluso en condiciones de ruido elevado [13].

Estos trabajos refuerzan la hipótesis de que los descriptores basados en correlación son altamente efectivos para destilar información oculta en distintos tipos de señales, y que su capacidad para capturar patrones estructurales puede transferirse con éxito al análisis computacional del habla patológica.

### 2.3.6. Extracción y análisis de características lingüísticas

En esta etapa final del proceso de extracción de características, se incorporaron también descriptores lingüísticos derivados de la transcripción automática del habla, con el objetivo de explorar su valor potencial en el diagnóstico computacional de patologías como la disartria y el Parkinson. Esta decisión se basó en resultados prometedores reportados por Lim et al. (2024), quienes demostraron que la combinación de descriptores acústicos y lingüísticos mejora la discriminación entre pacientes con Parkinson e individuos sanos, incluso en contextos multilingües [20].

Siguiendo ese enfoque, se extrajeron los siguientes descriptores a partir de la transcripción generada por la API de Google Speech-to-Text:

- **Speech Rate (Tasa de habla):** razón entre la duración del audio y la cantidad de caracteres transcritos. Valores más bajos indican una articulación más eficiente y fluida, lo que permite comparaciones consistentes entre audios de diferente duración.
- **Speech-to-Text Confidence Score (Confianza de la API):** medida entregada por Google que refleja la certeza del sistema en la precisión de su transcripción. Una mayor confianza suele correlacionarse con una mayor claridad fonológica.
- **Linguistic Perplexity (Perplejidad lingüística):** métrica que evalúa la improbabilidad del texto transcrito según un modelo de lenguaje. Valores altos podrían estar asociados a errores articulatorios o desorganización gramatical.

Estos descriptores fueron concebidos como un complemento a las características acústicas ya extraídas, apuntando a capturar aspectos relacionados con la inteligibilidad y eficiencia comunicativa de los pacientes.

## Limitaciones presupuestarias

Sin embargo, la implementación de esta etapa enfrentó una restricción importante: el uso de la API de Google Speech-to-Text está sujeto a un límite de crédito gratuito de 60 minutos de procesamiento. Esto obligó a limitar el análisis lingüístico a un subconjunto reducido de 471 audios, en contraste con los más de 15.000 audios procesados acústicamente. En consecuencia, los *features* lingüísticos presentan valores faltantes (NaN) en la mayoría del dataset, dificultando su uso directo en modelos entrenados sobre el conjunto completo.

## Estrategia adoptada ante la baja cobertura

Debido a la limitada disponibilidad de características lingüísticas, se optó por una estrategia conservadora: entrenar un modelo independiente utilizando únicamente el subconjunto de audios que sí contaba con estos descriptores, evitando así imputaciones o extrapolaciones que pudieran introducir sesgo o ruido.

Este modelo fue entrenado con el conjunto completo de características (acústicas y lingüísticas), pero únicamente sobre los 471 audios con transcripción válida. Aunque esta muestra es reducida y desbalanceada entre clases y datasets, permitió obtener resultados preliminares que ilustran el potencial diagnóstico de los descriptores lingüísticos propuestos en contextos multilingües, tal como sugiere Lim et al. (2024).

Este subconjunto estuvo compuesto por siete clases, resultantes de la combinación entre tres idiomas (inglés, español e italiano) y tres condiciones clínicas (hablantes sanos, con disartria y con Parkinson). Las clases incluidas fueron: *Healthy-English*, *Dysarthria-English*, *Parkinson-English*, *Healthy-Spanish*, *Parkinson-Spanish*, *Healthy-Italian* y *Parkinson-Italian*. Si bien todas las clases están representadas, la cantidad de ejemplos por clase varía considerablemente, lo que explica el desbalance observado.

## Consideraciones finales

Dado el tamaño reducido y el desbalance entre clases del subconjunto analizado, los resultados obtenidos al incluir los *features* lingüísticos deben interpretarse como preliminares y no representativos. A pesar de ello, se optó por incluirlos en esta investigación con fines exploratorios, considerando el respaldo bibliográfico y su potencial utilidad futura en entornos multilingües.

## 2.4. Modelos de Clasificación en Machine Learning

Los modelos de clasificación son algoritmos de aprendizaje automático cuyo objetivo es asignar elementos o muestras a categorías específicas en función de sus características. Dependiendo del número de clases, esta clasificación puede ser:

- **Binaria:** distingue entre dos clases, por ejemplo, voz sana vs. voz patológica.
- **Multiclase:** distingue entre tres o más clases, como diferentes tipos de patologías vocales.

Para lograr estas tareas, se utilizaron diversos modelos clásicos de aprendizaje automático,

conocidos por su robustez y capacidad de generalización. A continuación, se presenta una descripción general de los métodos utilizados:

## Support Vector Machine (SVM)

Un **SVM lineal**, también conocido como **SVC** (Support Vector Classifier), busca encontrar un hiperplano que separe las clases de forma óptima en el espacio de características, tal como se ve representado en la Figura 2.4. Este fue el modelo utilizado en la primera versión de RefractedSpeech, ya que presenta alta interpretabilidad y buena generalización con pocos datos.

SVM Binario - Frontera de decisión en 3D (PCA)

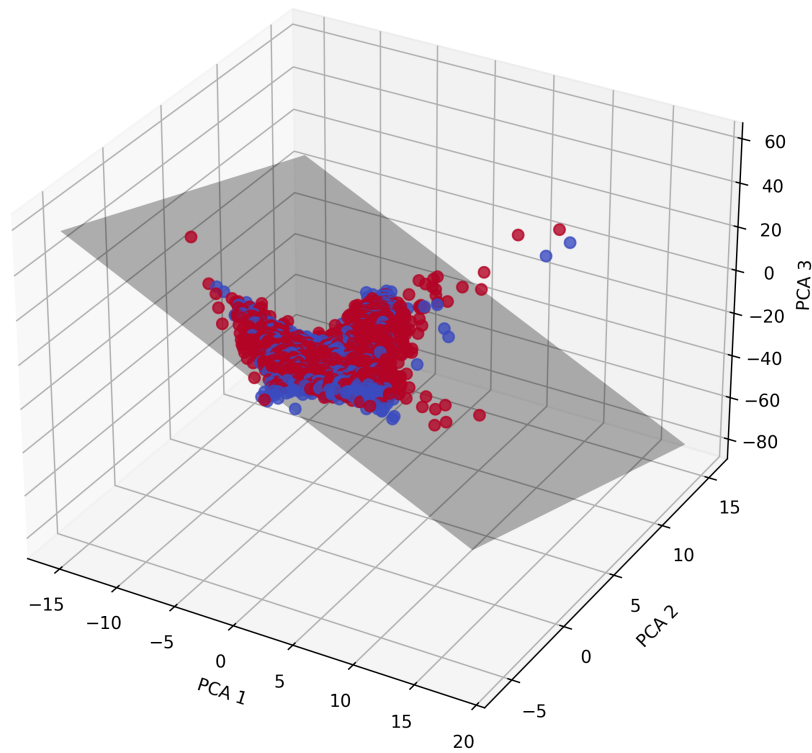


Figura 2.4: Visualización en 3D de la frontera de decisión generada por un modelo SVM binario lineal entrenado sobre los tres primeros componentes principales (PCA) de los datos. Los puntos corresponden a muestras proyectadas mediante PCA, tras ser preprocesadas con la misma pipeline usada por el modelo SVC-RBF (escalado y selección de características). Aunque este modelo lineal fue entrenado exclusivamente para fines de visualización, permite ilustrar cómo las clases se distribuyen en un espacio reducido. Se observa una importante superposición entre clases, lo que sugiere que la separación entre muestras no es linealmente evidente en las primeras tres dimensiones principales.

En cambio, una **SVM con kernel RBF** (Radial Basis Function) [36] permite separar clases no linealmente separables al proyectar los datos a un espacio de mayor dimensionalidad mediante un kernel gaussiano (ver Figura 2.5). Aunque más flexible, requiere mayor ajuste de hiperparámetros y mayor capacidad computacional.

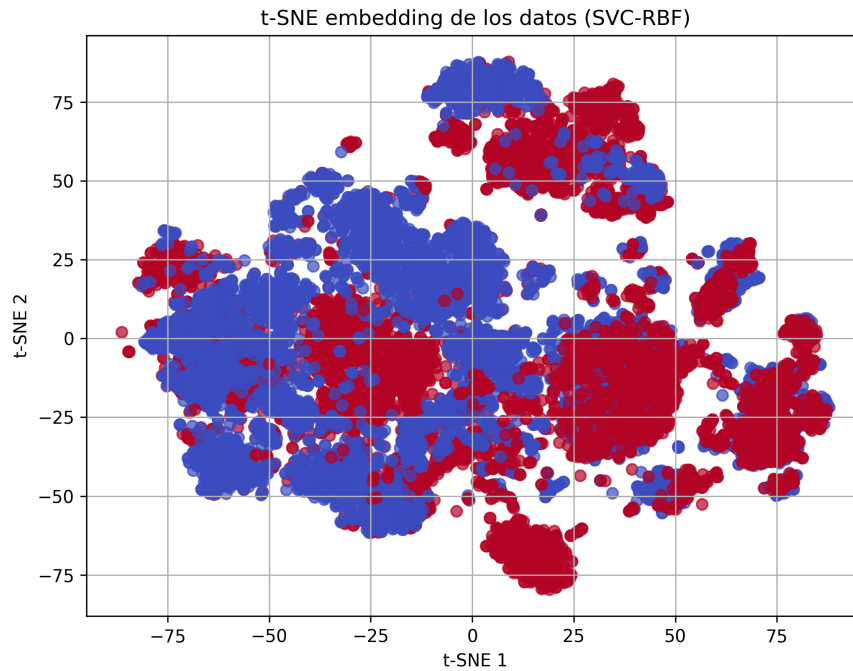


Figura 2.5: Distribución de las muestras tras proyección con t-SNE sobre dos dimensiones. Cada punto representa una muestra, coloreada según su clase verdadera (azul: sana, rojo: patológica). La proyección no lineal t-SNE busca preservar relaciones locales entre muestras, revelando clústeres dominados por una clase. Sin embargo, se observa una fuerte amalgama entre clases, lo que sugiere que, incluso con kernels no lineales (como en el modelo SVC-RBF), la separabilidad de las clases no es trivial.

## Random Forest (RF)

Los **Random Forest** son ensambles de árboles de decisión [6]. Cada árbol se entrena sobre una muestra diferente del conjunto de datos y aporta un voto para la clasificación final. Este enfoque reduce el sobreajuste y mejora la estabilidad del modelo.

## Redes Neuronales Multicapa (MLP)

Una **MLP** (Multilayer Perceptron) es un tipo de red neuronal feedforward compuesta por capas ocultas con funciones de activación no lineales (por ejemplo, ReLU, tanh o sigmoid).

Estas funciones permiten modelar relaciones complejas entre las características de entrada y las clases objetivo.

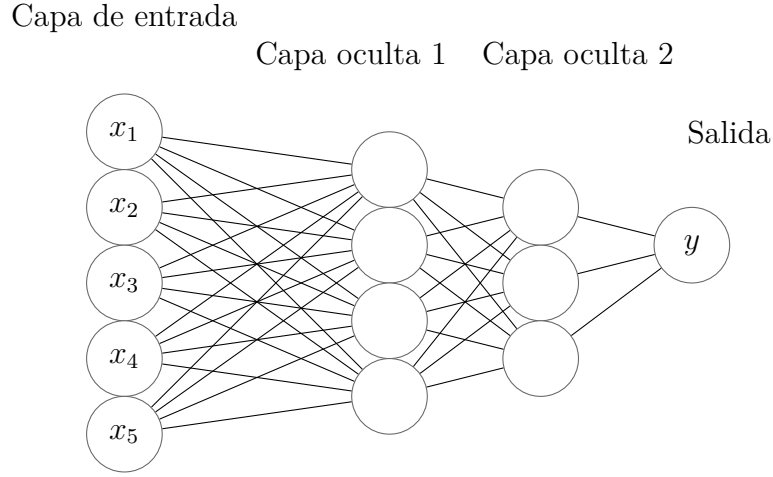


Figura 2.6: Ejemplo generado de red neuronal multicapa (MLP) con dos capas ocultas.

## Clasificación Binaria vs. Multiclase

En clasificación **binaria**, el modelo aprende a distinguir entre dos posibles resultados, utilizando por lo general una función de activación sigmoideal y una función de pérdida como entropía cruzada binaria.

En **clasificación multiclase**, se pueden utilizar estrategias como *uno contra todos* (OvR) o *uno contra uno*, junto con funciones de activación como **softmax**, que asigna una probabilidad a cada clase. La función de pérdida correspondiente suele ser la entropía cruzada categórica.

Una explicación técnica detallada de estos modelos, junto con su implementación específica y resultados obtenidos, se desarrolla en el capítulo metodológico de esta tesis.

Matemáticamente, la función de activación **sigmoideal** utilizada en clasificación binaria se define como:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

mientras que la función **softmax** usada en clasificación multiclase se expresa como:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$

donde  $K$  representa el número de clases posibles. Ambas funciones transforman las salidas del modelo en valores interpretables como probabilidades.

## Selección de características

En la práctica, los modelos no se entrenan sobre la totalidad de los descriptores extraídos, sino sobre un subconjunto optimizado mediante técnicas de selección de características. En este proyecto se utilizó la estrategia Recursive Feature Elimination con validación cruzada (RFECV) [1], la cual permite identificar las variables más relevantes para el rendimiento del modelo. Esta técnica no solo mejora la eficiencia computacional, sino que también facilita la interpretación clínica, al reducir el espacio de características a un conjunto más manejable y potencialmente significativo desde el punto de vista fisiopatológico. La implementación técnica de este proceso se detalla en el capítulo metodológico.

## 2.5. Estado del Arte y Enfoque del Proyecto

El diagnóstico automatizado de trastornos del habla ha sido abordado desde diversas perspectivas en las últimas décadas, en respuesta a las limitaciones asociadas a los métodos clínicos tradicionales. Si bien evaluaciones perceptuales realizadas por especialistas siguen siendo el estándar en muchos contextos, su naturaleza subjetiva y la variabilidad interevaluador comprometen la consistencia diagnóstica [19]. A ello se suman barreras de acceso asociadas a técnicas más especializadas pero invasivas, como la laringoscopia, las cuales requieren equipamiento y personal clínico calificado [14].

Frente a este panorama, el aprendizaje automático ha ofrecido herramientas prometedoras para el desarrollo de sistemas diagnósticos más objetivos, escalables y no invasivos. Según Brahmi et al. [5], modelos como Support Vector Machines (SVM) y redes neuronales profundas han demostrado una alta capacidad para modelar relaciones complejas en señales acústicas, especialmente cuando se emplean descriptores como los coeficientes cepstrales MFCC. Estos avances han permitido mejorar significativamente la precisión en la detección de patologías del habla bajo condiciones controladas.

Sin embargo, muchos de estos sistemas se han enfocado exclusivamente en escenarios binarios, donde se distingue entre habla sana y patológica, sin considerar la necesidad clínica de diferenciar entre múltiples condiciones específicas. En este contexto, Wu et al. [38] destacan el valor de los clasificadores multiclase, como SVM y GMM (Gaussian Mixture Models), para abordar tareas de clasificación más complejas que reflejen mejor la diversidad de alteraciones fonatorias presentes en la práctica clínica. Sus resultados sugieren que la clasificación multiclase puede aumentar la utilidad clínica de estos sistemas al permitir una detección y diferenciación más precisa de distintas patologías.

Paralelamente, la literatura reciente ha explorado cómo mejorar la generalización de estos modelos en contextos multilingües y multiculturales. Estudios como los de Vásquez-Correa et al. [37], ya discutidos en la sección *"Necesidad de generalización multi-patológica y multilingüe"*, han implementado técnicas de transferencia de aprendizaje para adaptar modelos entre idiomas, aprovechando corpora clínicos multilingües. Aunque sus resultados son prome-

tedores, las diferencias fonéticas y prosódicas entre lenguas siguen representando un obstáculo para una generalización robusta. En esta línea, Irizarry-Pérez et al. [16] proponen utilizar características acústicas compartidas entre idiomas como eje para desarrollar modelos bilingües más estables, sugiriendo que la búsqueda de descriptores fonéticamente universales es una vía estratégica aún en desarrollo.

En conjunto, la literatura evidencia que los modelos más efectivos para el diagnóstico computacional de patologías del habla serán aquellos capaces de integrar tres atributos clave: (1) la capacidad de clasificar múltiples patologías; (2) la generalización efectiva a distintos idiomas o acentos; y (3) una estructura suficientemente interpretable y robusta para adaptarse a condiciones clínicas reales, incluyendo grabaciones ruidosas o de baja calidad.

Bajo esta perspectiva, el presente trabajo propone el desarrollo de un sistema diagnóstico fonopatológico basado en descriptores acústicos optimizados, con énfasis en la robustez al ruido, la capacidad de adaptación multilingüe y la generalización multi-patológica. Este enfoque busca superar las limitaciones previamente descritas mediante una combinación de selección de características relevantes, clasificación multiclase e interpretación del modelo a través de técnicas explicativas, integrando distintos avances recientes bajo un marco unificado.

# Capítulo 3

## Propuesta de solución

### 3.1. Descripción General del Pipeline

El desarrollo de este sistema diagnóstico se estructuró en una serie de etapas que conforman el *pipeline* metodológico general, comenzando por la búsqueda de datos y terminando en el análisis interpretativo de las predicciones. Si bien el objetivo a largo plazo de esta investigación es identificar un conjunto de descriptores acústicos robustos que puedan generalizar entre distintas lenguas y patologías —incluso en contextos donde no se disponga de datos multilingües para entrenamiento— fue necesario, en esta etapa experimental, recopilar datasets en al menos tres idiomas distintos y con al menos dos patologías diferentes. Esta decisión respondió al propósito de validar la capacidad de generalización interlingüística y multi-patológica de los descriptores propuestos.

Una vez reunidos los datos, se procedió a la extracción de características acústicas. Esta etapa se llevó a cabo en dos fases: en primer lugar, se desarrolló e implementó la lógica de extracción utilizando un subconjunto reducido de los datos, con el fin de verificar la consistencia funcional del código y obtener una primera impresión del comportamiento de los descriptores. Posteriormente, se aplicó la misma lógica a la totalidad de los datos disponibles.

La selección de características se realizó en dos pasos. Primero, se llevó a cabo una eliminación de redundancia mediante una matriz de correlación de Pearson, que permitió remover variables altamente correlacionadas entre sí. Luego, sobre el conjunto resultante, se aplicó *Recursive Feature Elimination con Validación Cruzada* (RFECV), utilizando como modelos base un *Support Vector Machine* lineal y un *Random Forest Classifier*. El objetivo de esta etapa no fue solo mejorar el rendimiento del sistema, sino también reducir la complejidad computacional del entrenamiento y aumentar la interpretabilidad clínica al limitar la cantidad de características relevantes.

Para la etapa de clasificación, se llevaron a cabo experimentos con dos arquitecturas distintas. La primera, de tipo *one-tiered*, consistió en clasificadores multiclase directos en-

trenados sobre la totalidad del *dataset*. En esta configuración, además del SVM lineal y el Random Forest, se incluyeron dos modelos de *perceptrón multicapa* (MLP) entrenados sobre los conjuntos de características seleccionados por los dos clasificadores anteriores, respectivamente, y un SVC con *kernel* RBF entrenado sobre las características seleccionadas por el SVM lineal. Posteriormente, se implementó una segunda arquitectura de tipo *two-tiered*, en la cual se realizaba primero una clasificación binaria (voz sana vs patológica), y sobre el subconjunto patológico resultante se entrenaba un clasificador multiclase. En esta segunda arquitectura se utilizaron **únicamente tres modelos**: un SVM lineal, un SVC con *kernel* RBF, y un Random Forest. En todos los casos, excepto en el Random Forest, los modelos utilizaron el conjunto de características seleccionadas por el SVM lineal.

Finalmente, tras el entrenamiento de los modelos y el análisis de sus métricas de desempeño, se llevó a cabo un estudio interpretativo utilizando SHAP (*SHapley Additive exPlanations*) [22], con el objetivo de analizar la contribución individual de cada descriptor acústico en la decisión de clasificación del modelo. Esta última etapa permitió obtener una perspectiva más profunda sobre la relevancia y estabilidad de las características seleccionadas a lo largo del proceso.

## 3.2. Conjunto de Datos Utilizados

Para evaluar la capacidad de generalización interlingüística y multi-patológica del sistema propuesto, se recopilieron muestras de voz provenientes de siete grupos distintos, diferenciados tanto por idioma como por condición clínica. Estos fueron: **Dysarthria-English**, **Parkinson-English**, **Parkinson-Spanish**, **Parkinson-Italian**, **Healthy-English**, **Healthy-Spanish** y **Healthy-Italian**. Esta distribución permitió un análisis cruzado de las características acústicas en al menos tres idiomas distintos y en dos patologías neurológicas diferenciadas.

El conjunto **Dysarthria-English** se extrajo íntegramente del corpus **TORGO**, uno de los datasets más utilizados en la literatura para el estudio de disartria, y base del modelo original RefractedSpeech. El subconjunto **Parkinson-English**, en tanto, fue construido a partir de los datos presentados por Yu et al. [39], los cuales incluyen grabaciones de pacientes con Parkinson bajo un protocolo estandarizado y fueron obtenidos mediante colaboración clínica. El grupo **Healthy-English** corresponde a una mezcla de sujetos de control provenientes tanto del TORGO como del dataset de Parkinson-English.

Para el idioma español, se utilizaron los datos del corpus **NeuroVoz** [23], el cual proporciona grabaciones de hablantes castellanos con diagnóstico de Parkinson y de sujetos control, registrados bajo condiciones acústicas homogéneas. De este corpus se extrajeron los grupos **Parkinson-Spanish** y **Healthy-Spanish**.

Finalmente, los grupos **Parkinson-Italian** y **Healthy-Italian** fueron obtenidos a partir del estudio realizado por Dimauro et al. [10], en el cual se empleó el sistema *Voxtester* junto

con un protocolo de lectura fonéticamente balanceado en italiano. Este estudio incluye datos tanto de pacientes como de individuos sanos, permitiendo así su uso comparativo en tareas de clasificación.

Esta diversidad de idiomas y condiciones clínicas representa un desafío ideal para evaluar el grado de robustez y adaptabilidad del sistema, y constituye una de las bases metodológicas fundamentales de esta investigación.

### 3.3. Extracción y preprocesamiento de características

La extracción de características acústicas se realizó sobre la totalidad del conjunto de datos utilizando herramientas especializadas: *Parselmouth* para análisis de voz, *librosa* para procesamiento de audio, y scripts personalizados desarrollados en este trabajo. Se incluyeron:

- Descriptores tradicionales (*jitter*, *shimmer*, MFCCs)
- Propuestas innovadoras basadas en autocorrelación y correlación cruzada
- Parámetros prosódicos y formánticos

Adicionalmente, se incorporaron métricas derivadas de la transcripción automática del habla.

Cada archivo de audio fue procesado de forma independiente, generando un vector de características que se asoció a su respectiva clase de diagnóstico. Para mantener la consistencia de la matriz de entrada a los clasificadores, se implementaron mecanismos de limpieza y normalización de los datos. En particular, aquellos casos con valores faltantes (NaN) o valores nulos fueron tratados mediante imputación simple: se reemplazaron utilizando el promedio del mismo *feature* dentro del grupo al que pertenecía el audio (por ejemplo, paciente sano o con Parkinson). Esta estrategia permitió conservar información útil sin introducir sesgos artificiales.

Con el fin de identificar posibles redundancias entre características, se calcularon matrices de correlación de Pearson dentro de cada grupo de descriptores. En la Figura 3.1 se observa la matriz correspondiente a los parámetros acústicos de variabilidad de volumen, pausas y tono. La Figura 3.2 muestra la correlación entre los *landmarks* acústicos derivados de la detección de unigrama y bigrama de eventos fonatorios. La Figura 3.3 exhibe la correlación entre características de articulación vocal (formantes F1 y F2 para las vocales /a/, /i/ y /u/), junto a las métricas tVSA y VAI. Finalmente, la Figura 3.4 presenta la matriz de correlación entre los descriptores propuestos de autocorrelación, correlación cruzada, y sus combinaciones con *jitter* y *shimmer*.

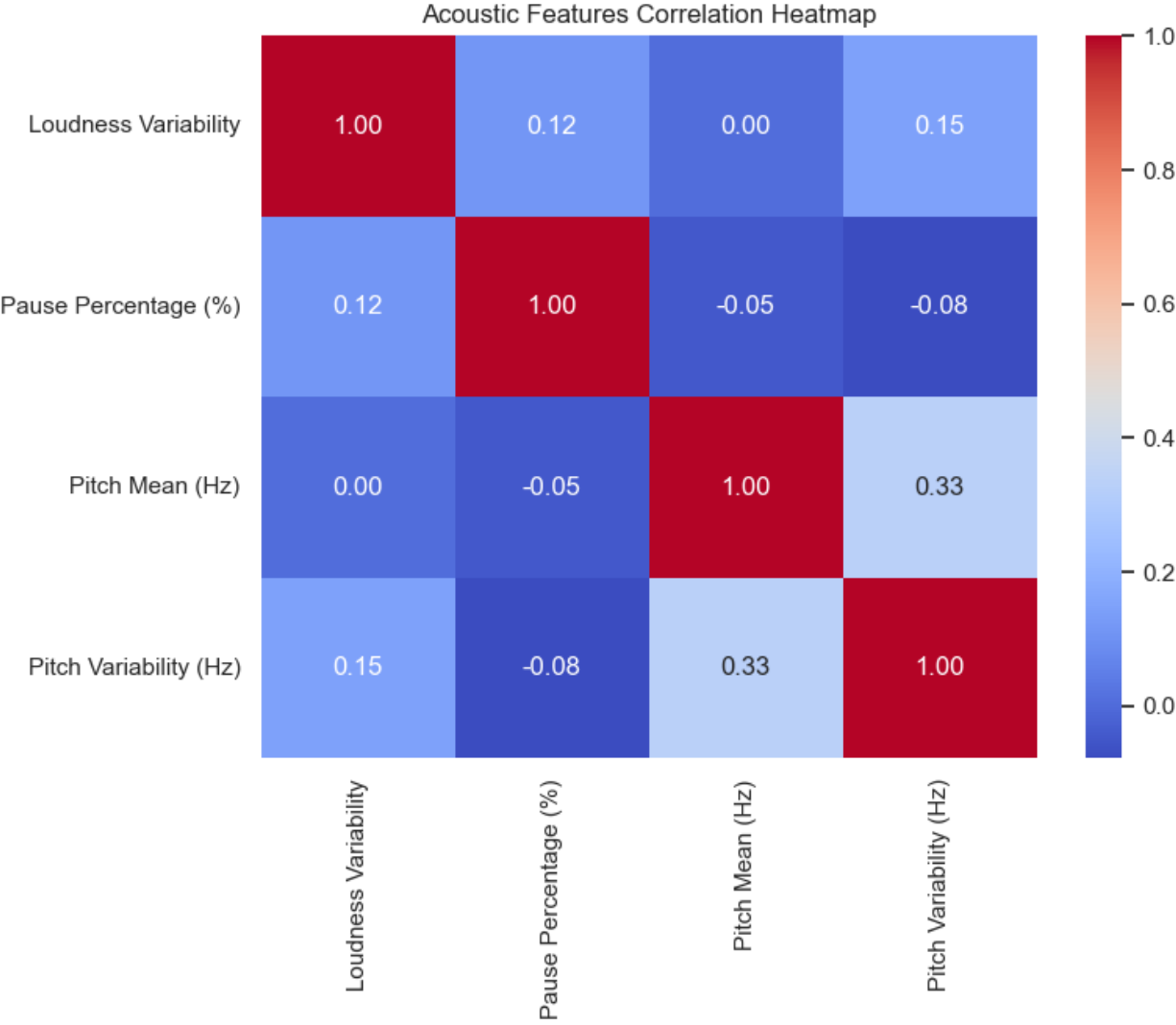


Figura 3.1: Matriz de correlación entre características acústicas prosódicas.

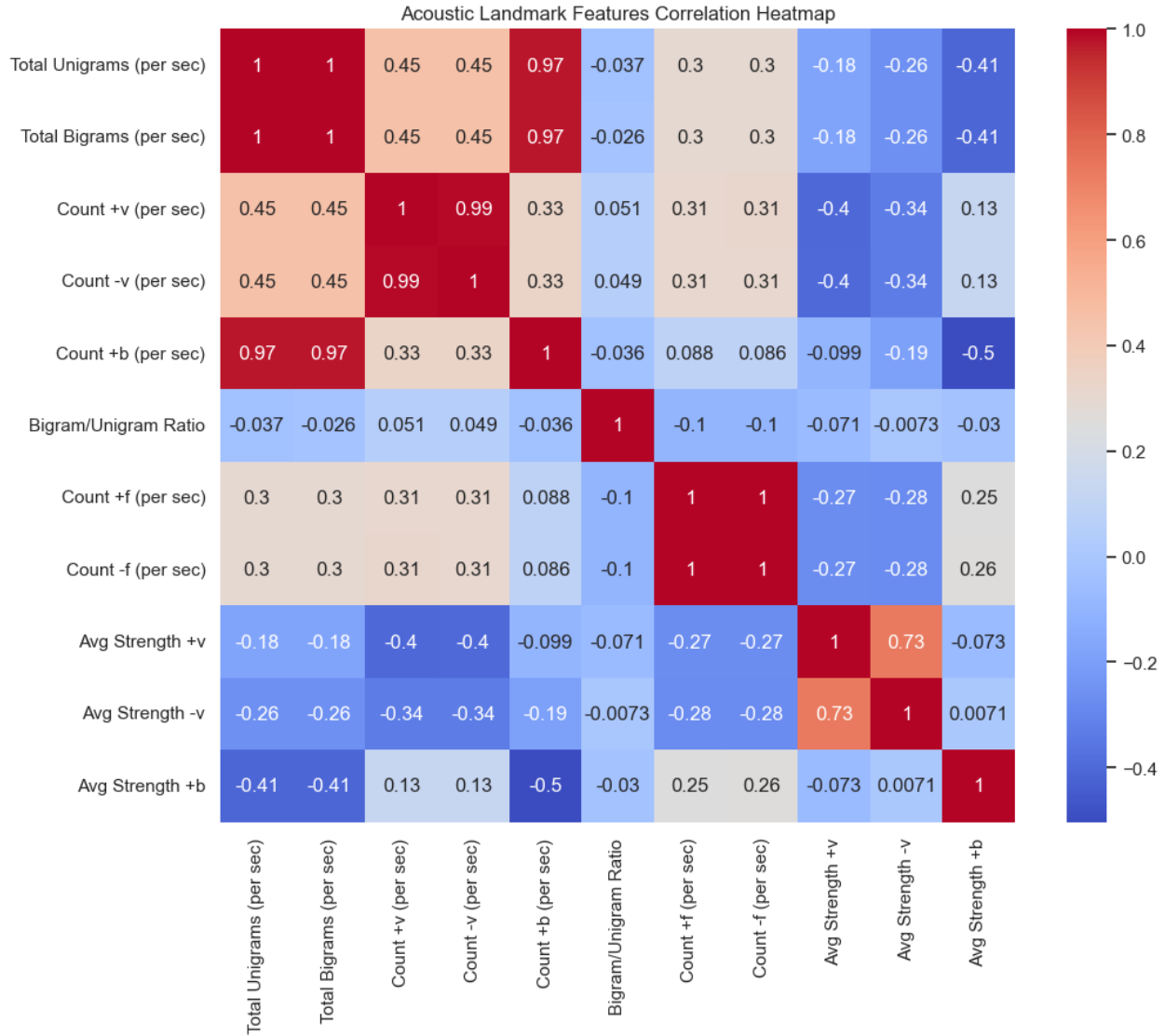


Figura 3.2: Matriz de correlación entre características acústicas derivadas de *landmarks*.

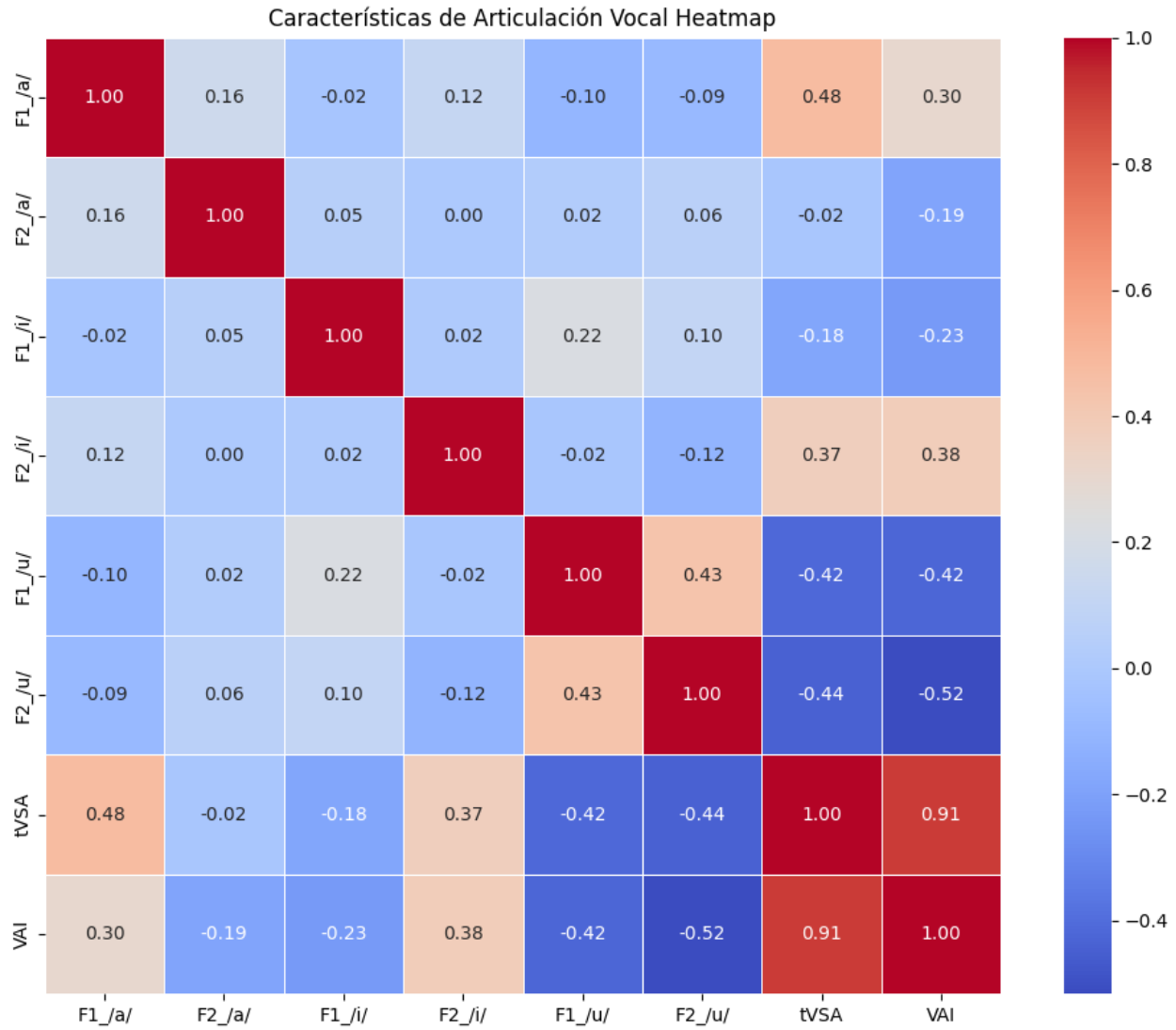


Figura 3.3: Matriz de correlación entre características de articulación vocal.

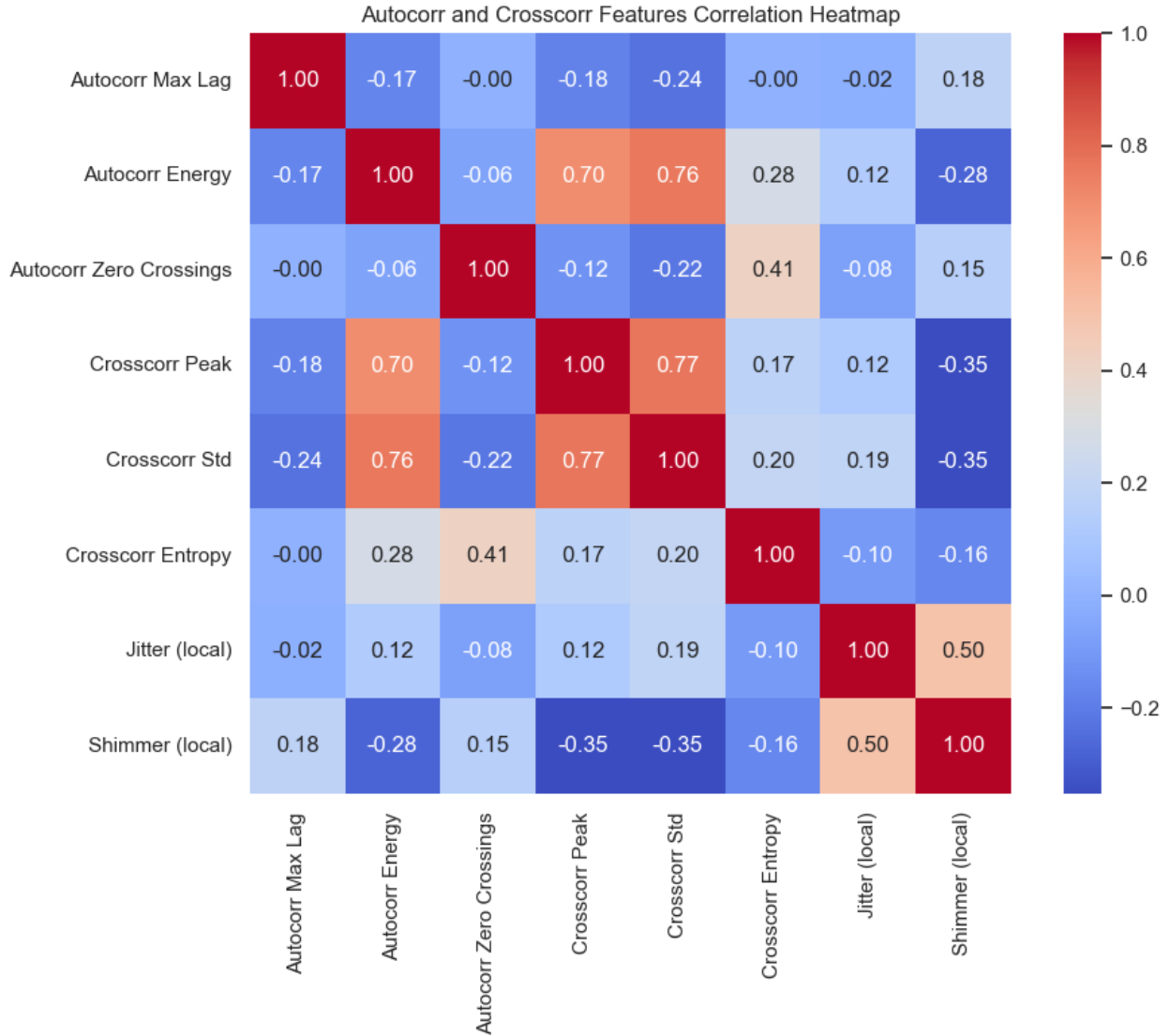


Figura 3.4: Matriz de correlación entre descriptores basados en autocorrelación, correlación cruzada, *jitter* y *shimmer*.

Aunque estas matrices permiten una inspección visual clara de la redundancia dentro de cada grupo de características, es importante destacar que el análisis de correlación no se realizó de manera global sobre todos los *features* simultáneamente. Esta decisión responde a una necesidad práctica: al tratarse de decenas de variables heterogéneas, una matriz de correlación total hubiese generado una visualización excesivamente compleja y poco interpretable. Por ello, se optó por analizar subconjuntos extraídos bajo la misma lógica o *pipeline*.

En la Tabla 3.1 se resumen las correlaciones más elevadas encontradas entre variables dentro de cada grupo, identificando posibles redundancias relevantes para el proceso posterior de selección.

Cuadro 3.1: Resumen de correlaciones altas entre características acústicas

| Característica 1         | Característica 2         | Correlación ( $r$ ) |
|--------------------------|--------------------------|---------------------|
| Total Unigrams (por seg) | Total Bigrams (por seg)  | 1.00                |
| Count +v (por seg)       | Count -v (por seg)       | 0.99                |
| Count +b (por seg)       | Total Unigrams (por seg) | 0.97                |
| tVSA                     | VAI                      | 0.91                |
| Crosscorr Peak           | Crosscorr Std            | 0.77                |
| Autocorr Energy          | Crosscorr Std            | 0.76                |
| Autocorr Energy          | Crosscorr Peak           | 0.70                |

Estas correlaciones serán tenidas en cuenta durante el proceso de eliminación de redundancias previo a la selección de características, con el objetivo de reducir la complejidad del modelo sin comprometer la diversidad informativa de los descriptores.

### 3.3.1. Descriptores basados en autocorrelación y correlación cruzada

Si bien este trabajo emplea una amplia variedad de descriptores acústicos tradicionales y contemporáneos, se ha optado por detallar formalmente únicamente aquellos que representan una contribución novedosa al campo: un conjunto de métricas diseñadas a partir de funciones de autocorrelación y correlación cruzada. La motivación para su diseño surgió a partir de la exposición del autor a conceptos de interferometría astronómica durante su trabajo en el observatorio ALMA, donde tuvo la oportunidad de interactuar con astrónomos e ingenieros especializados en correlación de señales. En dicho contexto, el uso de correlaciones cruzadas para detectar estructuras coherentes en señales de radio extremadamente débiles inspiró una analogía con los patrones temporales de la señal de voz patológica. Así, se propuso la extracción de métricas que capturen regularidades, fluctuaciones y simetrías internas de la señal, como posibles indicadores de alteraciones en la articulación del habla. Por ello, se presentan a continuación sus definiciones matemáticas y su lógica de implementación.

Dado un vector de audio  $y[n]$  de longitud  $N$ , se define su **autocorrelación** como:

$$R_{yy}[k] = \sum_{n=0}^{N-k-1} y[n] \cdot y[n+k], \quad \text{para } k \in [0, N-1] \quad (3.1)$$

A partir de esta función se extraen las siguientes métricas:

- **Autocorr Max Lag:**

$$\arg \max_{k \geq 20} R_{yy}[k] \quad (3.2)$$

Representa el primer pico más alto de la autocorrelación, ignorando los primeros 20 retardos para evitar detectar el pico trivial en  $k = 0$ .

■ **Autocorr Energy:**

$$R_{yy}[0] \quad (3.3)$$

Corresponde a la energía total de la señal, ya que la autocorrelación en  $k = 0$  equivale a la suma de los cuadrados de la señal.

■ **Autocorr Zero Crossings:**

$$\sum_{n=1}^{M-1} \mathbb{1}[R_{yy}[n-1] \cdot R_{yy}[n] < 0] \quad (3.4)$$

Esta métrica cuenta la cantidad de cambios de signo en la curva de autocorrelación, lo que puede reflejar irregularidades rítmicas.

Para los descriptores de **correlación cruzada**, se divide la señal  $y[n]$  en dos mitades:

$$y_1[n] = y[n], \quad n \in [0, \lfloor N/2 \rfloor - 1] \quad (3.5)$$

$$y_2[n] = y[n], \quad n \in [\lfloor N/2 \rfloor, N - 1] \quad (3.6)$$

La **correlación cruzada** se define como:

$$R_{y_1 y_2}[k] = \sum_n y_1[n] \cdot y_2[n + k] \quad (3.7)$$

De esta curva se extraen las siguientes métricas:

■ **Crosscorr Peak:**

$$\max_k R_{y_1 y_2}[k] \quad (3.8)$$

■ **Crosscorr Std:**

$$\sigma_{R_{y_1 y_2}} = \sqrt{\frac{1}{L} \sum_{k=1}^L (R_{y_1 y_2}[k] - \bar{R})^2} \quad (3.9)$$

Donde  $\bar{R}$  es la media de la señal de correlación cruzada.

■ **Crosscorr Entropy:** Dado un vector de probabilidad normalizado:

$$p_k = \frac{|R_{y_1 y_2}[k]|}{\sum_j |R_{y_1 y_2}[j]|} \quad (3.10)$$

la entropía se calcula como:

$$H = - \sum_k p_k \log_2(p_k) \quad (3.11)$$

A diferencia de algunas implementaciones que añaden un pequeño término  $\varepsilon$  para evitar logaritmos de cero, en este trabajo no se utilizó dicha corrección, ya que el vector  $p_k$  está definido a partir de la normalización del valor absoluto de la correlación cruzada, el cual nunca alcanza exactamente cero para señales reales en presencia de ruido. Esta decisión se tomó para evitar sesgos artificiales en la estimación de la dispersión informacional.

### 3.3.2. Selección de características

Antes de aplicar técnicas de selección de características, se realizó una proyección del espacio de descriptores mediante métodos de reducción de dimensionalidad, con el fin de explorar visualmente la complejidad estructural de los datos. Las Figuras 3.5 y 3.6 muestran las distribuciones resultantes al aplicar *t-Distributed Stochastic Neighbor Embedding* (t-SNE) y *Principal Component Analysis* (PCA) sobre el subconjunto de características acústicas de tipo landmark.

A pesar de que este subconjunto representa una fracción del total de descriptores extraídos, las proyecciones evidencian una alta superposición entre clases, junto con una distribución compleja y no lineal de los datos. Este comportamiento sugiere la presencia de redundancias internas y relaciones de alta dimensionalidad, lo cual refuerza la necesidad de aplicar estrategias de selección de características tanto para mejorar la interpretabilidad como para optimizar el rendimiento de los modelos de clasificación posteriores.

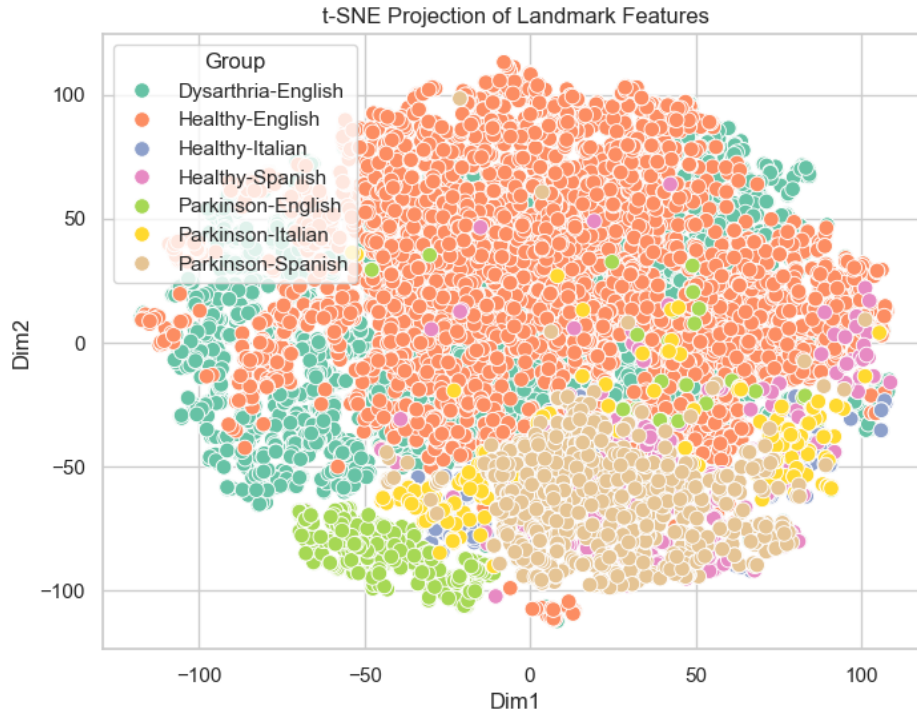


Figura 3.5: Proyección t-SNE del subconjunto de características acústicas de tipo landmark.

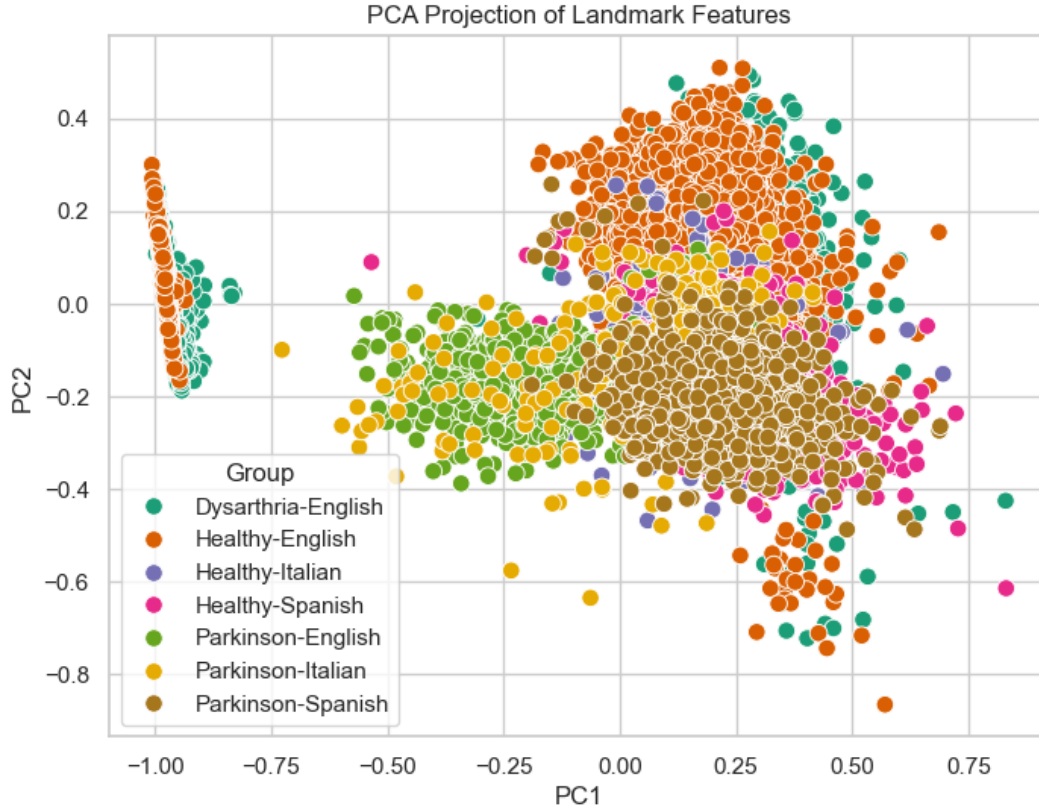


Figura 3.6: Proyección PCA del subconjunto de características acústicas de tipo landmark.

La selección de características es una etapa crítica para la reducción del espacio de búsqueda, la eliminación de redundancias, y la mejora del desempeño predictivo de los modelos. En este trabajo se implementaron dos enfoques complementarios: la eliminación de redundancias mediante análisis de correlación, y la selección recursiva de características mediante validación cruzada (RFECV).

**Redundancia basada en correlación de Pearson.** Como primer paso, se aplicó un criterio de filtrado (*filter method*) para eliminar pares de características altamente correlacionadas (coeficiente de Pearson superior a 0,95). Esta técnica se fundamenta en la idea de que si dos variables están fuertemente correlacionadas, una de ellas puede ser eliminada sin pérdida sustancial de información. Tal como describen Chandrashekar y Sahin [8], los métodos de filtrado permiten reducir la dimensionalidad del conjunto de datos de forma eficiente, sin requerir entrenamiento del modelo. Sin embargo, estos métodos no consideran la interacción entre variables, por lo que se complementan con un segundo enfoque.

**Selección recursiva con validación cruzada (RFECV).** La segunda etapa del proceso corresponde a un método de tipo *wrapper*, en donde el modelo de aprendizaje supervisado guía la selección iterativa de características. Se utilizó el algoritmo Recursive Feature Elimination with Cross-Validation (RFECV), que elimina recursivamente la característica menos relevante de acuerdo a la importancia asignada por el modelo en cada iteración. Este

procedimiento permite encontrar el subconjunto óptimo de características que maximiza el rendimiento del clasificador evaluado mediante validación cruzada. Tal como destacan Dewi y Chen [9], la combinación de métodos de ranking (como Random Forest) con técnicas de validación cruzada permite mejorar sustancialmente la precisión del modelo y reducir el error de predicción.

**Justificación del enfoque combinado.** La combinación de un enfoque de filtrado con uno basado en validación cruzada busca equilibrar eficiencia computacional y capacidad de generalización. Mientras la eliminación por correlación reduce la redundancia estructural, la RFECV permite seleccionar un subconjunto más informativo y no redundante desde el punto de vista del modelo. En conjunto, ambos enfoques buscan evitar el sobreajuste, minimizar la complejidad del modelo y mejorar la interpretabilidad del sistema.

**Evaluación de diferentes umbrales de correlación.** Con el fin de analizar el impacto de la eliminación de redundancias sobre el rendimiento del sistema, se realizaron tres experimentos independientes utilizando distintos umbrales para el coeficiente de correlación de Pearson: 0,95, 0,70 y 0,40. Este análisis buscó identificar el punto de equilibrio entre la eliminación de características efectivamente redundantes y la preservación de aquellas que, si bien correlacionadas, aportan información relevante para la clasificación.

El umbral alto (0,95) permitió conservar la mayoría de los descriptores, eliminando únicamente aquellas variables con redundancias extremas. El umbral medio (0,70) produjo una reducción más agresiva del espacio de características, mientras que el umbral bajo (0,40) tendió a eliminar incluso variables potencialmente informativas, afectando negativamente el rendimiento en algunos casos. Esta comparación permitió observar cómo el grado de tolerancia a la correlación incide directamente en la capacidad discriminativa del conjunto de features final.

Los resultados comparativos de estos tres escenarios se presentan más adelante en la sección de Resultados.

Para ilustrar el funcionamiento del algoritmo RFECV y su sensibilidad al clasificador base, se presentan a continuación tres visualizaciones generadas durante una fase preliminar del proyecto. La Figura 3.7 muestra la curva de validación cruzada utilizando LinearSVC como modelo base, mientras que la Figura 3.8 presenta los resultados obtenidos con Random Forest. Finalmente, la Figura 3.9 compara directamente ambas curvas, destacando las diferencias en la estabilidad y número óptimo de características seleccionadas. Si bien estas figuras no corresponden al conjunto de datos final, resultan útiles para comprender el comportamiento general del proceso de selección.

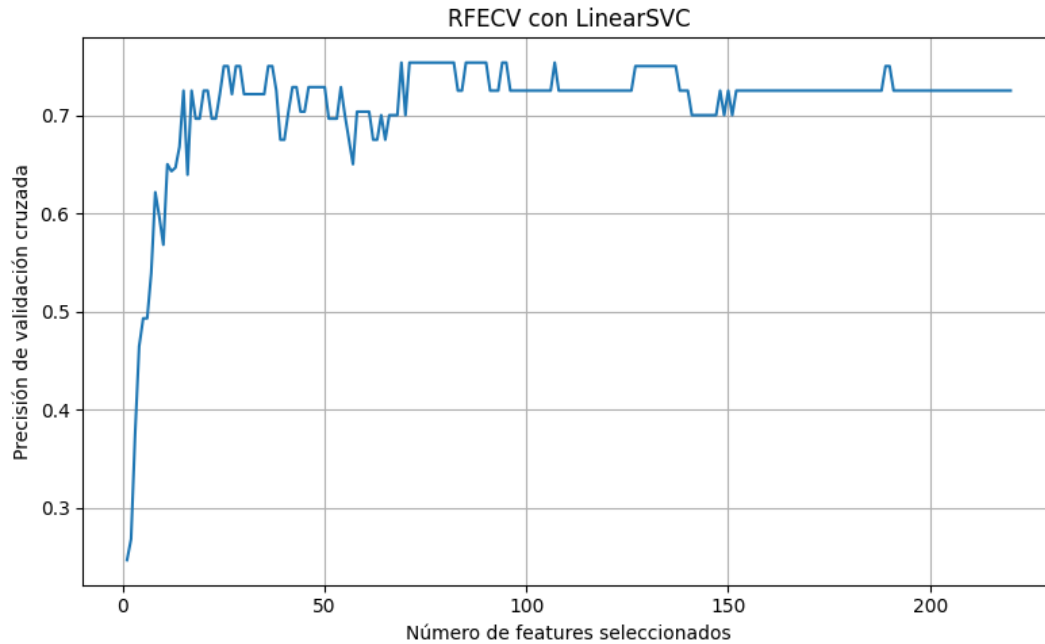


Figura 3.7: Curva de validación cruzada del algoritmo RFECV utilizando LinearSVC como modelo base. Resultados generados durante la fase exploratoria.

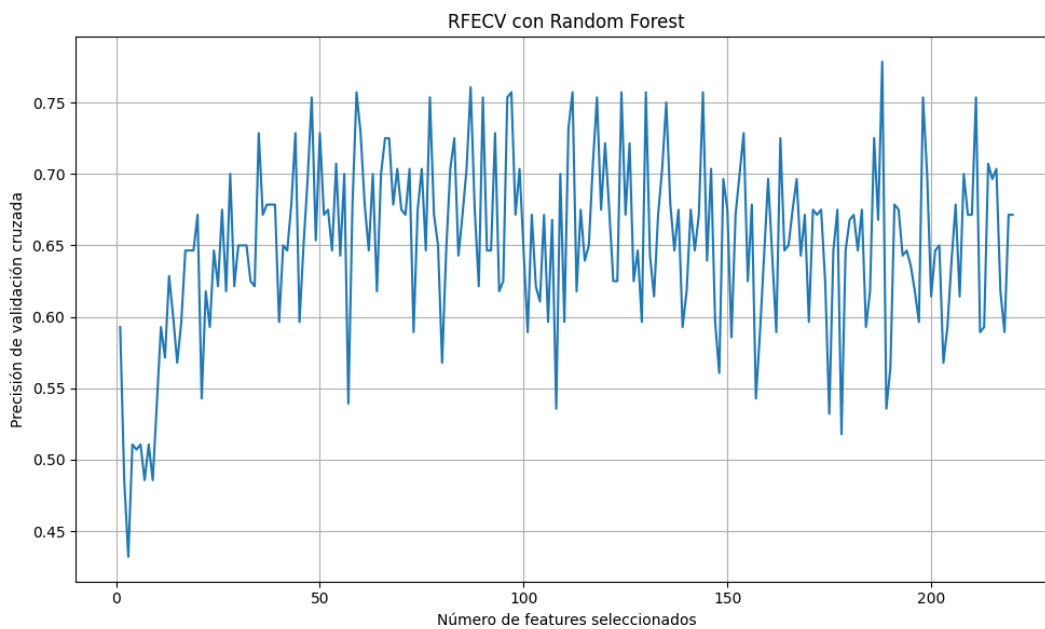


Figura 3.8: Curva de validación cruzada del algoritmo RFECV utilizando Random Forest como modelo base. Resultados generados durante la fase exploratoria.

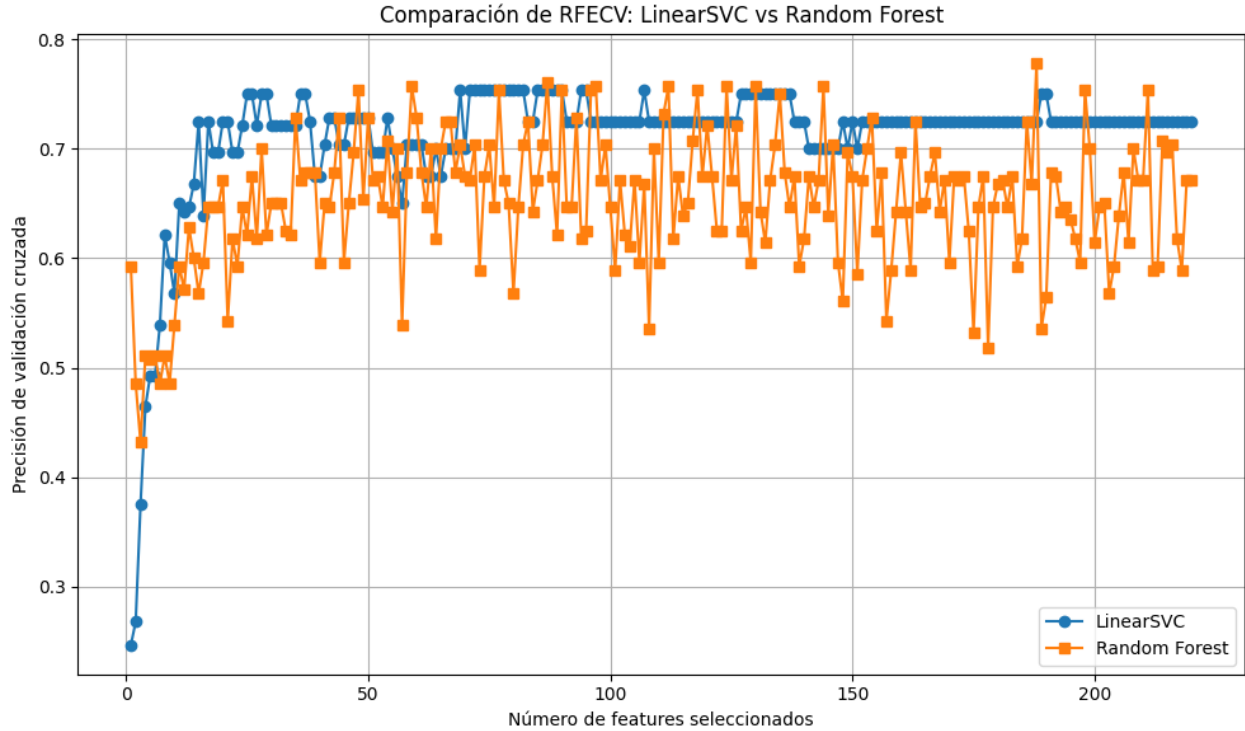


Figura 3.9: Comparación directa de la precisión de validación cruzada entre LinearSVC y Random Forest como clasificadores base en RFECV. Fase preliminar.

La combinación de eliminación de redundancia basada en correlación y selección recursiva mediante RFECV permitió construir un conjunto reducido y representativo de características, preservando aquellas con mayor poder discriminativo. Este enfoque híbrido aporta robustez y eficiencia al pipeline de clasificación, alineándose con los objetivos de generalización multi-lingüe y multi-patológico propuestos en este trabajo.

### 3.4. Modelos de clasificación y estrategias de entrenamiento

Para evaluar la eficacia de los descriptores acústicos propuestos, se implementaron dos enfoques complementarios de entrenamiento y clasificación: (1) un esquema directo de clasificación multiclase y (2) una estrategia escalonada en dos etapas (*two-tiered classification*). En ambos casos se utilizaron múltiples clasificadores con distintas configuraciones y técnicas de selección de características asociadas.

### 3.4.1. Modelos de clasificación utilizados

Se exploraron los siguientes algoritmos supervisados, seleccionados por su capacidad de generalización, robustez frente al ruido y diversidad de enfoques (lineales, no lineales y basados en árboles):

- **LinearSVC**: modelo lineal eficiente y ampliamente utilizado en tareas de clasificación de alta dimensionalidad. Se usó como base tanto para clasificación directa como para selección de características.
- **Random Forest Classifier**: conjunto de árboles de decisión que ofrece una medida interna de importancia de variables. Su tolerancia al sobreajuste y capacidad de modelar relaciones no lineales lo hacen especialmente útil en escenarios complejos.
- **SVC con kernel RBF**: clasificador no lineal que permite capturar patrones complejos mediante la proyección de los datos en espacios de mayor dimensión.
- **Multilayer Perceptron (MLP)**: red neuronal densa, utilizada para explorar modelos más expresivos, aunque con mayor riesgo de sobreajuste si no se dispone de regularización adecuada.

### 3.4.2. Clasificación multiclase directa

En este enfoque, los modelos intentan predecir directamente la clase específica de cada muestra de audio, entre un total de siete clases posibles, que combinan idioma y tipo de patología (e.g., *Healthy-English*, *Parkinson-Italian*, etc.). Este esquema busca evaluar la capacidad global del sistema para realizar una discriminación detallada sin etapas intermedias.

Para este experimento se probaron modelos con y sin selección de características previa. En los casos donde se aplicó selección, se utilizó RFECV con LinearSVC o la medida de importancia de Random Forest, dependiendo del clasificador. Los MLP fueron entrenados sobre los conjuntos reducidos generados por estos dos modelos.

### 3.4.3. Clasificación en dos etapas (two-tiered)

El segundo enfoque consistió en un pipeline secuencial dividido en dos fases:

1. **Clasificación binaria**: el primer modelo clasifica cada muestra como “patológica” o “no patológica”.
2. **Clasificación multiclase patológica**: aquellas muestras clasificadas como patológicas son luego ingresadas a un segundo modelo que predice la clase patológica específica.

Este enfoque refleja un escenario clínico más realista y ha sido respaldado recientemente por investigaciones en el área. Al-Ali et al.[2] proponen una arquitectura de dos etapas para clasificar casos de disartria, en la que se emplea primero una clasificación binaria para distinguir entre habla saludable y patológica, y posteriormente un modelo no supervisado para clasificar el grado de severidad. Inspirados en esta estrategia, nuestro trabajo adapta la lógica binaria-multiclase a un escenario multilingüe y multi-patológico, aplicando un pipeline secuencial que permite minimizar falsos positivos en el grupo saludable y optimizar la especialización del segundo clasificador (ver Figura 3.10).

Para cada una de las dos etapas del pipeline se aplicó selección de características independiente. En el caso de los modelos SVC (lineal o RBF), se utilizó RFECV con LinearSVC como selector. En el caso del Random Forest, se utilizó su propia métrica de importancia de características para generar los subconjuntos reducidos.

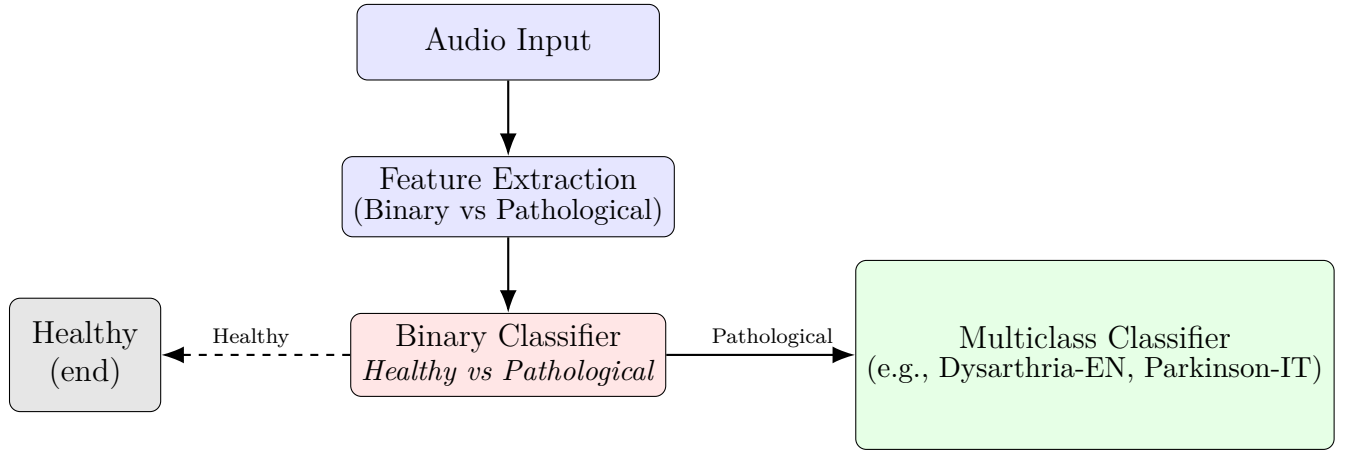


Figura 3.10: Esquema general del pipeline de clasificación en dos etapas: una primera fase binaria filtra los casos saludables, y una segunda fase multiclase identifica la patología específica.

### 3.5. Estrategias de evaluación

Con el fin de evaluar el rendimiento de los distintos modelos de clasificación implementados, se empleó un protocolo riguroso de validación y análisis métrico. Todas las pruebas se realizaron bajo un esquema de validación cruzada estratificada de 5 pliegues (*Stratified 5-Fold Cross Validation*), garantizando que cada clase estuviera proporcionalmente representada en cada partición.

### 3.5.1. Métricas de evaluación

Se utilizaron cuatro métricas fundamentales para evaluar el desempeño de los modelos, tanto a nivel global como por clase:

Cuadro 3.2: Métricas de evaluación utilizadas para la clasificación

| Métrica          | Fórmula                                                                                 | Descripción                                                                                                                 |
|------------------|-----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| <b>Accuracy</b>  | $\frac{TP + TN}{TP + TN + FP + FN}$                                                     | Proporción de predicciones correctas sobre el total de muestras evaluadas.                                                  |
| <b>Precision</b> | $\frac{TP}{TP + FP}$                                                                    | Proporción de verdaderos positivos entre todos los casos clasificados como positivos por el modelo.                         |
| <b>Recall</b>    | $\frac{TP}{TP + FN}$                                                                    | También conocido como sensibilidad. Mide la proporción de verdaderos positivos respecto al total de casos reales positivos. |
| <b>F1-score</b>  | $\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ | Media armónica entre precisión y recall. Útil en escenarios con clases desbalanceadas.                                      |

Estas métricas fueron computadas a nivel global por modelo, y también desagregadas por clase en los escenarios multiclase, lo que permitió analizar la consistencia del rendimiento entre distintos grupos patológicos e idiomas.

### 3.5.2. Criterios comparativos entre modelos

La evaluación se realizó de forma sistemática, aplicando cada modelo a múltiples configuraciones de datos:

- Conjunto completo de características vs. conjunto reducido (tras selección).
- Umbrales de correlación distintos para reducción de redundancia (0,95, 0,70, 0,40).
- Clasificación directa multiclase vs. clasificación escalonada (two-tiered).

De esta manera, se pudo comparar no solo el rendimiento absoluto de cada modelo, sino también su sensibilidad a la selección de características y al diseño arquitectónico del pipeline. La elección del mejor modelo se basó en la combinación de precisión global, consistencia por clase y estabilidad entre pliegues.

## 3.6. Interpretabilidad del modelo

Más allá del rendimiento cuantitativo, este trabajo busca también comprender qué características acústicas tienen mayor influencia en la toma de decisiones de los modelos de clasificación. Para ello, se aplicó el marco de interpretación basado en **SHAP (SHapley Additive exPlanations)** [22], una metodología que permite cuantificar el aporte de cada característica a la predicción final, tanto a nivel global como por muestra individual.

### 3.6.1. Cálculo de valores SHAP

Los valores SHAP fueron calculados para dos modelos seleccionados por su desempeño y relevancia estructural: **LinearSVC** y **Random Forest**, ambos entrenados sobre el conjunto reducido de características obtenidas tras la selección.

Se utilizaron los explicadores más adecuados para cada tipo de modelo:

- **LinearExplainer**, para el modelo **LinearSVC**, aprovechando su estructura lineal y eficiente.
- **TreeExplainer**, para el modelo **Random Forest**, el cual permite una interpretación exacta sobre estructuras de árboles.

El análisis se realizó en dos niveles:

- **Global**: para identificar las características más influyentes en la toma de decisiones general del modelo.
- **Local**: aplicado a muestras mal clasificadas, con el objetivo de detectar patrones que expliquen los errores de predicción.

Este enfoque de interpretabilidad permite aportar transparencia al proceso de clasificación, facilitar el análisis clínico y guiar el diseño de futuras mejoras del sistema. Para ilustrar gráficamente el funcionamiento de los distintos métodos de atribución de características, en la Figura 3.11 se muestra un ejemplo tomado de Lundberg y Lee [[21]]. En él se observa cómo SHAP mantiene la consistencia entre modelos a diferencia de otros enfoques como Gain, Split Count o Saabas.

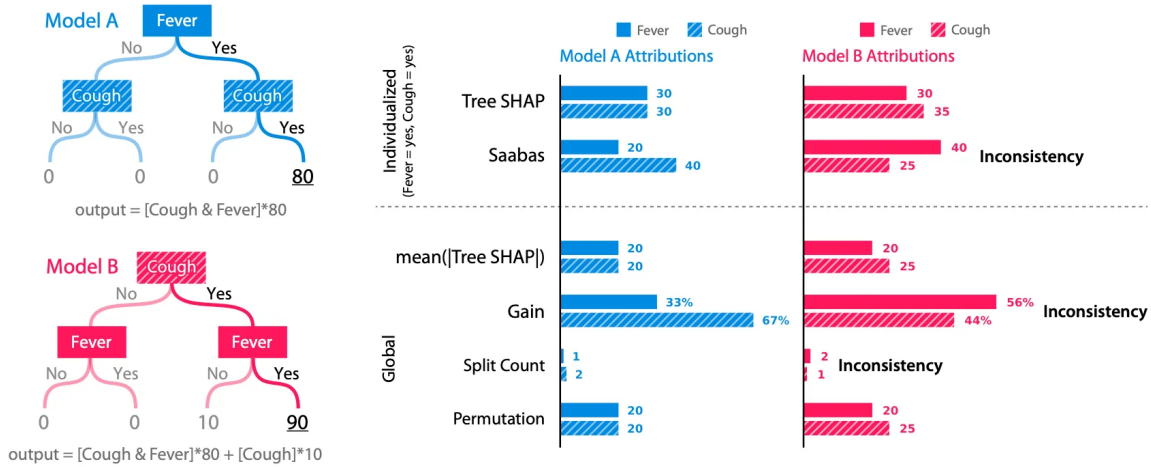


Figura 3.11: Ejemplo ilustrativo del funcionamiento de SHAP en modelos de árboles. La figura compara diferentes métodos de atribución de características, mostrando cómo SHAP mantiene la consistencia en comparación con métodos como Gain, Split Count y Saabas. Imagen tomada de Lundberg et al. [21].

# Capítulo 4

## Resultados Experimentales

### 4.1. Justificación del Enfoque Experimental sin MFCC

A pesar del uso extendido de los coeficientes cepstrales en escala Mel (MFCC) en tareas de clasificación de voz, este trabajo ha optado por relegar su uso a un rol comparativo. Esta decisión se sustenta en tres pilares: (1) la necesidad de explorar descriptores alternativos más interpretables, (2) el riesgo de que los MFCC dominen el espacio de características y opaquen otros efectos relevantes, y (3) la intención de construir modelos clínicamente transparentes. Si bien los MFCC ofrecen un rendimiento competitivo, los experimentos muestran que pueden ser prescindibles en favor de un conjunto más explicativo.

### 4.2. Comparación de métricas entre modelos

#### 4.2.1. Descripción General

En esta serie de experimentos se evaluó la capacidad discriminativa de distintos conjuntos de características para realizar una clasificación directa entre múltiples patologías. Se compararon configuraciones con todos los features disponibles (incluyendo MFCC), sin MFCC, y exclusivamente MFCC. En cada una se aplicaron tres umbrales diferentes para eliminar características altamente correlacionadas (0.95, 0.70 y 0.40). Además, se analizaron los efectos del uso o no de técnicas de selección de características (FS: Feature Selection; Raw: sin selección).

Se incluye también un modelo adicional: **LinearSVC (Linguistic Merged FS)**, que incorporó características lingüísticas extraídas automáticamente. Este modelo fue entrenado con un subconjunto reducido (457 audios) debido a limitaciones presupuestarias. Aunque mostró un rendimiento inferior, se incluye dada la relevancia potencial de estas características

en entornos multilingües y su valor exploratorio para estudios futuros.

## 4.2.2. Análisis de la clasificación multiclase directa

### Umbral de correlación de Pearson = 0.95

La Figura 4.1 muestra un heatmap comparativo con los resultados de la clasificación directa multiclase utilizando todos los features acústicos disponibles (arriba: con MFCC; abajo: sin MFCC) y un umbral de correlación de 0.95 para eliminar redundancias. Se presentan las métricas de accuracy, precisión, recall y F1-score para cada modelo y configuración de selección de características.

Con un umbral permisivo como 0.95, solo se eliminan variables extremadamente redundantes, lo que permite evaluar el rendimiento de los modelos en un escenario de alta dimensionalidad. Se observa que los modelos no lineales (Random Forest y MLP) logran desempeños sobresalientes, superando el 96 % de exactitud incluso sin MFCC, mientras que los modelos lineales se ven más afectados por la reducción de información espectral.

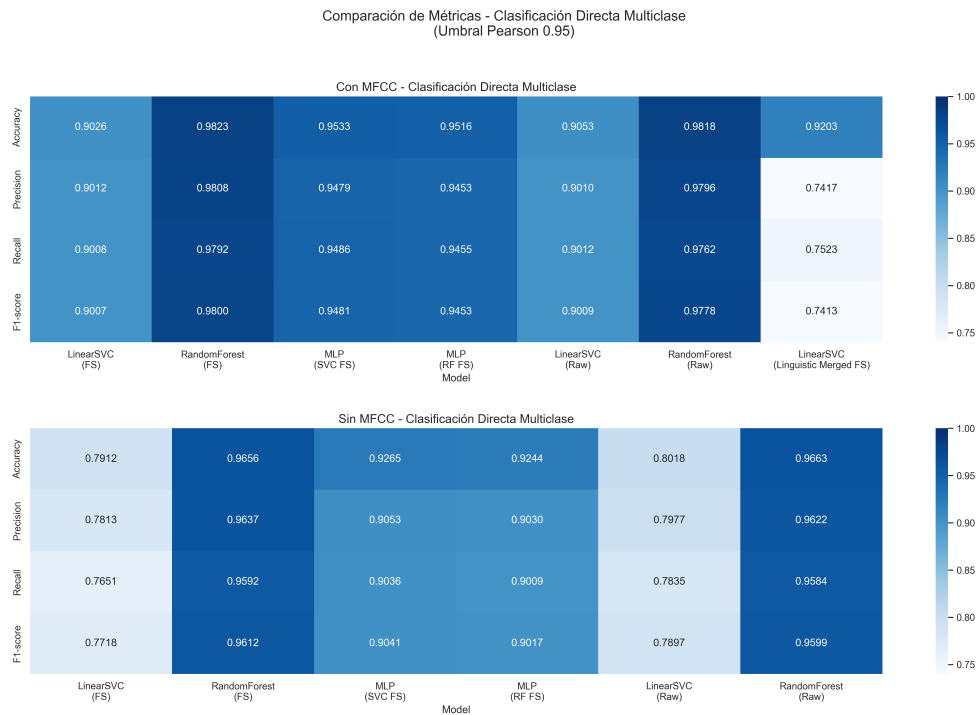


Figura 4.1: Heatmap de métricas para clasificación multiclase directa con (arriba) y sin MFCC (abajo). Umbral Pearson 0.95.

## Umbral de correlación de Pearson = 0.70

La Figura 4.2 presenta los resultados equivalentes usando un umbral más estricto de 0.70, que elimina un mayor número de características correlacionadas. Esto permite observar cómo la reducción moderada de dimensionalidad impacta el desempeño de los modelos.

En este escenario, los modelos no lineales mantienen su robustez, aunque muestran una ligera reducción en las métricas respecto al umbral más alto. Los modelos lineales (LinearSVC) se ven más afectados, especialmente cuando se aplicó selección de características, lo que evidencia su mayor dependencia de ciertas variables para lograr un buen desempeño.

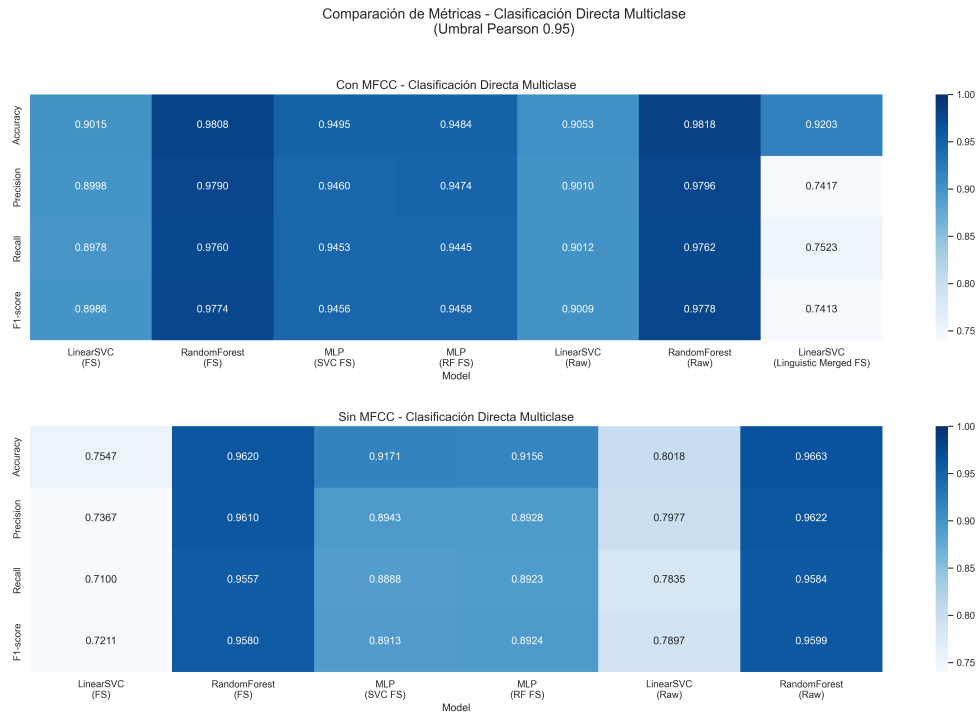


Figura 4.2: Heatmap de métricas para clasificación multiclase directa con (arriba) y sin MFCC (abajo). Umbral Pearson 0.70.

## Umbral de correlación de Pearson = 0.40

La Figura 4.3 ilustra los resultados más extremos, aplicando un umbral de correlación de 0.40. Esta reducción agresiva de características pone a prueba la resiliencia de los modelos ante la pérdida de información redundante y parcialmente relevante.

Bajo este umbral, la brecha de desempeño entre modelos no lineales y lineales se hace aún más marcada. Random Forest y MLP conservan métricas aceptables, mientras que LinearSVC con selección de características experimenta una caída pronunciada, reflejando la dificultad

de los modelos lineales para adaptarse a espacios muy reducidos de características.

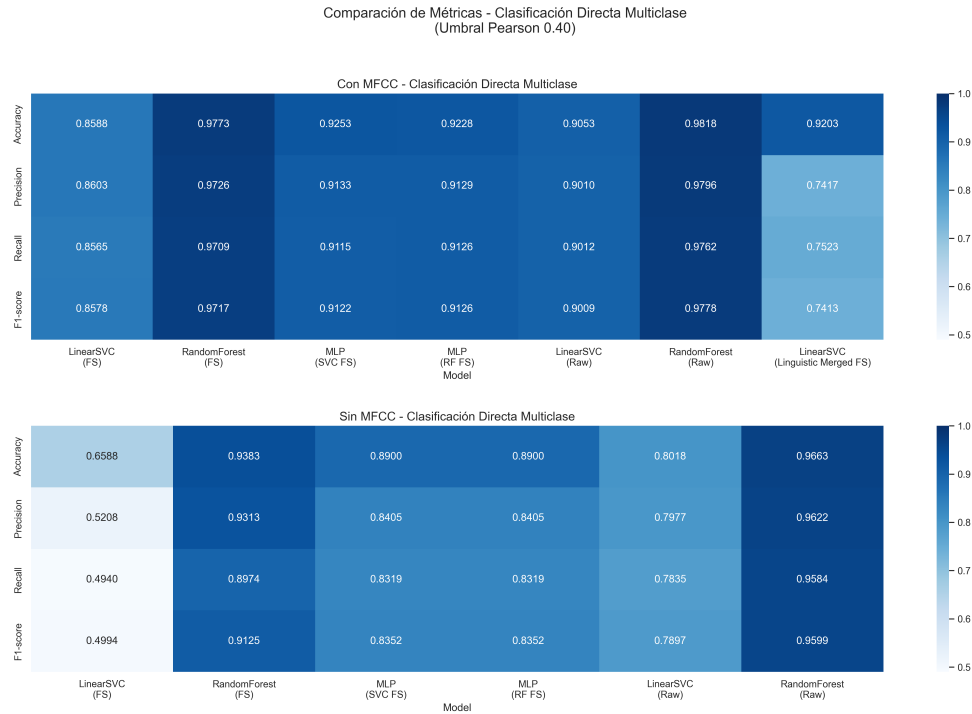


Figura 4.3: Heatmap de métricas para clasificación multiclase directa con (arriba) y sin MFCC (abajo). Umbral Pearson 0.40.

### Clasificación multiclase utilizando únicamente features MFCC. Umbral de correlación de Pearson = 0.95

La Figura 4.4 presenta un heatmap con los resultados obtenidos al entrenar modelos de clasificación multiclase utilizando exclusivamente los coeficientes MFCC, bajo un umbral de correlación de 0.95. Se muestran las métricas de accuracy, precisión, recall y F1-score para distintas configuraciones de modelo y selección de características.

Este experimento permite comparar directamente el poder discriminativo de los MFCC, considerados el estándar clásico en procesamiento de voz, frente al uso combinado o alternativo de descriptores acústicos personalizados. Si bien los modelos no lineales (Random Forest y MLP) siguen mostrando un rendimiento competitivo, los resultados globales son ligeramente inferiores a los obtenidos con todos los features combinados. Esto sugiere que, aunque los MFCC capturan información relevante, la integración de características adicionales puede aportar valor complementario.

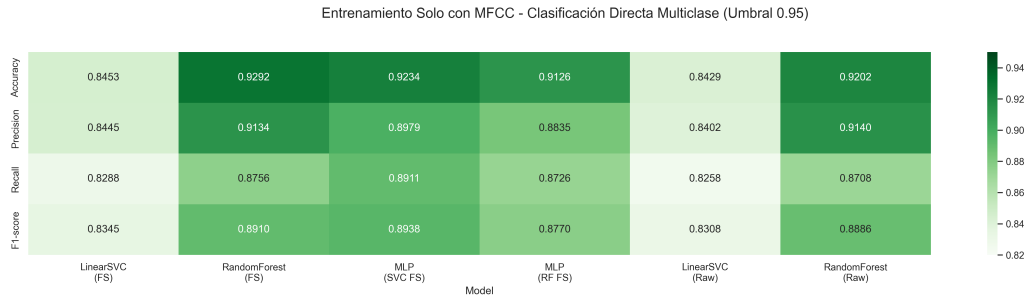


Figura 4.4: Heatmap de métricas para clasificación multiclase directa utilizando únicamente MFCC. Umbral Pearson 0.95.

### 4.2.3. Análisis de la clasificación mediante pipeline escalonada (two-tiered)

**Umbral de correlación de Pearson = 0.95**

La Figura 4.5 presenta los resultados de la arquitectura two-tiered con un umbral de correlación de 0.95, mostrando métricas para cada modelo y etapa (binaria y multiclase), tanto con MFCC (arriba) como sin MFCC (abajo). El umbral 0.95 es permisivo y conserva la mayoría de las variables originales, permitiendo observar el desempeño del pipeline en condiciones casi ideales. Destaca que Random Forest logra métricas superiores al 96 % incluso en la etapa binaria sin MFCC, mientras que en la multiclase, todos los modelos se acercan a la perfección (cerca o superiores al 99 %). Los modelos no lineales son particularmente robustos a la eliminación de MFCC y a la reducción mínima de features.



Figura 4.5: Heatmap de métricas para pipeline two-tiered. Arriba: con MFCC, abajo: sin MFCC. Umbral Pearson 0.95.

### Umbral de correlación de Pearson = 0.70

La Figura 4.6 muestra los resultados equivalentes con un umbral más estricto de 0.70, lo que elimina más variables redundantes del conjunto de entrenamiento. Se observa una ligera caída en las métricas, sobre todo en la etapa binaria y en los modelos lineales (LinearSVC). No obstante, tanto Random Forest como SVC-RBF se mantienen altamente precisos en ambas etapas y configuraciones, demostrando su robustez ante la reducción de características. La etapa multiclase sigue presentando métricas casi perfectas, validando la eficacia de los descriptores acústicos personalizados.



Figura 4.6: Heatmap de métricas para pipeline two-tiered. Arriba: con MFCC, abajo: sin MFCC. Umbral Pearson 0.70.

### Umbral de correlación de Pearson = 0.40

La Figura 4.7 muestra el caso más extremo, con un umbral agresivo de 0.40. Aquí, la cantidad de características se reduce drásticamente, permitiendo analizar la resiliencia de los modelos ante la pérdida masiva de información redundante.

En la etapa binaria, el modelo lineal (LinearSVC) sufre una disminución importante en todas las métricas, mientras que los modelos no lineales (SVC-RBF y Random Forest) mantienen un desempeño sobresaliente, especialmente Random Forest, que supera el 94 % de exactitud. En la etapa multiclase, el rendimiento se mantiene alto en todos los modelos, lo que demuestra que el conjunto reducido sigue siendo altamente discriminativo para diferenciar patologías específicas.



Figura 4.7: Heatmap de métricas para pipeline two-tiered. Arriba: con MFCC, abajo: sin MFCC. Umbral Pearson 0.40.

### 4.3. Impacto del Umbral de Correlación en la Arquitectura Two-Tiered

#### 4.3.1. Resultados en la Fase Binaria

Para profundizar en el comportamiento de los modelos de clasificación binaria, se presenta una serie de gráficos que muestran la evolución de las principales métricas de desempeño (*precision*, *recall*, *f1-score* y *accuracy* global) en función del umbral de correlación de Pearson utilizado para la selección de características. En este caso, se analizaron las clases **Healthy** y **Pathological** para los modelos LinearSVC, SVC-RBF y RandomForest, empleando exclusivamente características acústicas personalizadas (sin MFCC).

**Evolución de las métricas por clase y modelo** En las Figuras 4.8–4.10 se observa que, tanto para la clase **Healthy** como para **Pathological**, el modelo RandomForest exhibe consistentemente los valores más altos en todas las métricas evaluadas, incluso cuando se reduce el umbral de Pearson y, por lo tanto, la cantidad de características disponibles. Esta tendencia se mantiene para precisión, recall y f1-score, confirmando la notable robustez de los modelos no lineales ante la eliminación de variables correlacionadas.

SVC-RBF también presenta un rendimiento elevado y estable, aunque ligeramente inferior al de RandomForest. En contraste, el modelo lineal LinearSVC muestra una caída progresiva en sus métricas a medida que disminuye el umbral de correlación, lo que evidencia una mayor sensibilidad a la reducción de dimensionalidad. Esta diferencia de comportamiento resalta la capacidad de los modelos no lineales para capturar patrones complejos y adaptarse a conjuntos de características más restringidos.

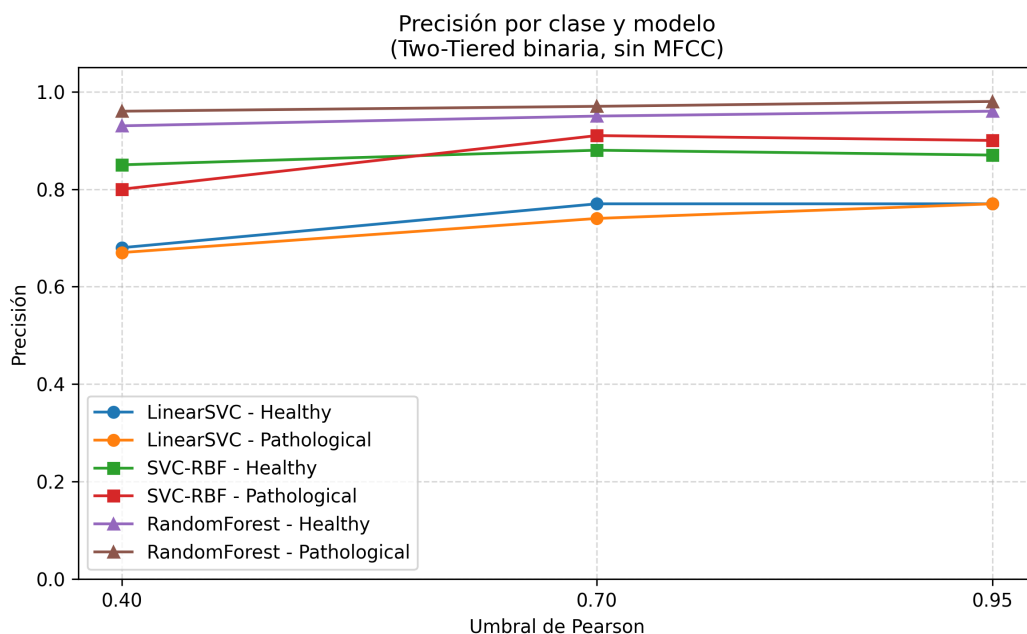


Figura 4.8: Evolución de la **precisión** por clase y modelo a través de distintos umbrales de correlación de Pearson (fase binaria, two-tiered sin MFCC).

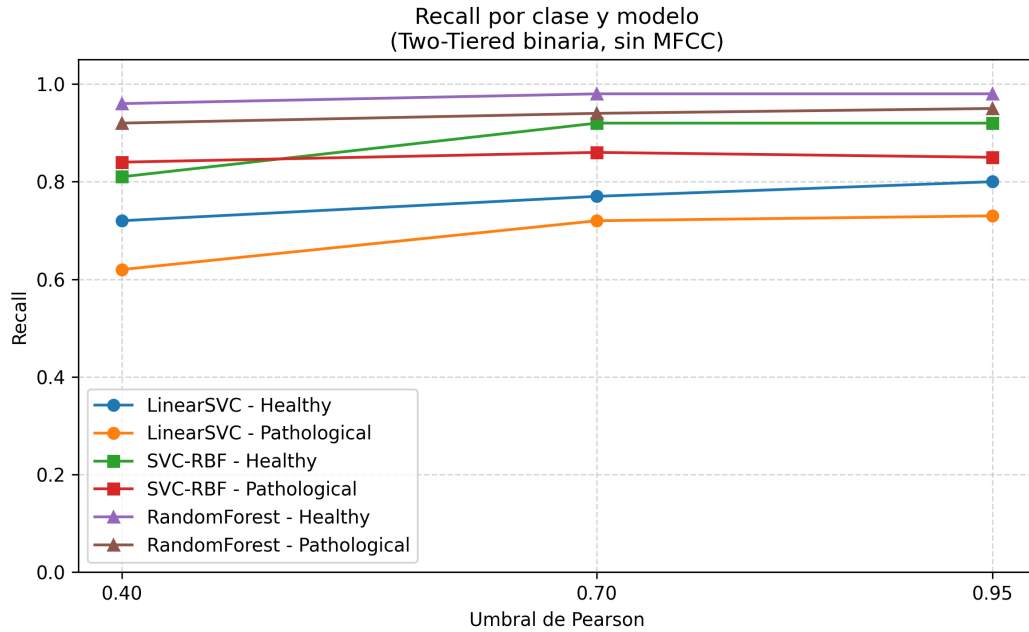


Figura 4.9: Evolución del **recall** por clase y modelo según el umbral de correlación de Pearson (fase binaria, two-tiered sin MFCC).

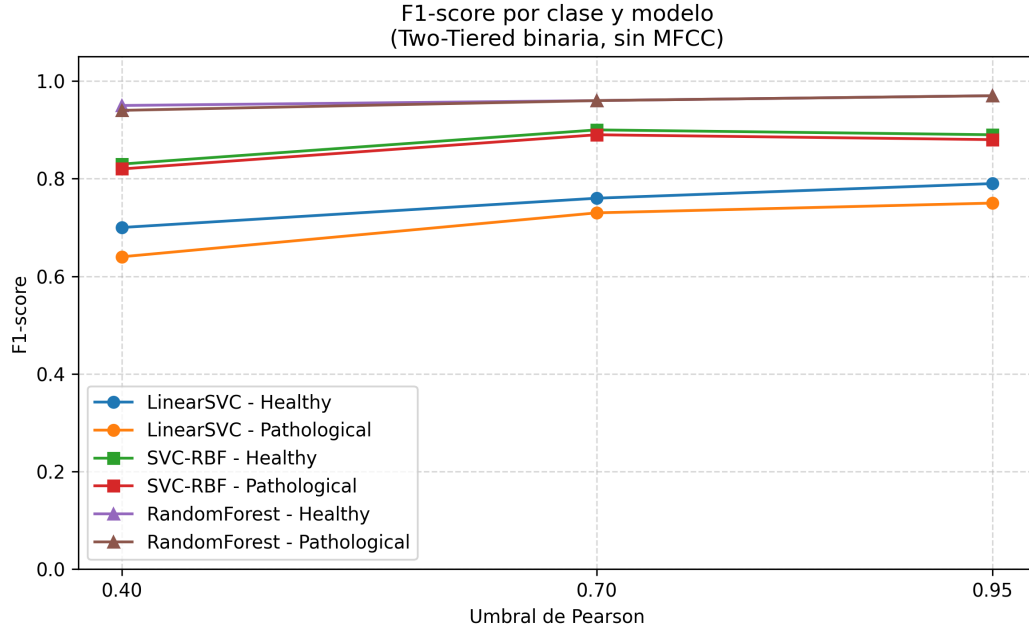


Figura 4.10: Evolución del **f1-score** por clase y modelo en función del umbral de correlación de Pearson (fase binaria, two-tiered sin MFCC).

**Comportamiento de la *accuracy* global** La Figura 4.11 muestra la **accuracy** global para cada modelo a través de los distintos umbrales. Se evidencia que RandomForest mantiene una exactitud muy superior al resto, mientras que SVC-RBF también muestra estabilidad y valores elevados. Por su parte, LinearSVC es notablemente más sensible a la reducción del umbral, experimentando una caída significativa en accuracy, en particular al pasar de 0.95 a 0.40.

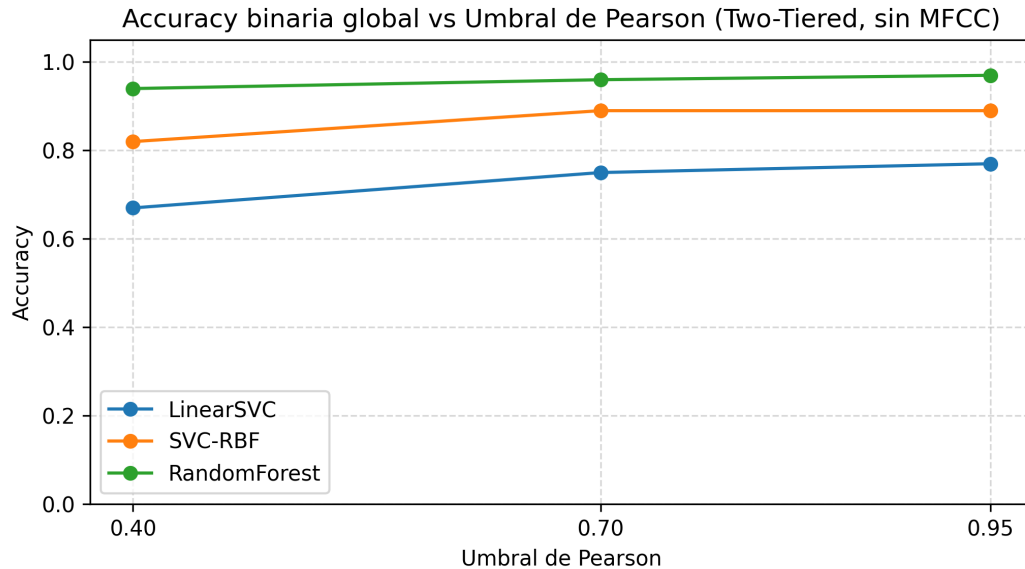


Figura 4.11: Accuracy binaria global para cada modelo en función del umbral de Pearson (Two-Tiered, sin MFCC).

**Análisis e implicancias** El análisis de estas curvas permite concluir que:

- **RandomForest** es sistemáticamente el modelo más robusto y preciso, con desempeño sobresaliente en ambas clases y para todos los umbrales evaluados. Esto lo posiciona como la mejor alternativa en contextos donde la resiliencia ante reducción de variables es prioritaria.
- **SVC-RBF** se comporta de manera similar a RandomForest, aunque con un leve descenso en métricas, y es significativamente más estable que LinearSVC.
- **LinearSVC** evidencia una marcada vulnerabilidad ante la eliminación de características correlacionadas, especialmente cuando el set de variables se vuelve muy reducido, lo que limita su aplicabilidad práctica en escenarios de alta selección de features.
- La reducción del umbral de Pearson tiene un impacto real sobre las métricas de desempeño, en especial para los modelos lineales, pero en general el efecto es moderado para los modelos no lineales, que mantienen resultados elevados incluso bajo configuraciones más restrictivas.

Este análisis gráfico refuerza la importancia de elegir modelos no lineales y descriptores acústicos robustos para la detección binaria automatizada de patologías del habla, especialmente en entornos multilingües y multipatológicos donde la dimensionalidad y la redundancia de los features pueden variar considerablemente.

### 4.3.2. Resultados en la Fase Multiclase

Tras la etapa binaria del pipeline two-tiered, la fase multiclase permite analizar la capacidad de los modelos para diferenciar entre diversas patologías específicas del habla y variantes idiomáticas. Para ello, se presentan a continuación las curvas de precisión, recall, f1-score y accuracy, evaluadas para cada clase y modelo en función del umbral de correlación de Pearson empleado en la selección de características acústicas (excluyendo MFCC).

**Evolución de las métricas por clase y modelo** Como se observa en las Figuras 4.12–4.14, la principal conclusión es la **estabilidad casi absoluta** de las métricas frente a la variación del umbral de Pearson. Tanto la precisión, el recall como el f1-score se mantienen virtualmente inalterados para todos los modelos y clases a través de los diferentes valores de umbral, lo que evidencia una notable resiliencia de los modelos —en particular RandomForest y SVC-RBF— y la alta calidad discriminativa de las características seleccionadas. En otras palabras, la reducción de variables redundantes no afecta el desempeño multiclase, a diferencia de lo que ocurre en la etapa binaria.

Estos resultados muestran que, una vez superado el filtro inicial binario, los modelos cuentan con suficientes variables informativas para mantener su rendimiento óptimo en la identificación específica de cada patología.

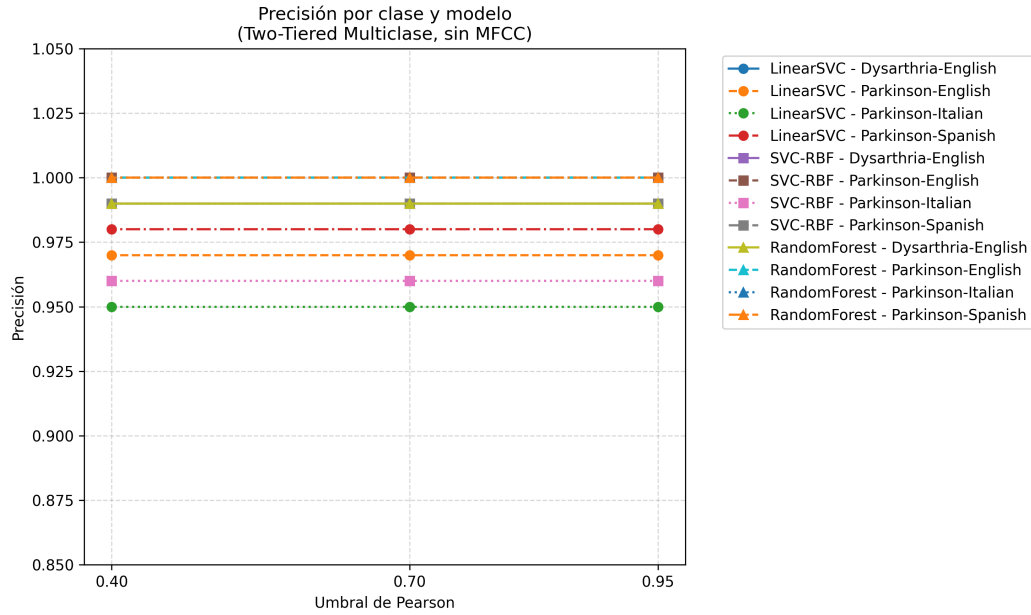


Figura 4.12: Evolución de la **precisión** por clase y modelo a través de distintos umbrales de correlación de Pearson (fase multiclase, two-tiered sin MFCC).

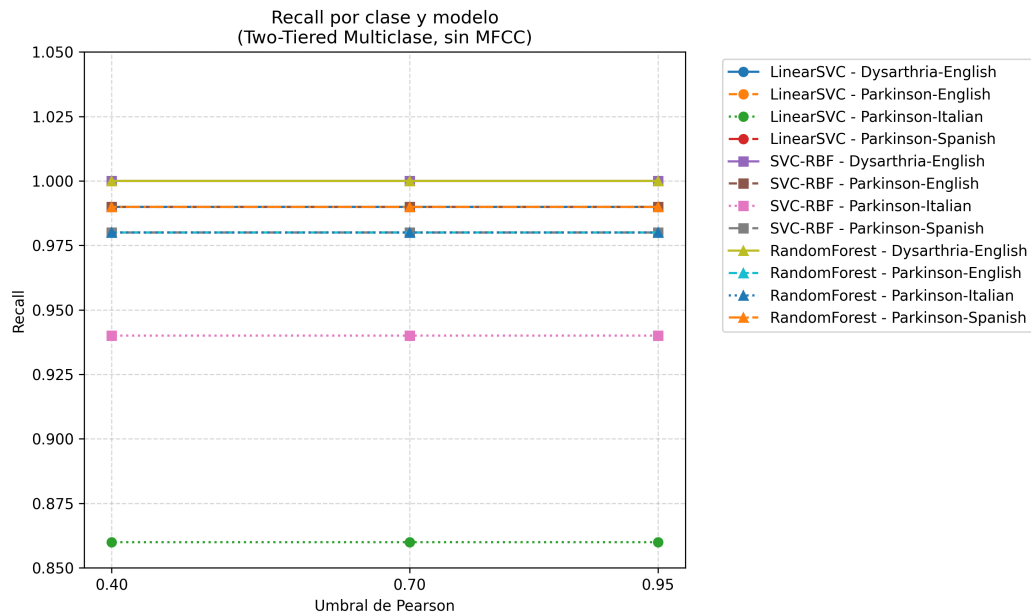


Figura 4.13: Evolución del **recall** por clase y modelo según el umbral de correlación de Pearson (fase multiclase, two-tiered sin MFCC).

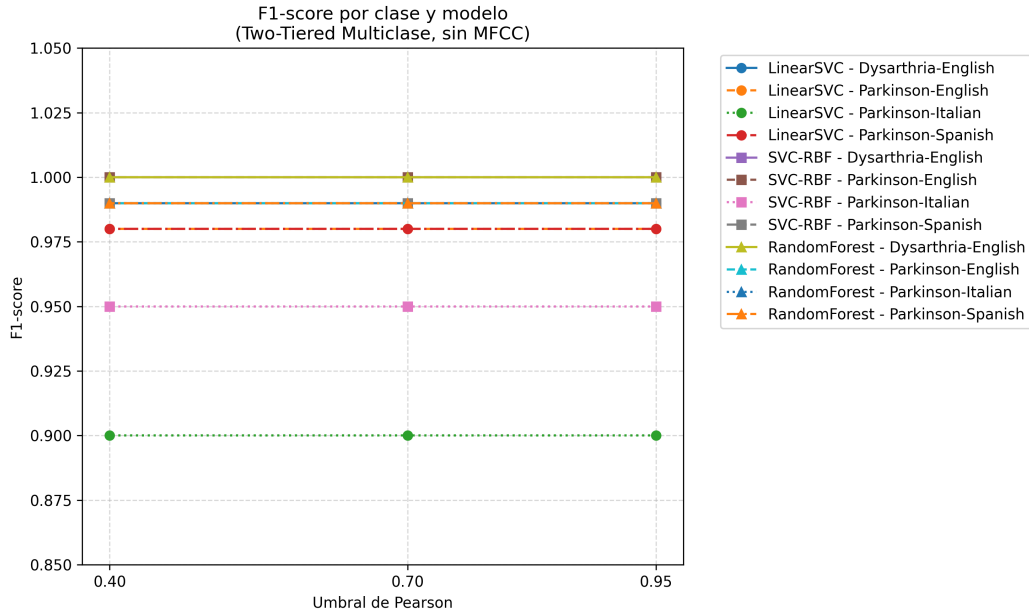


Figura 4.14: Evolución del **f1-score** por clase y modelo en función del umbral de correlación de Pearson (fase multiclase, two-tiered sin MFCC).

**Comportamiento de la *accuracy* global:** La Figura 4.15 muestra que la *accuracy* global para cada modelo permanece igualmente estable y en valores extremadamente altos (cerca al 98 % para LinearSVC, 99 % para SVC-RBF y 100 % para RandomForest), independientemente del umbral utilizado.

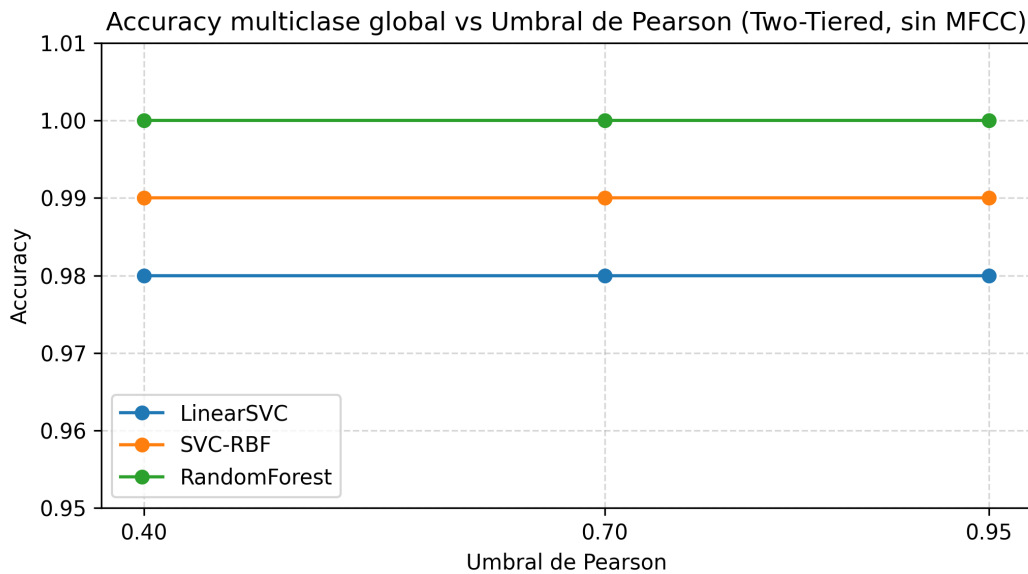


Figura 4.15: Accuracy multiclase global para cada modelo en función del umbral de Pearson (Two-Tiered, sin MFCC).

**Análisis por clase:** Si bien las métricas globales muestran una estabilidad sobresaliente, los gráficos permiten identificar diferencias sutiles entre las distintas clases. Específicamente, la clase **Parkinson-Italian** presenta los valores de precisión, recall y f1-score más bajos en comparación con el resto, lo que sugiere una mayor dificultad para su discriminación —posiblemente debido al menor tamaño muestral, variabilidad intra-clase o semejanzas acústicas con otras clases. Por el contrario, **Dysarthria-English** destaca como la clase más fácil de clasificar, con métricas prácticamente perfectas en todos los modelos y umbrales, evidenciando características acústicas altamente distintivas para esta patología en inglés.

**Implicancias:** Estos resultados reafirman la robustez de la arquitectura two-tiered para la clasificación multiclase en escenarios multilingües y multipatológicos, demostrando que la correcta selección y depuración de características permite mantener un rendimiento de excelencia, incluso bajo reducciones agresivas de dimensionalidad. Además, el análisis detallado por clase facilita la identificación de posibles desafíos diagnósticos en patologías específicas o en combinaciones idioma-patología, orientando futuros desarrollos y estudios clínicos.

### 4.3.3. Síntesis global

Una observación relevante a lo largo de los experimentos con la arquitectura two-tiered es que, si bien la disminución del umbral de correlación (de 0.95 a 0.40) reduce la cantidad de características retenidas, esto afecta principalmente la etapa binaria (sano vs patológico), provocando una caída notable en la precisión de los modelos lineales. Por el contrario, los modelos no lineales, como Random Forest y SVC-RBF, demuestran una robustez significativa frente a la reducción de dimensionalidad, manteniendo resultados notablemente altos incluso con un subconjunto de características más restringido.

En la etapa multiclase, se observa una estabilidad destacable: independientemente del umbral de correlación utilizado, modelos como Random Forest y SVC-RBF conservan métricas superiores al 97%, validando la calidad discriminativa de las características seleccionadas. Esta resiliencia subraya la importancia de emplear técnicas de selección de características sólidas, que permitan reducir la complejidad del modelo sin sacrificar precisión diagnóstica.

En conjunto, estos hallazgos reafirman el valor de combinar arquitecturas interpretables y selección exhaustiva de descriptores acústicos para construir sistemas de clasificación robustos, precisos y clínicamente aplicables en contextos multilingües y multipatológicos.

## 4.4. Selección de Características en la Two-Tiered Pipeline

En esta sección se presenta la comparación de la selección de características obtenida mediante RFECV para los distintos escenarios experimentales de la arquitectura two-tiered. Se muestran los resultados tanto para la fase binaria como multiclase, considerando entrenamientos con y sin coeficientes MFCC, y para los tres umbrales de correlación utilizados (**0.95**, **0.70** y **0.40**).

### 4.4.1. Sin MFCC

#### Fase Binaria

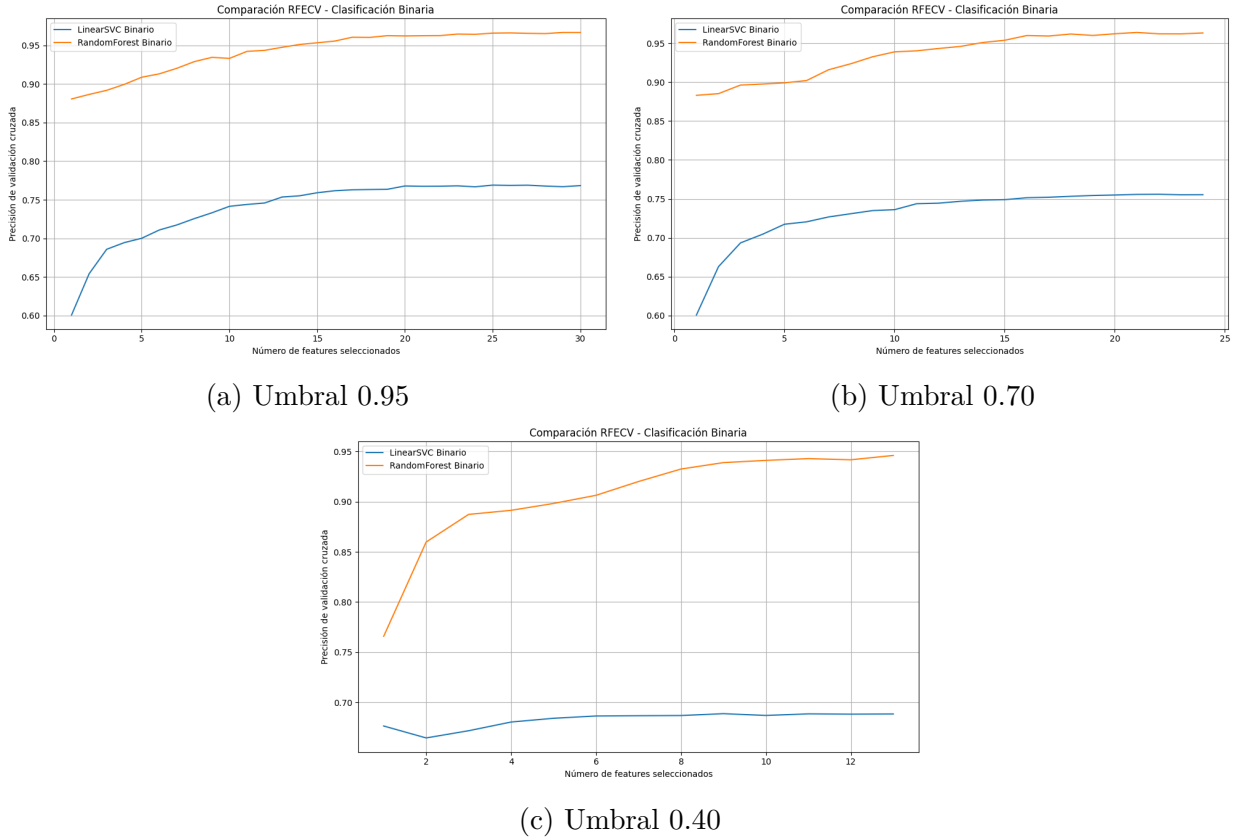


Figura 4.16: Comparación RFECV en la fase binaria sin MFCC. Se observa la precisión de validación cruzada para LinearSVC y RandomForest en función del número de características seleccionadas, bajo los tres umbrales de correlación.

## Fase Multiclase

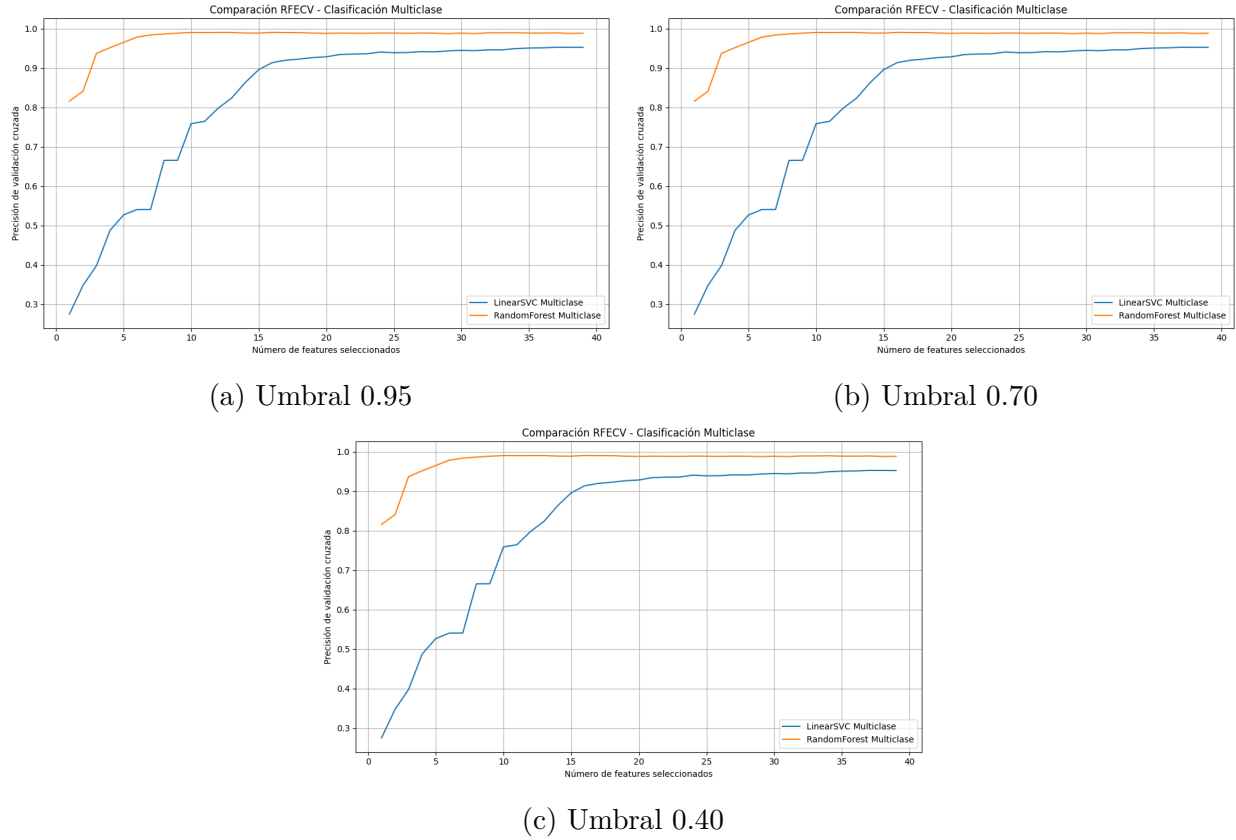


Figura 4.17: Comparación RFECV en la fase multiclase sin MFCC. Se observa la precisión de validación cruzada para LinearSVC y RandomForest en función del número de características seleccionadas, bajo los tres umbrales de correlación.

El análisis de la selección de características sin MFCC revela tendencias claras respecto al desempeño diferencial de los modelos evaluados. Como se puede observar en la Figura 4.16, en la fase binaria, RandomForest alcanza precisiones sobresalientes incluso con un número muy reducido de descriptores, evidenciando una notable robustez frente a la reducción del espacio de características provocada por umbrales de correlación más estrictos. Por el contrario, LinearSVC muestra una mayor dependencia respecto al número de features, especialmente en la tarea binaria, donde la precisión se ve más afectada a medida que disminuye el umbral. En la fase multiclase, la Figura 4.17 nos muestra que ambos modelos experimentan mejoras pronunciadas al incorporar los primeros features seleccionados, pero nuevamente RandomForest destaca por su capacidad para sostener altos niveles de precisión con pocos descriptores.

#### 4.4.2. Con MFCC

##### Fase Binaria

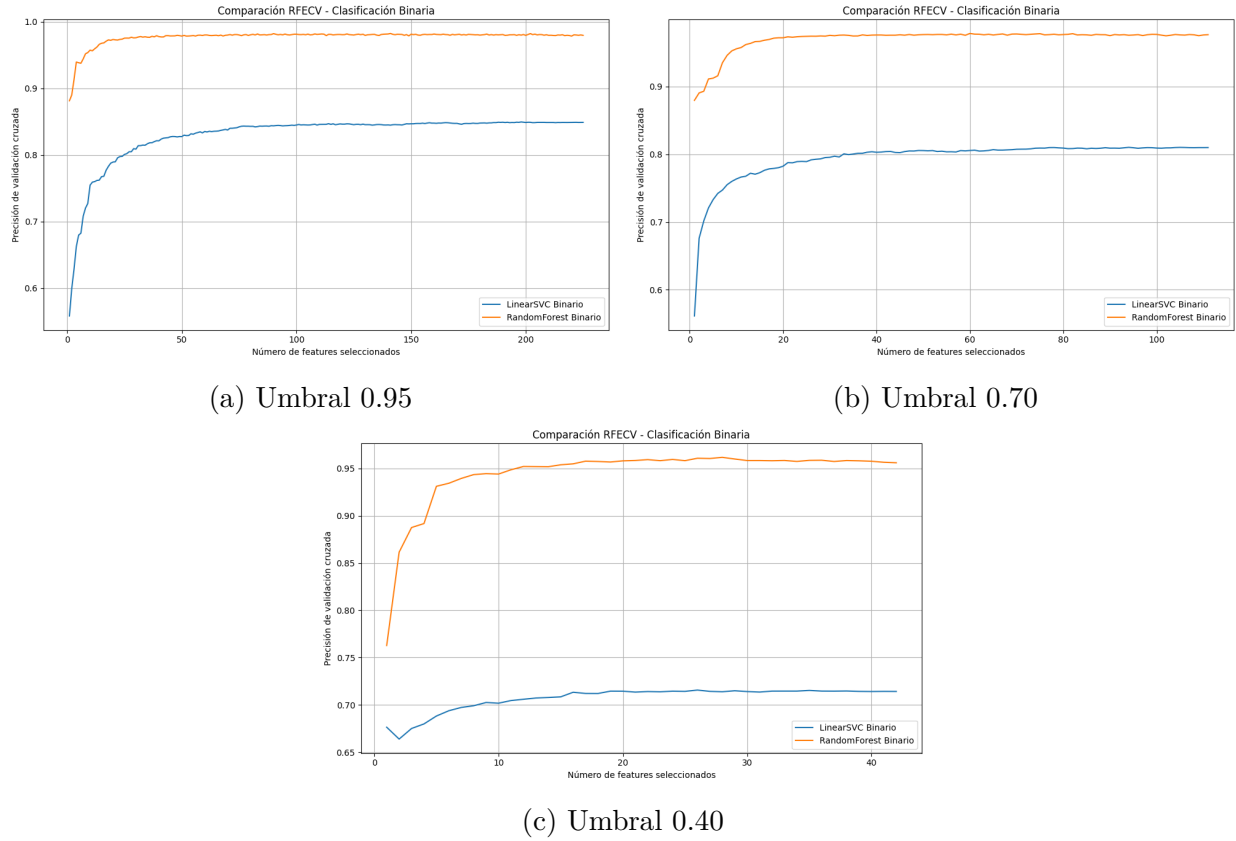


Figura 4.18: Comparación RFECV en la fase binaria con MFCC. Se observa la precisión de validación cruzada para LinearSVC y RandomForest en función del número de características seleccionadas, bajo los tres umbrales de correlación.

## Fase Multiclase

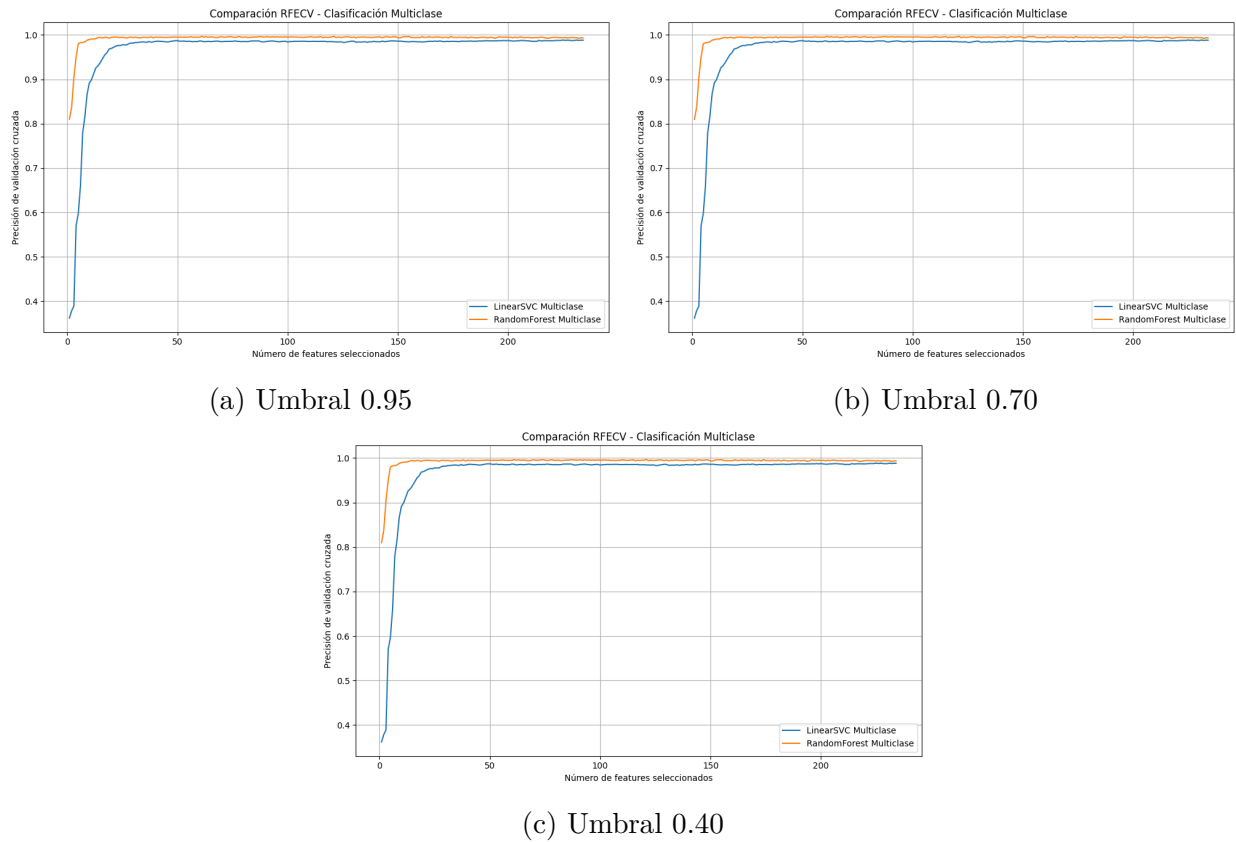


Figura 4.19: Comparación RFECV en la fase multiclase con MFCC. Se observa la precisión de validación cruzada para LinearSVC y RandomForest en función del número de características seleccionadas, bajo los tres umbrales de correlación.

Al incorporar los coeficientes MFCC en el proceso de selección de características, se observa una mejora sustancial en la precisión de los modelos, especialmente en la fase multiclase. Tanto LinearSVC como RandomForest alcanzan valores cercanos a la perfección con un número moderado de features, evidenciando la alta capacidad discriminativa de este conjunto combinado. La Figura 4.18 permite observar que en la fase binaria, RandomForest mantiene su robustez característica, logrando precisiones superiores al 95 % incluso bajo umbrales de correlación estrictos, mientras que LinearSVC muestra una dependencia mayor del número de características, especialmente cuando el set se reduce a menos de 40 features. Por su parte, la Figura 4.19 muestra que en la fase multiclase ambos modelos alcanzan rápidamente su máximo rendimiento con apenas 20 a 40 descriptores, lo que resalta el valor de la reducción de dimensionalidad para optimizar tanto la eficiencia como la interpretabilidad de los modelos.

## 4.5. Interpretabilidad de los Modelos: Análisis SHAP de la Selección de Características

En esta sección se presenta el análisis de interpretabilidad realizado mediante SHAP sobre los modelos entrenados sin coeficientes MFCC, considerando únicamente el umbral de correlación 0.95. La motivación detrás de este enfoque radica en priorizar la interpretabilidad clínica y la transparencia de los resultados, evitando la inclusión de descriptores cuyo significado acústico no es fácilmente interpretable.

### 4.5.1. Análisis SHAP para Clasificación Multiclase Directa (sin MFCC, umbral 0.95)

En esta sección se presenta el análisis de interpretabilidad realizado mediante SHAP sobre el modelo Random Forest entrenado sin coeficientes MFCC y con umbral de correlación 0.95. Se seleccionaron los gráficos más representativos para ilustrar las tendencias generales y específicas observadas en las distintas clases e idiomas.

#### Importancia relativa de los descriptores: comparación Random Forest vs. SVC lineal

El análisis SHAP permitió comparar la importancia relativa de los descriptores acústicos para dos modelos de referencia (Random Forest y SVC lineal) entrenados sin MFCC y con umbral de correlación 0.95. En las Figuras 4.20– 4.24, se observa que, aunque ambos modelos posicionan a los índices de espacio articulatorio (VAI y tVSA) como elementos clave en la discriminación de disartria en inglés, la relevancia de estos descriptores disminuye notoriamente en SVC, donde emergen con fuerza variables asociadas a la variabilidad temporal y energética, tales como *Autocorr Energy*, *Crosscorr Std* y *Avg Strength +b*.

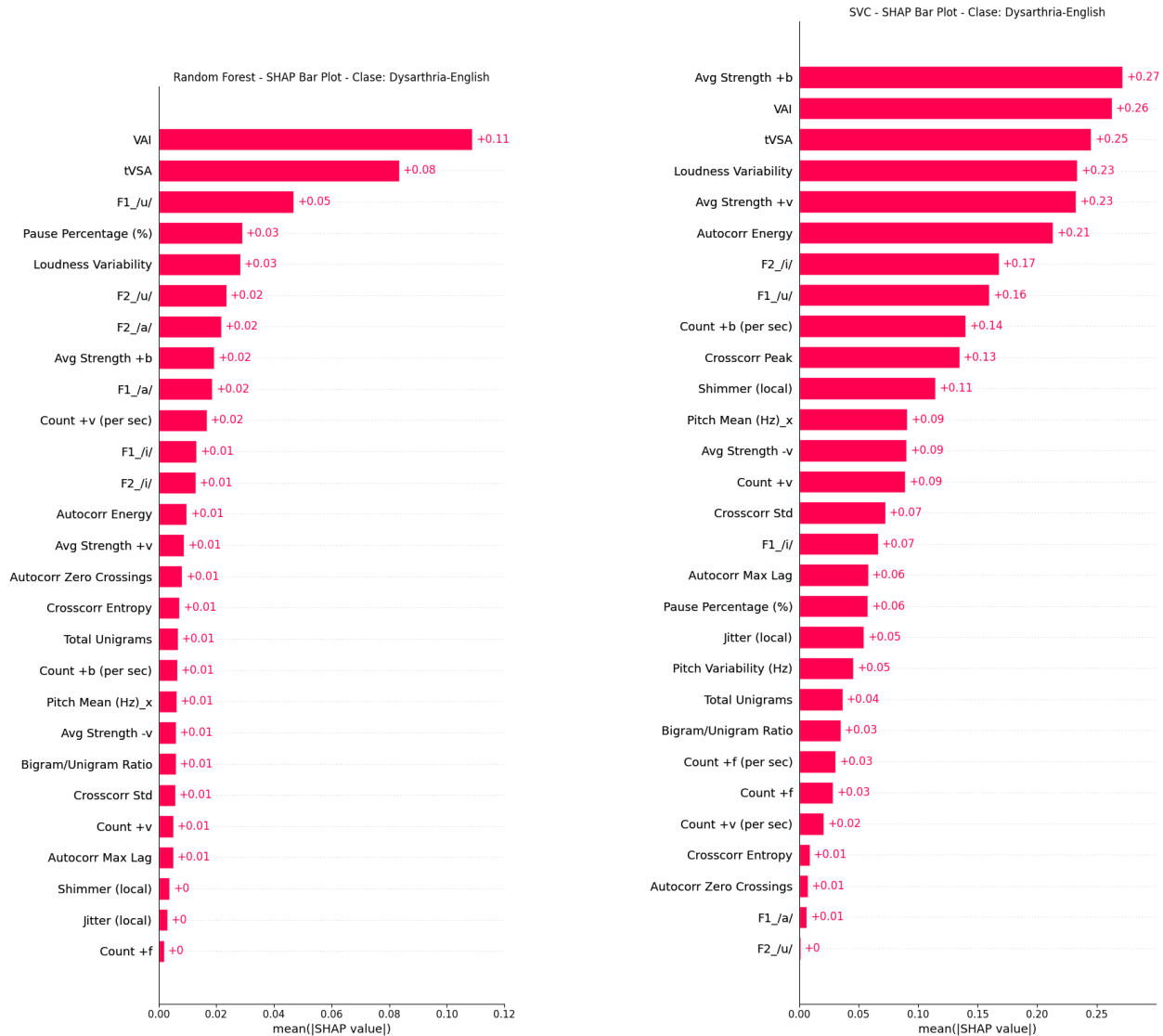
En el caso de Parkinson-Italian, el contraste es aún más marcado: mientras el Random Forest reparte la importancia entre VAI, tVSA y ciertas métricas de correlación, el SVC se apoya casi exclusivamente en la variabilidad de la señal (*Crosscorr Std*, *Count +b/v per sec*, *Bigram/Unigram Ratio*), sugiriendo un mecanismo de clasificación altamente dependiente de la fluctuación temporal de la voz.

Finalmente, en la clase Healthy-Spanish, Random Forest sigue privilegiando los descriptores articulatorios clásicos, pero el SVC otorga máxima relevancia a medidas de energía y variabilidad temporal, desplazando a VAI/tVSA al tercer lugar.

Estos resultados confirman que, aunque existe consenso respecto a la utilidad de los descriptores articulatorios en la clasificación robusta entre patologías, el tipo de modelo empleado puede alterar radicalmente el peso atribuido a las dimensiones temporales versus espectrales.

De manera consistente, jitter y shimmer muestran un impacto marginal, reafirmando su baja capacidad discriminativa interlingüística en este contexto.

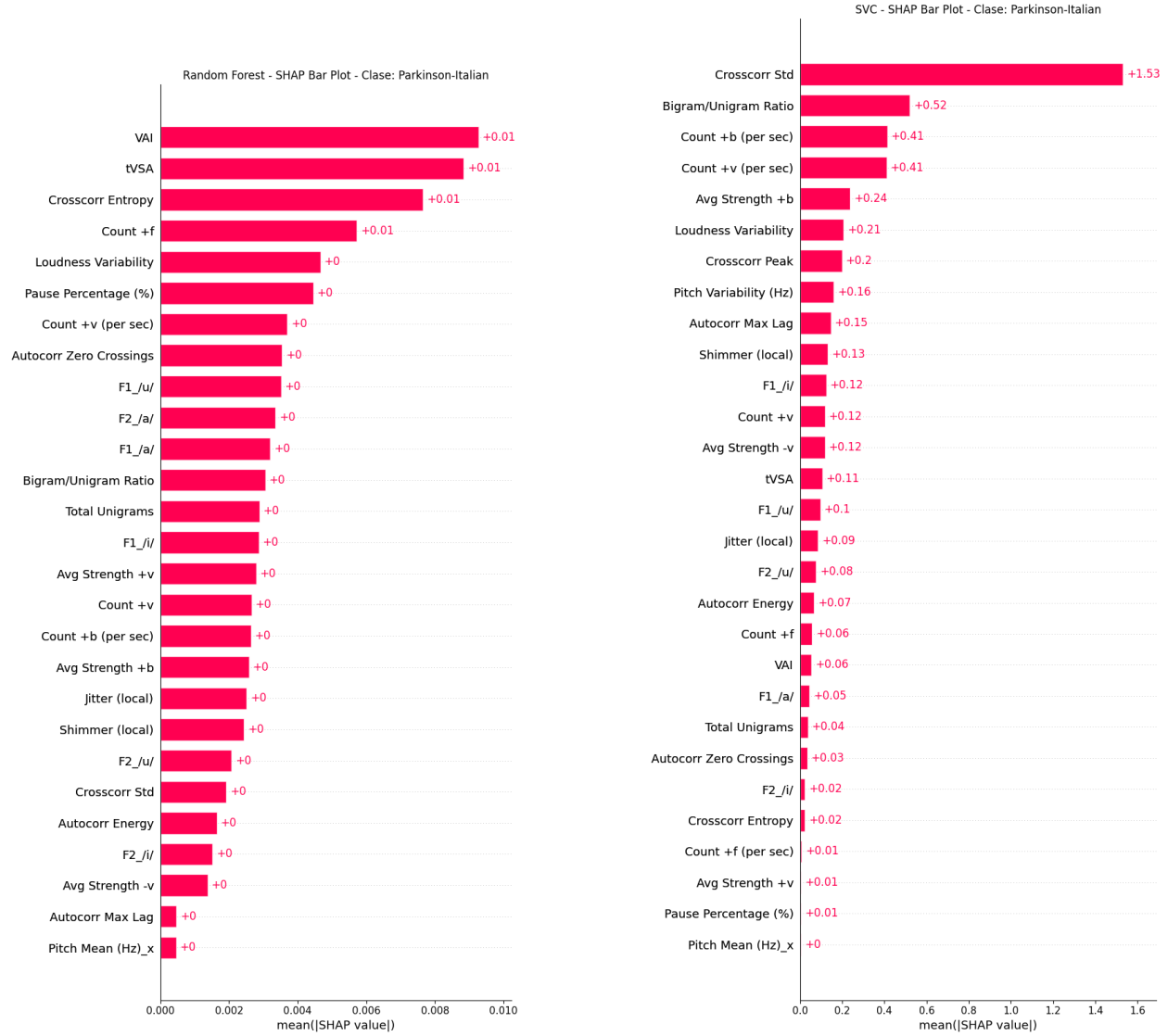
**Nota:** Las diferencias observadas en la dispersión de los valores SHAP (beeswarm plots) sugieren que el modelo lineal puede ser más susceptible a features con alta varianza, lo que puede tener implicancias en la generalización y estabilidad del sistema.



(a) Random Forest, Dysarthria-English

(b) SVC Lineal, Dysarthria-English

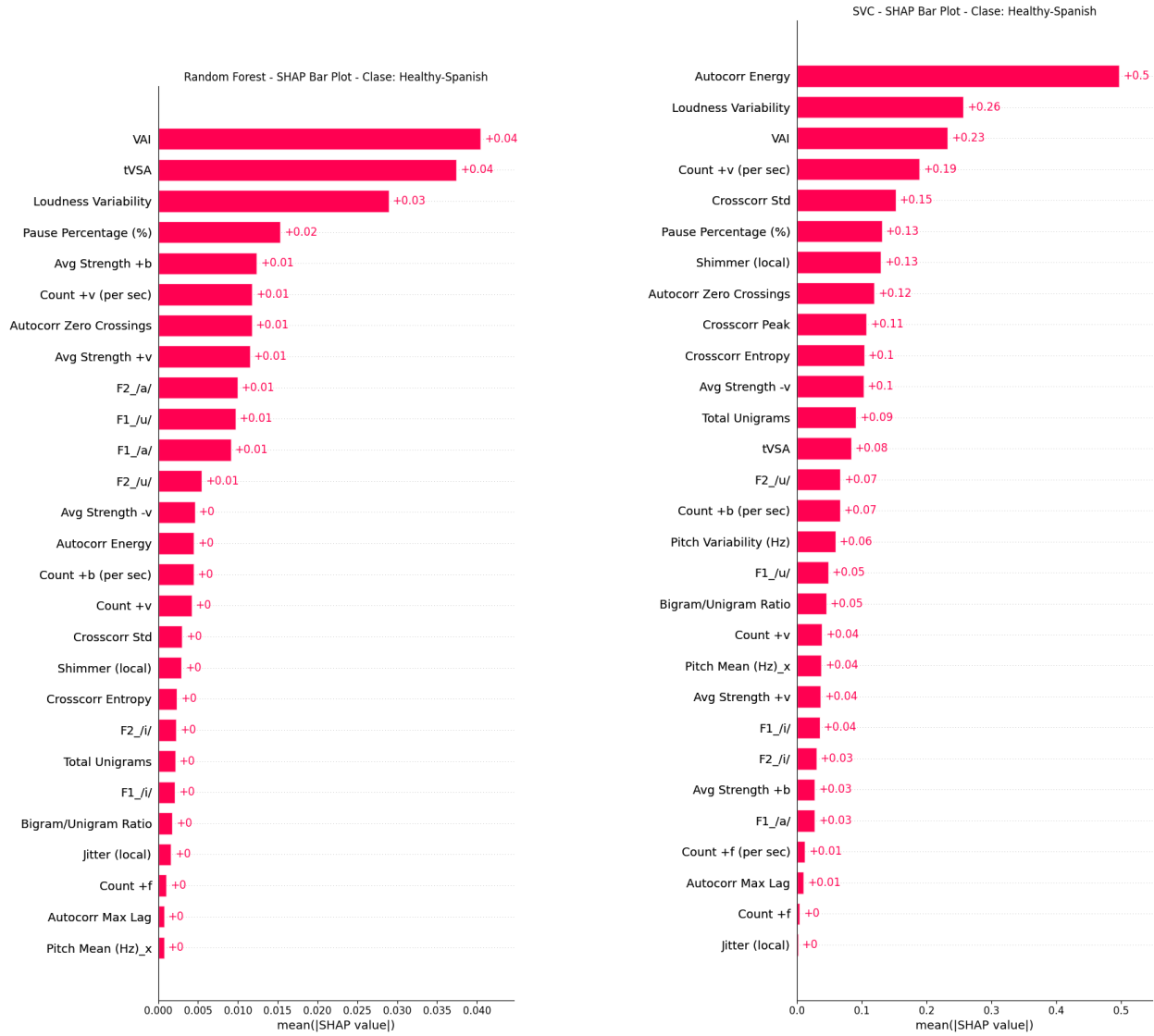
Figura 4.20: Comparación de importancia SHAP para la clase Dysarthria-English.



(a) Random Forest, Parkinson-Italian

(b) SVC Lineal, Parkinson-Italian

Figura 4.21: Comparación de importancia SHAP para la clase Parkinson-Italian.



(a) Random Forest, Healthy-Spanish

(b) SVC Lineal, Healthy-Spanish

Figura 4.22: Comparación de importancia SHAP para la clase Healthy-Spanish.

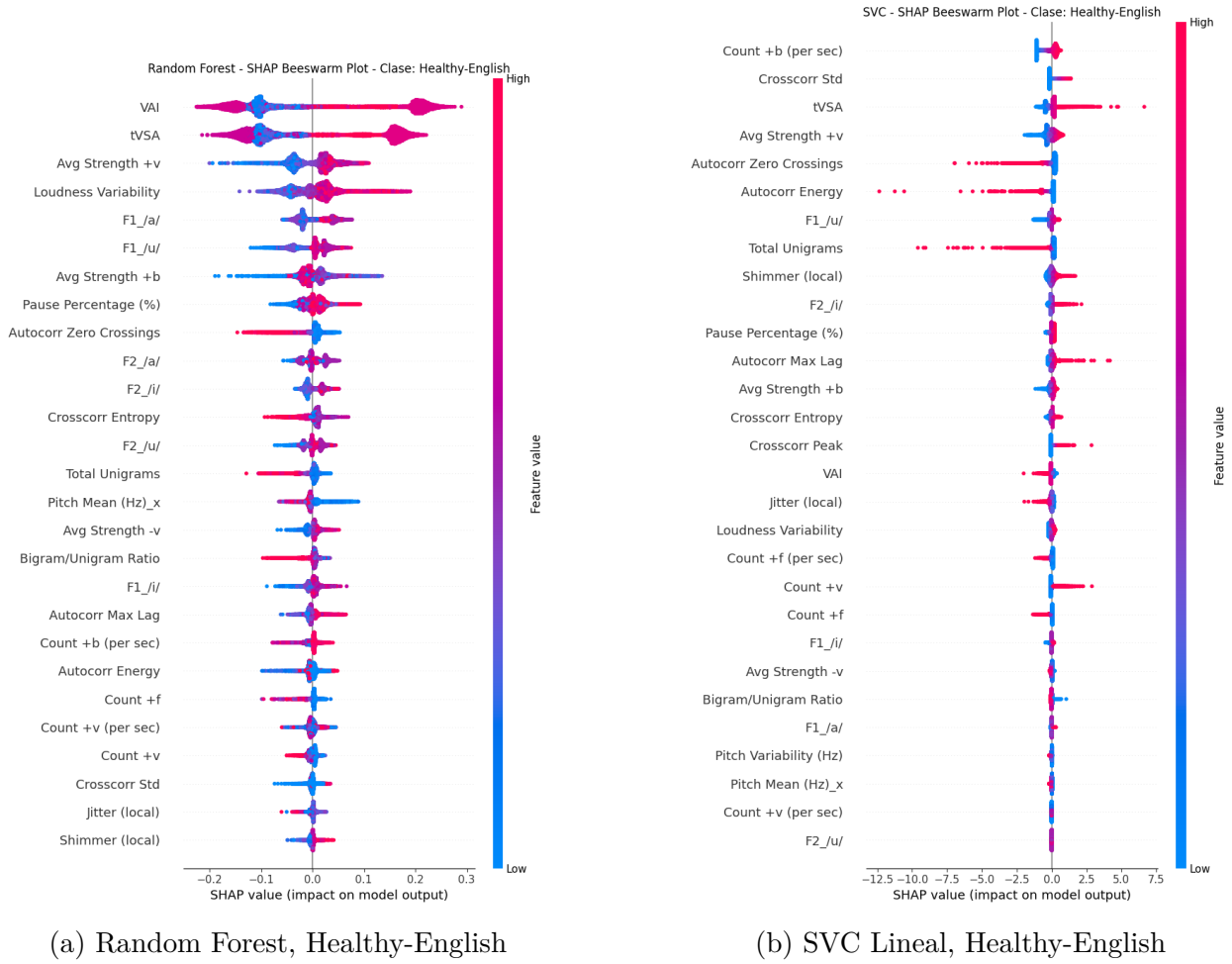


Figura 4.23: Beeswarm plots para Healthy-English: dispersión y perfil de impacto de los descriptores.

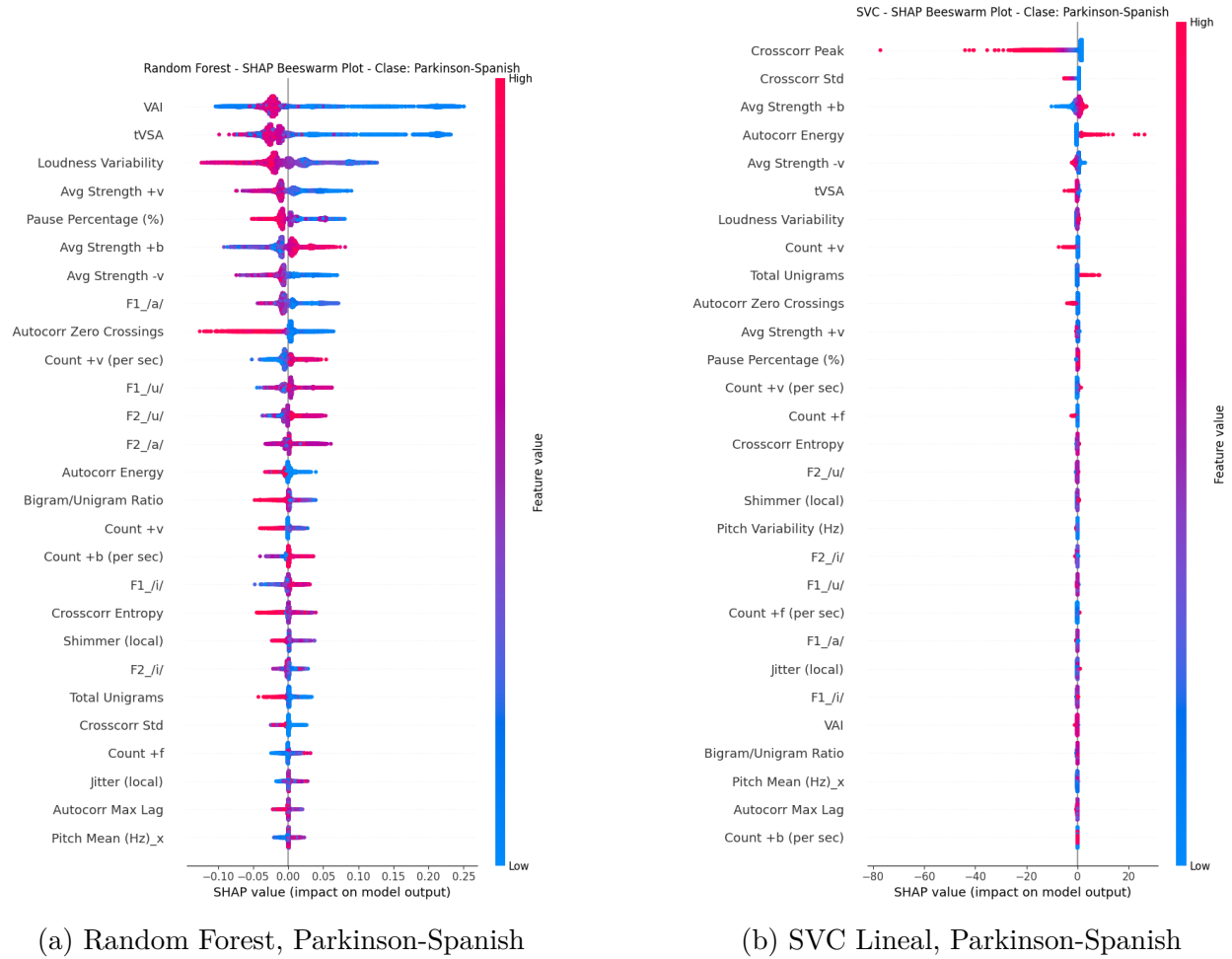


Figura 4.24: Beeswarm plots para Parkinson-Spanish: dispersión y perfil de impacto de los descriptores.

## Comparación de dependencias aprendidas: Random Forest vs. Linear SVC

Para profundizar en las diferencias de comportamiento entre modelos, la Figura 4.25 muestra una comparación directa entre los *dependence plots* de SHAP para la característica **Loudness Variability**, tanto en Random Forest como en Linear SVC, utilizando dos clases contrastantes: Dysarthria-English y Healthy-Spanish. En cada caso, el eje de color corresponde a la segunda variable más relevante para la interacción según el análisis de SHAP.

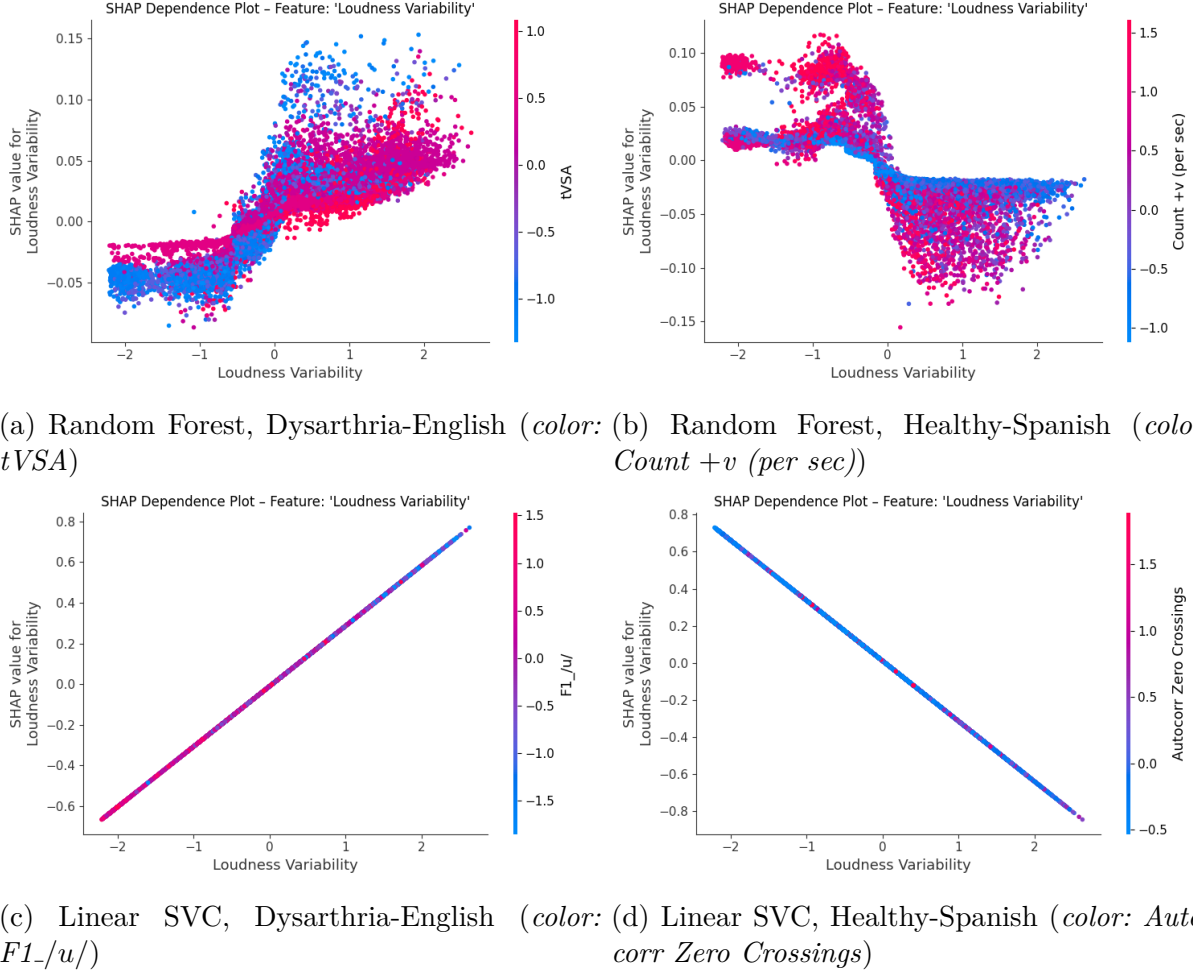


Figura 4.25: Comparación de *dependence plots* SHAP para la característica Loudness Variability en modelos Random Forest y Linear SVC. Cada plot representa una clase distinta, ilustrando cómo la importancia y las interacciones de la variable cambian según el modelo y la clase objetivo.

A nivel visual, se observa que el modelo Random Forest tiende a capturar relaciones no lineales más complejas, evidenciadas por las distribuciones curvas y dependientes

#### 4.5.2. Análisis SHAP para Clasificación mediante arquitectura two-tiered (sin MFCC), umbral 0.95

##### Importancia de características en la fase binaria: comparación entre modelos lineal y no lineal

Para la fase binaria de la arquitectura two-tiered, se analizó la importancia de las características utilizando SHAP tanto para el modelo *LinearSVC* como para *Random Forest*. Las

Figuras 4.26 y 4.27 presentan comparativamente los barplots y beeswarm plots para ambos modelos.

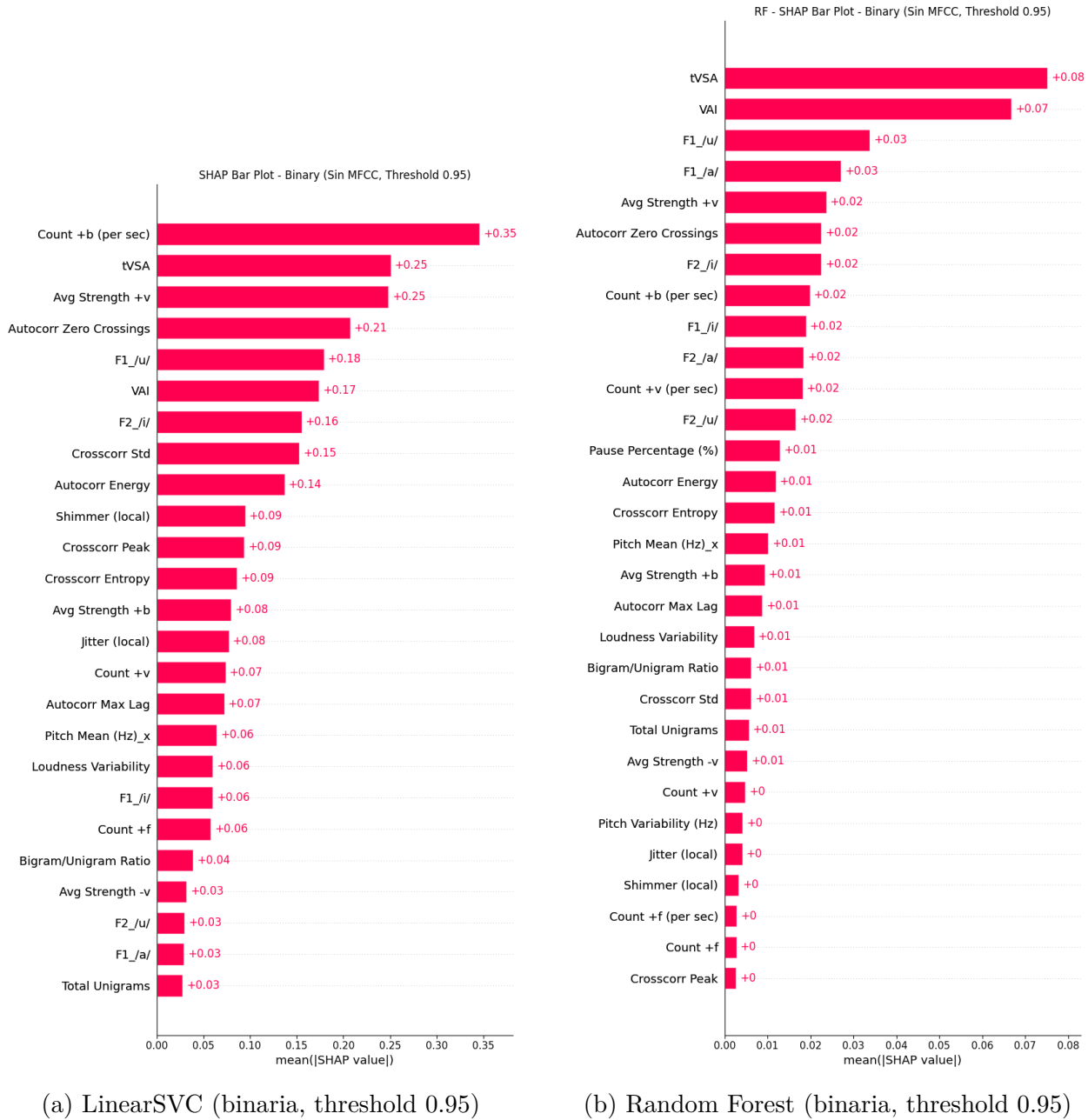
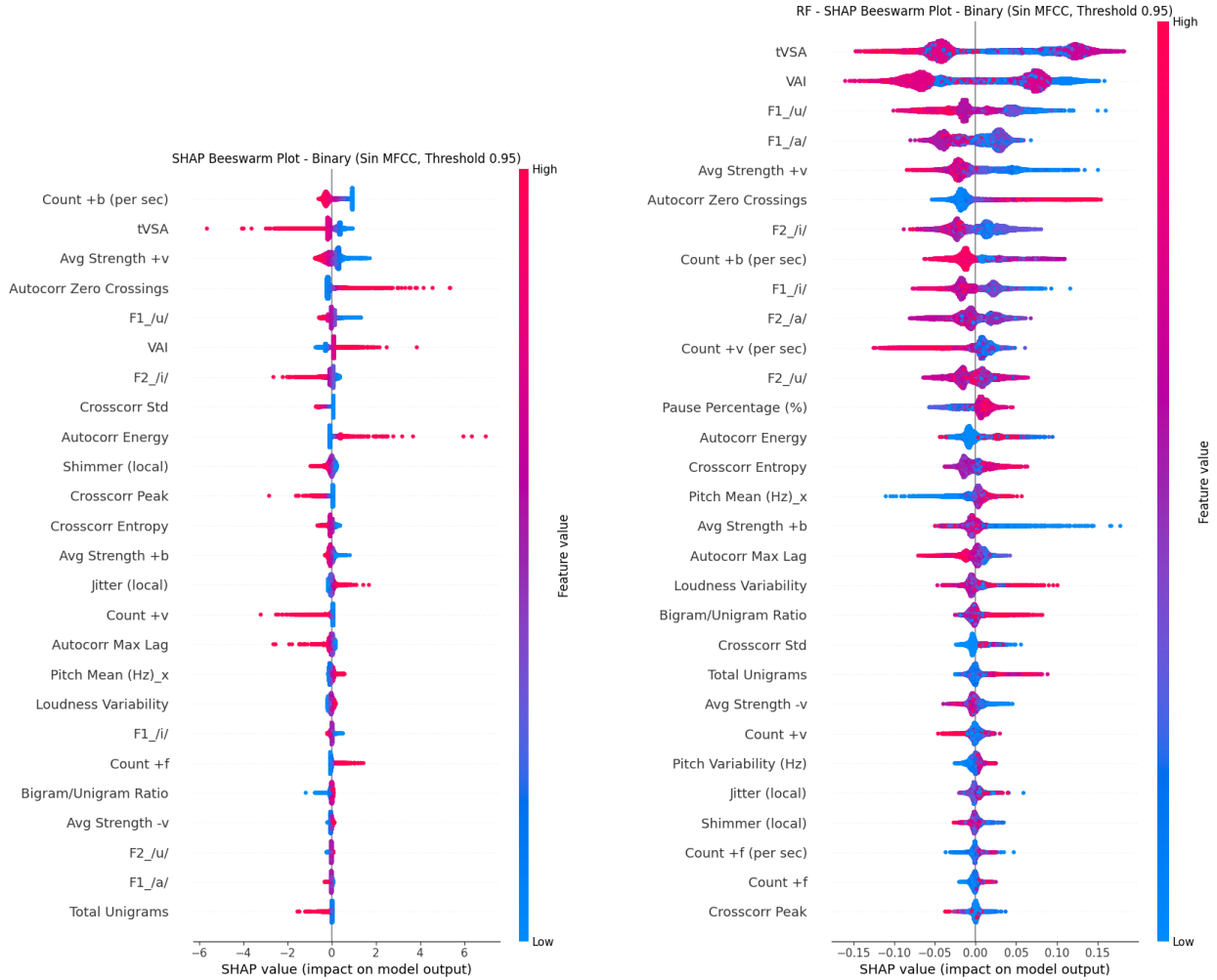


Figura 4.26: Barplots de valores medios absolutos de SHAP para la fase binaria, comparando LinearSVC y Random Forest.



(a) LinearSVC (binaria, threshold 0.95)

(b) Random Forest (binaria, threshold 0.95)

Figura 4.27: Beeswarm plots de SHAP para la fase binaria, comparando LinearSVC y Random Forest.

### Análisis comparativo:

El análisis de importancia de características revela diferencias notables en la manera en que los modelos lineal y no lineal aprovechan la información acústica:

- **LinearSVC:** El modelo lineal tiende a asignar mayor importancia a un grupo reducido de features dominantes, donde destacan *Count +b (per sec)*, *tVSA* y *Avg Strength +v*, con valores de SHAP considerablemente superiores al resto. Esto sugiere que la capacidad del modelo para separar clases depende fuertemente de unas pocas dimensiones linealmente separables.
- **Random Forest:** En contraste, el modelo de Random Forest distribuye la importancia de forma más homogénea entre un conjunto mayor de variables. Destacan *tVSA* y *VAI*,

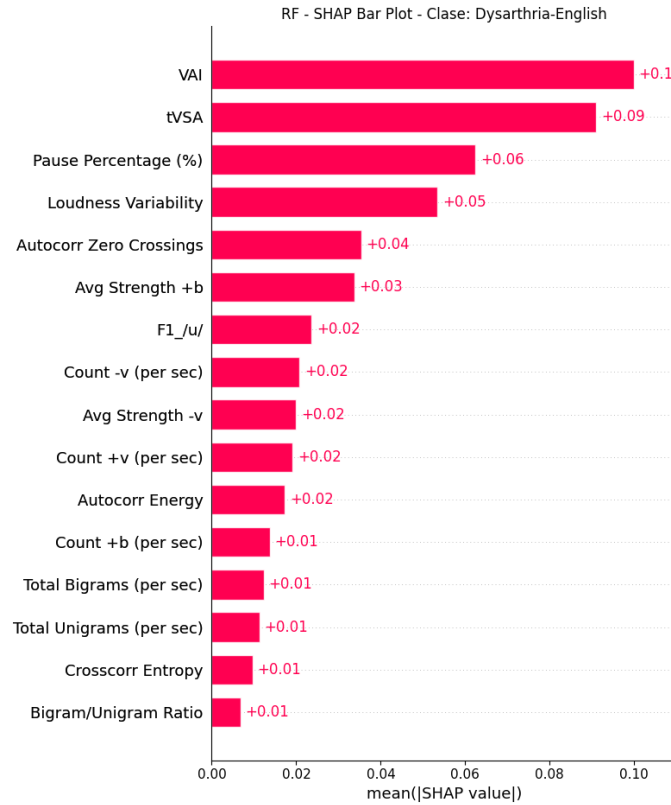
pero también adquieren relevancia múltiples descriptores formánticos ( $F1_u$ /,  $F1_a$ /,  $F2_i$ /) y parámetros derivados de autocorrelación y fuerza vocálica, lo que refleja la capacidad del modelo para capturar relaciones no lineales y combinaciones complejas de atributos.

- **Intersecciones y diferencias:** Es interesante notar que algunos descriptores relevantes para ambos modelos (por ejemplo,  $tVSA$ ,  $VAI$ ) muestran una mayor robustez inter-modelo, mientras que otros (como los relacionados a conteos de bigramas y ciertas energías de autocorrelación) adquieren importancia sólo en uno de los dos enfoques.
- **Beeswarm plots:** Las distribuciones de valores SHAP en los beeswarm plots muestran mayor dispersión y presencia de interacciones no lineales en Random Forest, versus el patrón más agrupado y direccional del LinearSVC. Esto refuerza la hipótesis de que la modelación no lineal permite capturar mejor la diversidad de patrones patológicos presentes en los datos.

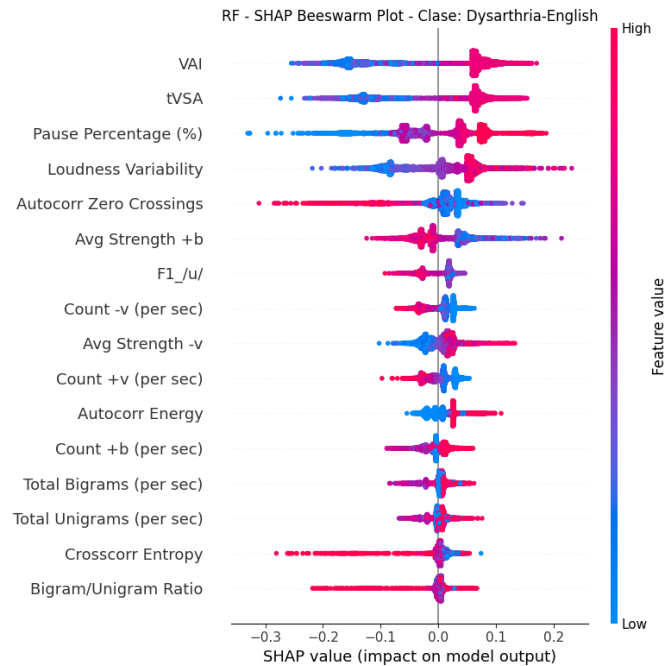
En síntesis, el uso conjunto de ambos enfoques confirma la complementariedad entre modelos lineales y no lineales para la identificación de descriptores acústicos robustos y generalizables en la tarea de discriminación binaria sano/patológico, pero evidencia que la riqueza de relaciones multivariadas es mejor capturada por modelos no lineales como Random Forest.

### **Análisis SHAP en la Fase Multiclase: Importancia y Universalidad de las Características**

En esta etapa se analizaron únicamente los resultados del modelo Random Forest, dado que este demostró la mayor capacidad para capturar relaciones no lineales y proporcionar una visión más rica de la relevancia y generalizabilidad de los descriptores acústicos a través de múltiples idiomas y patologías. Para cada clase (Dysarthria-English, Parkinson-Italian, Parkinson-Spanish y Parkinson-English), se presentan los gráficos SHAP barplot y beeswarm correspondientes (Figuras 4.28– 4.31), permitiendo así identificar tendencias compartidas y singularidades propias de cada subgrupo.

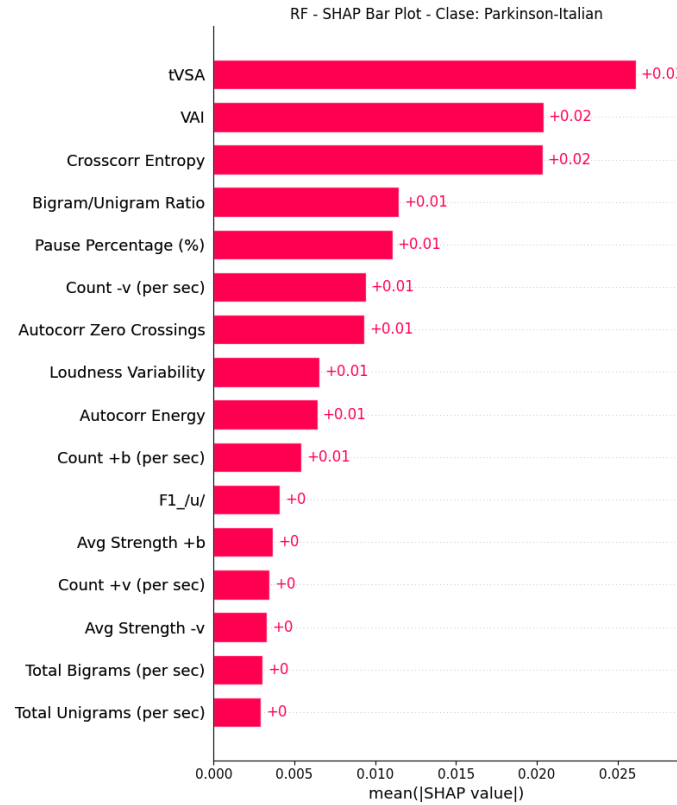


(a) Barplot - Dysarthria-English

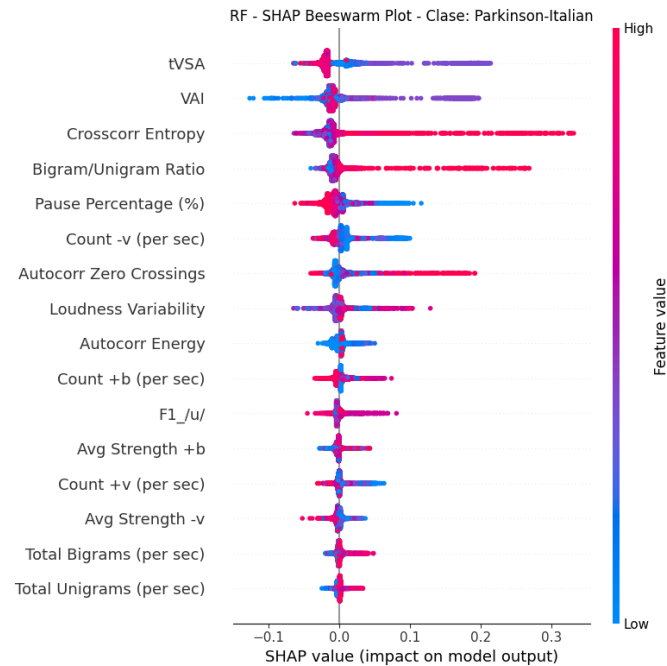


(b) Beeswarm - Dysarthria-English

Figura 4.28: Importancia de las características para la clase Dysarthria-English (fase multi-clase, Random Forest).

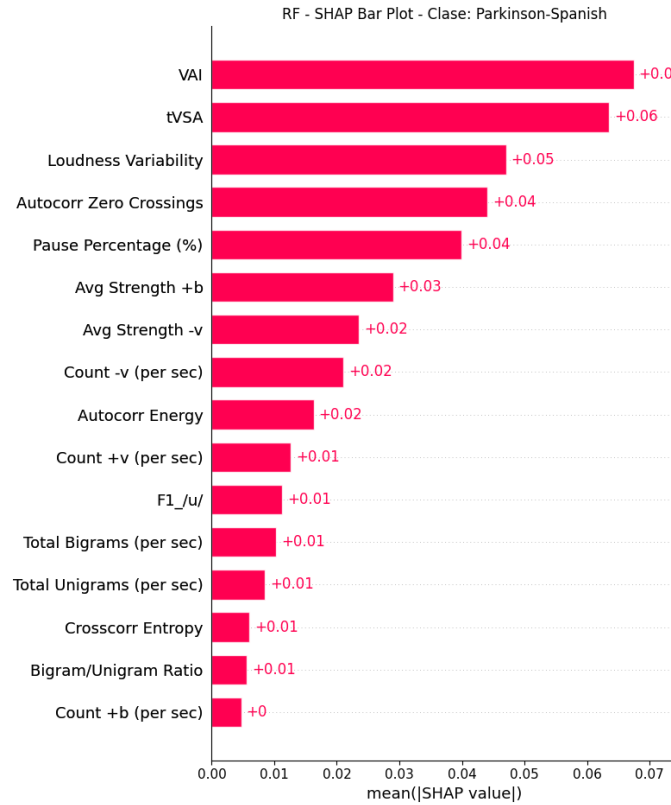


(a) Barplot - Parkinson-Italian

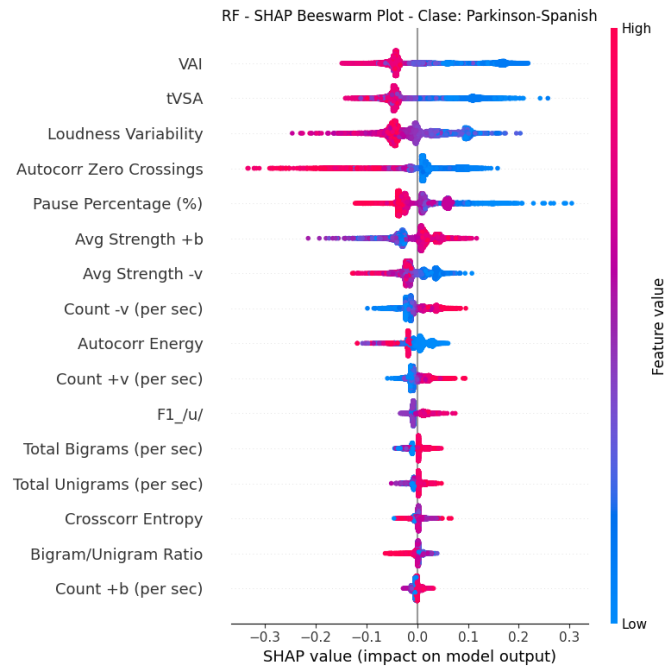


(b) Beeswarm - Parkinson-Italian

Figura 4.29: Importancia de las características para la clase Parkinson-Italian (fase multiclase, Random Forest).

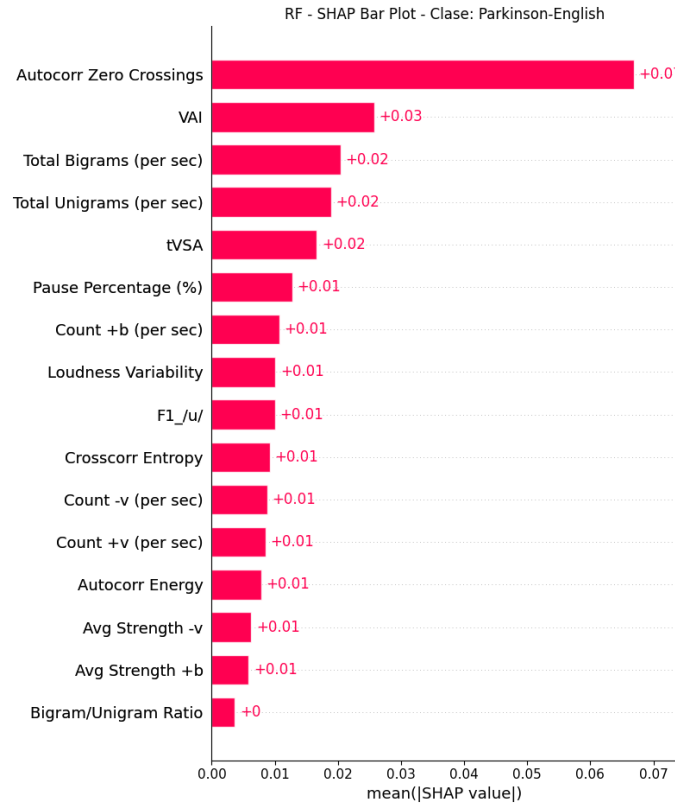


(a) Barplot - Parkinson-Spanish

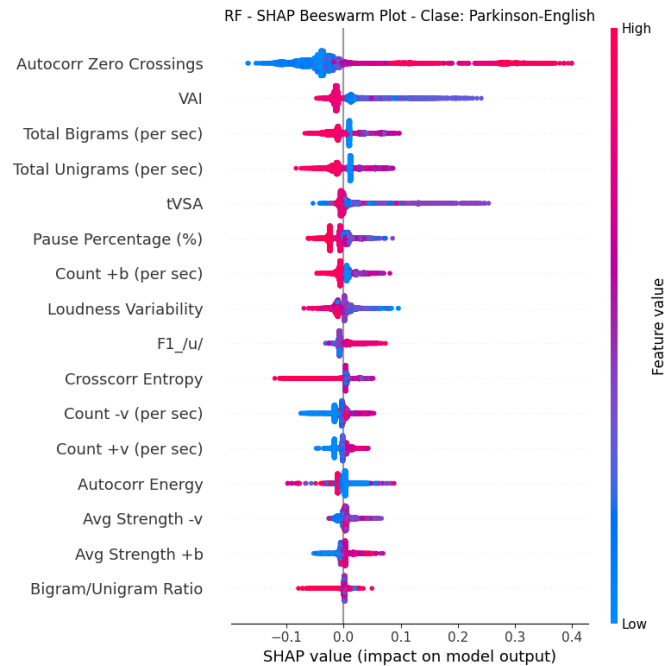


(b) Beeswarm - Parkinson-Spanish

Figura 4.30: Importancia de las características para la clase Parkinson-Spanish (fase multi-clase, Random Forest).



(a) Barplot - Parkinson-English



(b) Beeswarm - Parkinson-English

Figura 4.31: Importancia de las características para la clase Parkinson-English (fase multi-clase, Random Forest).

## Análisis global de los resultados:

El análisis detallado de los plots SHAP en la clasificación multiclase permite obtener conclusiones precisas sobre la relevancia, universalidad y especificidad de los descriptores acústicos, complementando los resultados de la clasificación binaria y orientando la selección futura de características:

- **Características robustas y generalizables:** Destacan consistentemente descriptores como *VAI* (Vowel Articulation Index), *tVSA* (total Vowel Space Area), *Pause Percentage (%)*, *Loudness Variability*, *F1-/u/* y *Autocorr Zero Crossings*. Estos se posicionan recurrentemente en los primeros lugares de relevancia en múltiples clases, demostrando que las medidas relacionadas con la variabilidad articulatoria, prosódica y temporal capturan patrones fonológicos altamente robustos y transferibles entre contextos lingüísticos y patológicos.
- **Especificidad por clase:** Algunos descriptores muestran un comportamiento diferencial según la clase. Por ejemplo, en *Dysarthria-English* resaltan particularmente el *VAI*, el *tVSA*, la *Loudness Variability* y el *Pause Percentage (%)*, reflejando que tanto las alteraciones prosódicas como la reducción del espacio vocálico son discriminantes claves en este contexto específico. Por otro lado, en las clases relacionadas con Parkinson (*Parkinson-Italian* y *Parkinson-English*), variables como *Crosscorr Entropy* adquieren relevancia destacada, apuntando hacia diferencias fonológicas y articulatorias particulares asociadas a esta enfermedad.
- **Variables lingüísticas basadas en conteos:** Los descriptores relacionados con la frecuencia y proporción de bigramas y unigramas, como *Total Bigrams (per sec)* y *Bigram/Unigram Ratio*, muestran contribuciones discriminativas moderadas, especialmente en las clases asociadas al Parkinson. Aunque estas variables no aparecen consistentemente en posiciones dominantes, sus valores SHAP son suficientemente relevantes como para considerar su inclusión selectiva en modelos específicos, particularmente en contextos lingüísticos o patológicos determinados.
- **Patrones específicos de dispersión SHAP:** Los gráficos tipo beeswarm revelan diferencias interesantes en la dispersión y distribución de los valores SHAP por clase. Algunas características, como *VAI* y *tVSA*, exhiben distribuciones marcadamente asimétricas con contribuciones dominantes claras en ciertas clases (por ejemplo, en *Dysarthria-English*). En contraste, clases asociadas al Parkinson tienden a mostrar una dispersión más equilibrada, sugiriendo que en estos escenarios el diagnóstico puede depender de un mayor número de descriptores con contribuciones moderadas pero significativas.
- **Importancia de los descriptores de correlación:** Los descriptores basados en correlación introducidos en este estudio —especialmente *Autocorr Zero Crossings* y *Autocorr Energy*— emergen como altamente competitivos, frecuentemente superando en relevancia a métricas clásicas como *Jitter*, *Shimmer* o *Pitch Mean*. Esto valida el aporte diferenciador del enfoque propuesto y destaca su potencial para mejorar modelos diagnósticos futuros.

En resumen, el análisis multiclase revela una clara jerarquía de descriptores acústicos, con un conjunto robusto y universalmente generalizable, junto a características específicas que aportan valor en contextos particulares. Esto reafirma la solidez metodológica del presente estudio y subraya su potencial para aplicaciones clínicas e investigación futura en detección de alteraciones fonológicas.

#### 4.5.3. Clasificación de universalidad multilingüe y multipatológica de cada feature

En esta subsección se presentan las conclusiones principales derivadas del análisis explicativo mediante SHAP aplicado a los descriptores acústicos manuales, considerando su relevancia tanto para la clasificación binaria (sano vs patológico) como su robustez frente a múltiples idiomas y patologías. La Tabla 4.1 resume de manera estructurada el nivel de universalidad de cada descriptor acústico, clasificándolos en cinco grandes categorías según la evidencia obtenida:

- **Universal clara:** El descriptor demuestra consistentemente alta importancia SHAP en múltiples patologías e idiomas evaluados, mostrando relevancia clara en la etapa binaria.
- **Universal con matices:** El feature mantiene relevancia en diversos contextos, pero presenta variabilidad, ya sea por diferencias en magnitud o por una importancia binaria menos consistente.
- **Específico por idioma:** Su utilidad es relevante solo en contextos lingüísticos específicos, siendo menos robusto en general.
- **Específico por patología:** Resulta especialmente discriminativo para una patología específica, pero pierde relevancia en otros contextos clínicos.
- **Poco relevante / No relevante:** El descriptor presenta una utilidad baja o nula en la mayoría de los contextos estudiados, exhibiendo consistentemente baja importancia SHAP.

En este marco, los descriptores **tVSA**, **VAI**, **Pause Percentage (%)**, **Autocorr Zero Crossings**, **F1\_/u/** y **Autocorr Energy** emergen claramente como universales, destacando en todos los idiomas y patologías con una alta relevancia explicativa en SHAP para los modelos Random Forest y LinearSVC.

Descriptores como **Loudness Variability**, **Avg Strength +b**, **Total Bigrams (per sec)** y **Count -v (per sec)** presentan una robustez considerable aunque con ciertos matices, mostrando algunas variaciones en magnitud o una relevancia menos marcada en la etapa binaria.

En contraste, features como **Avg Strength -v**, **Bigram/Unigram Ratio** y **Crosscorr Entropy** muestran claramente especificidad contextual, destacando en ciertas patologías o idiomas particulares. Por ejemplo, **Bigram/Unigram Ratio** es relevante principalmente en ciertos contextos lingüísticos, mientras que **Crosscorr Entropy** muestra relevancia específica hacia ciertas patologías como el Parkinson en italiano.

Un grupo considerable de features, especialmente formantes vocálicos como **F1\_/i/**, **F2\_/i/** y **F1\_/a/**, destacan principalmente en la etapa binaria, mostrando utilidad clara para distinguir casos sanos y patológicos, aunque con aportes limitados en diferenciación multiclase.

Finalmente, características como **Pitch Mean**, **Pitch Variability**, **Shimmer** y **Jitter** se confirman como poco relevantes o irrelevantes, exhibiendo una baja importancia SHAP en la mayoría de los contextos estudiados, lo que sugiere una limitada utilidad práctica en el marco evaluado.

Esta clasificación facilita la orientación estratégica en la selección de features para futuras investigaciones y aplicaciones clínicas, optimizando así la robustez, interpretabilidad y capacidad de generalización multilingüe y multipatológica.

| Feature                      | Dysarthria-English | Parkinson-English | Parkinson-Spanish | Parkinson-Italian | Utilidad Binaria | Universalidad            |
|------------------------------|--------------------|-------------------|-------------------|-------------------|------------------|--------------------------|
| tVSA                         | ✓                  | ✓                 | ✓                 | ✓                 | ✓                | Universal clara          |
| VAI                          | ✓                  | ✓                 | ✓                 | ✓                 | ✓                | Universal clara          |
| Pause Percentage (%)         | ✓                  | ✓                 | ✓                 | ✓                 | ≈                | Universal clara          |
| Loudness Variability         | ✓                  | ✓                 | ✓                 | ✓                 | ✗                | Universal con matices    |
| Autocorr Zero Crossings      | ✓                  | ✓                 | ✓                 | ✓                 | ✓                | Universal clara          |
| Avg Strength +b              | ✓                  | ≈                 | ✓                 | ✓                 | ≈                | Universal con matices    |
| Avg Strength +v              | ✗                  | ✗                 | ✗                 | ✗                 | ✓                | Solo útil en binaria     |
| Avg Strength -v              | ✓                  | ≈                 | ≈                 | ✗                 | ✗                | Específica por patología |
| F1_/a/                       | ✗                  | ✗                 | ✗                 | ✗                 | ✓                | Solo útil en binaria     |
| F2_/a/                       | ✗                  | ✗                 | ✗                 | ✗                 | ≈                | Poco relevante           |
| F1_/i/                       | ✗                  | ✗                 | ✗                 | ✗                 | ✓                | Solo útil en binaria     |
| F2_/i/                       | ✗                  | ✗                 | ✗                 | ✗                 | ✓                | Solo útil en binaria     |
| F1_/u/                       | ✓                  | ✓                 | ✓                 | ✓                 | ✓                | Universal clara          |
| F2_/u/                       | ✗                  | ✗                 | ✗                 | ✗                 | ≈                | Poco relevante           |
| Pitch Mean (Hz) <sub>x</sub> | ✗                  | ✗                 | ✗                 | ✗                 | ≈                | Poco relevante           |
| Pitch Variability (Hz)       | ✗                  | ✗                 | ✗                 | ✗                 | ✗                | No relevante             |
| Shimmer (local)              | ✗                  | ✗                 | ✗                 | ✗                 | ≈                | Poco relevante           |
| Jitter (local)               | ✗                  | ✗                 | ✗                 | ✗                 | ✗                | No relevante             |
| Total Unigrams (per sec)     | ≈                  | ✓                 | ✓                 | ≈                 | ✗                | Específica por patología |
| Total Bigrams (per sec)      | ✓                  | ✓                 | ✓                 | ≈                 | ✗                | Universal con matices    |
| Bigram/Unigram Ratio         | ≈                  | ≈                 | ≈                 | ✓                 | ✗                | Específica por idioma    |
| Count +b (per sec)           | ✓                  | ✓                 | ≈                 | ✓                 | ✓                | Universal clara          |
| Count +f (per sec)           | ✗                  | ✗                 | ✗                 | ✗                 | ✗                | No relevante             |
| Count +v (per sec)           | ✓                  | ✓                 | ✓                 | ✓                 | ≈                | Universal clara          |
| Count -f (per sec)           | ✗                  | ✗                 | ✗                 | ✗                 | ✗                | No relevante             |
| Count -v (per sec)           | ✓                  | ✓                 | ✓                 | ✓                 | ✗                | Universal con matices    |
| Total Unigrams               | ✗                  | ✗                 | ✗                 | ✗                 | ✗                | No relevante             |
| Total Bigrams                | ✗                  | ✗                 | ✗                 | ✗                 | ✗                | No relevante             |
| Avg Strength -b              | ✗                  | ✗                 | ✗                 | ✗                 | ✗                | No relevante             |
| Avg Strength +f              | ✗                  | ✗                 | ✗                 | ✗                 | ✗                | No relevante             |
| Avg Strength -f              | ✗                  | ✗                 | ✗                 | ✗                 | ✗                | No relevante             |
| Autocorr Max Lag             | ✗                  | ✗                 | ✗                 | ✗                 | ✗                | No relevante             |
| Autocorr Energy              | ✓                  | ✓                 | ✓                 | ✓                 | ✓                | Universal clara          |
| Crosscorr Peak               | ✗                  | ✗                 | ✗                 | ✗                 | ≈                | Poco relevante           |
| Crosscorr Std                | ✗                  | ✗                 | ✗                 | ✗                 | ✓                | Solo útil en binaria     |
| Crosscorr Entropy            | ≈                  | ✓                 | ≈                 | ✓                 | ≈                | Específica por patología |
| Pitch Mean (Hz) <sub>y</sub> | ✗                  | ✗                 | ✗                 | ✗                 | ✗                | No relevante             |

Cuadro 4.1: Universalidad y utilidad binaria de los descriptores acústicos manuales según la importancia SHAP y la evidencia multilingüe/multipatológica.

# Capítulo 5

## Conclusiones

La presente investigación tuvo como objetivo central evaluar la robustez y adaptabilidad de diversos descriptores acústicos en la detección automática de alteraciones fonológicas, con un enfoque que contempla la generalización entre múltiples patologías (disartria y Parkinson) y distintos contextos lingüísticos (inglés, español e italiano). La metodología propuesta, que combina técnicas avanzadas de selección de características con análisis interpretativos mediante valores SHAP, fue implementada sobre modelos de clasificación como *LinearSVC*, *Random Forest* y perceptrones multicapa (MLP), lo que permitió obtener conclusiones precisas y bien fundamentadas respecto al desempeño, relevancia y universalidad de los descriptores acústicos estudiados.

En primer lugar, se confirmó la existencia de un conjunto reducido pero altamente robusto de descriptores acústicos con capacidad de generalización sobresaliente. Variables como VAI (Vowel Articulation Index), tVSA (total Vowel Space Area), Pause Percentage (%), Loudness Variability, Autocorr Zero Crossings, F1\_/u/ y Autocorr Energy emergieron como universalmente relevantes, mostrando consistencia transversal en distintos escenarios idiomáticos y patológicos, tanto en clasificación binaria (sano vs. patológico) como en diferenciación multi-clase. Este hallazgo reafirma que los patrones articulatorios, prosódicos y temporales reflejan alteraciones neurológicas subyacentes de forma estable y reproducible, independiente del idioma o contexto específico, lo que posiciona a estos descriptores como prioritarios para futuros sistemas diagnósticos clínicos y multilingües.

En segundo lugar, se identificaron descriptores acústicos que, si bien presentan una relevancia destacada, requieren una aplicación contextualizada según el idioma o la patología particular. Variables como Avg Strength +b, Total Bigrams (per sec) y Count -v (per sec) evidenciaron robustez considerable, pero con variaciones importantes en la magnitud de su contribución entre clases y etapas clasificatorias. Del mismo modo, descriptores como Crosscorr Entropy, Bigram/Unigram Ratio y Avg Strength -v demostraron claras preferencias por contextos específicos, sugiriendo que su aplicación práctica en modelos diagnósticos deberá considerar cuidadosamente la patología y el contexto lingüístico de implementación.

En tercer lugar, este estudio aportó evidencia empírica significativa respecto al limitado desempeño de algunos descriptores acústicos clásicos en contextos multilingües y multipatológicos. Features tradicionalmente empleados en análisis clínicos de voz, tales como Pitch Mean, Pitch Variability, Jitter y Shimmer, resultaron en general poco relevantes o irrelevantes desde la perspectiva interpretativa de SHAP. Esto cuestiona la universalidad asumida de estos parámetros y sugiere la necesidad de reevaluar críticamente su inclusión en futuros sistemas diagnósticos automatizados, especialmente si se busca adaptabilidad y generalización interlingüística.

En cuarto lugar, la introducción novedosa de características basadas en autocorrelación y correlación cruzada—entre ellas, Autocorr Zero Crossings y Autocorr Energy—se validó como uno de los aportes más significativos de este trabajo, mostrando desempeños comparables o superiores respecto a métricas clásicas. Este resultado resalta el valor potencial de la transferencia de técnicas desde dominios técnicos (como la radioastronomía en ALMA) hacia aplicaciones biomédicas, ampliando así la gama de descriptores acústicos disponibles para estudios futuros y aplicaciones prácticas.

Finalmente, desde una perspectiva metodológica, la combinación eficaz de selección de características basada en correlación y métodos recursivos (RFECV), sumado al análisis explicativo mediante SHAP, proporcionó una estrategia robusta y replicable para investigar sistemáticamente la relevancia y estabilidad de grandes conjuntos de descriptores acústicos. La estrategia metodológica propuesta aquí no solo optimiza la eficiencia computacional, sino que también mejora la interpretabilidad clínica de los modelos, facilitando la toma de decisiones diagnósticas más transparentes e informadas.

En suma, los resultados obtenidos posicionan claramente la estrategia metodológica y los conjuntos de descriptores propuestos como una alternativa prometedora, robusta y altamente generalizable para la detección automatizada de alteraciones fonológicas. Este trabajo no solo aporta hallazgos teóricos importantes para la investigación en patologías del habla, sino que también abre oportunidades concretas para desarrollar soluciones prácticas en contextos clínicos reales, multilingües y multipatológicos. Se recomienda continuar explorando la transferencia de estas técnicas a nuevos contextos clínicos e idiomas adicionales, con el fin de consolidar la solidez y alcance generalizable del enfoque planteado en esta investigación.

# Bibliografía

- [1] Berke Akkaya. “The Effect of Recursive Feature Elimination with Cross-Validation Method on Classification Performance with Different Sizes of Datasets”. En: *Proceedings of the Conference on Data Science and Applications*. 2021.
- [2] Afnan S. Al-Ali et al. “Integrating binary classification and clustering for multi-class dysarthria severity level classification: a two-stage approach”. En: *Cluster Computing* 28.136 (2025). DOI: 10.1007/s10586-024-04748-1.
- [3] Rohit Bose, Kaniska Samanta y Soumya Chatterjee. “Cross-Correlation Based Feature Extraction from EMG Signals for Classification of Neuro-muscular Diseases”. En: *2016 International Conference on Intelligent Control Power and Instrumentation (ICICPI)*. IEEE, 2016, págs. 1-6.
- [4] Suzanne Boyce, Harriet Fell y Joel MacAuslan. “SpeechMark: Landmark Detection Tool for Speech Analysis”. En: *Proceedings of Interspeech 2012*. Accessed: 2025-07-24. Portland, OR, USA: ISCA, 2012, págs. 1263-1266.
- [5] Z. Brahmi et al. “Exploring the Role of Machine Learning in Diagnosing and Treating Speech Disorders: A Systematic Literature Review”. En: *Psychology Research and Behavior Management* 17 (2024), págs. 2205-2232.
- [6] Leo Breiman. “Random Forests”. En: *Machine Learning* 45.1 (2001), págs. 5-32. URL: <https://link.springer.com/article/10.1023/A:1010933404324>.
- [7] C. Bugueño et al. *Guía terapéutica para la rehabilitación de la disartria hipocinética secundaria a enfermedad de Parkinson*. Seminario de investigación para optar al título de Licenciado en Fonoaudiología, Universidad de Valparaíso, sede San Felipe. 2022.
- [8] Girish Chandrashekar y Ferat Sahin. “A survey on feature selection methods”. En: *Computers & Electrical Engineering* 40.1 (2014), págs. 16-28.

- [9] C. Dewi y R.-C. Chen. “Random forest and support vector machine on features selection for regression analysis”. En: *International Journal of Innovative Computing, Information and Control* 15.6 (2019), págs. 2027-2037.
- [10] G. Dimauro et al. “Assessment of Speech Intelligibility in Parkinson’s Disease Using a Speech-To-Text System”. En: *IEEE Access* 5 (2017), págs. 22199-22208.
- [11] R. González y J. Bevilacqua. “Las disartrias: clasificación, evaluación y diagnóstico diferencial”. En: *Revista Hospital Clínico Universidad de Chile* 23.1 (2012), págs. 24-30.
- [12] I. Guyon y A. Elisseeff. “An introduction to variable and feature selection”. En: *Journal of Machine Learning Research* 3 (2003), págs. 1157-1182.
- [13] Zhengjun He et al. “Multiple fault feature extraction using autocorrelation of logarithmic envelope analysis spectrum for gearbox diagnosis”. En: *Measurement* 159 (2020), pág. 107780.
- [14] S. Hegde et al. “A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders”. En: *International Journal of Speech Technology* (2018).
- [15] Abner Hernandez, Sunhee Kim y Minhwa Chung. “Prosody-Based Measures for Automatic Severity Assessment of Dysarthric Speech”. En: *Applied Sciences* 10.19 (2020), pág. 6999.
- [16] C. D. Irizarry-Pérez et al. “A cross-linguistic approach to treating speech sound disorders in bilingual children”. En: *Clinical Linguistics & Phonetics* (2023).
- [17] Yannick Jadoul, Bill Thompson y Bart de Boer. “Introducing Parselmouth: A Python interface to Praat”. En: *Journal of Phonetics* 71 (2018), págs. 1-15. DOI: <https://doi.org/10.1016/j.wocn.2018.07.001>.
- [18] G. Jhavar, P. Nagraj y P. Mahalakshmi. “Speech disorder recognition using MFCC”. En: *2016 International Conference on Communication and Signal Processing (ICCSP)*. IEEE. 2016, págs. 246-249.
- [19] R. D. Kent et al. “Acoustic studies of dysarthric speech: Methods, progress, and potential”. En: *Journal of Communication Disorders* 32.3 (1999), págs. 141-186.
- [20] Wee Shin Lim et al. “A cross-language speech model for detection of Parkinson’s disease”. En: *Journal of Neural Transmission* (2024).

- [21] Scott M Lundberg, Gabriel G Erion y Su-In Lee. “Consistent individualized feature attribution for tree ensembles”. En: *arXiv preprint arXiv:1802.03888* (2019).
- [22] Scott M. Lundberg y Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. <https://arxiv.org/abs/1705.07874>. 2017.
- [23] Janaina Mendes-Laureano et al. “NeuroVoz: a Castillian Spanish corpus of parkinsonian speech”. En: *Scientific Data* (2024). Available online.
- [24] Alan V Oppenheim y Ronald W Schafer. *Discrete-Time Signal Processing*. 3rd. Pearson, 2009. ISBN: 9780131988422.
- [25] R. Pestana-Nobles et al. “Cross-Correlation and Fractal Analysis in the Images Diatoms Symmetry”. En: *Applied Sciences* 13.8 (2023), pág. 4909.
- [26] S. Pinto et al. “Dysarthria in Parkinson’s disease”. En: *The Lancet Neurology* 3.9 (2004), págs. 547-556.
- [27] James Robert y Eric Joanis. *Pydub*. <https://github.com/jiaaro/pydub>. Accessed: 2025-07-24.
- [28] G. Sanguedolce et al. “Universal speech disorder recognition: Towards a foundation model for cross-pathology generalisation”. En: *38th Conference on Neural Information Processing Systems (NeurIPS)*. 2024.
- [29] Shabda Team. *Shabda Documentation*. Accessed May 2025. Shabda Project. 2018.
- [30] SinestesiaML. *Refracted Speech - Feria de Software 2024*. 2024. URL: <https://refracted-speech.feriadesoftware.cl/> (visitado 23-07-2025).
- [31] Parwinder Pal Singh y Pushpa Rani. “An approach to extract feature using MFCC”. En: *IOSR Journal of Engineering* 4.8 (2014), págs. 21-25.
- [32] Sabine Skodda, Wenke Visser y Uwe Schlegel. “Vowel articulation in Parkinson’s disease”. En: *Journal of Voice* 25.4 (2011), págs. 467-472.
- [33] Na Sripriya et al. “Non-intrusive Technique for Pathological Voice Classification using Jitter and Shimmer”. En: *IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP)*. IEEE, 2017.

- [34] T. Sunderajan y S. V. Kanhere. “Speech and language delay in children: Prevalence and risk factors”. En: *Journal of Family Medicine and Primary Care* 8.5 (2019), págs. 1642-1646.
- [35] K. Suthar et al. “Feature engineering and machine learning for computer-assisted screening of children with speech disorders”. En: *PLOS Digital Health* 1.5 (2022).
- [36] Karl Thurnhofer-Hemsi et al. *Radial Basis Function Kernel Optimization for Support Vector Machine Classifiers*. <https://arxiv.org/abs/2007.08233>. Accessed: 2025-07-24. 2020.
- [37] J. C. Vásquez-Correa et al. “Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages”. En: *Pattern Recognition Letters* 150 (2021), págs. 272-279.
- [38] Yuanbo Wu et al. “Voice Pathology Detection and Multi-classification Using Machine Learning Classifiers”. En: *International Conference on Optoelectronic Science and Engineering* (2020).
- [39] Qing Yu, Yi Ma y Yongfu Li. “Enhancing Speech Recognition for Parkinson’s Disease Patient Using Transfer Learning Technique”. En: *J. Shanghai Jiao Tong Univ. (Sci.)* 27.1 (2022), págs. 90-98. DOI: 10.1007/s12204-021-2376-3.
- [40] B. J. Zyski y B. E. Weisiger. “Identification of dysarthria types based on perceptual analysis”. En: *Journal of Communication Disorders* 20.5 (1987), págs. 367-378.