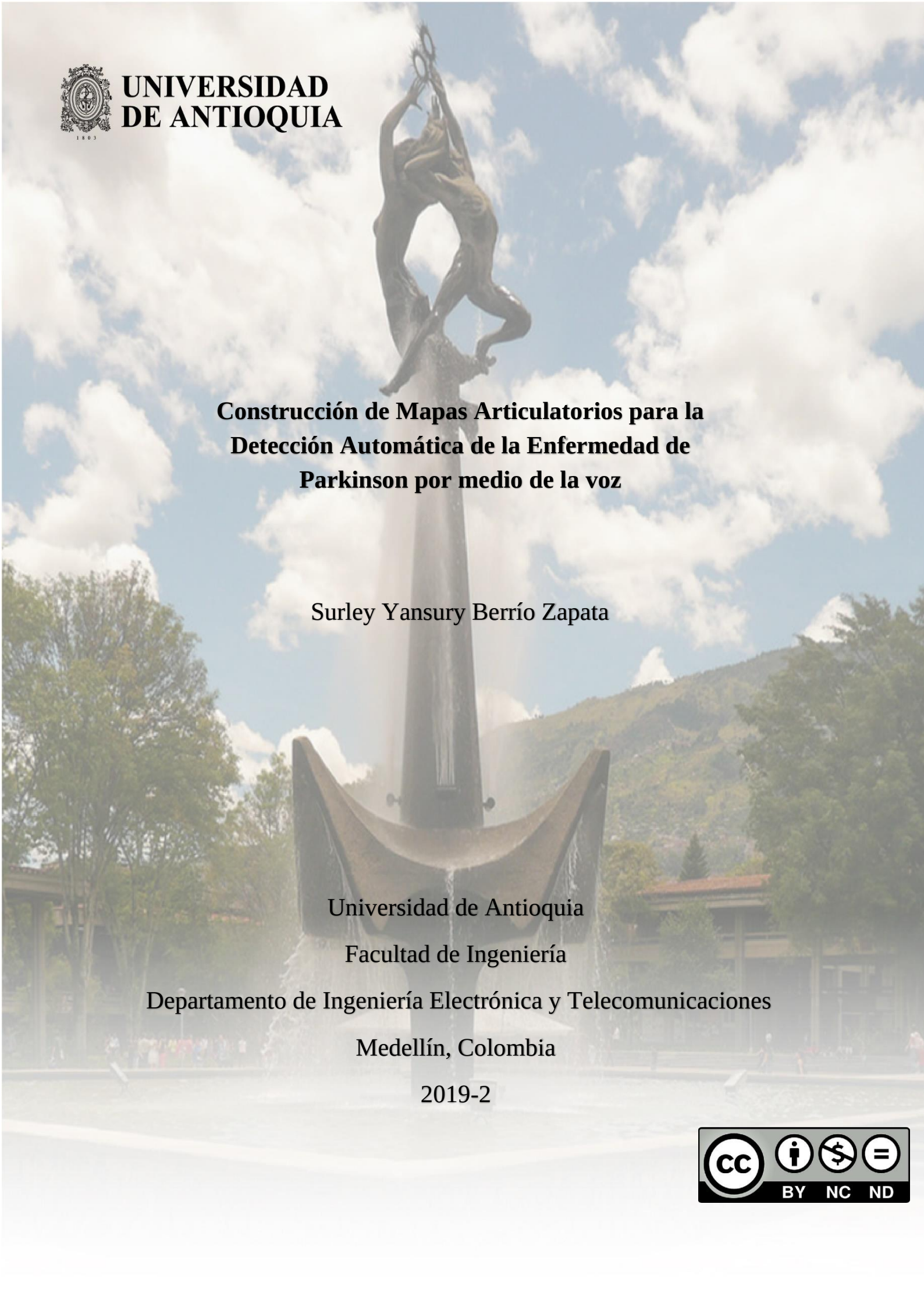

UNIVERSIDAD DE ANTIOQUIA

Construcción de Mapas Articulatorios para la Detección Automática de la Enfermedad de Parkinson por medio de la voz

Surley Yansury Berrío Zapata

Universidad de Antioquia

Facultad de Ingeniería

Departamento de Ingeniería Electrónica y Telecomunicaciones

Medellín, Colombia

2019-2

Construcción de Mapas Articulatorios para la Detección Automática de la Enfermedad de Parkinson por medio de la voz

Surley Yansury Berrío Zapata

Trabajo de grado presentado como requisito para optar al título de:
Ingeniera de Telecomunicaciones

Asesor:
PhD. Juan Rafael Orozco Arroyave

Co-Asesor:
MSc. Tomás Arias Vergara

Línea de Investigación:
Procesamiento digital de señales y análisis de patrones
Grupo de Investigación en Telecomunicaciones Aplicadas
GITA

Universidad de Antioquia
Facultad de Ingeniería
Departamento de Ingeniería Electrónica y Telecomunicaciones
Medellín, Colombia
2019-2

Agradecimientos

Mi sincera gratitud para mi asesor Juan Rafael Orozco Arroyave por sus ideas, orientación y apoyo en la realización de este proyecto. También agradezco a Tomás Arias y a mis compañeros del grupo de investigación GITA, Sebastián Guerrero, Felipe López, Nicanor García, Felipe Parra, Felipe Gómez, y muy especialmente a mis compañeros y amigos Cristian David Ríos y Daniel Escobar por creer en mí, por darme ánimos y por su colaboración durante todo este tiempo. Sin duda, el conocimiento de todos ustedes, experiencia y apoyo, contribuyó de manera significativa e hizo posible el desarrollo de este trabajo.

A mi querida Universidad de Antioquia, infinitas gracias por permitirme concluir este sueño e inspirarme a crecer en todos los ámbitos no sólo académicos, ha sido una lámpara en este largo camino. Gracias a todos mis profesores y compañeros de carrera que aportaron su conocimiento y compañía durante estos años. Gracias al proyecto CODI PR19-2-11 por la ayuda para este proyecto. También quiero agradecer a Konecta, en especial a Andrés Felipe Maya y a Diana Catalina Velásquez, por brindarme la oportunidad de hacer parte de sus equipos de trabajo, han sido de gran apoyo estos últimos semestres.

Finalmente, mi gratitud y reconocimiento a mi madre Silvia Zapata, a quien debo lo que soy y quien me enseñó a no rendirme a pesar de las circunstancias; su amor, paciencia y comprensión me sostuvieron todo este tiempo. Gracias a mis hermanos en especial a Angela Berrío, quien siempre creyó en mí y me dió ánimos cuando me faltaron fuerzas para continuar. También agradezco a Gustavo Muñoz por su apoyo, sus consejos y por motivarme a seguir adelante en la consecución de mis objetivos académicos.

Índice

Índice de figuras

Índice de tablas

Resumen

La enfermedad de Parkinson (EP), es una enfermedad neurológica progresiva que afecta el sistema motor, entre los síntomas más frecuentes se encuentran rigidez muscular, temblor y trastornos de la voz. La representación acústica del habla ha sido el foco de investigaciones sobre la voz patológica; la extracción de los rasgos más importantes de las señales de voz a través del uso de algoritmos computacionales, técnicas de estimación y extracción de características, así como diferentes modelos para la clasificación, han permitido la detección de esta enfermedad neurodegenerativa. El presente trabajo propone realizar un análisis fonológico de la señal de voz, para la construcción de mapas articulatorios que permita la detección de la EP y que además sirva de apoyo para los especialistas de la voz, en determinar una posible terapia del habla. La estrategia usada en el presente trabajo consistió en analizar 100 grabaciones de audio (50 EP y 50 controles sanos). Los participantes leyeron un texto, balanceado fonológicamente, que contiene todos los sonidos del español hablado en Colombia. Se extrajeron características articulatorias y además fonológicas través de dos herramientas PhonVoc y Phonet. Luego, a través de alineamiento forzado, se realizó un etiquetado a nivel de fonema para agrupar las clases fonológicas y posteriormente, se implementaron dos técnicas de aprendizaje automático para la clasificación de pacientes vs controles, máquinas de vectores de soporte y árboles aleatorios; para ambos algoritmos se realizó optimización de parámetros con una validación cruzada con k=10. Los experimentos muestran un acierto de 90 % en la clasificación de pacientes con EP vs controles con la clase Vocales y aciertos superiores al 80 % para las clases Nasales, Fricativas sordas, Oclusivas sonoras, *Back, Coronal, Voice, Open, High* y *Low*; lo que indica que esta propuesta es una alternativa adecuada tanto para la detección automática de la EP como para la evaluación del déficit en la articulación de los fonemas contenidos en las clases fonológicas. En la etapa final del proyecto, se construyeron mapas articulatorios usando GMM's (*Gaussian Mixture Models*), que modelaron las dos poblaciones (pacientes , controles) y agruparon las clases fonológicas para su análisis y visualización en un espacio de dos dimensiones.

1. Introducción

Los especialistas de la voz hablada trabajan en forma multidisciplinaria en la observación clínica (anatomía y fisiología del aparato fonador) e investigación en el fenómeno acústico para abordar los aspectos importantes que conllevan a una adecuada clasificación de los diferentes problemas vocales [1]. Algunas afecciones que influyen en el funcionamiento del aparato fonador, son las enfermedades neurológicas, entre ellas la EP, que es la segunda enfermedad neurodegenerativa más frecuente en el mundo, después del Alzheimer, afecta aproximadamente el 2 % de la población mayor de 65 años [2] y se caracteriza por la pérdida gradual de los receptores dopaminérgicos en el sistema nervioso central, propiamente en la sustancia nigra del cerebro, afectando en principio la capacidad motora del paciente. Entre los síntomas más comunes se encuentran rigidez muscular, temblor, bradicinecia, trastornos posturales, alteraciones en la escritura, disartria hipocinética [3], entre otros. En la producción del habla, los mecanismos de emisión sonora se ven afectados por la rigidez de los diferentes nervios que participan en la modulación de la voz, provocando trastornos en la fonación - proceso mediante el cual se produce aire en los pulmones, para hacer vibrar los pliegues vocales y generar los sonidos vocálicos - y la articulación que tiene que ver con el movimiento de los diferentes órganos articuladores para modificar el sonido y pronunciar de manera correcta determinado fonema; siendo estas dos alteraciones fonación y la articulación, el principal objeto de estudio en la detección de la EP [4] [5]. En un análisis fonético-acústico de la señal de voz, se realiza el estudio de las características físicas del sonido, el análisis fonético-articulatorio se basa en la forma como se produce el sonido y un análisis fonológico se centra en el estudio exclusivo de las unidades mínimas de sonido/fonemas de una lengua concreta [6].

1.1. Análisis articulatorio en voces de pacientes con enfermedad de Parkinson

La función articulatoria puede ser evaluada y analizada en sonidos tanto sonoros como sordos así como en las transiciones entre ellos; una reducción de la amplitud, velocidad, precisión y variabilidad de los movimientos de los órganos articuladores (labios, lengua y mandíbula), evidencian un déficit en la articulación y conduce a una producción imprecisa de vocales y consonantes [7] [8]. En estudios previos, se han realizado evaluaciones y análisis de la articulación vocal en pacientes con EP a través de la pronunciación de las vocales sostenidas, así como de la articulación de consonantes con

otro tipo de tareas del habla, incluyendo aquellas dependientes del texto como frases leídas y/o texto leído; tareas de habla espontánea o monólogo; y tareas para evaluar los parámetros de diadococinesia (DDK) tras la repetición de las sílabas "pa-ta-ka", y observar la función articulatoria de los fonemas oclusivos sordos (/p/,/t/,/k/) al unirse con el sonido vocálico sonoro /a/ [9] [10] [11] [12] [13]. Algunos autores concluyen en sus estudios que la tarea de producción del habla puede ser un factor importante para la evaluación de la disartria, indicando que el habla, en pacientes con EP, es menos inteligible cuando habla espontáneamente que cuando lee un texto elaborado, lo que puede conllevar a una planificación articulatoria [14] [15].

En la caracterización, se extraen principalmente las frecuencias del primer (F1) y segundo (F2) formantes, ya que estos están relacionados con la activación de los músculos articulatorios específicos de la lengua y los labios [16] [17] [18]; mientras que la frecuencia F1 varía inversamente con la altura de la lengua, la frecuencia F2 varía directamente con el avance o posición que adopte la lengua (punto y modo de articulación) [19]. Los valores de F1 y F2 se usan también para hallar medidas como el área de espacio vocálico (VSA: *Vowel Space Area*) [20] [21], que describe entre otras, la calidad de la voz del hablante y que en personas con disartria -característica de pacientes con EP- se evidencia una reducción en su medida [22] [23] [24] [25] [26] [27] [28]. El índice de articulación vocal (VAI: *Vowel Area Index*) también es una medida del nivel de disartria al igual que su valor inverso, conocido como la tasa de centralización de formantes (FCR: *Formant Centralization Ratio*), que miden y minimizan los efectos de la variabilidad entre hablantes para maximizar la sensibilidad a la centralización de formantes [29] [30]. En [31] los autores analizaron la voz de 20 pacientes con EP, diagnosticados de manera temprana, y 15 controles sanos. Las tareas del habla utilizadas fueron fonaciones sostenidas de las vocales /a/, /i/, /u/; frases leídas, texto leído, y monólogo. El conjunto de características fue conformado por F1, F2, VSA, VAI y la relación entre los segundos formantes de las vocales /i/ y /u/ (F2i /F2u). Los resultados sugieren que la fonación sostenida es una tarea inapropiada para investigar la articulación de las vocales en la EP, mientras que el monólogo fue el más sensible para la clasificación entre pacientes con EP y controles con un acierto mayor del 80 %. Luego, en [32] se analizaron los efectos en la articulación de las vocales en pacientes con EP, posterior al tratamiento de la voz (LSVT: *Lee Silverman Voice Treatment*) [33]. Se consideraron grabaciones de voz de 14 controles y 29 pacientes con EP donde sólo 15 recibieron LSVT. Las vocales fueron extraídas de las palabras *key, stew* y *Bobby* que se encontraban inmersas en las frases leídas. La extracción de características fue a través de primer (F1) y segundo (F2) formante de las

vocales /a/, /i/, /u/ y la relación entre los segundos formantes de las vocales /i/ y /u/ (F2i /F2u) y VSA. Se concluye que la terapia con un solo foco -aumento de la sonoridad- tiene un efecto positivo en las funciones articulatorias. Es necesario futuros estudios para determinar los beneficios de esta terapia a largo plazo. En [34] se modelaron diferentes déficits articulatorios de personas con EP, para este propósito, los autores consideraron para el conjunto de características, evaluar seis aspectos articulatorios como calidad vocal, la coordinación de la actividad laríngea y supralaríngea, la precisión de la articulación consonante, el movimiento de la lengua, el debilitamiento de la oclusión y la sincronización del habla. La base de datos está formada por 24 personas con EP y 22 controles, balanceada en edad, donde cada participante realizó tareas DDK con la repetición de las sílabas "pa-ta-ka". Los fonemas fueron etiquetados manualmente y luego los autores implementaron un clasificador con SVM (*Support Vector Machine*), reportando que el mejor resultado de clasificación entre EP y controles logró una tasa de acierto del 88 %. Además, de acuerdo con los resultados de esta metodología, los autores descubrieron que la articulación imprecisa de consonantes es el indicador más poderoso de disartria relacionada con la EP. En [35] los autores proponen una estrategia para modelar a través de una red neuronal convolucional, el déficit articulatorio de los pacientes con EP relacionado con la capacidad para iniciar/ detener la vibración de las cuerdas vocales a partir de dos representaciones tiempo-frecuencia de la señal: la transformada de Fourier de tiempo corto y una representación de onda continua para evaluar y analizar características acústicas de la señal en las transiciones de los segmentos sonoros y no sonoros. En este estudio se usaron bases de datos en tres diferentes idiomas: español (50 pacientes, 50 controles), alemán (88 pacientes, 88 controles) y checo (20 pacientes, 15 controles), donde los participantes realizaron las repeticiones de las sílabas /pa-ta-ka/. Se realizó la clasificación con SVM pacientes vs sanos, obteniendo resultados de aciertos de 85.4 % en la base de datos en español, 70.7 % en la base de datos en Alemán y 89.2 % en la base de datos en Checo. De manera similar en [36] se analizaron características acústicas de segmentos sonoros y no sonoros para la detección automática de la EP. Las grabaciones fueron obtenidas bajo condiciones de ruido no controlado de 28 participantes, 14 pacientes y 14 controles capturando seis oraciones y un texto leído por participante. Su estrategia se basó en aplicar una técnica de realce del habla (SE: *Speech enhancement*) para mejorar la calidad de las señales, además realizaron una caracterización de los segmentos sonoros y no sonoros de acuerdo con las tareas de voz. La clasificación de pacientes vs controles se realizó con SVM y se reportaron aciertos para los segmentos sonoros entre el 64 % y el 86 % y para los segmentos no sonoros aciertos entre el 78 % y el 99 %, concluyeron además que el algoritmo SE había mejorado

en 11 puntos porcentuales los aciertos de los segmentos no sonoros, mientras que los valores de acierto disminuyeron para los segmentos sonoros al aplicar la técnica. En las metodologías antes mencionadas, se realizó un análisis articulatorio en pacientes con EP, sin embargo, no se realizó un análisis fonológico a través de una evaluación de fonemas específicos con criterios de punto y modo de articulación.

En otras investigaciones, se examinó la función articulatoria en pacientes con EP teniendo en cuenta un estudio fonético y también fonológico. Los modos de fonación, describen según propiedades laríngeas y supralaríngeas propias de un individuo, las emisiones del habla o fuerzas musculares que actúan en los pliegues vocales que caracterizan la calidad de la voz (VQ: *Voice Quality*); se clasifican en fonación modal (fonación normal o habitual de un hablante) y fonaciones no modales: voz de hálito (*Breathy voice*), voz dura (*Harsh voice*), voz de falsete (*Falsetto*), voz rota (*Creaky voice*) y voz susurrada (*Whispery voice*) [37]. En [38] los autores realizan una caracterización de la calidad de la voz en pacientes con EP midiendo el deterioro del habla de los pacientes en las dimensiones fonación y articulación caracterizada por presentar una voz ronca y con sonido semi-suspendido. Esta metodología se basó en un análisis tiempo-frecuencia de fonaciones no modales (*breathy*, *creaky*, *tense*, *falsetto y harsh*) de 50 pacientes con EP y 50 controles sanos nativos de Colombia; se extrajeron características fonológicas con rasgos de pronunciación propias del español por punto y modo de articulación a través del software [39]; posteriormente realizaron medidas estadísticas sobre los valores obtenidos de las características fonológicas tanto de los pacientes como de los controles sanos para calcular el nivel de afectación en las fonaciones no modales de cada población. Los autores concluyeron que la mayor parte del espectro de VQ de la EP está compuesto por un 30 % de voz de hálito, 23 % voz crujiente, y 20 % voz tensa, las tres medidas más esperadas en la valoración de la disartria hipocinética en la EP. En [40], el autor propone un análisis de la señal de voz con extractor de i-vector para evaluar y monitorear el estado neurológico de las personas con EP. En esta metodología se extrajeron características que modelan diferentes dimensiones del habla: fonación, articulación, prosodia y fonología; luego se extrajeron i-vectores de las características de las señales de prueba de 50 pacientes con EP y 50 controles sanos con las que se entrena un extractor de i-vector. Finalmente, se realizó una comparación entre los i-vectores de referencia y unos i-vectores de prueba utilizando la distancia coseno. Se utilizaron varias tareas del habla como DDK con la repetición rápida de las sílabas "pa-ta-ka", frases, texto leído y monólogo. Los experimentos muestran que esta metodología tiene un acierto en la clasificación de hasta el 80 %, una correlación de Spearman de hasta 0,54 con el estado neu-

rológico y de hasta 0,72 con el nivel de disartria del paciente. En [41] [42] [43] los autores informaron una articulación imprecisa en los fonemas oclusivos los cuales se producían como fonemas fricativos, evidenciando un déficit articulatorio producto de una incorrecta elevación de la lengua y una constricción inadecuada en la producción de los fonemas oclusivos y fricativos.

En [44], el autor presenta una metodología con agrupación de fonemas, utilizando la técnica de alineamiento forzado para el etiquetado de seis clases fonológicas: vocales, oclusivas, fricativas, africadas, líquidas y nasales. En este trabajo se busca encontrar un modelo para cada clase fonológica usando un GMM-UBM (*Gaussian Mixture Model - Universal Background Model*) donde el UBM es generado a partir del corpus Albayzin [45] y otros tres corpus con habla parkinsoniana (Neurovoz [46], GITA [47], CzechPD [31]) son usados para la adaptación de los GMM. En los corpus GITA (50 EP, 50 C) y Neurovoz (47 EP, 32 C) se usaron DDK con la repetición de las sílabas "pa-ta-ka", frases leídas y monólogo; para el corpus CzechPD (20 EP, 14 C), sólo se tuvo en cuenta los DDK a través de la repetición de las sílabas "pa-ta-ka". Las señales fueron caracterizadas usando la primera y segunda derivada de la técnica Rasta-plp (*RelAtive SpecTrAl - perceptual linear predictive*), la cual está basada en el espectro de tiempo corto del habla. Para la clasificación con GMM-UBM se presentaron tres enfoques "raw-phon", "phon-phon" y "phon-raw". Los resultados arrojan un acierto en clasificación pacientes vs controles entre el 77 % y el 81 % usando el enfoque "raw-phon"; un acierto entre el 77 % y el 86 % con "phon-phon" y un acierto entre el 77 % y el 94 % con "phon-raw". El autor informa que los resultados sugieren que las oclusivas, vocales y fricativas (en ese orden), son los segmentos acústicos más relevantes para la detección de EP, además el empleo de frases leídas produjo modelos más consistentes y precisos que el uso de monólogos o tareas DDK. Se destaca de este trabajo que es de las primeras metodologías en utilizar GMM-UBM por clase fonológica para la clasificación de pacientes con EP y controles.

En las diferentes metodologías analizadas, a excepción de la última, no se realizó una evaluación de fonemas específicos o agrupamiento de estos con criterios de punto y modo de articulación. Lo cual servirá en el futuro para orientar la terapia de los pacientes tal que mejore los movimientos articulatorios y por ende la producción de fonemas específicos.

1.2. Hipótesis

Es posible medir y evaluar características fonológicas de punto y/o modo de

articulación para fonemas específicos del Español producidos por pacientes en enfermedad de Parkinson y personas sanas.

1.3. Objetivos

1.3.1. Objetivo general

Construir mapas articulatorios considerando punto y/o modo de articulación mediante algoritmos de inteligencia artificial para la clasificación automática de personas con enfermedad de Parkinson y personas sanas.

1.3.2. Objetivos específicos

1. Evaluar al menos dos herramientas computacionales que permitan el etiquetado automático de las diferentes categorías fonológicas en grabaciones de la población objetivo.

2. Aplicar métodos de análisis de patrones que utilicen las etiquetas fonológicas y permitan la clasificación automática de personas con enfermedad de Parkinson y personas sanas.

3. Implementar al menos dos métodos que permitan una visualización apropiada de la distribución de las características fonológicas en espacios de dos dimensiones y que además permitan observar la separabilidad entre las dos clases (pacientes vs sanos).

4. Evaluar la capacidad de generalización del sistema a través de la técnica de validación cruzada con k-particiones, así como evaluar el desempeño de la metodología propuesta mediante diferentes métricas de rendimiento del sistema tales como acierto, especificidad, sensibilidad, curvas ROC (*Receiver Operating Characteristic*) y AUC (*Area Under the ROC Curve*).

1.4. Contribución de este trabajo

En este trabajo se propone una metodología para la construcción de mapas articulatorios que consiste en un análisis articulatorio y fonológico del habla tal que permite la clasificación automática de personas con EP vs sanos. Además de los experimentos de clasificación, se presenta una metodología, basada en técnicas de análisis de patrones y métodos de agrupamiento, que permite una fácil visualización e interpretación de los resultados.

2. Marco teórico

A continuación se presentan todos los fundamentos y conceptos teóricos en los cuales se apoya este trabajo.

2.1. Proceso de producción del habla

La producción de los sonidos del habla es el proceso mediante el cual una representación lingüística - el mensaje - se convierte en una onda sonora por acción del aparato fonador. En términos físicos, el aparato fonador se compone de tres partes fundamentales: Los pulmones (generador de energía), la laringe (sistema vibratorio) y las cavidades supraglóticas (cavidad de resonancia) [6].

Así, el proceso de producción de los sonidos del habla, inicia con una corriente de aire que proviene de los pulmones, la cual atraviesa la laringe y al paso por la glotis, pone en vibración las cuerdas vocales si estas se encuentran juntas, dando lugar a la fonación o producción de la voz. Luego, el aire llega a las cavidades supraglóticas y, según la posición de la úvula, se reparte entre la cavidad nasal y bucal, o se concentra únicamente en la cavidad bucal donde es modificado mediante el proceso conocido como articulación. Los órganos que componen las cavidades supraglóticas son el paladar (cuya función es servir de barrera a la corriente de aire); la lengua (que con su posición y punto de apoyo define el modo de articulación del sonido); los dientes y los labios (que cumplen también con la función de servir de barrera física al sonido, modificándolo, según su movimiento).

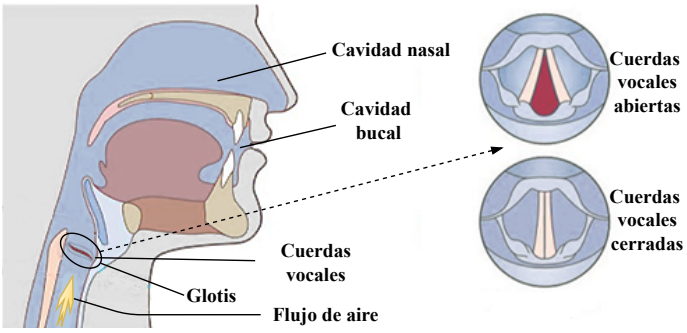

Figura 1. Aparato Fonador. [a]

[a] Figura adaptada de Logopedia, 2018. https://ugrlogop.wordpress.com/fotos-graficos-e-ilustraciones

2.2. Fonética

La fonética es la rama de la lingüística que estudia la variación articulatoria y acústica de los sonidos del habla. Tiene como objetivo determinar el modo en que los sonidos del habla se emplean con fines comunicativos en las lenguas naturales y qué mecanismos intervienen en su producción y percepción. Los sonidos del habla pueden considerarse desde el punto de vista de su producción en el emisor, como fonética articulatoria, debido a los movimientos coordinados de los diferentes articuladores, para emitir un sonido con determinada estructura lingüística y de su transmisión, como fonética acústica, que se encarga de las propiedades físicas (onda sonora) de los sonidos del habla [50].

2.2.1. Clasificación articulatoria de los sonidos del habla

Partiendo de la configuración de cada una de las partes del aparato fonador para la producción de los sonidos del habla, Ladefoged [50] realiza una descripción de estos sonidos bajo los siguientes criterios de clasificación:

✓ En función del estado de la glotis:

Tabla 1. Clasificación articulatoria de los sonidos del habla en función del estado de la glotis

Estado de la glotis	***Clasificación***
Glotis abierta, sin vibración de las cuerdas vocales.	Sordo
Glotis cerrada, con vibración de las cuerdas vocales, que se mantienen juntas por la acción de los cartílagos aritenoideos.	Sonoro

✓ En función del punto de articulación:

Tabla 2. Clasificación articulatoria de los sonidos consonánticos en función del punto de articulación

Punto de articulación y contacto entre los articuladores	***Clasificación***
Labio superior y labio inferior	Bilabial
Labio inferior e incisivos superiores	Labiodental
Ápice o dorso de la lengua y la parte posterior de los incisivos superiores	Dental
Ápice de la lengua situado entre los incisivos inferiores y superiores	Interdental
Elevación del ápice de la lengua hacia la parte posterior de los alvéolos	Retroflejo
Parte anterior de la lengua y alvéolos	Alveolar
Parte anterior del dorso de la lengua y paladar duro	Palatal
Parte posterior de la lengua y velo del paladar	Velar
Parte posterior de la lengua y úvula	Uvular
Raíz de la lengua y pared faríngea	Faringeo
Cierre de la glotis	Glotal

Tabla 3. Clasificación articulatoria de los sonidos vocálicos en función del punto de articulación

Punto de articulación para los sonidos vocálicos	***Clasificación***
La lengua ocupa la región delantera o zona del paladar duro	Anterior o Palatal
La lengua ocupa la zona intermedia cubierta por el mediopaladar	Central
La lengua ocupa la región posterior o zona del paladar blando	Posterior o Velar

- ✓ En función de los procesos oronasales o también conocido como modo de articulación:

Tabla 4. Clasificación articulatoria de los sonidos consonánticos en función del modo de articulación

Modo de articulación para los sonidos consonánticos	***Clasificación***
El velo del paladar se eleva provocando que la úvula obstruya el paso del aire hacia la cavidad nasal	Oral
El velo del paladar desciende y el aire puede pasar libremente hacia la cavidad nasal	Nasal
Cierre completo del paso del aire	Oclusivo
Estrechamiento del canal de paso del aire de modo que se produce una turbulencia	Fricativo
Cierre completo del paso del aire seguido de un estrechamiento del canal que produce una turbulencia	Africado
Aproximación de dos articuladores sin que se llegue a dar un estrechamiento tal que se produzca una turbulencia	Aproximante
Vibración repetida de un articulador contra otro	Vibrante Múltiple
Contacto momentáneo de un articulador móvil con uno fijo	Vibrante Simple
Salida del aire por ambos lados de la cavidad bucal	Lateral
Salida del aire por el centro de la cavidad bucal	Central

Tabla 5. Clasificación articulatoria de los sonidos vocálicos en función del modo de articulación

Modo de articulación para los sonidos vocálicos	***Clasificación***
La lengua está muy próxima al paladar duro o al paladar blando	Alta o Cerrada
La lengua no está ni muy próxima ni muy separada de la bóveda de la cavidad bucal	Media
La lengua se separa totalmente del paladar y se encuentra en el límite máximo de alejamiento	Baja o Abierta

2.3. Fonología

La fonología es la rama de la lingüística que se ocupa del estudio de los fonemas o unidad mínima de sonido de una determinada lengua, que tiene carácter distintivo y produce diferencias de significado.Los fonemas vocálicos son aquellos que no encuentran ningún tipo de obstáculos en la cavidad bucal cuando el sonido es representado y en el caso de los fonemas consonánticos encuentran algún tipo de obstáculo a la hora de ejecutar el sonido. Cada idioma contiene su propio sistema fonológico o alfabeto fonético, en el caso del idioma Español hablado en Colombia, está conformado por 23 diferentes fonemas, 5 vocales y 18 consonantes [51]. En la Tabla 6 y la Tabla 7 se describen estos fonemas con su punto y modo de articulación, así mismo, en la Figura 2 se observa el punto de articulación para la producción de estos.

Tabla 6. Fonemas consonánticos de uso del español hablado en Colombia. **Sor**: Sordos, **Son**: Sonoros

MODO DE ARTICULACION	**PUNTO DE ARTICULACION**												
		Bilabiales		Labiodentales		Dentales		Alveolares		Palatales		Velares	
	Sonoridad	Sor	Son	Sor	Son	Sor	Son	Sor	Son	Sor	Son	Sor	Son
	Oclusivos	p	b			t	d					k	g
	Fricativos			f				s			ʝ	x	
	Africados									ʧ			
	Laterales							l			ʎ		
	Nasales		m						n		ɲ		
	Vibrantes								ɾ				
									r				

Tabla 7. Fonemas vocálicos de uso del español hablado en Colombia.

MODO DE ARTICULACION	PUNTO DE ARTICULACION		
	Anterior	Central	Posterior
Alta	i		u
Media	e		o
Baja		a	

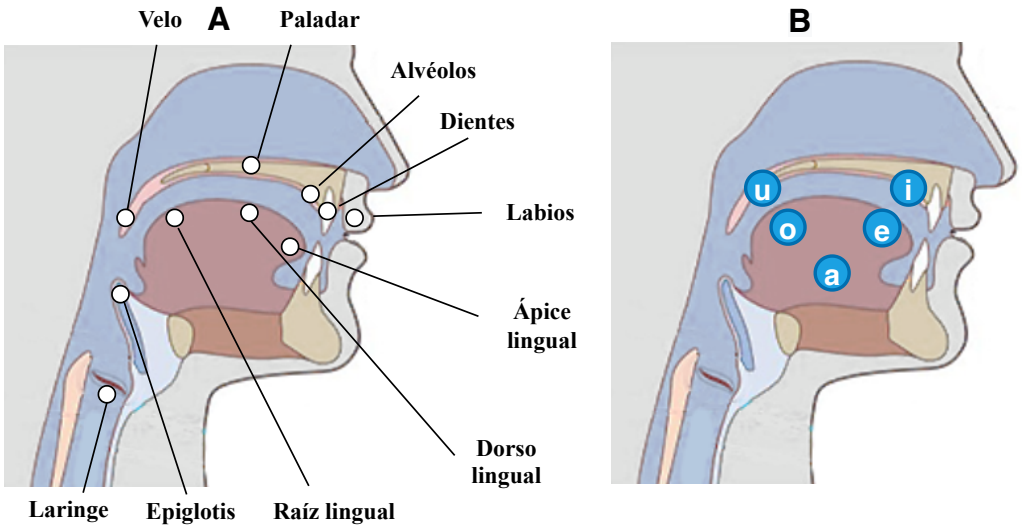

Figura 2. Punto de Articulación. **A** Consonantes, **B** Vocales

En la presente propuesta se agruparon los fonemas por punto y modo de articulación en las siguientes clases fonológicas:

- **Vocales:** /a/,/e/,/i/,/o/,/u/
- **Nasales:** /m/,/n/,/ɲ/ (grafema: ñ)
- **Fricativas sordas:** /f/, /s/ (grafemas: s,c,z), /x/ (grafemas: g, j)
- **Fricativas sonoras:** /ʝ/ (grafema: y)
- **Africadas:** /ʧ/ (grafema: ch)
- **Oclusivas sordas:** /p/,/t/,/k/
- **Oclusivas sonoras:** /b/, /d/, /g/
- **Laterales:** /l/, /ʎ/ (grafema: ll)
- **Vibrantes:** /r/, /ɾ/ (grafema: rr)

2.4. Alineamiento Forzado

Los sistemas de alineación forzada son muy utilizados para el entrenamiento de modelos acústicos y reconocimiento del habla. Esta técnica consiste en alinear los fonemas y su ocurrencia en el tiempo dentro de cada archivo de audio y asignarle las marcas de tiempo [52].

Un sistema automático de reconocimiento del habla (ASR: *Automatic Speech Recognition*) procesa las señales de voz para modelar el lenguaje hablado basados en técnicas como HMM *(Hiden Markov Model)* y redes neuronales profundas que describen los patrones acústicos que faciliten la clasificación de los segmentos o niveles semánticos de la voz y convertir en texto las palabras habladas [53]. SARHA (Sistema Automático de Reconocimiento del HAbla) es un ASR desarrollado por el grupo de investigación en telecomunicaciones aplicadas GITA para efectos de investigación. El sistema se desarrolló usando una base de datos recolectada por el grupo; se grabaron 103 hablantes quienes leyeron diez frases diferentes y repitieron diez veces cada frase. Se capturaron 10300 grabaciones para una duración total de diez horas de audio; el 90 % de los datos fueron utilizados para entrenar el sistema y el 10 % restante para validar su desempeño. El sistema se entrenó utilizando un modelo de lenguaje trigrama (palabra de tres letras) y un modelo acústico basado en HMM y GMM de tres estados, siendo el trifonema la unidad mínima para incluir el contexto. El reconocedor cuenta con una tasa de acierto de aproximadamente

90 %.Esta herramienta fue utilizada en este trabajo para obtener los tiempos de inicio y fin de cada uno de los fonemas y etiquetar las clases fonológicas; para tal fin, se ingresó cada audio con la transcripción de su correspondiente texto leído. En la Figura 3 se muestra un ejemplo de etiquetado de fonemas.

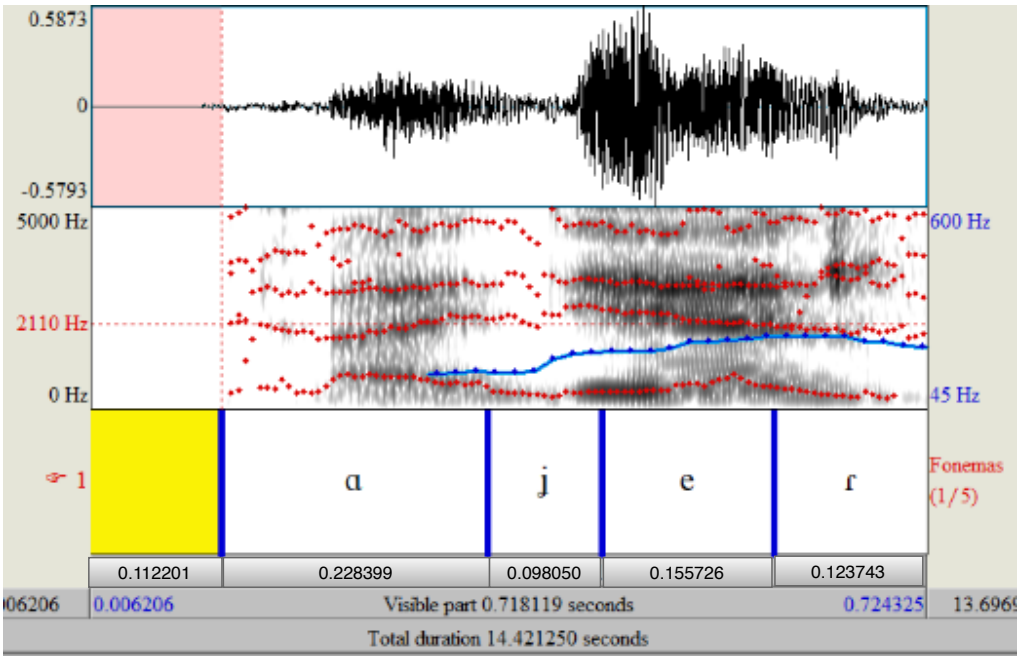

Figura 3. Forma de onda y espectrograma de un paciente con EP y tiempos de los fonemas pronunciando la primera palabra del texto leído [a-y-e-r].

2.5. Reconocimiento de patrones

En ciencias de la computación, un patrón del habla se representa usando valores de características vectoriales de los rasgos significativos que, luego de asignarle una etiqueta de clase, se le aplican en la etapa de reconocimiento, técnicas de clasificación con algoritmos de aprendizaje automático para realizar tareas discriminatorias, predictivas o explicativas. En la literatura sobre la voz patológica en pacientes con Parkinson, el reconocimiento de patrones ha mostrado excelentes resultados en este campo de aplicación [54], los valores estimados de características acústicas entre las más usadas MFCC's, frecuencia fundamental, frecuencias formantes, entre otras; forman el vector

de características que representan la señal a analizar. En la Figura 4 se muestra la metodología implementada en este estudio, siguiendo la estructura de las diferentes etapas de un sistema clásico de reconocimiento de patrones.

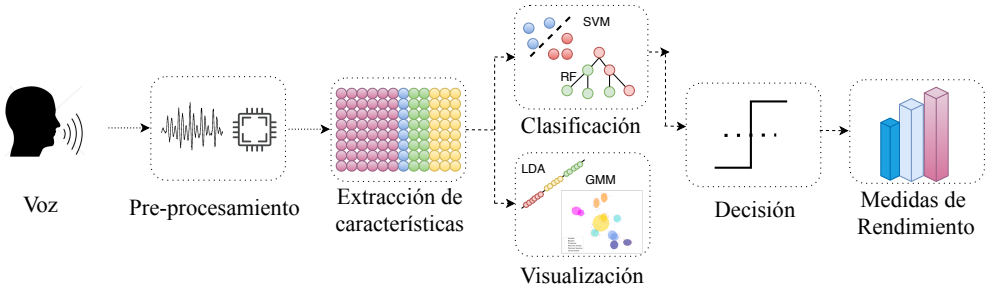

Figura 4. Diagrama de un sistema de reconocimiento de patrones con la metodología implementada en este estudio.

2.5.1. Adquisición de los datos

El corpus (PC-GITA) fue construido por el Grupo de Investigación en Telecomunicaciones Aplicadas GITA y presentado a la comunidad científica en 2014 [47]. Esta base de datos contiene grabaciones de 100 hablantes nativos del español de Colombia, 50 pertenecen a personas diagnosticadas con EP y el resto son controles sanos, 25 hombres y 25 mujeres en cada grupo. La edad de los hombres con EP va desde 33 hasta 77 años con un promedio y variación de (62.2±11.2), la edad de las mujeres con EP se encuentran en el rango de 44 a 75 años con un promedio y variación de (60.1±7.8). El rango de edad para las hombres en controles sanos es de 31 a 86 años (media 61.2±11.3) y el de las mujeres es de 43 a 76 años (media 60.7±7.7). Las grabaciones fueron capturadas con una tasa de muestreo de 44100 Hz y 16 bits de resolución, usando un micrófono dinámico omnidireccional (Shure, SM 63L) y una tarjeta de audio profesional (M-Audio, Fast Track C400). Los pacientes fueron diagnosticados con EP por Neurólogos y etiquetados de acuerdo con la escala UPDRS *(Unified Parkinson's Disease Rating Scale)* [55], escala perceptual utilizada para evaluar el estado neurológico de los pacientes. Ver Tabla 8 con los detalles del corpus.

Los participantes de las grabaciones, realizaron diferentes tareas del habla que incluían:

- **Evaluación diadococinética (*DDK : Diadochokinetic*).** Realizaron la repetición rápida de las sílabas /pa-ta-ka/, /pa-ka-ta/, /pe-ta-ka/, /pa/,/ta/,/ka/.

- **Evaluación de habla espontánea.** Los participantes hablaron acerca de lo que hacen en su día normal.

- **Frases cortas.** Leyeron en voz alta diez oraciones cortas aisladas.

- **Texto leído.** Leyeron en voz alta un texto fonéticamente balanceado. Esta fue la tarea de lenguaje elegida para el análisis de voz en el presente trabajo, ya que contiene todos los sonidos hablados en Colombia. Se presenta como un diálogo entre un paciente (**P**) y un médico (**M**) descrito como sigue:

  - **P:** Ayer fui al médico.
  - **M:** Qué le pasa? Me preguntó.
  - **P:** Yo le dije: Ay doctor! Donde pongo el dedo me duele.
  - **M:** Tiene la uña rota?.
  - **P:** Sí.
  - **M:** Pues ya sabemos qué es. Deje su cheque a la salida.

Tabla 8. Edad, UPDRS y tiempo (t) tras el diagnóstico de EP. Para el caso de controles sanos, sólo se proporciona la edad.

Hombres				**Mujeres**			
EP			**Controles**	**EP**			**Controles**
Edad	UPDRS	t	Edad	Edad	UPDRS	t	Edad
81	5	12	86	75	52	3	76
77	92	15	76	73	38	4	75
75	13	1	71	72	19	2,5	73
75	75	16	68	70	23	12	68
74	40	12	68	69	19	12	65
69	40	5	67	66	28	4	65
68	14	1	67	66	28	4	64
68	67	20	67	65	54	8	63
68	65	8	67	64	40	3	63
67	28	4	65	62	42	12	63
65	32	12	64	61	21	4	63
65	53	19	63	60	29	7	62
64	28	3	63	59	40	14	62
64	45	3	62	59	71	17	61
60	44	10	60	58	57	1	61
59	6	8	59	57	41	37	61
57	20	0,4	56	57	61	17	60
56	30	14	55	55	30	12	58
54	15	4	55	55	43	12	57
50	53	7	54	55	30	12	57
50	19	17	51	55	29	43	55
48	9	12	50	54	30	7	55
47	33	2	42	51	38	41	50
45	21	7	42	51	23	10	50
33	51	9	31	49	53	16	49

2.5.2. Preprocesamiento

En esta etapa se acondicionan los datos para ser procesados y analizados adecuadamente. En el presente estudio, sólo fue necesario submuestrear los audios a 16000 Hz ya que las herramientas que se utilizan para la estimación de características fonológicas, así lo requerían.

2.5.3. Extracción de características de la señal de voz

2.5.3.1. Características fonológicas

Estas características son extraídas usando dos herramientas computacionales, una encontrada en la literatura y otra que fue desarrollada en el grupo GITA:

PhonVoc(*Phonectic and Phonological Vocoding*):

PhonVoc es un modelo entrenado con grabaciones de sonido en idioma inglés[1], que utiliza 15 redes neuronales profundas (DNN: *Deep Neural Network*) para extraer las características fonológicas a partir de características acústicas de corta duración [39]. El vector de los parámetros fonológicos $\mathbf{z}_n = \left[z_n^1, ..., z_n^k, ..., z_n^K\right]^T$, consiste en K probabilidades posteriores de características fonológicas. Los posteriores fonológicos se calculan mediante un banco de DNN paralelos, cada uno estimando el posterior z_n^k como probabilidades de que ocurra la k-ésima característica fonológica (versus no ocurrencia). Las estimaciones a posteriori $p(c_k \mid x_n)$ también son $0 \leqslant p(c_k \mid x_n) \leqslant 1$, $\forall k$, pero $max \sum_{k=1}^{K} p(c_k \mid x_n) = K$. Sólo unas pocas clases están activas durante una señal de corta duración $\sum_{k=1}^{K} p(c_k \mid x_n) << K$, lo que da como resultado un vector disperso $\mathbf{z}_n$.

El uso de las probabilidades fonológicas posteriores puede considerarse como un esquema paralelo a través de k diferentes canales independientes. Los 15 posteriores fonológicos o características fonológicas inferidas por la DNN se detallan en la Tabla 9.

Tabla 9. Lista de clases fonológicas extraídas con PhonVoc.

Posterior Fonológico	Definición
Vocalic	Las cuerdas vocales vibran y no hay constricción en el tracto vocal.
Consonantal	Indica que hay una obstrucción del tracto vocal.
High	El cuerpo de la lengua está por encima de su posición neutral.
Back	La lengua se retrae desde su posición neutral.
Low	Una posición más baja que la neutral del cuerpo de la lengua.
Anterior	Se refiere a una obstrucción delante de la región palato-alveolar de la boca.
Coronal	La cuchilla de la lengua se eleva desde su posición neutral.
Round	Se refiere a los labios entrecerrados.
Rising	Diferencia los diptongos de los monoptongos.
Tense	Indica vocales estresadas.
Voice	Las cuerdas vocales están vibrando.
Continuant	Diferencia las plosivas de las no plosivas.
Nasal	Indica un velum descendido donde el aire escapa por la nariz.
Strident	Se refiere a sonidos con más energía en componentes de alta frecuencia.
Silence	Indica que no hay voz en esa trama.

[1] https://github.com/idiap/phonvoc

Phonet:

Phonet es una herramienta desarrollada en GITA, que de manera similar a PhonVoc calcula las probabilidades a posteriori de las clases fonológicas, pero a través de 18 redes neuronales bidireccionales recurrentes paralelas y usando una base de datos en idioma Español para su entrenamiento [2]. Las redes neuronales son entrenadas con una función de pérdida para equilibrar las clases en el proceso de entrenamiento [56]. Los factores de peso w_i para cada clase $i = \{1, ..., C\}$, se definen en base al porcentaje de muestras del conjunto de entrenamiento que pertenecen a cada clase. Las características extraídas se muestran en la Tabla 10.

Tabla 10. Lista de clases fonológicas extraídas con Phonet.

Posterior Fonológico	Fonemas
Vocalic	/a/, /e/, /i/, /o/, /u/
Consonantal	/b/, /ʧ/, /d/, /f/, /g/, /x/, /k/, /l/,/ʎ/, /m/, /n/, /p/, /r/, /ɾ/, /s/, /t/
Back	/a/, /o/, /u/
Anterior	/e/, /i/
Open	/a/, /e/, /o/
Close	/i/, /u/
Nasal	/m/, /n/
Stop	/p/, /b/, /t/, /k/, /g/, /ʧ/, /d/
Continuant	/f/, /b/, /ʧ/, /d/, /s/, /g/, /ʎ/, /x/
Lateral	/l/
Flap	/ɾ/
Trill	/r/
Voice	/a/, /e/, /i/, /o/, /u/, /b/, /d/, /l/,/m/, /n/, /r/, /g/, /ʎ/
Strident	/f/, /s/, /ʧ
Labial	/m/, /p/, /b/, /f/
Dental	/t/, /d/
Velar	/k/, /g/, /x/
Pause	/sil/

2.5.3.2. Características fonéticas (acústicas y articulatorias)

Estas características constituyen los métodos clásicos de representación de señales de voz. Típicamente los sistemas de reconocimiento automático del habla son alimentados con estas características. A continuación se presentan los dos grupos más conocidos en la literatura:

Frecuencias Formantes: Las Frecuencias Formantes son bandas de frecuencia donde se concentra la mayor parte de la energía sonora de un sonido y que permiten identificar los sonidos del habla humana [57]. Para hallarlos, el tracto vocal es modelado como un sistema de todos los polos,

[2] https://jcvasquezc.github.io/software/phonet/

donde las frecuencias formantes corresponde a los polos del sistema como se muestra en la ecuación 1:

$$H(z) = \frac{G}{1 - \sum_{K=1}^{p} a_k z^{-k}} \tag{1}$$

El formante de frecuencia más baja se denomina F1; el segundo, F2; el tercero, F3, etc. Normalmente sólo se necesitan los dos primeros para caracterizar una vocal, sobre todo en las lenguas con menos de seis vocales. En la producción de las vocales, las vibraciones de aire desembocan en tres puntos cardinales: garganta [a], paladar [i] y labios [u] [58]. La representación del punto y modo de articulación de las vocales da lugar a lo que se emplea en Fonética como el triángulo vocal basado en la cavidad de resonancia y hace uso de dos dimensiones: un eje vertical correspondiente al grado de abertura vocal (F1) que clasifica las vocales en altas, medias y baja; y un eje horizontal, correspondiente con la posición de la lengua (F2) que clasifica las vocales dentro de los rasgos anterior, posterior y central. Cuanta mayor abertura tenga una vocal (cuanta más baja está la lengua), mayor es la frecuencia en que aparece el F1 y cuanto más anterior es una vocal (cuanto más hacia adelante está posicionada la lengua), mayor es el F2. En la Figura 5, se muestra un ejemplo de triángulo vocal, elaborado con el programa Praat [3].

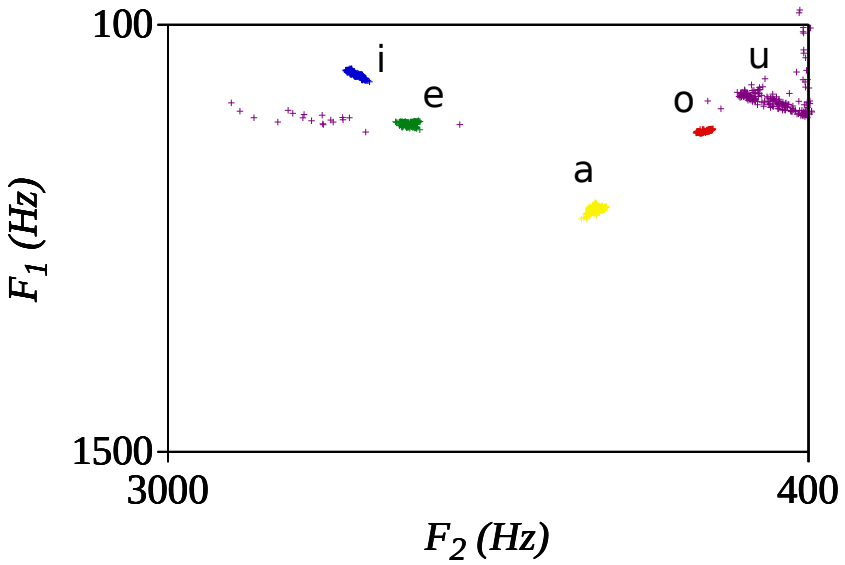

Figura 5. Representación de formantes con valores vocálicos del Español, de un hablante sano de género masculino.

[3] https://www.fon.hum.uva.nl/praat/

Coeficientes Cepstrales en las frecuencias de Mel (MFCC): Los MFCC es la técnica más usada para el reconocimiento automático de voz ya que modela la manera en que se produce el discurso, teniendo en cuenta todos los órganos que intervienen en la producción del habla como tracto vocal y los órganos articuladores. Se basa en las frecuencias que percibe el oído humano tratando de imitar cómo llega el sonido a la cóclea, así la escala de Mel, emplea un banco de 26 filtros que se asemeje a las bandas en las que el oído percibe las frecuencias [57]. Los coeficientes se pueden hallar en los siguientes pasos, asumiendo que $x(n)$ es la señal de voz de entrada:

1. Calcular el espectro de la energía:

$$\bar{x}(n) = \sum_{n=0}^{N_w-1} x(n)W(n)e^{-j2\pi nk/N_w} \quad ; \quad 0 \leqslant k \leqslant N_w \tag{2}$$

   Donde N_w corresponde al tamaño en ms de la ventana de Hanning dada por:

$$W(n) = \beta_w \left(0{,}5 - 0{,}5\cos\left(\frac{2\pi n}{N_w - 1}\right)\right) \quad ; \quad 0 \leqslant k \leqslant N_w \tag{3}$$

   β_w es un factor de normalización definido de modo que el valor cuadrático medio de la ventana sea la unidad. El espectro de la energía está dado por:

$$X_k = |\bar{x}(k)|^2 \quad ; \quad 0 \leqslant k \leqslant K \tag{4}$$

   Donde K se toma igual a $N_w/2$ porque solo se considera la mitad del espectro.

2. Calcular la energía en cada segmento

$$E_j = \sum_{k=0}^{K-1} (\Phi)_j(k)X_k \quad ; \quad 0 \leqslant j < J \tag{5}$$

   Donde J, generalmente igual a 26, es el número de filtros triangulares $(/Phi)_j$ utilizados, con la siguiente restricción:

$$\sum_{k=0}^{K-1} (\Phi)_j = 1 \quad ; \quad \forall j \tag{6}$$

3. Calcular el MFCC:

$$c_m = \beta_c \sum_{j=0}^{J-1} \cos\left(m\frac{\pi}{J}\left(j+0{,}5\right)\right) \log_{10}\left(E_j\right) \tag{7}$$

Esta última ecuación también puede mostrarse como un producto escalar entre el vector de energía espectral logarítmica y un vector de factores de ponderación V_m dados por

$$V_m = \left\{\cos\left(m\frac{\pi}{J}\left(j+0{,}5\right)\right) \mid \quad 0 \leqslant j < J\right\} \tag{8}$$

lo que lleva a la siguiente ecuación:

$$c_m = \beta_c \sum_{j=0}^{J-1} V_m \log_{10}\left(E_j\right) \tag{9}$$

El factor de amplificación,β_c, que acomoda el rango dinámico de los coeficientes cm, depende de valor del factor de normalización β_w .En general, solo se conservan los primeros 15 valores de c_m.

La fórmula de conversión para convertir de Hertz a mels viene dada por la Ecuación 10:

$$mel = 2595 \log_{10}\left(1+\frac{f}{700}\right) = 1127 \ln\left(1+\frac{f}{700}\right) \tag{10}$$

2.5.4. Decisión

2.5.4.1. Algoritmos de clasificación

Máquina de soporte vectorial (*Support vector Machine*, SVM) El método de clasificación-regresión Máquinas de Vector Soporte es un algoritmo de aprendizaje supervisado que discrimina puntos de muestras usando un hiperplano de separación a través de unos vectores de soporte o puntos más cercanos al hiperplano, los cuales definen y maximizan el margen entre las clases [54]. En el proceso de encontrar el hiperplano óptimo, una SVM de margen suave permite cierta flexibilidad, es decir, que algunos puntos se ubiquen en el lado incorrecto del hiperplano, tal que el clasificador tenga mayor capacidad predictiva al aplicarlo a nuevas observaciones. La función de decisión de una SVM de margen suave está dada por la Ecuación 11:

$$y_n \cdot (\mathbf{w}^T \phi(\mathbf{x}_n) + b) \geq 1 - \xi_n, \;\; n = 1, 2, 3, \cdots, N \tag{11}$$

Donde ξ_n es una variable de holgura que penaliza la cantidad de errores permitidos en el proceso de optimización. $y_n \in \{-1, +1\}$ son las etiquetas de clase, $\phi(\mathbf{x}_n)$ es una función kernel para transformar el espacio de características original $\mathbf{x}$ en un espacio de dimensión mayor para encontrar una solución lineal del problema. El hiperplano de separación está definido por el vector de pesos $\mathbf{w}$ y el valor de sesgo b.

Para hallar el hiperplano que clasifique adecuadamente la mayoría de los puntos, se resuelve un problema de optimización convexa y se define en la Ecuación 12, donde el hiperparámetro C controla la compensación entre ξ_n y el ancho del margen. Las muestras $\mathbf{x}_n$ que satisfacen la condición de igualdad en la Ecuación 11 se llaman vectores de soporte ($\mathbf{x}_m$).

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \tfrac{1}{2} ||\mathbf{w}||^2 + C \textstyle\sum_{n=1}^{N} \xi_n \\ \text{subject to} \quad & y \cdot (\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n \\ & \xi_n \geq 0 \end{aligned} \tag{12}$$

En el presente trabajo se considera la función de kernel Gaussiano $\phi(\mathbf{x}_n) = \mathrm{e}^{-\gamma^2 ||\mathbf{x}_n - \mathbf{x}_m||^2}$, donde el hiperparámetro γ controla el ancho de banda del kernel, cuyo valor óptimo puede encontrarse mediante validación cruzada.

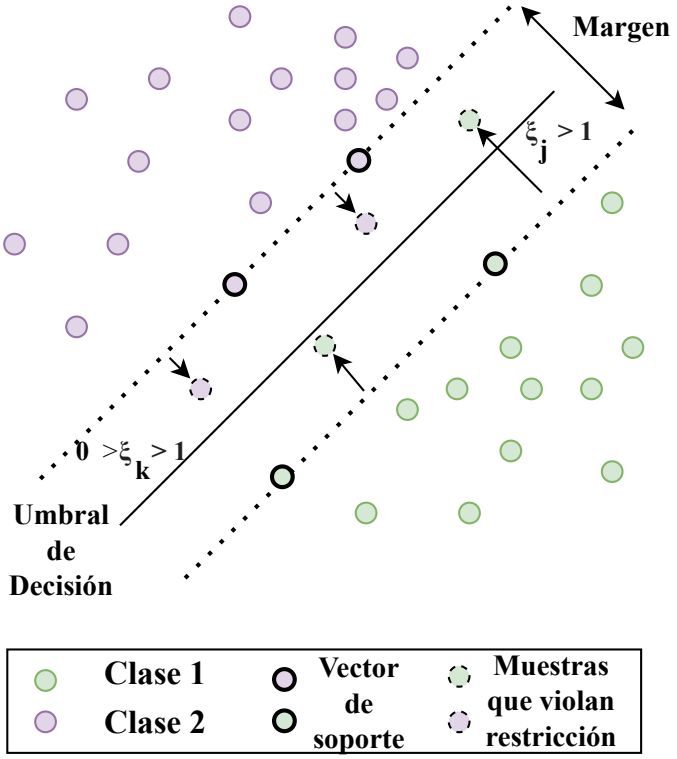

Figura 6. Ejemplo de un clasificador SVM de margen suave

Bosques Aleatorios(*Random Forest*, RF) Random Forest es una técnica de aprendizaje automático muy utilizada para regresión y clasificación. Este algoritmo se fundamenta en cultivar muchos árboles donde cada uno, clasifica un objeto con determinadas características emitiendo un voto para que finalmente el algoritmo tome la decisión con la clasificación que obtenga la mayoría de votos. Es definido en [59] y consiste entonces en una familia de clasificadores $h(X \mid \Theta_1)$, ... , $h(X \mid \Theta_k)$ basados en un árbol de clasificación con parámetros Θ_k elegidos aleatoriamente de un modelo de vector aleatorio Θ. Para la clasificación final f(x) (que combina los clasificadores $\{h_k(x)\}$), cada árbol vota por la clase más popular en la entrada x y la clase con la mayoría de votos gana.

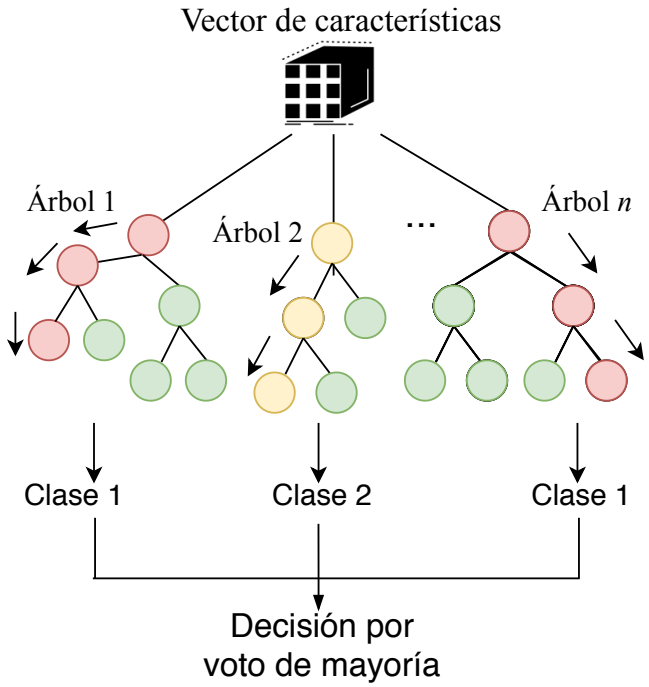

Figura 7. Ejemplo esquema de un clasificador RF.

2.5.4.2. Visualización de datos y reducción de dimensionalidad

Las técnicas de visualización de datos permiten a través de un gráfico, observar la distribución de los datos en el espacio con el fin de analizar su comportamiento y posteriormente tomar decisiones. Cuando el conjunto de datos utilizados contiene alta dimensionalidad, y se requiere visualizar la información de estos, es necesario recurrir a técnicas para reducir su dimensión. En la literatura se encontró que las más utilizadas para reducción de dimensión y selección de características son análisis de componentes principales (ACP) y análisis lineal discriminante. Para esta metodología ambas técnicas se probaron sin embargo, se decidió usar la última ya que mostraba una mejor separabilidad de las clases.

Análisis Lineal Discriminante (*Linear Discriminant Analysis*, LDA)

El análisis discriminante lineal, es una técnica de aprendizaje supervisado utilizado en estadística y en inteligencia artificial que consiste, según el criterio de Fisher [60] y con el fin de resolver problemas de reconocimiento de patrones, en encontrar una combinación lineal de características que ayude

a discriminar entre dos o más clases, esto es, hallando una nueva variable llamada discriminante, que minimiza la varianza de cada clase y maximiza la separación de las medias de las clases garantizando la máxima dispersión entre estas. Por otro lado, la combinación resultante puede ser utilizada para la reducción de dimensiones antes de la posterior clasificación.

El objetivo de LDA es proyectar el conjunto de datos original en un espacio de menor dimensión mediante un vector w que garantice la máxima separación de las clases. Se tiene $\boldsymbol{X} = \{x_1, x_2, ...x_N\}$ tuplas d-dimensionales, donde x_i es la i-ésima muestra, que se etiquetan en c clases de esta forma $\boldsymbol{X} = \{w_1, w_2, ...w_c\}$, donde cada clase tiene n_i muestras, N es el número total de muestras y se calcula de la siguiente manera:

$$N = \sum_{i=1}^{c} n_i \tag{13}$$

El discriminante lineal de Fisher viene dado por el vector $\boldsymbol{W}$ necesario para encontrar $y_i = \boldsymbol{W}^T \boldsymbol{x}_i$, estas serán las proyecciones en baja dimensión de cada una de las tuplas. Para garantizar la mayor distancia entre las clases y la menor dispersión intra clases, se maximiza la función objetivo $J(\boldsymbol{W})$:

$$\text{Max.} J(\boldsymbol{W}) = \frac{\boldsymbol{W}^T \boldsymbol{S}_B \boldsymbol{W}}{\boldsymbol{W}^T \boldsymbol{S}_W \boldsymbol{W}} \tag{14}$$

Donde $\boldsymbol{S}_B$ es la matriz que garantiza la distancia de separación entre las clases $(m_i - m)$ la cual se calcula de la siguiente manera:

$$(\boldsymbol{m}_i - \boldsymbol{m})^2 = \left(\boldsymbol{W}^T \boldsymbol{\mu}_i - \boldsymbol{W}^T \boldsymbol{\mu}\right)^2 = \boldsymbol{W}^T (\boldsymbol{\mu}_i - \boldsymbol{\mu}) (\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \boldsymbol{W} \tag{15}$$

Donde $\boldsymbol{m}_i$ es la proyección de la media de la i-ésima clase $(\boldsymbol{m}_i = \boldsymbol{W}^T \boldsymbol{\mu}_i)$, $\boldsymbol{m}$ es la proyección de la media total de todas las clases $(\boldsymbol{m} = \boldsymbol{W}^T \boldsymbol{\mu})$, $\boldsymbol{W}$ es la matriz de transformación de LDA, $\boldsymbol{\mu}_j(1 \text{ x } d)$ es la media de la i-ésima clase $\left(\boldsymbol{\mu}_j = \frac{1}{n_j} \sum_{\boldsymbol{x}_i \in w_j} \boldsymbol{x}_i\right)$, $\boldsymbol{\mu}(1 \text{ x } d)$ es la media total de todas las clases $\left(\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i = \frac{1}{c} \sum_{j=1}^{c} \boldsymbol{\mu}_j\right)$ y c es el número total de clases.

El término $(\boldsymbol{\mu}_i - \boldsymbol{\mu}) (\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$ representa la distancia de separación entre la media de la i-ésima clase $(\boldsymbol{\mu}_i)$ y la media total $(\boldsymbol{\mu})$. Así la variación total entre las clases se calcula de la siguiente manera:

$$\boldsymbol{S}_B = \sum_{i=1}^{c} n_i (\boldsymbol{\mu}_i - \boldsymbol{\mu}) (\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \tag{16}$$

$\boldsymbol{S}_W$ es la matriz que posee la dispersión en cada una de las clases a proyectar y busca minimizar la varianza dentro de la clase, es decir, la diferencia

entre la media proyectada ($\boldsymbol{m}_j$) y las muestras proyectadas de cada clase $\boldsymbol{W}^T\boldsymbol{x}_i$.

$$\sum_{x_i \in w_j} \left(\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{m}_j\right)^2 = \sum_{x_i \in w_j} \left(\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{\mu}_j\right)^2 \tag{17}$$

También se puede expresar de la siguiente manera:

$$\sum_{x_i \in w_j} \boldsymbol{W}^T \left(\boldsymbol{x}_i - \boldsymbol{\mu}_j\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_j\right)^T \boldsymbol{W} = \sum_{x_i \in w_j} \boldsymbol{W}^T \boldsymbol{S}_{W_j} \boldsymbol{W} \tag{18}$$

$\boldsymbol{S}_{W_j}$ se puede calcular:

$$\boldsymbol{S}_{W_j} = \boldsymbol{l}_j^T * \boldsymbol{l}_j = \sum_{i=1}^{n_j} \left(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_j\right)\left(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_j\right)^T \tag{19}$$

Donde $\boldsymbol{x}_{ij}$ es la i-ésima muestra en la j-ésima clase, $\boldsymbol{l}_j$ es el centro de los datos de la j-ésima clase ($\boldsymbol{l}_j = \boldsymbol{w}_j - \boldsymbol{\mu}_j = \{\boldsymbol{x}_i\}_{i=1}^{n_j} - \boldsymbol{\mu}_j$). La variación total dentro de las clases se puede calcular con la Ecuación 20:

$$\boldsymbol{S}_W = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_j\right)\left(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_j\right)^T \tag{20}$$

Luego de calcular la varianza entre clases ($\boldsymbol{S}_B$) y la varianza dentro de una clase ($\boldsymbol{S}_W$), la matriz de transformación ($\boldsymbol{W}$) puede ser calculada como un problema de optimización con la ecuación 21:

$$\boldsymbol{S}_B\boldsymbol{W} = \lambda \boldsymbol{S}_W\boldsymbol{W} \tag{21}$$

Donde λ es el valor propio de la matriz $\boldsymbol{W}$. Si $\boldsymbol{S}_W$ es una matriz no singular, se puede desarrollar con las operaciones de valores propios para la matriz $\boldsymbol{S}_W^{-1}\boldsymbol{S}_B$:

$$\boldsymbol{S}_W^{-1}\boldsymbol{S}_B\boldsymbol{W} = \lambda\boldsymbol{W} \tag{22}$$

Se sustituye el resultado en $J(\boldsymbol{W})$:

$$J(\boldsymbol{W}) = \frac{\boldsymbol{W}^T\boldsymbol{S}_B\boldsymbol{W}}{\boldsymbol{W}^T\boldsymbol{S}_W\boldsymbol{W}} = \lambda_k \frac{\boldsymbol{W}^T\boldsymbol{S}_B\boldsymbol{W}_k}{\boldsymbol{W}^T\boldsymbol{S}_W\boldsymbol{W}_k} = \lambda_k \quad con \quad k = 1...d \tag{23}$$

Del cual se tiene un $\boldsymbol{V}_k$ vector propio k de valor propio λ_k, es decir, se debe tener un vector propio asociado a un valor propio para maximizar la función objetivo. Cada vector propio representa un eje del espacio LDA y el valor propio asociado representa la información que contiene el vector propio

y que refleja la capacidad para discriminar entre las diferentes clases.

Para implementar LDA, se siguen los siguientes pasos:

1. Calcular los vectores medios d - dimensionales para las diferentes clases del conjunto de datos.
2. Calcular las matrices de covarianza (entre clases y dentro de las clases).
3. Calcular los vectores propios (*eigenvectores*) y sus correspondientes valores propios (*eigenvalores*) para las matrices de covarianza.
4. Ordenar los vectores propios de forma decreciente en relación a los valores propios y seleccionar los k vectores propios con los valores propios más grandes para formar una matriz $W \in \mathbb{D} \times \mathbb{K}$ (donde cada columna representa un vector propio).
5. Utilizar la matriz $W \in \mathbb{D} \times \mathbb{K}$ de vectores propios para transformar las muestras en un nuevo subespacio. Para esto se recurre a la expresión matemática $Y = X \cdot W$.

Cada muestra de la matriz de características tiene una etiqueta que indica la clase a la que pertenece, y a su vez, las etiquetas son utilizadas para hallar la matriz de covarianza de cada clase para encontrar las componentes con máxima variación en los datos.

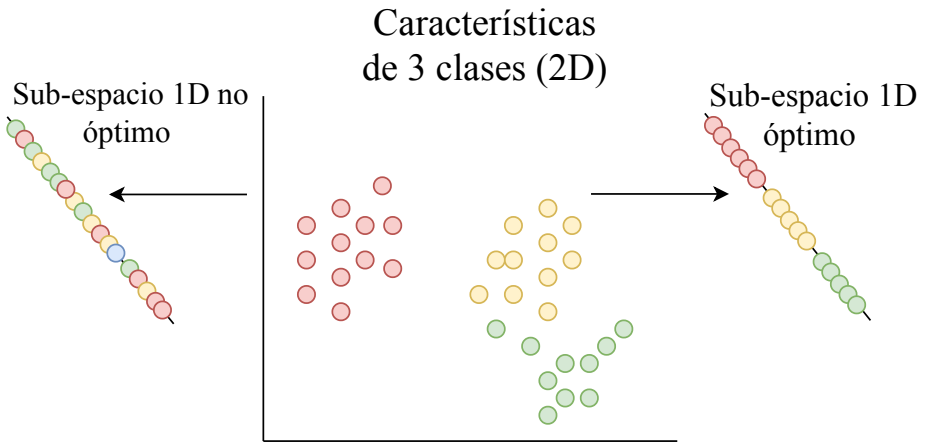

Figura 8. Análisis Lineal Discriminante.

Modelo de Mezclas Gaussianas (*Gaussian Mixture Model*, GMM)
Un Modelo de mezclas Gaussianas, es un modelo probabilístico que busca representar una población a partir de una combinación lineal de un número finito de distribuciones de probabilidad Gaussianas, donde cada distribución busca modelar una sub-población, para que en conjunto la mezcla de Gaussianas modele toda población en general [61].

Un variable escalar aleatoria y continua, x, tiene una distribución normal o Gaussiana si su función de densidad de probabilidad (FDP) es:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \mathcal{N}(x;\mu,\sigma) \tag{24}$$

Donde μ y σ son la media y la desviación estándar de la variable aleatoria x respectivamente. Esta definición se puede extender a la distribución normal de múltiples variables. Si $\boldsymbol{x} = (x_1, x_2, ..., x_D)^\intercal$ es un vector aleatorio normal, también llamado variable aleatoria Gaussiana multivariada o vectorial, este cumple una distribución Gaussiana si:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{25}$$

Donde $\boldsymbol{\mu}$ es el vector de medias y $\boldsymbol{\Sigma}$ es la matriz de covarianzas de $\boldsymbol{x}$.

Las distribuciones Gaussianas son comunmente usadas en el campo de la ingeniería, no solo por sus características computacionales altamente deseables, sino también, por su capacidad de describir de una manera adecuada los datos obtenidos de fenómenos naturales del mundo real, gracias a la ley de los grandes números.

Una distribución general y más completa a la anterior, es la distribución de mezclas Gaussianas. Un vector aleatorio sigue un distribución de mezclas Gaussianas si su FDP puede ser descrita así:

$$\begin{aligned} p(\boldsymbol{x}) &= \sum_{m=1}^{M} \frac{c_m}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma_m}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_m})^T \boldsymbol{\Sigma_m}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_m})\right] \\ &= \sum_{m=1}^{M} c_m N(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \end{aligned} \tag{26}$$

Donde c_m es el peso asociado a la m-ésima componente Gaussiana y cumple $\sum_{m=1}^{M} c_m = 1$. Para la construcción de los mapas articulatorios, se hace uso de los GMM para modelar cada población (pacientes y controles), donde cada modelo contiene un número de gaussianas igual a las clases fonológicas,

esto es, una gaussiana por cada clase fonológica. Se elige el tipo de covarianza diagonal y aunque es un método de agrupamiento o de aprendizaje no supervisado, se utilizan las etiquetas de clase para que el algoritmo inicialice cada gaussiana desde su media y así dibujar las elipses que permiten visualizar los mapas en un espacio de dos dimensiones.

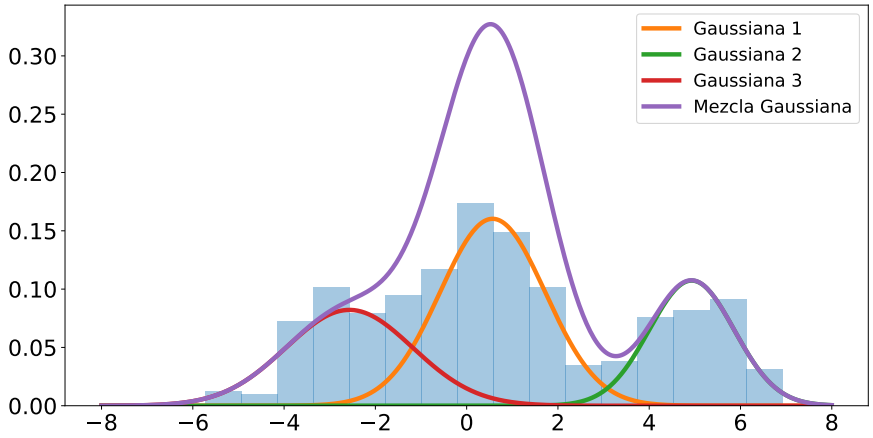

Figura 9. Modelamiento con un GMM de tres componentes.

✓ Estimación de parámetros

Como se vio anteriormente a una distribución Gaussiana de una variable aleatoria multivariada, se le asocia una media $\boldsymbol{\mu}$ y una matriz de covarianzas $\boldsymbol{\Sigma}$. Las cuales construyen el conjunto de parámetros de la distribución. En una distribución de mezclas Gaussianas, adicionalmente, se tiene el parámetro C_m, por ende el conjunto de parámetros es $\boldsymbol{\Theta} = \{c_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}$, para cada componente Gaussiana m, perteneciente a la mezcla.

Los parámetros $\boldsymbol{\mu}_m$ y $\boldsymbol{\Sigma}_m$ definen que sub-población es modelada por la distribución, el parámetro c_m define la contribución de la distribución en el modelo general. En principio estos parámetros son desconocidos y generalmente se estiman mediante el uso del algoritmo de maximización de esperanza EM (*Estimation Maximization*) [62]. Al problema de estimar estos parámetros se le conoce como aprendizaje.

Algoritmo de maximización de esperanza

El algoritmo EM es una de las técnicas más usada para estimar los parámetros de una mezcla cuando se tiene un numero fijo de Gaussianas. Este itera, alternando entre una etapa de esperanza o etapa E y una etapa de maximización o etapa M [54].

Para la primera iteración el conjunto de parámetros $\boldsymbol{\Theta}$ puede ser inicializado de forma aleatoria, o mediante un método de agrupamiento como el K-medias. Esto debido a que en la etapa E, es necesario tener un conjunto de parámetros iniciales [54].

En la etapa E se busca encontrar $h_m^{(j)}(t)$, la cual es la probabilidad de que un dato t pertenezca a una Gaussiana m en la iteracion j. Esta probabilidad se define como

$$h_m^{(j)}(t) = \frac{c_m^{(j)} \mathcal{N}(\boldsymbol{x}^{(t)}; \boldsymbol{\mu}_m^{(j)}, \boldsymbol{\Sigma}_m^{(j)})}{\sum_{i=1}^{M} c_i^{(j)} \mathcal{N}(\boldsymbol{x}^{(t)}; \boldsymbol{\mu}_i^{(j)}, \boldsymbol{\Sigma}_i^{(j)})}, \tag{27}$$

en donde, $\mathcal{N}(\boldsymbol{x}^{(t)}; \boldsymbol{\mu}_m^{(j)}, \boldsymbol{\Sigma}_m^{(j)})$ es la distribución de probabilidad de la Gaussiana m, evaluada en el dato t, en la iteracion j y $c_m^{(j)}$ es el peso de la distribución m, en la iteración j. Luego, de encontrar $h_m^{(j)}(t)$ se continua con la etapa de maximización o etapa M.

En la etapa M se busca encontrar un nuevo conjunto Θ, tal que maximice la probabilidad anterior. Para ello se usan las Ecuaciones:

$$c_m^{(j+1)} = \frac{1}{N} \sum_{t=1}^{N} h_m^{(j)}(t), \tag{28}$$

$$\boldsymbol{\mu}_m^{(j+1)} = \frac{\sum_{t=1}^{N} h_m^{(j)}(t) \boldsymbol{x}^{(t)}}{\sum_{t=1}^{N} h_m^{(j)}(t)}, \tag{29}$$

$$\boldsymbol{\Sigma}_m^{(j+1)} = \frac{\sum_{t=1}^{N} h_m^{(j)}(t) [\boldsymbol{x}^{(t)} - \boldsymbol{\mu}_m^{(t)}][\boldsymbol{x}^{(t)} - \boldsymbol{\mu}_m^{(t)}]^\intercal}{\sum_{t=1}^{N} h_m^{(j)}(t)}. \tag{30}$$

La Ecuación 28 determinará el peso de la distribución m para la siguiente iteración. Como puede ver el peso dependerá de la cantidad de datos que tengan una alta probabilidad de pertenecer a la $m-$*ésima* distribución. Aunque cada dato aporta para aumentar el peso, dicho aumento será proporcional a la probabilidad de que el dato sea modelado por la distribución m. Las Ecuaciones 29 y 30 determinan el $\boldsymbol{\mu}_m$ y $\boldsymbol{\Sigma}_m$ de la próxima iteración. En cada iteración se busca que cada Gaussiana modele de mejor forma los datos en donde $h_m^{(j)}(t)$ presenta un valor más alto.

Finalmente el algoritmo itera entre la etapa E y la etapa M hasta que la diferencia entre $h_m^{(j+1)}(t) - h_m^{(j)}(t)$ de cada distribución, no supere cierto umbral de convergencia.

2.5.5. Medidas de Rendimiento

2.5.5.1. Matriz de confusión En un sistema de reconocimiento de patrones, la matriz de confusión es una herramienta que permite conocer y evaluar el desempeño de un algoritmo de clasificación [63]. Cada fila de la matriz representa las instancias en una clase predicha, mientras que cada columna representa las instancias en una clase real (o viceversa); esto permite ver si el sistema está confundiendo las diferentes clases o resultados en la clasificación y en qué medida. La Tabla 11 muestra una matriz de confusión para sistemas de clasificación biclase.

Tabla 11. Matriz de confusión.

	Clase verdadera	
Clase estimada	Clase 0	Clase 1
Clase 0	TP	FP
Clase 1	FN	TN

- ✓ **Verdadero Positivo** (**TP**: *True Positive*): Es o porcentaje de instancias de clase 0 que el el sistema clasifica correctamente como pertenecientes a la clase 0.
- ✓ **Falso Negativo** (**FN**: *False Negative*): Es el número o porcentaje de instancias de clase 0 que el el sistema clasifica incorrectamente como pertenecientes a la clase 1.
- ✓ **Falso Positivo** (**FP**: *False Positive*): Es el número o porcentaje de instancias de clase 1 que el el sistema clasifica incorrectamente como pertenecientes a la clase 0.
- ✓ **Verdadero Negativo** (**TN**: *False Positive*): Es el número o porcentaje de instancias de clase 1 que el el sistema clasifica correctamente como pertenecientes a la clase 1.

En base a estos términos de la matriz de confusión, se definen una serie de métricas que permiten medir el rendimiento del sistema:

✓ **Acierto** (**Acc**: *Accuracy*): Es la cantidad de predicciones positivas que fueron clasificados correctamente por el sistema.

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \tag{31}$$

✓ **Sensibilidad** (***Recall o Sensitivity***): Se trata de los casos positivos que el algoritmo ha clasificado correctamente. Cuando los valores se presentan en porcentaje, la sensibilidad coincide con la tasa de verdaderos positivos (**TPR**: *True Positive Rate*).

$$Sensitivity = \frac{TP}{TP + FN} \tag{32}$$

✓ **Especificidad** (***Specificity***): Se trata de los casos negativos que el algoritmo ha clasificado correctamente. Cuando los valores se presentan en porcentaje, la especificidad coincide con la tasa de verdaderos negativos (**TNR**: *True Negative Rate*).

$$Specificity = \frac{TN}{TN + FP} \tag{33}$$

2.5.5.2. Curva ROC (*Receiver Operating Characteristic*) La curva ROC es una representación gráfica de la tasa de los verdaderos positivos (TPR: *True Positive Rate*) y la tasa de los falsos positivos (FPR: *False Positive Rate*) o (1-Especificidad), que permite evaluar la capacidad discriminante de un clasificador. Un espacio ROC se define por TPR y FPR como ejes x e y respectivamente como se observa en la Figura 10, donde la TPR mide la capacidad del clasificador de detectar los casos positivos correctamente, de entre todos los casos positivos disponibles en la prueba a medida que se desplaza el umbral (línea vertical de la gráfica de la izquierda) y la FPR define cuántos resultados positivos son incorrectos de entre todos los casos negativos disponibles durante la prueba mientras se desplaza el umbral. La línea diagonal punteada en la gráfica de la derecha se le llama línea de *no-discriminación* y corresponde con un ejemplo típico de adivinación aleatoria. Los puntos por encima de la diagonal representan los buenos resultados de clasificación (mejor que el azar). Con las curvas ROC se conoce el rendimiento de un modelo de clasificación biclase a través del cálculo del área bajo la curva (**AUC**: *Area Under the ROC Curve*) que representa el grado de separabilidad de las clases. El AUC varía entre 0.5 y 1; entre más cercano sea el valor a 1, mejor será el desempeño del modelo para distinguir entre la clase

0 y la clase 1.

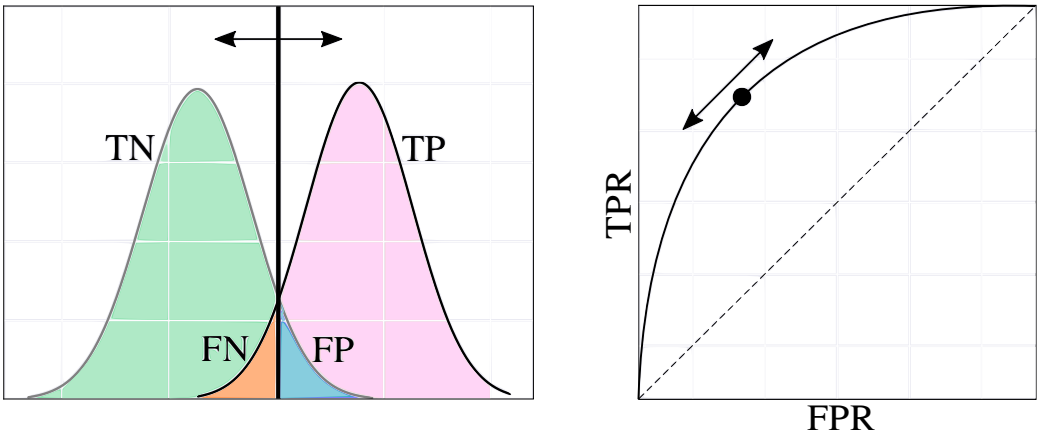

Figura 10. Construcción curva ROC.

2.5.5.3. Validación Cruzada de K-particiones La validación cruzada de K-particiones o *K-fold cross-validation*, es una técnica utilizada en inteligencia artificial para validar los resultados de los modelos generados. Los datos de muestra se dividen en K subconjuntos, uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. En este estudio se realizó una validación cruzada con K=10 para la optimización de los hiperparámetros en la implementación de los algoritmos de clasificación, además se repitió el experimento diez veces, con el fin de dar una mayor generalidad al modelo.

3. Metodología

La metodología propuesta en este trabajo consiste en la construcción de mapas articulatorios para la detección automática de personas con EP. El conjunto de características está conformado por probabilidades posteriores fonológicas estimadas con PhonVoc o con Phonet; se adicionaron 12 características MFCC y las frecuencias de los dos primeros formantes de la señal acústica (F1 y F2), extraídos mediante algoritmos en Python. Con SARHA se realizó el alineamiento forzado para extraer los tiempos donde se encontraban cada uno de los fonemas y etiquetar las diferentes clases fonológicas.

Se estimaron además medidas estadísticas (curtosis, asimetría, media y desviación estándar) con el objetivo de obtener un vector de características por cada hablante.

Un aspecto importante antes de la clasificación, es la normalización de los datos o proceso mediante el cual se ajustan los valores utilizados por diferentes escalas para que, en una nueva escala, sea posible comparar datos procedentes de diferentes muestras o poblaciones. La normalización de los datos se realizó usando la función StandardScaler de la librería scikit-learn de Python para eliminar la media y escalar la desviación estándar a uno. Posteriormente, se implementaron los algoritmos para la clasificación pacientes vs controles, SVM de margen suave con función de kernel gaussiano y Random Forest. Para la partición del conjunto de datos de entrenamiento y prueba, se usó la técnica de validación cruzada de K-particiones (K=10), además se realizaron diez iteraciones con el fin de darle generalidad al sistema.

El proceso de optimización de los modelos de clasificación se realizó con la librería de Python GridSearchCV; para SVM se variaron los hiperparámetros C (margen de costo) y γ (ancho de banda del kernel gaussiano) y para RF, se variaron los hiperparámetros max-depth (profundidad máxima del árbol) y n-estimators(número de estimadores); para el resultado de cada iteración se eligió la moda de los hiperparámetros y el promedio para las medidas de rendimiento acierto, sensibilidad, especificidad y AUC en ambos métodos de clasificación.

Visualización mapas articulatorios

Se usó LDA para mostrar la separabilidad de las clases en un espacio de dos dimensiones y GMM's de covarianza diagonal para visualizar los datos. Se construye un modelo GMM para los controles y un modelo GMM para los pacientes, donde cada modelo tiene un número de gaussianas igual al número de clases fonológicas, esto es, una gaussiana por cada clase fonológica que modela la distribución de estas clases y permite la visualización de los mapas articulatorios que describe la capacidad que tienen tanto los controles como los pacientes con EP para articular los fonemas. En la Figura 11 se muestra un ejemplo de mapa articulatorio de hablantes sanos con las clases fonológicas agrupadas.

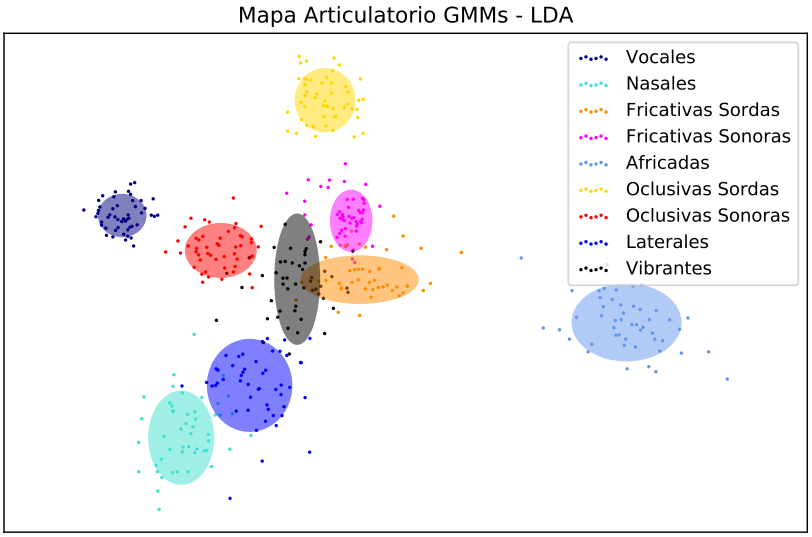

Figura 11. Ejemplo mapa articulatorio de hablantes sanos.

4. Experimentos y Resultados

Este conjunto de experimentos tenía como finalidad evaluar las herramientas PhonVoc y Phonet y comprobar su idoneidad en la extracción de características fonológicas, para medir la articulación de fonemas propios del Español, producidos por pacientes con EP y personas sanas.

Al realizar una comparación entre los dos modelos de clasificación implementados, SVM y RF, ambos mostraron un buen rendimiento, sin embargo, en términos generales, SVM mostró los mejores aciertos. Los resultados que se muestran a continuación son el promedio de las 10 iteraciones realizadas, además de usar la técnica de validación cruzada con k=10.

4.1. Experimento 1: Evaluación PhonVoc y Phonet con características fonológicas por punto y modo de articulación para clasificar entre pacientes con EP y controles.

Este primer experimento contiene la metodología del presente estudio, en la cual se extrajeron las características fonológicas teniendo en cuenta todos los posteriores fonológicos para el agrupamiento de las nueve clases por punto y modo de articulación, además de los MFCC y F1,F2.

Resultados PhonVoc
La Tabla 12 muestra los resultados de la clasificación de pacientes con EP vs controles. Se observa que, con SVM, las clases fonológicas (vocales, nasales, fricativas sordas y oclusivas sonoras) obtuvieron aciertos mayores al 80 % , donde la clase vocales alcanzó el mejor resultado con un acierto en promedio de 90.4 %. También se puede observar que la clase africadas no obtuvo un buen desempeño en la clasificación con un acierto en promedio de 67.9 %. Con RF el resultado mayor al 80 % fue para la clase vocales con un acierto de 86.6 % y el desempeño más bajo también lo obtuvo la clase africadas con un acierto de 69.1 %. Los resultados muestran que las clases fonológicas con fonemas principalmente sonoros, a excepción de las fricativas sordas, son las que reflejan un mayor déficit para articular los fonemas en la forma como sale el aire, pero principalmente en el punto de articulación dental y alveolar. Así mismo, el resultado de la clase africadas muestra que se produce un cierre completo para el paso del flujo del aire y luego se libera de forma rápida con el fonema siguiente estrechado el canal que produce la turbulencia, lo que sugiere un déficit articulatorio menor para la producción de estos fonemas, sin embargo, debe tenerse en cuenta el punto de articulación y la posición de la lengua del fonema inmediatamente anterior, lo que pudo

facilitar la producción de éste. Estos resultados también se pueden observar en la Figura 12, donde la curva ROC describe la capacidad discriminativa de los clasificadores para cada una de las clases a través de los valores de AUC. Estos resultados sugieren que las clases Vocales y Oclusivas sonoras permiten una clasificación entre enfermos y sanos con una mayor medida.

Tabla 12. Resultados experimento 1. Clasificación pacientes vs controles, usando PhonVoc. **Ac**: Acierto, **AUC**: Area Under the ROC Curve, **Sen**:Sensibilidad, **Esp**:Especificidad. ($\mu \pm \sigma$) μ: media σ: desviación estándar.

Clases fonológicas	**Clasificador**	**Ac (%)**	**AUC**	**Sen (%)**	**Esp (%)**
Vocales	**SVM**	**90.4 ± 1.11**	**0.96**	**91.8 ± 1.66**	**89.0 ± 1.84**
	RF	86.6 ± 1.62	0.95	87.6 ± 1.49	85.6 ± 2.65
Nasales	SVM	82.1 ± 1.22	0.88	82.6 ± 3.10	81.6 ± 2.50
	RF	76.2 ± 1.98	0.75	77.8 ± 2.44	74.6 ± 3.90
Fricativas sordas	SVM	80.9 ± 1.45	0.89	80.2 ± 3.16	81.6 ± 2.50
	RF	78.6 ± 2.80	0.87	79.0 ± 3.92	78.2 ± 3.84
Fricativas sonoras	SVM	75.3 ± 1.49	0.83	74.6 ± 2.97	76.0 ± 2.53
	RF	74.5 ± 2.53	0.82	75.6 ± 1.95	73.4 ± 4.00
Africadas	SVM	67.9 ± 2.55	0.76	78.8 ± 5.88	57.0 ± 3.61
	RF	69.1 ± 3.20	0.78	70.6 ± 5.14	67.6 ± 3.55
Oclusivas sordas	SVM	69.4 ± 2.37	0.77	58.0 ± 3.10	80.8 ± 5.31
	RF	72.2 ± 2.95	0.81	71.0 ± 4.31	73.4 ± 2.37
Oclusivas sonoras	SVM	80.8 ± 5.31	0.90	80.8 ± 5.31	84.4 ± 3.20
	RF	78.7 ± 2.32	0.87	76.2 ± 4.23	81.2 ± 2.99
Laterales	SVM	76.4 ± 2.54	0.85	73.6 ± 3.20	73.6 ± 3.20
	RF	77.5 ± 2.87	0.89	79.6 ± 3.77	75.4 ± 4.29
Vibrantes	SVM	75.9 ± 1.76	0.81	74.0 ± 2.83	77.8 ± 3.16
	RF	71.3 ± 4.40	0.80	69.8 ± 6.35	72.8 ± 5.45

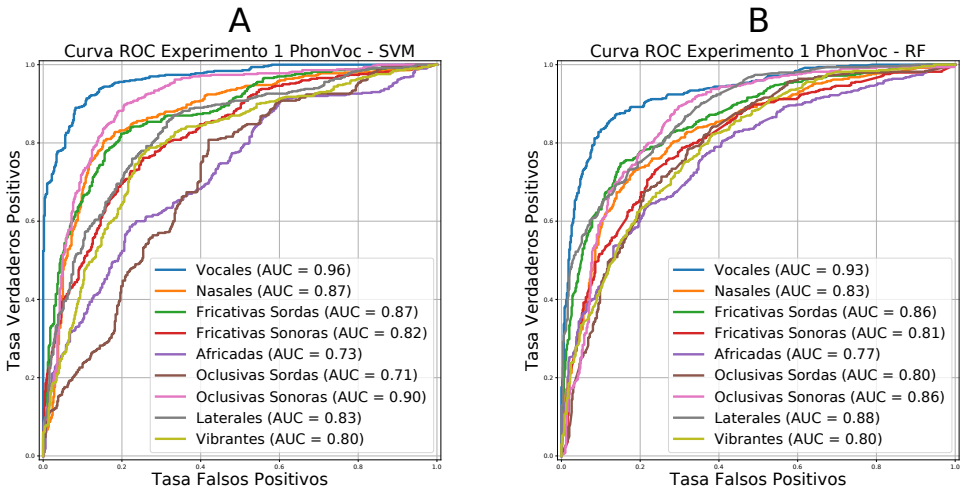

Figura 12. Curvas ROC Experimento 1 PhonVoc. **A** SVM **B** RF

Las burbujas visualizadas en la Figura 13 muestran la manera en que se agrupan los fonemas según el punto y modo de articulación. Se presentan dos burbujas por cada color indicando que una es para sanos y otra es para enfermos, los puntos rojos y los puntos azules señalan sanos y enfermos respectivamente. Con la construcción de estos mapas articulatorios, se evidencia que no todos los tipos de articulación son buenos para clasificar entre sanos y enfermos por el solape que se presenta entre las burbujas de las dos poblaciones. Sin embargo, en las oclusivas sonoras representadas por los fonemas /b d g / y las oclusivas sordas /p t k/ se aprecia un menor solape de las burbujas lo que puede estar relacionado con la dificultad de los pacientes para iniciar la sonoridad de los fonemas oclusivos, como ya se ha analizado en estudios anteriores [64] [65].

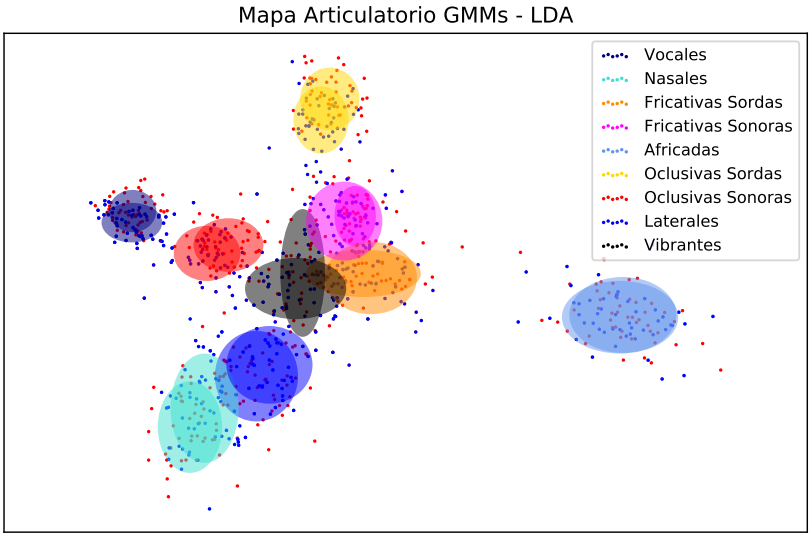

Figura 13. Mapa articulatorio pacientes y sanos. Experimento 1 - PhonVoc

Resultados Phonet

Usando Phonet para la extracción de características fonológicas, se describen los resultados de la clasificación pacientes vs controles con SVM y RF en la Tabla 13. Con SVM, las clases fonológicas (vocales, oclusivas sonoras y fricativas sordas) en ese orden, logran aciertos superiores al 80 %; el mejor resultado fue para las vocales con un acierto de 89.8 %, mientras que las clases (africadas y laterales) arrojaron resultados con aciertos de 60.6 % y 67.7 % respectivamente. Con RF, la clase vocales obtuvo un acierto en promedio de 84.5 % y las clases con el resultado más bajo siguen siendo las clases africadas con acierto de 66.9 % y laterales con acierto de 68.3 %. Llama la atención el resultado de la clase laterales representada por el fonema /l/ ya que su punto de articulación es alveolar y de la manera que se presenta en el texto, la antecede y/o precede un fonema vocálico (/a/,/e/,/i/), lo cual según se muestra en los resultados, son los fonemas con una mayor dificultad para su producción en los pacientes con EP; sin embargo, PhonVoc mostró resultados con diez puntos porcentuales por encima para esta misma clase, siendo más coherente con los resultados. Las medidas de AUC en la Figura 14 muestran que Vocales tiene una mayor capacidad discriminativa que las demás clases.

Tabla 13. Resultados experimento 1. Clasificación pacientes vs controles, usando Phonet. **Ac**: Acierto, **AUC**: Area Under the ROC Curve, **Sen**:Sensibilidad, **Esp**:Especificidad. ($\mu \pm \sigma$) μ: media σ: desviación estándar.

Clases fonológicas	Clasificador	Ac (%)	AUC	Sen (%)	Esp (%)
Vocales	**SVM**	**89.8 ± 1.47**	**0.97**	**91.0 ± 2.24**	**88.6 ± 2.01**
	RF	84.5 ± 1.74	0.93	84.2 ± 2.74	84.8 ± 1.59
Nasales	SVM	75.3 ± 2.15	0.86	73.2 ± 2.23	77.4 ± 3.35
	RF	72.2 ± 1.72	0.79	67.4 ± 3.8	77.0 ± 3.25
Fricativas sordas	SVM	80.6 ± 2.54	0.90	81.2 ± 2.71	80.0 ± 4.00
	RF	80.1 ± 3.17	0.89	80.2 ± 3.94	80.0 ± 4.73
Fricativas sonoras	SVM	71.5 ± 1.20	0.80	74.4 ± 2.15	68.6 ± 2.54
	RF	70.4 ± 2.93	0.81	73.0 ± 6.27	67.8 ± 4.14
Africadas	SVM	60.6 ± 1.11	0.72	78.8 ± 3.60	42.4 ± 2.65
	RF	66.9 ± 3.53	0.73	70.6 ± 6.63	63.2 ± 4.91
Oclusivas sordas	SVM	74.3 ± 1.62	0.84	68.8 ± 2.23	79.8 ± 3.52
	RF	72.4 ± 1.90	0.81	67.8 ± 3.02	77.0 ± 3.37
Oclusivas sonoras	SVM	81.4 ± 2.84	0.89	75.8 ± 5.17	76.8 ± 3.37
	RF	79.1 ± 3.44	0.89	78.8 ± 2.40	79.4 ± 5.44
Laterales	SVM	67.7 ± 2.72	0.74	51.2 ± 2.56	84.2 ± 4.69
	RF	68.3 ± 3.63	0.78	63.8 ± 3.40	72.8 ± 4.66
Vibrantes	SVM	73.4 ± 1.56	0.82	69.0 ± 1.34	77.8 ± 3.03
	RF	70.6 ± 3.26	0.76	71.0 ± 3.37	70.2 ± 4.51

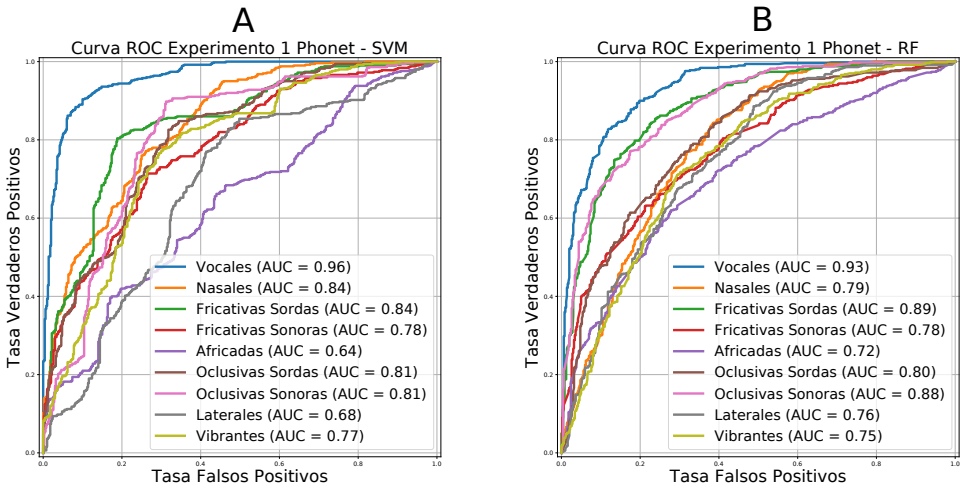

Figura 14. Curvas ROC Experimento 1 Phonet. **A** SVM **B** RF

En la Figura 15 se observa un solape entre las burbujas de pacientes y sanos, así como un solape entre clases, sin embargo, el análisis del mapa arti-

culatorio es muy similar al realizado con PhonVoc para los fonemas oclusivos y africados.

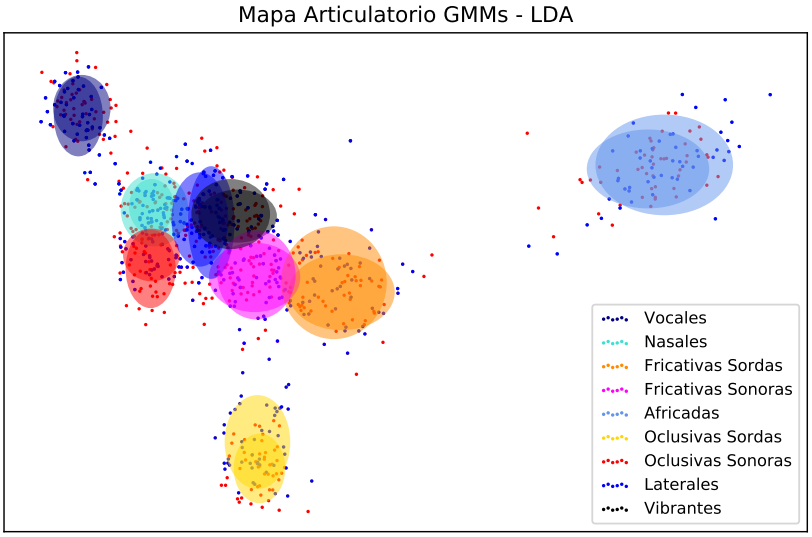

Figura 15. Mapa articulatorio pacientes y sanos. Experimento 1 - Phonet

4.2. Experimento 2: Evaluación PhonVoc y Phonet con características fonológicas por punto de articulación para clasificar entre pacientes con EP y controles.

Para este segundo experimento, se evalúan los posteriores fonológicos pero teniendo en cuenta sólo aquellos categorizados por punto de articulación, según el criterio de cada una de las herramientas PhonVoc y Phonet. Se adicionan también MFCC y F1,F2.

Se aclara que la clase *Anterior* es categorizada de manera diferente; con PhonVoc se agrupan en esta clase los fonemas /b d g f l m n p t s/, mientras que Phonet agrupa los fonemas /e i/.

Resultados PhonVoc

En la Tabla 14 se observan que SVM proporciona resultados entre el 69.1 % (clase *Anterior*) y el 81.8 % (Clase *Coronal*). Con RF se observa un rendimiento muy similar a SVM; para la clase *Anterior* muestra un acierto del 70.2 % y para la clase *Coronal* un acierto del 81.2 %. Estos resultados de la clase *Coronal* sugieren la dificultad de los pacientes con EP, para levantar la

parte frontal de la lengua desde su posición neutral y articular los fonemas con punto de articulación dental, alveolar y postalveolar. Para la clase *Anterior* y *Round* se evidencia un déficit menor al articular los fonemas que pertenecen a estas clases. Estos resultados también son observados en la curva ROC de la Figura 16, donde se aprecia una afectación en mayor medida de la sensibilidad en estas últimas clases donde decrece la curva.

Tabla 14. Resultados experimento 2. Clasificación pacientes vs controles, usando PhonVoc. **Ac**: Acierto, **AUC**: Area Under the ROC Curve, **Sen**:Sensibilidad, **Esp**:Especificidad. ($\mu \pm \sigma$) μ: media σ: desviación estándar.

Clases fonológicas	Clasificador	Ac (%)	AUC	Sen (%)	Esp (%)
Back	SVM	77.4 ± 2.01	0.89	77.0 ± 3.38	77.8 ± 3.52
	RF	80.0 ± 2.28	0.86	77.8 ± 3.02	82.2 ± 3.84
Anterior	SVM	69.1 ± 2.74	0.77	72.6 ± 2.97	65.6 ± 4.36
	RF	70.2 ± 3.05	0.78	70.4 ± 3.97	70.0 ± 4.09
Coronal	**SVM**	**81.8 ± 1.40**	**0.92**	**83.0 ± 2.72**	**80.6 ± 2.37**
	RF	81.2 ± 1.83	0.91	82.4 ± 2.49	80.0 ± 3.46
Round	SVM	69.9 ± 1.81	0.78	71.2 ± 2.04	68.6 ± 2.54
	RF	70.3 ± 1.79	0.78	71.4 ± 3.23	69.2 ± 4.39

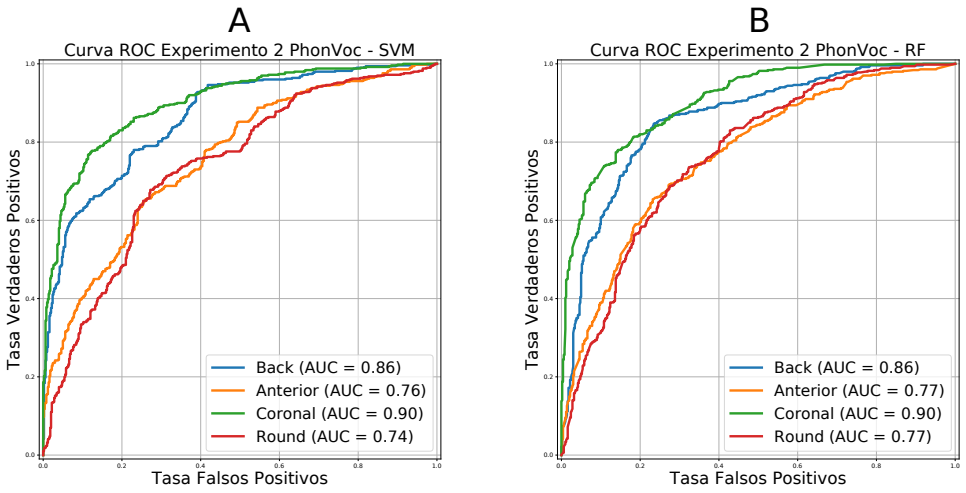

Figura 16. Curvas ROC Experimento 2 PhonVoc. **A** SVM **B** RF

EL mapa articulatorio de la Figura 17 muestra una separación completa entre clases, pero se sigue evidenciando un solape entre pacientes y controles. Los fonemas con punto de articulación *Anterior* y *Round* indican una produc-

ción imprecisa de estos fonemas, contrario a lo observado con los algoritmos de clasificación.

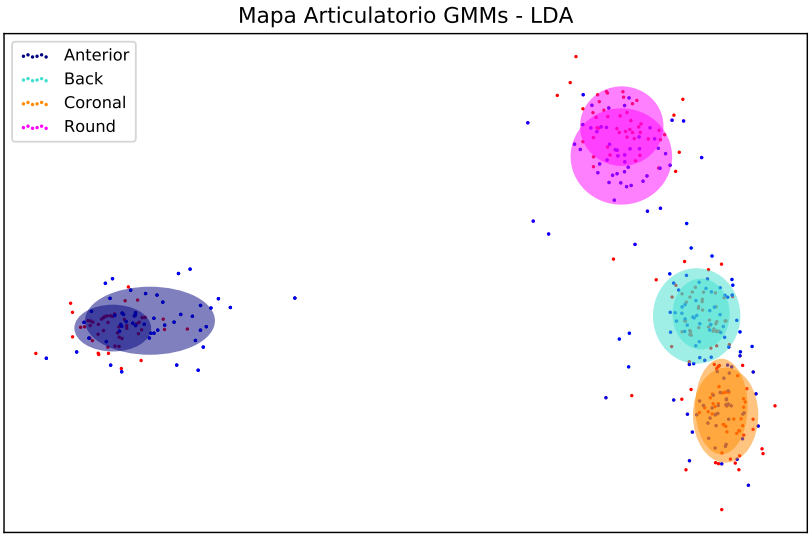

Figura 17. Mapa articulatorio pacientes y sanos. Experimento 2 - PhonVoc

Resultados Phonet

Los resultados de la clasificación pacientes vs controles en la Tabla 15, muestran que con SVM se obtuvo aciertos entre el 73.4 % (clase *Velar*) y el 89.8 % (clase *Back*). RF para estas mismas dos clases muestra un acierto del 72.1 % y un acierto del 85.1 % respectivamente. Los resultados de la clase *Back* pueden estar en relación con la dificultad de los pacientes con EP para retraer la lengua desde su posición neutral y producir estos fonemas vocálicos. Así mismo, la clase *Anterior* con acierto de 81.6 % con SVM y 79.4 % con RF, sugiere una articulación imprecisa de estos fonemas sonoros y que además implican levantar o retraer la lengua desde su posición neutral. Las medidas de AUC en la Figura 18 dan prueba de la capacidad de estas dos clases para discriminar entre enfermos y sanos.

Tabla 15. Resultados experimento 2. Clasificación pacientes vs controles, usando Phonet. **Ac**: Acierto, **AUC**: Area Under the ROC Curve, **Sen**:Sensibilidad, **Esp**:Especificidad. ($\mu \pm \sigma$) μ: media σ: desviación estándar.

Clases fonológicas	Clasificador	Ac (%)	AUC	Sen (%)	Esp (%)
Back	**SVM**	**89.8 ± 2.27**	**0.98**	**89.6 ± 3.2**	**90.0 ± 2.97**
	RF	85.1 ± 1.81	0.94	86.0 ± 2.0	84.2 ± 3.15
Anterior	SVM	81.6 ± 1.91	0.91	81.6 ± 1.91	81.6 ± 1.91
	RF	79.4 ± 1.68	0.86	78.0 ± 3.68	80.8 ± 2.22
Labial	SVM	77.1 ± 2.34	0.86	75.2 ± 2.86	79.0 ± 2.41
	RF	72.9 ± 3.56	0.81	72.6 ± 5.06	73.2 ± 4.48
Dental	SVM	75.3 ± 1.90	0.87	69.4 ± 1.56	81.2 ± 3.37
	RF	74.8 ± 1.93	0.85	69.8 ± 3.94	79.8 ± 4.14
Velar	SVM	73.4 ± 2.15	0.84	69.8 ± 2.27	77.0 ± 3.49
	RF	72.1 ± 2.84	0.82	71.2 ± 4.57	73.0 ± 2.86

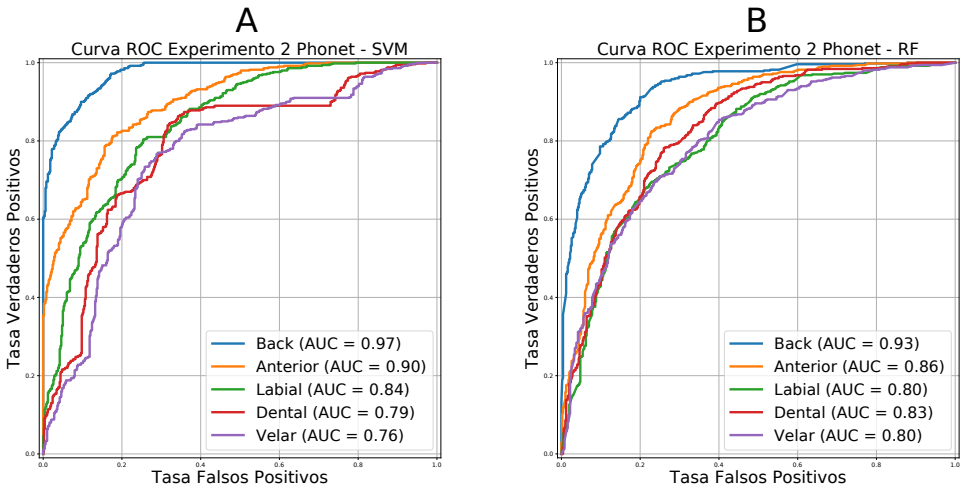

Figura 18. Curvas ROC Experimento 2 Phonet. **A** SVM **B** RF

Los fonemas con punto de articulación *Dental* y *Labial* se sobreponen en la Figura 19. Así mismo, los fonemas con punto de articulación *Velar* muestra un solape entre muestras de pacientes con EP y controles, sugiriendo una buena capacidad articulatoria al levantar la lengua hacia este punto de articulación, contrario a lo obtenido con los experimentos de clasificación, sin embargo, se debe tener en cuenta que el espacio de características con las que se entrenaron y probaron los clasificadores fue transformado con LDA a un espacio de dos dimensiones para esta etapa de visualización.

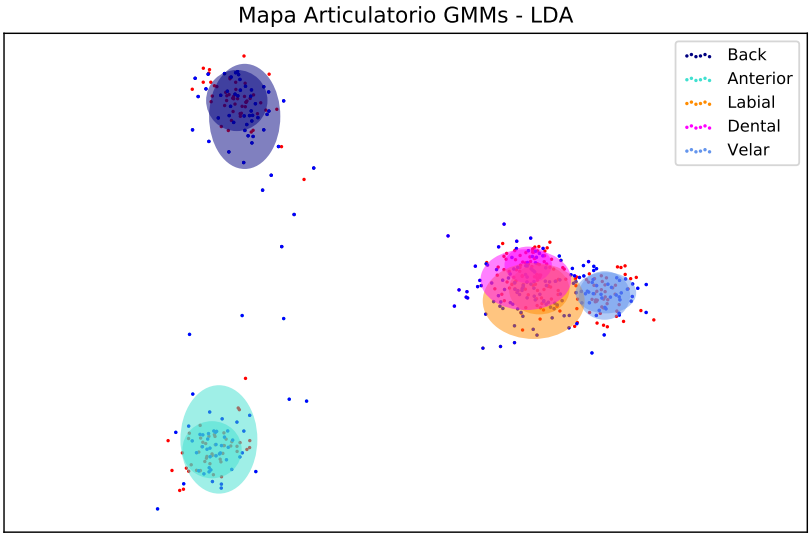

Figura 19. Mapa articulatorio pacientes y sanos. Experimento 2 - Phonet

4.3. Experimento 3: Evaluación PhonVoc y Phonet con características fonológicas por modo de articulación para clasificar entre pacientes con EP y controles.

Este tercer experimento se realiza evaluando aquellos posteriores fonológicos estimados por modo de articulación según lo determinado por cada una de las herramientas, además de los MFCC y F1,F2 que se adicionaron en el conjunto de características.

Resultados PhonVoc

Los resultados en la clasificación pacientes vs controles en la Tabla 16, muestran con SVM que los fonemas sonoros principalmente vocálicos, obtienen los mejores resultados (*High*, *Low*, *Vocalic* y *Voice*) con aciertos superiores al 80 %; el resultado más bajo fue para la clase *Nasal* con un acierto del 68.3 %. Con RF se observan resultados entre el 68.1 % (clase *Strident*) y 80.3 % (Clase *Low*). Así mismo, estos resultados se contrastan con la curva ROC de la Figura 20 con un buen rendimiento a nivel del parámetro AUC. Llama la atención en este experimento, el bajo rendimiento en la clase*Nasal*, revisando el Experimento 1, esta misma clase de fonemas nasales obtiene un acierto del 82.1 %. Esto podría sugerir, que si se tienen en cuenta todos los posteriores fonológicos, serían más los rasgos distintivos que permiten la discriminación

entre enfermos y sanos, debido tal vez a los efectos coarticulatorios con la producción de los fonemas inmediatamente anterior y posterior a los fonemas nasales. Sin embargo, sería necesario realizar más experimentos y contrastar los resultados con la base teórica para dar claridad a este resultado.

Tabla 16. Resultados experimento 3. Clasificación pacientes vs controles, usando PhonVoc. **Ac**: Acierto, **AUC**: Area Under the ROC Curve, **Sen**:Sensibilidad, **Esp**:Especificidad. ($\mu \pm \sigma$) μ: media σ: desviación estándar.

Clases fonológicas	Clasificador	Ac (%)	AUC	Sen (%)	Esp (%)
High	**SVM**	**82.1 ± 2.12**	**0.86**	**83.8 ± 3.16**	**80.4 ± 3.32**
	RF	76.8 ± 2.52	0.84	78.0 ± 4.28	75.6 ± 2.65
Low	SVM	81.5 ± 2.11	0.91	83.4 ± 3.35	79.6 ± 3.77
	RF	80.3 ± 2.49	0.90	82.0 ± 2.68	78.6 ± 3.69
Nasal	SVM	68.3 ± 3.58	0.76	67.4 ± 4.57	69.2 ± 4.02
	RF	69.0 ± 2.86	0.76	70.0 ± 3.57	68.0 ± 4.0
Stop	SVM	70.1 ± 2.30	0.78	70.8 ± 3.37	69.4 ± 2.84
	RF	72.5 ± 3.26	0.79	73.6 ± 4.88	71.4 ± 5.06
Continuant	SVM	71.8 ± 2.14	0.84	71.4 ± 1.80	72.2 ± 4.24
	RF	73.7 ± 2.68	0.82	72.4 ± 3.32	75.0 ± 3.25
Strident	SVM	72.1 ± 2.62	0.79	76.2 ± 4.77	68.0 ± 4.0
	RF	68.1 ± 2.58	0.74	74.6 ± 3.90	61.6 ± 2.49
Vocalic	SVM	81.4 ± 1.91	0.90	81.4 ± 1.8	81.4 ± 2.84
	RF	79.0 ± 2.36	0.88	79.8 ± 2.89	78.2 ± 2.6
Voice	SVM	81.3 ± 2.37	0.92	83.6 ± 2.8	79.0 ± 3.38
	RF	79.5 ± 2.61	0.86	77.2 ± 3.48	81.8 ± 2.08

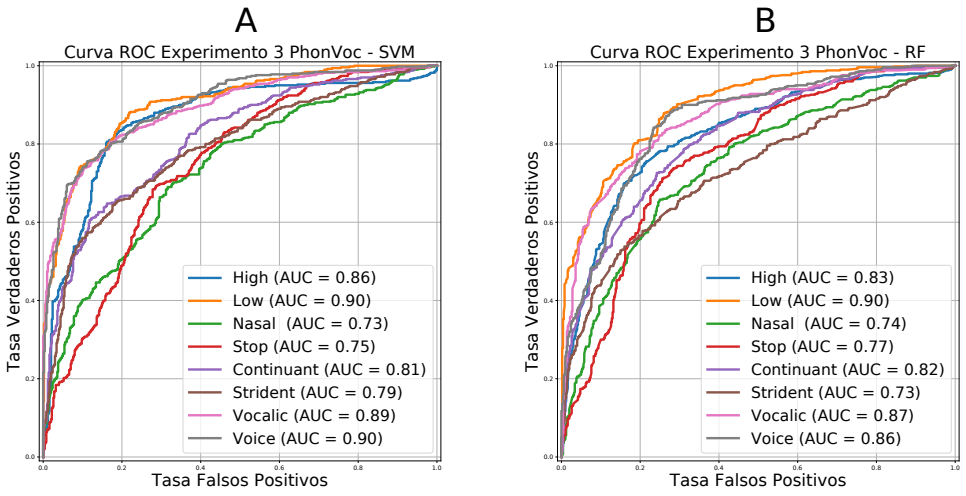

Figura 20. Curvas ROC Experimento 3 PhonVoc. **A** SVM **B** RF

De manera similar que en el experimento 1, los mapas articulatorios muestran una separación entre clases a excepción de *Voice* y *Continuant* como se observa en la Figura 21. Nuevamente, se evidencia que los fonemas oclusivos (*Stop*) muestran rasgos distintivos más notorios por la dificultad que presentan los pacientes con EP para producirlos.

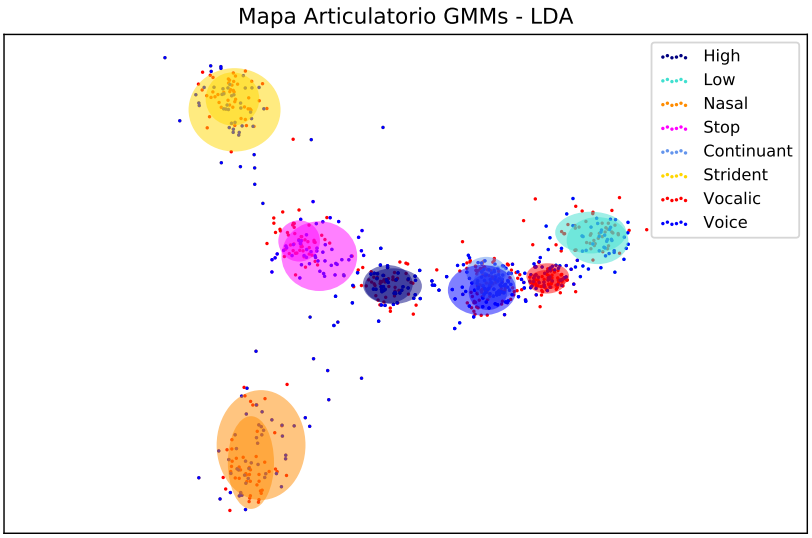

Figura 21. Mapa articulatorio pacientes y sanos. Experimento 3 - PhonVoc

Resultados Phonet

Usando Phonet se muestran los resultados en la Tabla 17. Se observa que, de manera similar que con PhonVoc, las clases con fonemas sonoros especialmente vocálicos arrojaron resultados de acierto mayores al 80 % (*Vocalic*, *Voice* y *Open*); el mejor resultado lo obtuvo la clase *Vocalic* con un acierto del 90.0 % mientras que las clases *Trill* y *Lateral*, mostraron resultados muy bajos con aciertos del 58.7 % y 67.7 % respectivamente. La clase *Trill* contiene el fonema vibrante doble / ɾ/ y sugiere una articulación más adecuada que las demás clases, al hacer vibrar la lengua dos o más veces al unirse con el paladar. También se debe tener en cuenta la posición de la lengua al pronunciar el fonema inmediatamente anterior. La clase *Lateral* obtuvo un acierto del 67.7 % igual al obtenido en el experimento 1 con esta misma herramienta, por lo que se hace el mismo análisis presentado allí. En la Figura 22 se ob-

serva que, a nivel de AUC las clases con mejor resultado, sugieren una mejor capacidad para discriminar entre enfermos y sanos. Se evidencia cómo decae la curva para la clase *Lateral*, su valor de AUC se encuentra por debajo de un intervalo medio para una adecuada discriminación.

Tabla 17. Resultados experimento 3. Clasificación pacientes vs controles, usando Phonet. **Ac**: Acierto, **AUC**: Area Under the ROC Curve, **Sen**:Sensibilidad, **Esp**:Especificidad. ($\mu \pm \sigma$) μ: media σ: desviación estándar.

Clases fonológicas	Clasificador	Ac (%)	AUC	Sen (%)	Esp (%)
Nasal	SVM	75.9 ± 2.62	0.87	73.8 ± 3.94	78.0 ± 4.73
	RF	73.3 ± 2.00	0.79	66.6 ± 2.53	73.5 ± 2.67
Stop	SVM	77.3 ± 1.73	0.89	72.6 ± 2.01	82.0 ± 2.97
	RF	76.1 ± 3.69	0.78	73.4 ± 5.06	78.8 ± 3.37
Continuant	SVM	74.1 ± 1.58	0.87	78.2 ± 3.40	70.0 ± 1.26
	RF	74.4 ± 2.37	0.84	73.2 ± 2.56	75.6 ± 4.36
Strident	SVM	74.7 ± 2.97	0.84	73.0 ± 4.31	76.4 ± 4.88
	RF	74.9 ± 1.81	0.84	75.0 ± 4.58	74.8 ± 4.11
Lateral	SVM	67.7 ± 2.19	0.76	52.2 ± 2.44	83.2 ± 4.92
	RF	67.9 ± 3.98	0.77	65.2 ± 3.48	70.6 ± 5.51
Flap	SVM	75.0 ± 1.41	0.81	72.6 ± 2.84	77.4 ± 1.80
	RF	70.3 ± 2.41	0.77	70.0 ± 3.22	70.6 ± 3.10
Trill	SVM	58.7 ± 3.10	0.63	55.8 ± 3.28	61.6 ± 4.88
	RF	59.7 ± 2.49	0.63	64.4 ± 4.36	55.0 ± 4.12
Open	SVM	86.7 ± 1.79	0.94	86.6 ± 2.97	86.8 ± 2.99
	RF	85.7 ± 2.28	0.93	85.8 ± 3.28	85.6 ± 2.80
Close	SVM	77.4 ± 2.20	0.87	76.6 ± 3.58	78.2 ± 2.44
	RF	79.7 ± 2.57	0.88	80.0 ± 2.19	79.4 ± 4.38
Vocalic	**SVM**	**90.0 ± 1.61**	**0.96**	**92.0 ± 2.37**	**88.0 ± 1.55**
	RF	85.5 ± 2.01	0.94	84.2 ± 2.74	86.8 ± 3.48
Voice	SVM	86.2 ± 2.14	0.95	84.8 ± 3.49	87.6 ± 2.94
	RF	85.9 ± 1.44	0.93	83.6 ± 2.65	88.2 ± 2.27

Las burbujas del mapa articulatorio de la Figura 23 no muestra resultados muy concluyentes, las clases se muestran muy cercanas entre sí y solapadas. La clase *Stop* parece mostrar posibilidad de separación entre pacientes y sanos, sin embargo, no se puede determinar con certeza ya que la clase *Continuant* se encuentra superpuesta.

En todos los mapas articulatorios evaluados, se observaron las burbujas con fonemas principalmente vocálicos muy cercanas entre sí, como era de esperarse, sin embargo, el modelo GMM logró agruparlas por separado. También las burbujas de los pacientes se mostraron más grandes en todos las figuras revelando una mayor dispersión de sus muestras.

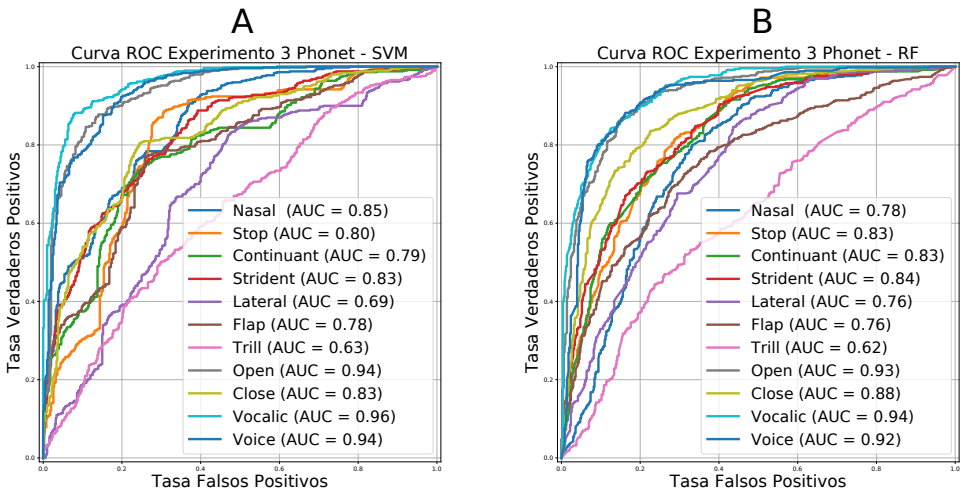

Figura 22. Curvas ROC Experimento 3 Phonet. **A** SVM **B** RF

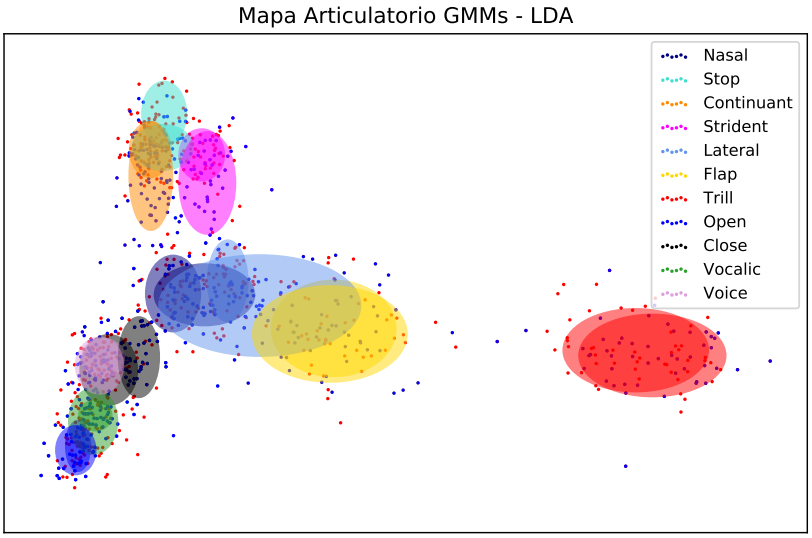

Figura 23. Mapa articulatorio pacientes y sanos. Experimento 3 - Phonet

5. Conclusiones

Este trabajo evaluó las herramientas PhonVoc y Phonet para comprobar su idoneidad en la estimación de características fonológicas para la clasificación entre pacientes con EP y personas sanas. Se implementaron los algoritmos de clasificación SVM y RF y aunque ambos clasificadores obtuvieron un rendimiento similar, SVM mostró los mejores resultados en términos de acierto y medidas de AUC; estas y otras métricas registradas, son el promedio de una validación cruzada con k=10 además de iterar diez veces los experimentos.

En el experimento 1 evaluando punto y modo de articulación, PhonVoc mostró el mejor resultado para la clase Vocales con acierto de 90.4 % y AUC de 0.96, además las clases Nasales, Fricativas sordas y Oclusivas sonoras registraron aciertos superiores al 80 % y AUC por encima de 0.87; con Phonet se obtuvo resultados muy similares para estas clases a excepción de la clase Nasales. En el experimento 2, la clase *Back* y usando Phonet mostró un acierto de 89.8 % y AUC de 0.98. Finalmente, en el experimento 3 y de manera similar al anterior, se eligieron únicamente las clases con posterior fonológico según el modo de articulación. Los resultados con Phonet, indicaron aciertos superiores al 86 % y AUC superior a 0.93, para las clases *Vocalic*, *Voice* y *Open*; el acierto para *Vocalic* fue de 90 % y AUC de 0.95. Con PhonVoc se registraron aciertos mayores al 81 % y AUC mayores a 0.85 para las clases *High*, *Low*, *Vocalic* y *Voice*.

El conjunto de experimentos realizados, indica de manera general, que las clases fonológicas conformadas por fonemas sonoros, principalmente aquellos que incluyen los fonemas vocálicos, evidencian una articulación imprecisa de los fonemas, con puntos de articulación dental, alveolar, anterior y posterior que implican levantar o retraer la lengua desde su posición neutral.

Los diferentes resultados mostraron que, con la metodología propuesta a través un análisis articulatorio y fonológico haciendo uso de Phonet y PhonVoc para agrupar clases fonológicas, es posible la detección automática de la EP con ciertas clases. Además, con los mapas articulatorios considerando punto y modo de articulación, es posible visualizar la separabilidad de las diferentes clases fonológicas e interpretar la capacidad principalmente de los pacientes, de articular los fonemas del Español. El análisis fonológico, con agrupamiento de fonemas, representa un apoyo para los neurólogos en el diagnóstico de la EP y para los especialistas de la voz patológica, en determinar una posible terapia del habla, de acuerdo con los modos y puntos de articulación de los fonemas que evidencian una mayor dificultad para su producción.

Los mapas articulatorios implementados con GMM por punto y modo de

articulación, mostraron la capacidad del modelo para agrupar las diferentes clases de fonemas y visualizarlas en un espacio de dos dimensiones, sin embargo, se evidenció que no todos los tipos de articulación son buenos para clasificar entre sanos y enfermos. Los fonemas oclusivos en la mayoría de sus figuras, mostraron un menor solape entre las dos poblaciones analizadas lo que indica la dificultad que tienen los pacientes con EP en la producción de estos fonemas que han sido analizados ampliamente desde otros enfoques.

Referencias

[1] Jackson-Menaldi, M.C.A, (2002), La voz patológica, Buenos Aires-Argentina, Editorial Médica Panamericana S.A (pp 121-226).

[2] De Rijk, M. D., Launer, L. J., Berger, K., Breteler, M. M., Dartigues, J. F., Baldereschi, M., ... Hofman, A. (2000). Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. Neurology, 54(11 Suppl 5), S21-3.

[3] Darley, F. L., Aronson, A. E., Brown, J. R. (1969). Differential diagnostic patterns of dysarthria. Journal of speech and hearing research, 12(2), 246-269.

[4] Skodda, S., Visser, W., Schlegel, U. (2011). Vowel articulation in Parkinson's disease. Journal of voice, 25(4), 467-472.

[5] Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Daqrouq, K., Skodda, S., ... Nöth, E. (2016). Automatic detection of Parkinson's disease in running speech spoken in three different languages. The Journal of the Acoustical Society of America, 139(1), 481-500.

[6] Llisterri, J. (1996). Los sonidos del habla. In C. Martín Vide (Ed.), Elementos de lingüística. (pp. 67-128). Barcelona: Octaedro.

[7] Logemann J, Fisher H, Boshes B, Blonsky E. Frequency and concurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. J Speech Hear Disord. 1978; 43 47-57.

[8] Ackermann H, Ziegler W. Articulatory deficits in Parkinsonian dysarthria. J Neurol Neurosurg Psychiatry. 1991; 54 1093-1098.

[9] Hegde, M. N., Freed, D. (2011). Assessment of communication disorders in adults. Plural Publishing.

[10] Icht, M., Ben-David, B. M. (2014). Oral-diadochokinesis rates across languages: English and Hebrew norms. Journal of Communication Disorders, 48, 27-37.

[11] Tjaden, K., Watling, E. (2003). Characteristics of diadochokinesis in multiple sclerosis and Parkinson's disease. Folia Phoniatrica et Logopaedica, 55(5), 241-259.

[12] Kumar, S., Kar, P., Singh, D., Sharma, M. (2018). Analysis of diadochokinesis in persons with Parkinson's disease. Journal of Datta Meghe Institute of Medical Sciences University, 13(3), 140.

[13] Goberman, A. M., Coelho, C. (2002). Acoustic analysis of Parkinsonian speech I: Speech characteristics and L-Dopa therapy. NeuroRehabilitation, 17(3), 237-246.

[14] Kempler, D., Van Lancker, D. (2002). Effect of speech task on intelligibility in dysarthria: A case study of Parkinson's disease. Brain and language, 80(3), 449-464.

[15] Levelt, W. J. M. (1989). Speaking: From Intention to Articulation. ABradford Book (The MIT Press, Cambridge, MA), pp. 1–545.

[16] Ladefoged, P. (1967). Three areas of experimental phonetics.

[17] Eriksson, A., Traunmüller, H. (2002). Perception of vocal effort and distance from the speaker on the basis of vowel utterances. Perception psychophysics, 64(1), 131-139.

[18] Kent, R. D., Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. Journal of communication disorders, 74, 74-97.

[19] Kent, R., Weismer, G., Kent, J. Vorperian, H., and Duffy, J. (1999).“Acoustic studies of dysarthric speech: methods, progress, and potential,” J. Commun. Disord. 32, 427–445.

[20] Delattre, P. (1948). Un triangle acoustique des voyelles orales du français. French review, 477-484.

[21] Joos, M. (1948). Acoustic phonetics. Language, 24(2), 5-136.

[22] Ziegler, W., Cramon, D. V. (1983). Vowel distortion in traumatic dysarthria: Lip rounding versus tongue advancement. Phonetica, 40(4), 312-322.

[23] Lansford, K. L., Liss, J. M. (2014). Vowel acoustics in dysarthria: Speech disorder diagnosis and classification. Journal of Speech, Language, and Hearing Research.

[24] Bradlow, A. R., Kraus, N., Hayes, E. (2003). Speaking clearly for children with learning disabilities. Journal of Speech, Language, and Hearing Research.

[25] Ferguson, S. H., Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. The Journal of the Acoustical Society of America, 112(1), 259-271.

[26] Goberman, A. M., Elmer, L. W. (2005). Acoustic analysis of clear versus conversational speech in individuals with Parkinson disease. Journal of Communication Disorders, 38(3), 215-230.

[27] Johnson, K., Flemming, E., Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. Language, 505-528.

[28] Tjaden, K., Lam, J., Wilding, G. (2013). Vowel acoustics in Parkinson's disease and multiple sclerosis: Comparison of clear, loud, and slow speaking conditions. Journal of Speech, Language, and Hearing Research.

[29] Sapir, S., Ramig, L., Spielman, J., Fox, C. (2009). Acoustic metrics of dysarthric vowel articulation: Comparison with vowel space area in Parkinson's disease and healthy aging: Th-14. Movement Disorders, 24.

[30] Sapir, S., Ramig, L., Spielman, J., Fox, C. (2010). Formant centralization ratio (FCR) as an acoustic index of dysarthric vowel articulation: Comparison with vowel space area in Parkinson disease and healthy aging. Journal of Speech, Language, and Hearing Research, 53, 114-125.

[31] Rusz, J., Cmejla, R., Tykalova, T., Ruzickova, H., Klempir, J., Majerova, V., ... Ruzicka, E. (2013). Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task. The Journal of the Acoustical Society of America, 134(3), 2171-2181.

[32] Sapir, S., Spielman, J. L., Ramig, L. O., Story, B. H., Fox, C. (2007). Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: acoustic and perceptual findings. Journal of Speech, Language, and Hearing Research.

[33] Ramig, L. O., Pawlas, A. A., Countryman, S. (1995). The Lee Silverman Voice Treatment: A practical guide for treating the voice and speech disorders in Parkinson disease. National Center for Voice and Speech.

[34] Novotný, M., Rusz, J., Čmejla, R., Růžička, E. (2014). Automatic evaluation of articulatory disorders in Parkinson's disease. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(9), 1366-1378.

[35] Vásquez-Correa, J. C., Orozco-Arroyave, J. R., Nöth, E. (2017). Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease. In INTERSPEECH (pp. 314-318).

[36] Vásquez-Correa, J. C., Arias-Vergara, T., Orozco-Arroyave, J. R., Vargas-Bonilla, J. F., Arias-Londoño, J. D., Nöth, E. (2015). Automatic detection of Parkinson's disease from continuous speech recorded in non-controlled noise conditions. In Sixteenth Annual Conference of the International Speech Communication Association.

[37] Bartholomew, W. T. (1934). A physical definition of «good voice-quality» in the male voice. The Journal of the Acoustical Society of America, 6, 25–33.

[38] Cernak, M., Orozco-Arroyave, J. R., Rudzicz, F., Christensen, H., Vásquez-Correa, J. C., Nöth, E. (2017). Characterisation of voice quality of Parkinson's disease using differential phonological posterior features. Computer Speech Language, 46, 196-208.

[39] Cernak, M., Garner, P. N. (2016). PhonVoc: A phonetic and phonological vocoding toolkit. In Interspeech (No. CONF 2016).

[40] Garcia, N., Orozco-Arroyave, J. R., Luis Fernando, D. H., Dehak, N., Nöth, E. (2017, August). Evaluation of the Neurological State of People with Parkinson's Disease Using i-Vectors. In Interspeech (pp. 299-303).

[41] Canter, G. J. (1965). Speech characteristics of patients with Parkinson's disease: III. Articulation, diadochokinesis, and over-all speech adequacy. Journal of Speech and Hearing Disorders, 30(3), 217-224.

[42] Logemann, J. A., Fisher, H. B. (1981). Vocal tract control in Parkinson's disease. Journal of Speech and Hearing Disorders, 46(4), 348-352.

[43] Weismer, G. (1984). Articulatory characteristics of Parkinsonian dysarthria: Segmental and phrase-level timing, spirantization, and glottal-supraglottal coordination. The dysarthrias: Physiology, acoustics, perception, management, 101-130.

[44] Moro-Velazquez, L., Gomez-Garcia, J. A., Godino-Llorente, J. I., Grandas-Perez, F., Shattuck-Hufnagel, S., Yagüe-Jimenez, V., Dehak, N. (2019). Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's Disease. Scientific Reports, 9(1), 1-16.

[45] Moreno Bilbao, M. A., Poig, D., Bonafonte Cávez, A., Lleida, E., Llisterri, J., Mariño Acebal, J. B., Nadeu Camprubí, C. (1993). Albayzin speech database: Design of the phonetic corpus. In EUROSPEECH 1993: 3rd European Conference on Speech Communication and Technology: Berlin, Germany: September 22-25, 1993 (pp. 175-178). . EUROSPEECH.

[46] Moro-Velazquez, L., Gomez-Garcia, J. A., Godino-Llorente, J. I., Villalba, J., Rusz, J., Shattuck-Hufnagel, S., Dehak, N. (2019). A forced gaussians based methodology for the differential evaluation of Parkinson's Disease by means of speech processing. Biomedical Signal Processing and Control, 48, 205-220.

[47] Orozco-Arroyave, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Gonzalez-Rátiva, M. C., Nöth, E. (2014, May). New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. In LREC (pp. 342-347).

[48] Ackermann, H., Hertrich, I., Hehr, T. (1995). Oral diadochokinesis in neurological dysarthrias. Folia phoniatrica et logopaedica, 47(1), 15-23.

[49] Vásquez-Correa, J. C., Klumpp, P., Orozco-Arroyave, J. R., Nöth, E. (2019). Phonet: A Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech. Proc. Interspeech 2019, 549-553.

[50] Ladefoged, P. (1971). Preliminaries to linguistic phonetics. University of Chicago Press.

[51] Hieronymus, J. L. (1993). ASCII phonetic symbols for the world's languages: Worldbet. Journal of the International Phonetic Association, 23, 72.

[52] Gorman, K., Howell, J., Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. Canadian Acoustics, 39(3), 192-193.

[53] Yu, D., Deng, L. (2016). AUTOMATIC SPEECH RECOGNITION. Springer london limited.

[54] Bishop, C. M. (2006). Pattern recognition and machine learning.springer.

[55] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. (2003). The unified Parkinson's disease rating scale (UPDRS): status and recommendations. Movement Disorders, 18(7), 738-750.

[56] Vásquez-Correa, J. C., Klumpp, P., Orozco-Arroyave, J. R., Nöth, E. (2019). Phonet: A Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech. Proc. Interspeech 2019, 549-553.

[57] L. R. Rabiner and R. W. Schafer, Introduction to digital speech processing, vol. 1. Hanover, MA: Now Publishers Inc., 4 ed., 2007.

[58] Fant, Gunar (1960). Mouton, The Hague., ed. Acoustic theory of speech production.

[59] Brieman, L. (2001). Random forest machine learning. 45: 5-32.

[60] Fisher, R. (1936). Linear discriminant analysis. Ann. Eugenics, 7, 179.

[61] Reynolds, D. A. (2009). Gaussian Mixture Models. Encyclopedia of biometrics, 741.

[62] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

[63] Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

[64] Argüello-Vélez, P., Arias-Vergara, T., González-Rátiva, M. C., Orozco-Arroyave, J. R., Nöth, E., Schuster, M. E. (2020, September). Acoustic Characteristics of VOT in Plosive Consonants Produced by Parkinson's Patients. In International Conference on Text, Speech, and Dialogue (pp. 303-311). Springer, Cham.

[65] Nöth, E., Rudzicz, F., Christensen, H., Orozco-Arroyave, J. R., Chinaei, H. (2016, August). Remote monitoring of neurodegeneration through speech. In Final Presentation of the Third Frederick Jelinek Memorial Summer Workshop (JSALT).