Análisis de la capacidad articulatoria en la voz de Pacientes con la Enfermedad de Parkinson

Manuel Alejandro Andrade Franco

Asesor:
Tomás Arias Vergara

Trabajo de investigación presentado como requisito parcial para optar por el título de:
Ingeniero de Telecomunicaciones

Grupo de Investigación en Telecomunicaciones Aplicadas – GITA

Universidad de Antioquia
Facultad de Ingeniería, Departamento de Ingeniería Electrónica y Telecomunicaciones
Medellín, Colombia
2018

Tabla de contenidos

Agradecimientos

Gracias a mi familia por todo la compañía y por estar siempre ahí, en especial a mi madre Marina Franco por su preocupación constante por mi bienestar, a mi hermano Andrés Andrade por que a pesar de sus ocupaciones siempre ha sido un gran apoyo en mi vida y en general a todos mis demás hermanos y familiares.

Un agradecimiento al grupo de investigación GITA, en especial al Prof. MSc.-Ing. Tomás Arias Vergara por todo el acompañamiento brindado durante el desarrollo de esta tesis de pregrado, por la colaboración e inducción a los conceptos relacionados con el análisis del habla. Al Prof. Dr.-Ing. Juan Rafael Orozco Arroyave por abrirme las puertas del grupo GITA en la línea de análisis de patrones y a todos los compañeros de dicha línea que siempre estuvieron atentos a solucionar las inquietudes y aportarme conocimientos con relación a entender mejor los conceptos desarrollados en esta tesis.

Resumen

En este trabajo se desarrolló una metodología que permite detectar problemas en la voz de pacientes con la enfermedad del Parkinson. Fueron evaluados 100 individuos, 50 afectados con la enfermedad de Parkinson y 50 hablantes sanos. Para esto se consideraron grabaciones de voz con tareas diseñadas para analizar la transición de un sonido a otro (coarticulación). La idea central fue extraer diferentes características acústicas que entregaran información acerca de la dificultad de los pacientes para articular diferentes sonidos, las cuales se basaron principalmente en 17 energías extraídas de las bandas de Bark, los coeficientes cepstrales de Mel y los momentos espectrales. Se propusieron diferentes ejercicios de clasificación basados en métodos de aprendizaje de máquina, principalmente en máquinas de soporte vectorial, K - vecinos más cercanos y redes neuronales, buscando encontrar la mejor tasa de acierto para la diferenciación entre hablantes sanos y pacientes. Se logró una tasa de acierto de hasta un 76,00% con AuC de 0,81, en la diferenciación automática, para el análisis de segmentos VOT.

Capítulo 1

Introducción

1.1 Motivación

La enfermedad de Parkinson (EP), es la segunda enfermedad neurodegenerativa más frecuente en el mundo, después del Alzheimer. De acuerdo a un estudio realizado en 10 de la naciones más pobladas en el mundo, en el 2005 el número de personas con EP estaba entre 4.1 millones y 4.6 millones, y se estima que este número incrementará a 8.7-9-3 millones en 2030 (Dorsey, 2007). En el caso de Colombia, la prevalencia de la EP es de 4.4 por cada 1000 habitantes, en todo el territorio nacional (Sánchez, 2004). Para el departamento de Antioquia, se detectó una prevalencia de 30 por cada 100.000 habitantes (Pradilla, 2003). La EP es un enfermedad neurodegenerativa caracterizada por la pérdida progresiva de células dopaminérgicas en la substantia nigra en la parte media del cerebro (Hornykiewicz, 1998). El Parkinson afecta principalmente las capacidades motoras de los pacientes. Entre los síntomas motores más comunes se encuentran la rigidez, inestabilidad postural, bradicinesia y temblor. Adicionalmente, la EP afecta los músculos y articulaciones involucradas en el proceso de producción de habla (Logemann, 1978). La mayoría de los síntomas son controlados con medicación, sin embargo, no existe evidencia clara que indiquen los efectos positivos de dicho tratamiento para controlar los problemas en la voz (Skodda, 2010).

Se estima que los trastornos del habla afectan entre el 60% - 80%, de los pacientes diagnosticados con EP. Entre los síntomas más comunes se incluyen voz monótona, bajo tono de la voz, tendencia a desvanecerse al final de la fonación y marcadas pausas para respiración entre sílabas, lo que dificulta la identificación del estado emocional del paciente y sus intenciones al expresarse (Martínez, 2010). Cuando las dificultades de tipo motoras están en un nivel muy avanzado, el habla es una de las facultades que se ven más afectadas y la rehabilitación se hace mucho más complicada, al perder la capacidad de comunicarse con el paciente. Por estas razones, existe interés en desarrollar métodos computacionales automáticos para ayuda en el diagnóstico y terapia de pacientes con EP. Actualmente, los neurólogos se basan en la historia clínica, pruebas físicas y neurológicas para diagnosticar la EP. Sin embargo, la EP es evaluada una sola vez durante la cita médica, lo cual puede resultar inadecuado para tomar decisiones acerca de la medicación y/o terapia de los pacientes. Dado que la EP es una enfermedad degenerativa, sin un tratamiento adecuado, la capacidad de comunicación oral de un paciente se puede perder parcial o totalmente (Goetz, 2008).

1.2 Dificultades del habla en pacientes con la EP

1.2.1 Diagnóstico clínico

El estado neurológico de los pacientes con la EP, es evaluado de acuerdo a diferentes escalas, entre las que se incluyen UPDRS, MDS-UPDRS, H&Y (Hoehn y Yahr). La escala UPDRS es un sistema diseñado para la medición, clasificación y seguimiento del avance de la EP, consta de 4 etapas. En la primer etapa se evalúa dominio mental, conductual y de ánimo, en la segunda etapa se evalúan actividades de la vida cotidiana, en la tercer etapa se realiza una evaluación motora y en la cuarta etapa se tienen en cuenta complicaciones motoras. Al final de la sumarización del puntaje adquirido en las diferentes etapas, el valor máximo es total incapacidad y el valor mínimo normalidad de los parámetros (Prieto, 2015). Posteriormente la MDS realiza una modificación a la UPDRS, por lo que es nombrada (MDS-UPDRS), es la escala más utilizada en la actualidad. Se continúa con la estructura de 4 partes para la medición del avance de la enfermedad, pero se incluyen elementos de tipo no motor. En la primer etapa se evalúan experiencias no motoras de la vida diaria, en la segunda etapa se evalúan experiencias motoras, en la tercer etapa se realiza un examen motor y en la cuarta etapa se evalúan complicaciones motoras (Goetz, 2008). Por último H&Y hace una determinación de parámetros como: características sintomatológicas, extensión de las afecciones y el nivel de discapacidad física en que se encuentra el paciente. Aunque es actualmente usada, se tiene más como medida descriptiva y no tanto como medida de avance de la enfermedad, pues tiene diferentes falencias especialmente en lo que se refiere a falta de linealidad y demasiado valor a las medidas de inestabilidad postural, sobre otras manifestaciones motoras (Prieto, 2015).

En este trabajo sólo se considera la sección 3 de la escala MDS-UPDRS (MDS-UPDRS-III), dado que es la sección en la que se incluye una evaluación de las capacidades motoras de los pacientes. Esta sección cuenta con 33 ítems para la evaluación de capacidades motoras pero solamente uno de estos ítems considera el habla. Dado que el proceso de la producción del habla involucra diferentes órganos para la producción de sonidos como: vocales, consonantes (Benesty, 2007); es interesante el análisis de la producción del habla, pues ofrece diferentes variables de análisis como la estabilidad en la vibración de los pliegues vocales, contenido de energía, capacidad articulatoria, entre otras.

1.2.2 Afectaciones de la voz en pacientes con EP

Algunas de las dificultades en cuanto al habla que se presentan en los pacientes con la EP, son la disartria hipocinética y la disfagia orofaríngea. En la disartria hipocinética la voz de los la personas con EP se caracteriza por tener una intensidad reducida del volumen de la voz (hipofonía), reducción del rango de movimientos articulatorios (articulación hipocinética), inflexión del tono (hipoprosodia), tartamudeos, voz entrecortada y ronca. En la disfagia orofaríngea se presentan

dificultades como la rigidez en el habla, la bradicinesia, temblor, poca coordinación de los músculos de la estructura oral y faríngea (Herrero, 2011).

1.2.3 Modelado de problemas en la voz

La EP afecta distintos aspectos del habla de los pacientes, tales como la fonación, articulación, prosodia e inteligibilidad (Orozco, 2016). La fonación es la capacidad de controlar la respiración y expulsar el aire de los pulmones, para hacer vibrar las cuerdas vocales y así producir sonidos. Entre las principales dificultades que se presentan en la fonación se tiene un inadecuado cierre e inclinación de las cuerdas vocales (Hanson, 1984), lo que genera problemas en la estabilidad y periodicidad de la vibración de las cuerdas vocales. La fonación en EP ha sido investigada en (Tsanas, 2014), se analizan 14 pacientes con el método LSVT (Lee Silverman voice treatment), midiendo características como la perturbación, contenido de ruido y linealidad. Se analiza únicamente la información de la vocal sostenida \a/, teniendo una precisión del 90%, cuando se califica entre emisiones “aceptables” e “inaceptables”. La articulación es la capacidad de controlar correctamente los diferentes articuladores (lengua, labios, mandíbula, velo, entre otros), en el tiempo correcto y con la energía apropiada, para la producción clara y distinta de las palabras. Los problemas de articulación se vinculan principalmente con reducción de la amplitud y velocidad en el movimiento de los labios, la lengua y la mandíbula (Ackermann, 1991), lo que genera reducción de la capacidad articulatoria para la producción de vocales y generación del habla continua (Skodda, 2011). La prosodia brinda información acerca de diferentes fenómenos fónicos cómo: entonación, acentuación, ritmo, velocidad de habla, entre otros. Las dificultades en el habla de pacientes con EP, incluyen una disminución en el volumen y bajas variaciones de tono, que se relaciona con la frecuencia de la vibración de las cuerdas vocales (Ho, 1999; Darley, 1969). En (Bocklet, 2013) se realiza un estudio de las características de fonación en 88 pacientes con EP vs 88 hablantes sanos, hablantes de alemán nativo. Llegaron a un 81,9% de precisión en la discriminación entre pacientes con EP y hablantes sanos. En (Orozco, 2016), se realizó una discriminación de pacientes con EP y hablantes sanos basándose en la fonación, la articulación y la prosodia, mediante el uso de diferentes tareas diadococinéticas (DDK). Se consideraron tres idiomas diferentes (español, alemán y checo). Se basaron principalmente en el análisis de segmentos sonoros y segmentos sordos. La precisión informada estuvo en el rango de 85% a 99% dependiendo del idioma.

En (Novotný, 2014) se tiene un modelamiento de 6 diferentes dificultades articulatorias basadas en el análisis de segmentos VOT (del inglés Voice Onset Time), como lo son: calidad vocálica, coordinación de la actividad laríngea y supra laríngea, precisión de la articulación consonante, movimiento de la lengua, debilitamiento de la oclusión y tiempo del habla. Se analizan 24 pacientes con EP, hablantes nativos de checo, pronunciando las sílabas /pa-ta-ka/. Llegaron a un 88% de precisión en la discriminación entre pacientes con EP y hablantes sanos. En (Montaña, 2018) se genera un análisis basado también en segmentos VOT, enfocado en la repetición de las sílabas /ka/. Se analizan 27 pacientes

diagnosticados con EP y 27 controles sanos. Llegaron a un 92.2% de precisión en la discriminación entre pacientes con EP y hablantes sanos.

1.3 Objetivos

Desarrollar una metodología que permita analizar la capacidad articulatoria del habla en pacientes con enfermedad de Parkinson.

1.3.1 Objetivos Específicos

- Implementar diferentes características articulatorias que permitan detectar problemas en la voz de pacientes con EP.
- Implementar un algoritmo para la detección automática del VOT.
- Analizar la capacidad articulatoria en pacientes con EP considerando el VOT.
- Evaluar el desempeño de la metodología propuesta por medio de diferentes medidas utilizadas en aprendizaje de máquina.

1.4 Contribución al análisis de voz de hablantes con EP

En este trabajo se propone desarrollar una metodología que permita detectar problemas en la voz de pacientes con EP. Para esto se considerarán grabaciones de voz con tareas diseñadas para analizar la transición de un sonido a otro (coarticulación). La idea es extraer diferentes características acústicas que entreguen información acerca de la dificultad de los pacientes para articular diferentes sonidos. Para esto se propone analizar segmentos VOT, los cuales permiten analizar diversos signos de la disartria en la EP como lo son la imprecisión en las consonantes, la calidad de la voz, la coordinación de la actividad laríngea y supralaríngea, debilitamiento de la lengua y sincronización del habla. Dado que los segmentos VOT están definidos como el instante de tiempo que tarda en iniciar la sonorización después del estallido de la consonante, la metodología propuesta busca encontrar los instantes en que se produce: la oclusiva[1] (ráfaga inicial), el inicio de la vibración del pliegue vocal y la finalización de la vocal.

Posteriormente utilizando los segmentos VOT se extraen diversas características tanto en tiempo como en frecuencia basadas principalmente en los coeficientes cepstrales de la escala de frecuencia de Mel, las energías de Bark y los formantes vocálicos. Para explorar el nivel de discriminación que se puede lograr entre pacientes con EP y hablantes sanos, por medio de las características extraídas, se finaliza con la aplicación de métodos de aprendizaje de máquina.

[1] Sonido consonántico obstruyente que se produce por la detención del flujo del aire y su posterior liberación. Desde el punto de vista articulatorio, se les llama oclusivas debido al cierre de los órganos articulatorios, que se produce durante su pronunciación (Navarro, 2004).

El esquema de esta tesis es el siguiente. El **capítulo 2** presenta un marco teórico, que explica de forma clara y concisa los distintos conceptos teóricos que son utilizados como base para el desarrollo de la metodología propuesta. En el **capítulo 3** se describe el proceso generado para el desarrollo de la metodología propuesta. En el **capítulo 4** se muestran los resultados obtenidos y se realiza un análisis de los mismos. Finalmente en el **capítulo 5** se generan conclusiones y se proponen trabajos futuros.

Capítulo 2

Marco Teórico

2.1 Proceso de generación del habla

Para poder entender cómo la EP puede afectar la capacidad de una persona para expresarse, es necesario hacer un repaso de la base teórica que sustenta el proceso que produce el habla.

La capacidad de habla en el ser humano tiene su base en el cerebro, la corteza cerebral está compuesta por diversas partes, entre ellas, las que se encargan de la capacidad premotora (planeación) y motora (ejecución). El movimiento de los músculos es planeado en la corteza pre-motora, la cual se encarga de proyectar la "orden" (Orozco, 2016). Este conjunto de instrucciones es enviado a los núcleos motores del tronco del encéfalo y la médula espinal, llevando la información a los músculos de la garganta, lo que permite la producción de la voz (Gutiérrez, 2013).

La producción de la voz es el resultado de una serie de órdenes dadas por el sistema nervioso central, generando una actividad coordinada de la musculatura laríngea, torácica, abdominal y las estructuras resonadoras y articuladoras. Las neuronas del cerebro viajan en forma de una pequeña señal eléctrica que pasa por los nervios de la columna vertebral, hasta llegar al músculo necesario para completar la acción deseada (Gutiérrez, 2013).

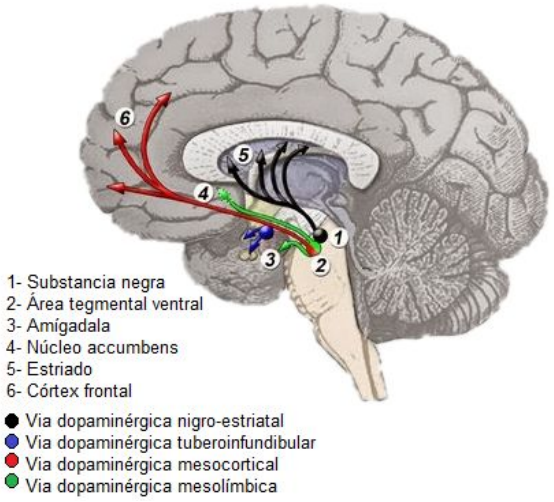


Figura 2.1: Vías dopaminérgicas en el cerebro (Muzio, 2013).

2.2 Clasificación de señales de voz

Según las necesidades en el proceso de comunicación, el habla humana puede producir distintos tipos de sonidos. A su vez estos tipos de sonidos, producen diversas características, que deben ser analizadas por separado. Existen dos tipos principales de señales de voz:

2.2.1 Señal de voz sonora

Los sonidos sonoros tienen una naturaleza cuasi-periódica en el dominio del tiempo y una estructura armónica fina en el dominio de la frecuencia, provocada por la vibración de las cuerdas vocales. Se caracterizan por generar segmentos energéticos debido a que el aire encuentra poca obstrucción al pasar por el tracto vocal. La periodicidad de los segmentos sonoros tiene una característica clave en el análisis de voz, que es el pitch o frecuencia fundamental. Se identifica como la periodicidad de los picos de amplitud en la forma de onda. Las frecuencias pitch de hombres y mujeres normalmente están en el rango de 50 - 250 Hz y 120 - 500 Hz, respectivamente (Rabiner, 1993).

2.2.2 Señal de voz sorda

Los sonidos no sonoros tienen una estructura sin periodicidad en el dominio del tiempo y un espectro más compensado en el dominio de la frecuencia. Se caracterizan por tener un comportamiento en forma de ruido aleatorio (Rabiner, 1993).

2.2.3 Transiciones onset y transiciones offset

Las transiciones onset y las transiciones offset (transiciones on/off), consisten básicamente en los segmentos de voz en los que la señal pasa de un espacio sonoro a uno sordo y de uno sordo a uno sonoro, respectivamente. Estos segmentos han sido utilizados usualmente en el estado del arte para el análisis de voz.

2.3 Análisis de características articulatorias del habla

El análisis de habla se convierte en un método efectivo para el monitoreo del inicio y progreso de la EP, así como para el chequeo del progreso de los distintos tratamientos (Postuma, 2012). Las diferentes pruebas para el análisis del habla incluyen entre otras, la repetición rápida de determinadas sílabas, la fonación sostenida, diversas lecturas y el monólogo hablado libremente para evaluar el alcance de las manifestaciones del habla. Los audios grabados suelen someterse a métodos tradicionales, como evaluación de la presión sonora, frecuencia

fundamental, frecuencia de formantes; que permiten determinar el rendimiento del habla (Baumgartner, 2001).

Las alteraciones articulatorias generadas por la EP, pueden ser evidentes por medio de la repetición de tareas DDK (Ackerman,1997). Este ejercicio es ampliamente preferido porque requiere complejos movimientos que involucran diversos articuladores (labios, mandíbula y lengua), durante una tarea con estructura definida, que contribuye a la reducción de la complejidad en el procesamiento de datos y ofrece una variedad de características relevantes para el análisis de voz (Fletcher, 1972). La pronunciación más típica de DDK, consiste en la repetición de las sílabas /pa/-/ta/-/ka/.

2.3.1 Escala de Bark

Se refiere a los rangos de frecuencia que corresponden a las regiones de la membrana basilar al ser estimuladas a frecuencias específicas. Se halla a partir de las bandas críticas correspondientes al oído humano, cada banda crítica se refiere a 1 Bark, variando entre 1 y 24 Bark. Esta escala es aproximadamente lineal por debajo de los 500Hz y logarítmica en frecuencias más altas. En la ecuación (1) se presenta la fórmula para el cálculo de la escala de Bark, relacionada con la frecuencia lineal.

$$Bark(f) = 13tan^{-1}(0.0076f) + 3.5tan^{-1}\quad((\frac{f}{7500})^2) \tag{1}$$

Donde f es la frecuencia en la escala lineal y $Bark(f)$ es la frecuencia resultante en la escala de Bark (Villa, 2012).

2.3.2 Coeficientes cepstrales en la escala de frecuencia de Mel (MFCC)

Se definen como el cepstrum real de una señal por ventanas de corta duración derivadas del espectro de la FFT. El cálculo de MFCC se realiza aplicando una serie de filtros pasa-banda equiespaciados en la escala de frecuencias de Mel, con una escala de frecuencias lineal por debajo de 1 kHz y logarítmica por encima de este valor, con igual número de muestras por debajo y por encima. La salida de cada filtro representa la energía de la señal dentro de la banda de paso de dicho filtro (Villamil, 2015). En la ecuación (2) se presenta la fórmula para el cálculo de la escala de Mel, en la ecuación (3) se presenta la expresión para el cálculo de los filtros en la escala de Mel.

$$Mel(f) = 2595 * log_{10}(700 + \frac{f(hz)}{700}) \tag{2}$$

Donde f es la frecuencia en la escala lineal y $Mel(f)$ es la escala en la frecuencia de Mel (Villa, 2015).

$$f_{Mel}(n) = \frac{K}{Fs}B^{-1}[B(f_l + n\frac{B(f_h - B(f_l))}{N+1})] \tag{3}$$

2.3.3 Formantes

Los formantes son frecuencias pico del espectro de voz para sonidos sonoros, en torno al cual se concentra la mayor parte de la energía de la señal. Existen muchos formantes, pero los primeros tres son esenciales para obtener una representación adecuada del tracto vocal. Los dos primeros llevan la mayor parte de la potencia del sonido mientras que el tercero tiene un efecto relevante en la inteligibilidad, aspecto importante para la comprensión del mensaje hablado. Los formantes se ubican en todas las vocales y en algunas consonantes, por el reparto que se hace en la energía, permiten la clasificación y categorización de la voz (San Juan, 2013).

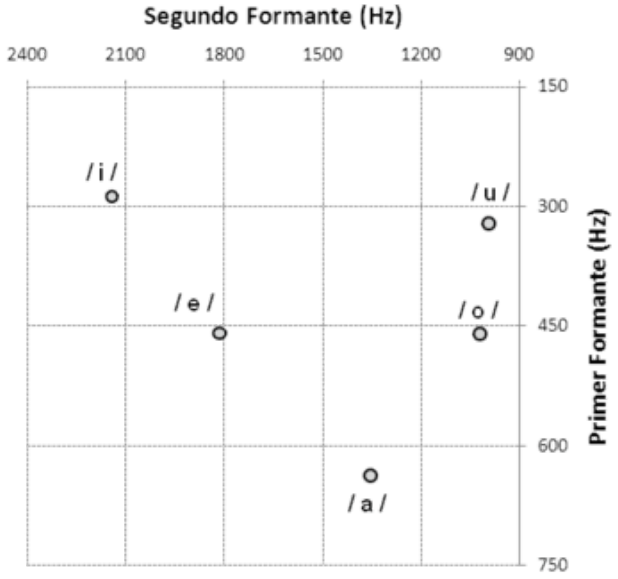


Figura 2.2: Triángulo vocálico en castellano, una representación gráfica de las diferencias de articulación en las vocales mediante la posición relativa de la lengua (Bradlow, 1995).

	Hombres		Mujeres	
Vocal	**F1 [Hz]**	**F2 [Hz]**	**F1[Hz]**	**F2[Hz]**
/a/	657	1215	664	1168
/e/	454	1995	492	2252
/i/	265	2318	241	2835
/o/	475	888	511	981
/u/	294	669	243	629

Tabla 2.1: Distribución de formantes vocálicos en español, para hombres y mujeres (Martínez, 2003).

2.3.4 Momentos espectrales

El espectro es representado por cuatro parámetros μ_n, que se encargan de codificar sus propiedades básicas. Este planteamiento es tomado desde la estadística y dichos parámetros consisten en la media, la desviación estándar, la asimetría y la curtosis (Paredes, 2017).

La media es considerada el valor representativo de la distribución de coeficientes y puede ser identificado como el valor medio de la energía, describe la tendencia central.

$$\mu_1 = (1/n) * \sum_{i=1}^{n}(x_i) \tag{4}$$

La desviación estándar es una medida estadística de dispersión que describe la tendencia a mantener picos frecuenciales.

$$\mu_2 = (1/n) * \sqrt{\left(\sum_{i=1}^{n} \bar{x} - x_i\right)^2} \tag{5}$$

La asimetría es el tercer momento estadístico central y está relacionado con la existencia de valores atípicos.

$$\mu_3 = \frac{(1/n) * \sum_{i=1}^{n}(\bar{x} - x_i)^3}{\left[(1/n-1) * \sum_{i=1}^{n}(\bar{x} - x_i)^2\right]^{3/2}} \tag{6}$$

La curtosis describe la persistencia a concentrarse los datos de la distribución en el valor medio de la energía.

$$\mu_4 = \frac{(1/n) * \sum_{i=1}^{n}(\bar{x} - x_i)^3}{\left[(1/n-1) * \sum_{i=1}^{n}(\bar{x} - x_i)^2\right]^2} - 3 \tag{7}$$

2.3.5 Voice Onset Time

Uno de los indicadores más comunes de disartria en la EP, es la imprecisión en la coordinación de las consonantes, la cual puede ser evaluada por medio VOT (Liberman, 1958). Consiste en el análisis de la duración del período de tiempo entre la producción de una consonante plosiva, i.e. /p/, /b/, /t/, /d/, /k/ o /g/ (se identifica en el espectrograma por una barra de explosión y se marca acústicamente por una ráfaga de ruido) y el inicio de la vibración del pliegue vocal (se identifica en el espectrograma de banda ancha por una fina estría y se marca acústicamente por el inicio de la sonorización periódica). El VOT toma el tiempo que tardan los repliegues vocales en comenzar a vibrar en relación con la retirada del obstáculo que genera la consonante en las cavidades supraglóticas, e implica la coordinación temporal entre

la articulación oral en el fin de una consonante y los mecanismos laríngeos involucrados en el proceso necesario para producir la vibración de las cuerdas vocales. Por tal motivo el VOT es un índice fiable para el análisis de la coordinación laríngea y supralaríngea (Martínez, 2010).

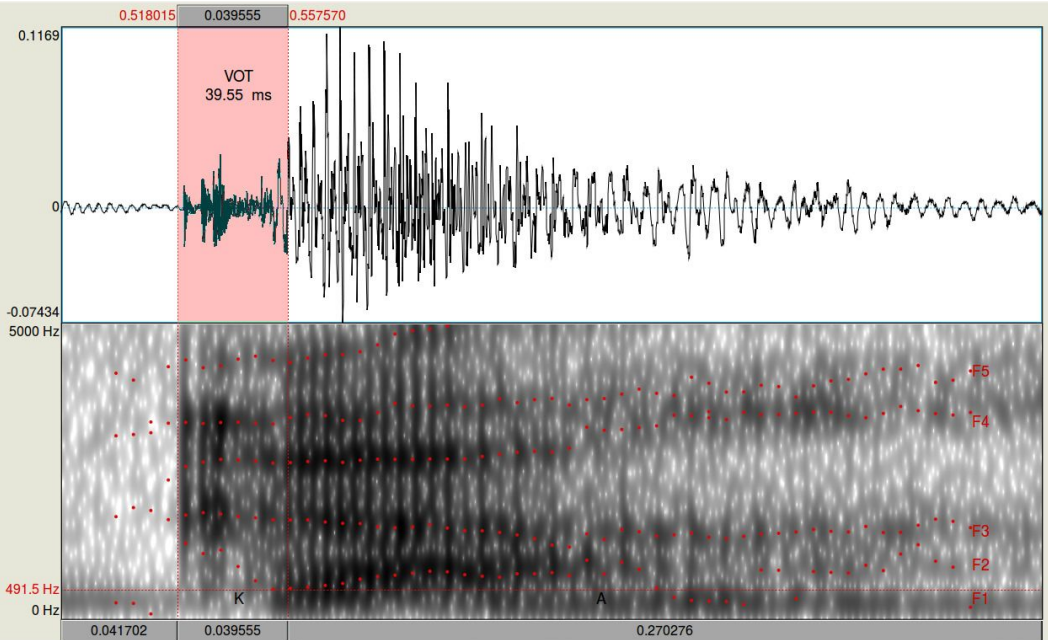


Figura 2.3 : Oscilograma y espectrograma de banda ancha de una consonante oclusiva /k/ seguida de la vocal /a/, donde se puede apreciar un VOT de 39,55 ms y la correspondencia de los formantes con las frecuencias indicadas en la tabla (1). Última sílaba de la repetición de /pe/-/ta/-/ka/, extraída desde el programa Praat[2] (Herramienta de uso libre para el análisis fonético del habla).

Los conceptos y métricas que utiliza el VOT para su desarrollo, pueden ser rastreadas en la investigación desde finales del siglo XIX cuando en (Adjarian, 1899), se estudiaron las paradas en la voz de hablantes Armenios, se caracterizaron dos momentos: el momento en que la consonante estalla cuando el aire se libera de la boca (explosión) y cuando la laringe comienza a vibrar. Sin embargo, el concepto de VOT surge en la década de 1960, cuando como se describe en (Lin, 2011), se inició un debate acerca de qué atributo fonético permitiría distinguir efectivamente paradas sonoras y no sonoras. Se analizaron atributos como: sonoridad, aspiración y fuerza articulatoria. En esta búsqueda de un mejor atributo fonético, con una mejor capacidad discriminativa en 1964 surge (Lisker,1964). Un estudio que se basó en el análisis de características acústicas de voz, midiendo la sonoridad de paradas iniciales en once lenguajes. Los resultados revelaron que entre los catorce parámetros que podrían afectar la clasificación sonora / no sonora de la voz, el VOT fue el más efectivo.

2.4 Métodos de aprendizaje de máquina

[2] http://www.fon.hum.uva.nl/praat/
Última visita 22/10/2018.

Cuando un sistema ha sido modelado y se tienen un conjunto de características con las que se espera determinar la presencia o ausencia de la patología, se pasa a la etapa de clasificación del sistema que proporcionará un límite de decisión entre las diferentes clases. Para este fin se utilizan métodos de aprendizaje de máquina. En este trabajo se utilizan tres métodos comúnmente usados en el estado del arte, por presentar muy buenas tasas de reconocimiento en el análisis de patrones: Máquinas de soporte vectorial, K-vecinos cercanos y Redes neuronales (Li Zuo, 2013).

2.4.1 Máquinas de soporte vectorial (SVM)

Una SVM es un método de aprendizaje de máquina que permite clasificar muestras, pertenecientes a dos poblaciones diferentes (clases o grupos), por medio de un hiperplano de separación ubicado de forma óptima entre los dos grupos, es decir, dicho hiperplano está ubicado de forma tal que se garantiza la máxima separación posible entre dos clases. Una muestra es clasificada en uno de los dos grupos, dependiendo de su ubicación respecto al hiperplano de separación.

Para elegir el hiperplano de separación, tan solo se consideran aquellas muestras de la clase de entrenamiento que se encuentran en las fronteras de la nueva muestra, estas muestras de frontera se convierten en los vectores de soporte.

En la búsqueda de linealizar de forma más óptima para tener un mayor ajuste del hiperplano, cada muestra es clasificada según la forma de separación de los datos, para lo cual existen tres métodos: separables linealmente, cuasi-separables linealmente y no separables linealmente (Carmona, 2013).

2.4.1.1 SVM para clasificación de margen separable linealmente

Se dice que una SVM es de margen separable linealmente, cuando el hiperplano de separación se puede definir como una función lineal. Como se visualiza en la figura 2.4, existen infinitos hiperplanos que pueden separar ambas clases de forma lineal, sin error. Por tal motivo surge el concepto de margen óptimo, definido como el punto medio entre las clases y que equidista de sus vectores de soporte.

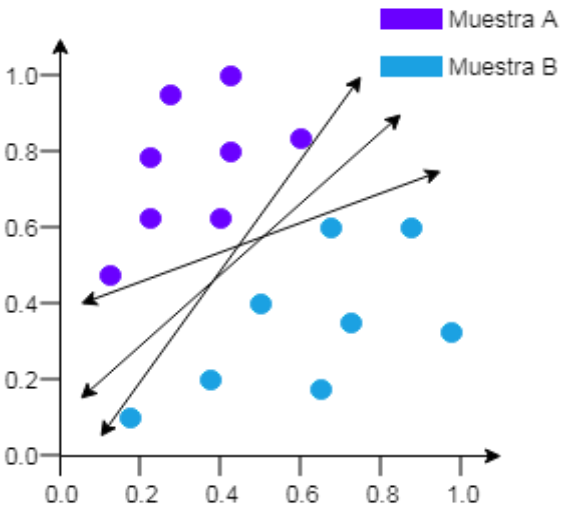


Figura 2.4: Distintos planos de separación, para función separable linealmente en un espacio bidimensional.

Dado un conjunto separable de ejemplos $S = \{(x1, y1), ..., (xn, yn)\}$, donde $x_i \varepsilon R^d$ e $y_i \varepsilon \{+1, -1\}$, se puede definir un hiperplano de separación como una función lineal que es capaz de separar dicho conjunto sin error:

$$D(x) = (w_1 x_1 + ... + w_d x_d) + b = < w, x > + b \quad (8)$$

Donde w y b son coeficientes reales. El hiperplano de separación cumplirá las siguientes restricciones para todo x_i del conjunto de ejemplos:

$$< w, x > + b \geq 0 \text{ si } y_i = +1 \quad (9)$$

$$< w, x > + b \leq 0 \text{ si } y_i = -1 \qquad i = 1, ..., n \quad (10)$$

De forma reducida:

$$y_i (< w, x_i > + b) \geq 0 \quad (11)$$

2.4.1.2 SVM para clasificación de margen cuasi-separable linealmente

La estrategia para este tipo de problemas es relajar el grado de separabilidad del conjunto de entrenamiento, permitiendo que haya errores de clasificación en algunos de los ejemplos. Según el planteamiento matemático, una muestra es no separable sino cumple la ecuación (12). Pueden darse diversas condiciones para determinar que la clasificación de la SVM es cuasi-separable como se visualiza en la figura 2.5.

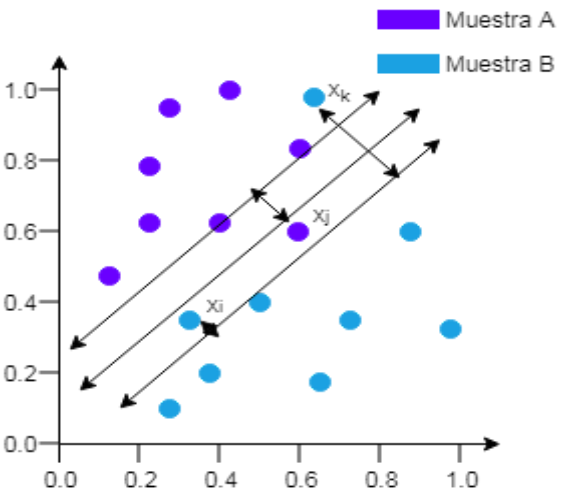


Figura 2.5: Distintos casos en los que las muestras son no-separables. En Xi la muestra está en el espacio correcto pero no supera el margen de decisión del vector de soporte. En X_j y X_k las muestras se ubican en el espacio de la clase contraria.

Ecuación de margen de separabilidad:

$$y_i(< w + x_i >) + b \geq 1 \qquad i = 1, ..., n \tag{12}$$

La idea para abordar este nuevo problema es introducir en la condición de separación del hiperplano, un conjunto de variables reales positivas, llamadas variables de holgura, $\xi_i, i = 1, ..., n,$ que van a indicar el número muestras no separables que se van a admitir en el clasificador.

$$y_i\,(< w, x_i > + b) \geq 1 - \xi_i\,, \quad \xi_i \geq 0, \; i = 1, ..., n, \tag{13}$$

2.4.1.3 SVM para clasificación de margen no separable linealmente

Se tiene que los casos comúnmente usados no suelen ser separables linealmente, pues en un sistema real influyen múltiples factores que sesgan la calidad de la información y por ende dificultan su diferenciación. Para calcular este tipo de hiperplanos se debe hacer corresponder con un punto en el espacio transformado de características f a cada ejemplo de entrada x, siendo $\Phi : x \rightarrow f$. Para $\Phi(x) = [\phi_1(x), ..., \phi_m(x)]$ y $\exists\, \phi_i(x), \; i = 1, ..., m,$ tal que $\phi_i(x)$ es una función no lineal. Se busca finalmente construir un hiperplano de función lineal en este nuevo espacio. La función de decisión, en el espacio de características viene dado por:

$$D(x) = (w_1\phi_1(x) + ... \; + w_m\phi_m(x)) \;\; = \;\; < w, \Phi(x) > \tag{14}$$

Para expresar una función en su forma dual, se usa la siguiente ecuación:

$$D(x) = \sum_{i=1}^{n} \alpha_i^* < x, x_i > \;\; + b^* \tag{15}$$

Finalmente se tiene una función de decisión en su forma dual, combinando la ecuación (15) y la ecuación (16):

$$D(x) = \sum_{i=1}^{n} \alpha_i^* y_i K(x, x_i) \quad (16)$$

Donde $K(x, x_i)$ se denominará función Kernel. Por definición es una función $K : X * X \rightarrow R$, usada para asignar a cada par de elementos de entrada X, un valor real correspondiente al producto escalar de las imágenes de dichos elementos en un nuevo espacio f (espacio de características), se tiene:

$$K(x, x_i) = \ < \Phi(x), \Phi(x)' > \ = (\phi_1(x)\phi_1(x)' + ... + \phi_m(x)\phi_m(x)') \quad (17)$$

Existen diversos tipos de funciones Kernel

Kernel lineal:

$$K(x, x') = \ < x, x' > \quad (18)$$

Kernel polinómico:

$$K_p(x, x') = [\gamma < x, x' > + \tau]^p \quad (19)$$

Kernel gaussiano:

$$K(x, x') = \ exp(-\gamma ||x, x'||^2), \ \gamma > 0 \quad (20)$$

2.4.2 K - Vecinos cercanos

Los Vecinos más cercanos o K-NN (Del inglés K-Nearest Neighbors), es un método de clasificación supervisada, usado para estimar la función de densidad $f(x, c_j)$ de las predictoras x por cada c_j. Este método estima el valor de la densidad de la probabilidad, de que un elemento x pertenece a una clase c_j, a partir de la información proporcionada por el sistema.

En el reconocimiento de patrones K-NN es usado como un método de clasificación de muestras, basado en un entrenamiento mediante ejemplos cercanos al espacio de K-NN. Parte de un conjunto de muestras conocidas, llamado conjunto de entrenamiento. Los vectores de entrenamiento se encuentran en un espacio característico multidimensional, descrito en términos de p términos, considerando q clases para clasificar. Con valores para i con $1 \leq i \leq n$, se representan por el vector p - dimensional $x_i = (x_{1i}, x_{2i}, x_{3i}..., x_{pi} \ \varepsilon \ x)$.

Un punto es asignado a la clase c si esta es la clase más frecuente entre los k ejemplos de entrenamiento más cercanos. Para determinar la distancia entre los puntos p, se tienen distintas ecuaciones como: distancia Euclidiana, distancia

Chebyshev, distancia Manhattan & distancia Kullback-Leibler. La más comúnmente usada es la distancia Euclidiana, que mide en línea recta la distancia más corta posible entre dos puntos.

$$D_{EUC}(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ri}, x_{rj})^2} \quad (21)$$

La fase de entrenamiento consiste en almacenar los vectores característicos y las etiquetas de las clases de entrenamiento. Dada una muestra desconocida y un conjunto de entrenamiento, se consideran las muestras circundantes a la nueva muestra y se asigna a la clase que contiene la mayoría de sus k vecinos más cercanos.

La elección del K juega un rol importante en esta técnica, pues determina el número de vecinos a considerar, para la clasificación de una nueva muestra. La mejor elección del K depende básicamente de los datos. Valores grandes de K reducen el efecto de ruido en la clasificación, pero crean límites entre clases parecidas. Generalmente se debe elegir un valor impar para evitar empates en la clasificación(Arriagada, 2015).

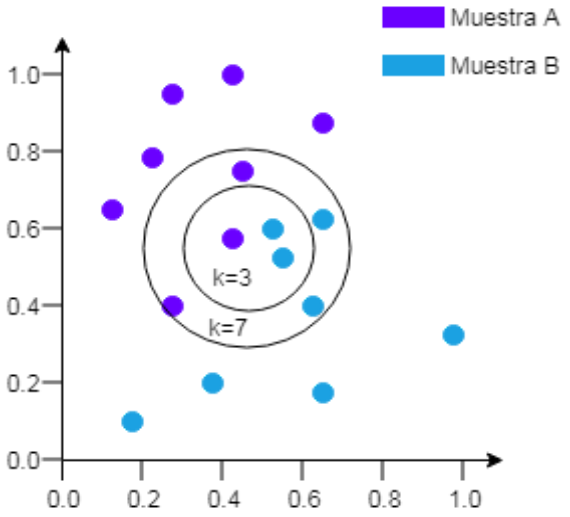


Figura 2.6: Representación del proceso de selección de K-NN para una nueva muestra con K=3 y K=7

2.4.3 Redes Neuronales

Las redes neuronales artificiales (RNA) son un método de aprendizaje de máquina, que se basa en redes interconectadas masivamente en paralelo con elementos adaptativos y con organización jerárquica, las cuales buscan interactuar con objetos del mundo real, de la misma forma que lo hace el sistema nervioso biológico.

Entre las principales ventajas que tiene el uso de RNA en reconocimiento de patrones se tienen: la capacidad de aprendizaje adaptativo, la capacidad de autoorganización, la capacidad de generalización ante datos irrelevantes, operación en tiempo real, entre otras.

El funcionamiento de RNA está basado en capas, normalmente se utilizan tres capas para el planteamiento de una RNA. Los datos ingresan por la capa de entrada, pasan a través de las capas ocultas y salen por la capa de salida.

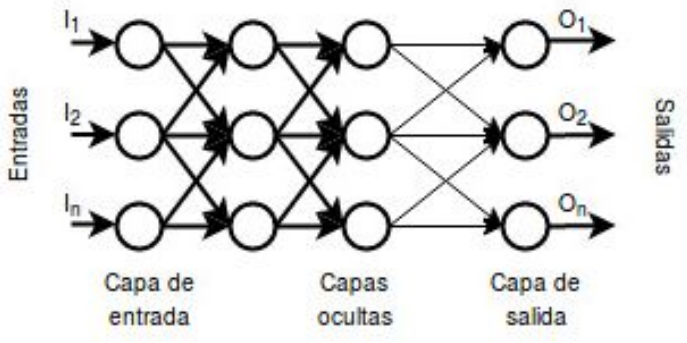


Figura 7: funcionamiento básico de una RNA totalmente conectada. Los datos ingresan por la capa de entrada, se hace el procesamiento propio de la RNA en las capas ocultas y finalmente entrega la mejor combinación de neuronas en la capa de salida.

La función de entrada trata los valores como si fueran uno solo, lo que recibe el nombre de entrada global, es representada en la ecuación (22).

$$input_i = (in_{i1}, w_{i1}) * (in_{i2}, w_{i2}) * \ldots * (in_{in}, w_{in}) \tag{22}$$

Donde: $*$ representa el operador adecuado (por ejemplo: máximo, sumatoria, productoria, entre otras), n representa al número de entradas a la neurona, N_i y w_i al peso. Los valores de entrada se multiplican por los pesos ingresados con anterioridad a la neurona, por tal motivo los pesos pueden cambiar la influencia de los valores de entrada.

Del mismo modo en que una neurona biológica puede estar activada y no activada, las neuronas artificiales tienen estados de activación. Por tal motivo se utiliza la función de activación para transformar la entrada global en un valor de activación, cuyo rango va de 0 a 1, pues la neurona puede estar inactiva (0 o -1) o activa (1). La función de activación es una función de entrada global (gin_i) menos el umbral (Θ_i).

Las funciones de activación más comúnmente usadas son:

Función lineal:

$$f(x) = \begin{cases} -1 & \text{con} \quad x \leq \frac{-1}{a} \\ a * x & \text{con} \quad \frac{-1}{a} \leq x \leq \frac{1}{a} \\ 1 & \text{con} \quad x \geq \frac{1}{a} \end{cases} \tag{23}$$

Los valores de salida por medio de la función de activación serán $a * (gin_i - \Theta_i)$ cuando el argumento de $gin_i - \Theta_i$ esté comprendido entre $(\frac{-1}{a}, \frac{1}{a})$. Por encima o por debajo de esta zona se fija la salida como 1 o -1, respectivamente.

Función sigmoidea:

$$f(x) = \frac{1}{1+e^{-gx}} \quad \text{con} \quad x = gin_i - \Theta_i \tag{24}$$

Los valores de salida que proporciona la función sigmoidea van de 0 a 1. Al modificar el valor de g se ve afectada la pendiente de la función de activación.

Función tangente hiperbólica:

$$f(x) = \frac{e^{gx}+e^{-gx}}{e^{gx}+e^{-gx}} \quad \text{con} \quad x = gin_i - \Theta_i \tag{25}$$

Los valores de salida que proporciona la función tangente hiperbólica van de -1 a 1. Al modificar el valor de g se ve afectada la pendiente de la función de activación.

Las neuronas ocultas están internas en la red y pueden estar distribuidas de distintas formas, lo que va a determinar junto con el número de neuronas seleccionadas la topología de la RNA y la capacidad de clasificación que tendrá entre las distintas clases.

La función de salida es el último componente necesario en las RNA y determina qué valor se transfiere a las neuronas vinculadas. Cuando el valor de salida resultante $i(out_i)$ está por debajo de un umbral determinado, ninguna salida se pasa a la neurona subsiguiente, los valores permitidos suelen estar entre [0,1] o [-1,1].

El entrenamiento de una RNA se basa principalmente en modificar sus pesos en respuesta a los datos de entrada. Los cambios de peso que se producen determinan el aprendizaje de la red y se traducen en destrucción, modificación y creación de conexiones entre neuronas (Matich, 2001).

2.5 Medidas de desempeño: Tasa de acierto y curva ROC

Este tipo de mediciones son bastante útiles principalmente en el análisis de enfermedades y diagnósticos clínicos. Hacen parte de un sistema de clasificación supervisado llamado matriz de confusión, en el que se combinan diversos grupos para diferenciar características específicas según el área de análisis, por tal motivo, para su realización es necesario plantear dos grupos principales, uno llamado clase de referencia, que es el grupo que va a ser analizado y clase de control, que es el grupo contra el que se va a comparar la clase de referencia, al plantearse dos

grupos se le llama matriz de confusión biclase. Los elementos de la matriz de confusión se definen de la siguiente manera:

- TP (True Positive, Verdadero Positivo): Son todas las muestras de clase de referencia que fueron clasificadas correctamente.
- FP (False Positive, Falso Positivo): Son todas las muestras de clase de referencia que fueron clasificadas incorrectamente (clasificadas como pertenecientes a la clase de control).
- TN (True Negative, Verdadero Negativo): Son todas las muestras de clase de control que fueron clasificadas correctamente.
- FN (False Negative, Falso Negativo): Son todas las muestras de clase de control que fueron clasificadas incorrectamente (clasificadas como pertenecientes a la clase de referencia).

La tasa de acierto se refiere al porcentaje de muestras clasificadas correctamente en todo el sistema, tanto en la clase de referencia como en la clase de control. La ecuación (26) presenta la fórmula para el cálculo del porcentaje de la tasa de acierto:

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} * 100 \quad (26)$$

La sensibilidad es el porcentaje de muestras clasificadas correctamente en la clase de referencia, en este proyecto en particular, indica la capacidad del sistema para detectar hablantes patológicos. La ecuación (27) presenta la fórmula para el cálculo del porcentaje de sensibilidad:

$$SEN = \frac{TP}{TP+FP} * 100 \quad (27)$$

La especificidad es el porcentaje de muestras clasificadas correctamente en la clase de control, indica la capacidad del sistema para detectar personas sanas. La ecuación (28) presenta la fórmula para el cálculo del porcentaje de especificidad:

$$ESP = \frac{TN}{TN+FN} * 100 \quad (28)$$

La curva ROC permite visualizar el desempeño de un sistema biclase. Cruza en el eje vertical la tasa de verdadero positivo con la tasa de falso positivo en el eje horizontal. Se grafica tomando diversos pares de valores de sensibilidad contra 1-ESP, para diversos criterios de decisión que varían según la patología.

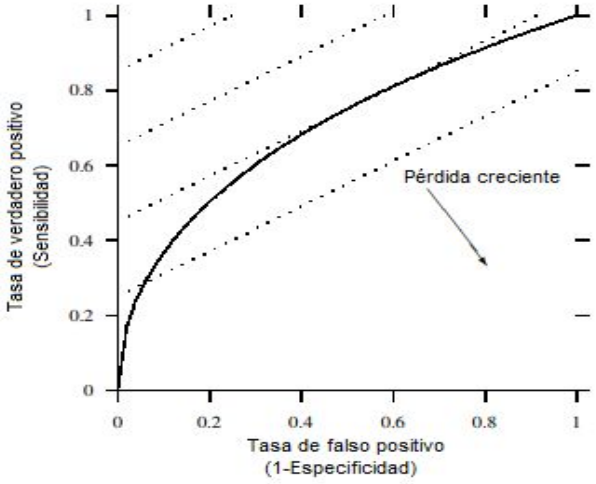


Figura 2.8: Representación de la Curva ROC

Gráficamente, cuanto más próximo al punto cartesiano (0,1) se encuentre la curva, mejor será el resultado del diagnóstico automático, pues se tienen más posibilidades de poder discernir entre enfermedad y no enfermedad (Webb, 2002). El AUC es la medida de desempeño de la curva ROC y varía entre 0 y 1, donde los valores más cercanos a 1 indican un buen desempeño en el sistema de clasificación. En aplicaciones biomédicas, el AUC se califica en la siguiente escala: AUC<0.70 es malo, 0.70< AUC <0.80 es justo, 0.80 < AUC < 0.90 es bueno y AUC>0.90 es excelente (Swets, 2000).

Capítulo 3

Metodología

El principal objetivo de este trabajo fue desarrollar una metodología que permitiera analizar la capacidad articulatoria del habla en pacientes con EP. Para validar el cumplimiento del enfoque propuesto, este trabajo fue realizado en coordinación con expertos neurólogos. Particularmente, siguiendo las recomendaciones del grupo de investigación de Neurociencias (Grupo de Neurociencias de Antioquia, Universidad de Antioquia, Medellín, Colombia.). Adicionalmente, este proyecto cuenta con la experiencia del grupo de investigación GITA (Universidad de Antioquia) y el laboratorio de reconocimiento de patrones (Universidad de Erlangen-Nürnberg) para el análisis de habla.

3.1 Base de datos

La base de datos consiste en grabaciones de voz capturadas por el grupo GITA (PC-GITA), la cual cuenta con un total de 100 de personas, 50 pacientes con EP y 50 hablantes sanos (25 hombres y 25 mujeres en cada grupo). La edad de los pacientes con EP oscila entre los 33 y 81 años (media 61,14 ± 9,61), mientras que la edad de los hablantes sanos oscila entre los 31 a 86 años (media 60,9 ± 9,46). Las grabaciones contienen la voz de las personas mencionadas, quienes repitieron determinadas vocales, palabras y oraciones de carácter DDK. Dichas grabaciones fueron capturadas en una cabina insonorizada en la fundación clínica Noel y cuentan con los respectivos consentimientos firmados por los hablantes, necesarios para el tratamiento y análisis de las grabaciones y de datos personales. Las señales de voz fueron capturadas con un micrófono omnidireccional a una frecuencia de muestreo de 44100 Hz y con una resolución de 16 bits. Los pacientes estaban en estado ON durante la sesión de grabación, es decir, no más de 3 horas después de la medicación de la mañana. Ninguno de los hablantes en el grupo de los “sanos” presenta síntomas asociados con la EP ni ninguna otra enfermedad neurológica.

	Pacientes con EP		**Hablantes sanos**	
	Hombres	**Mujeres**	**Hombres**	**Mujeres**
Número de sujetos	25	25	25	25
Edad $(\mu \pm \sigma)$	61.3±11.4	60.7±7.3	60.5±11.6	61.4±7.0

Rango de edad	33-81	49-75	31-86	49-76
Duración de la enfermedad $(\mu \pm \sigma)$	8.7±5.8	12.6±11.6	-	-
MDS-UPDRS-III $(\mu \pm \sigma)$	37.8±22.1	37.6±14.1	-	-

Tabla 3.1: Información de la base de datos PC-GITA

3.2 Extracción de características

3.2.1 Preprocesado

Los datos contenidos en las grabaciones, a pesar de ser adquiridos en ambientes controlados, pueden tener variaciones entre pacientes, como la distancia del hablante al micrófono o el volumen de la voz. Por tal motivo se normalizan las amplitudes de la señal de audio entre -1 y 1, para que todas queden en el mismo rango de valores.

Los audios para el análisis contienen la repetición rápida de las palabras /pa/-/ta/-/ka/, /pa/-/ka/-/ta/ y /pe/-/ta/-/ka/. Al no conocer el tiempo de duración, ni el número de repeticiones que generó cada persona en su respectiva grabación, fue necesario aplicar técnicas de segmentación para especificar el tiempo de duración de cada fonema. Dichas técnicas se basan en la utilización de filtros de Gabor sobre los espectrogramas, para detectar los cambios de energía correspondientes al cambio entre silencio-plosiva y vocal-plosiva, como se visualiza en la figura 3.1.

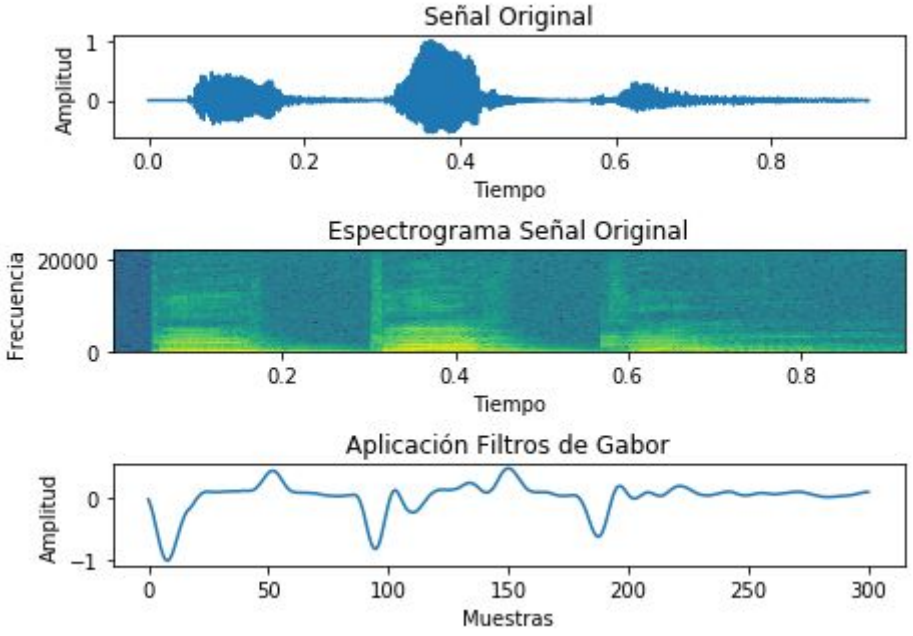


Figura 3.1: Señal de audio para la sílaba /pe/-/ta/-/ka/. Aplicación de técnicas de segmentación basadas en filtros de Gabor, vemos como una técnica que se utiliza para analizar texturas en procesamiento de imágenes, puede ser aplicada para detectar los bordes de separación en el paso

de una sílaba a otra, en función de extraer los segmentos VOT. Los picos máximos representan el instante de tiempo de la señal en los que se pasa de una sílaba a otra.

En los dominios del tiempo y la frecuencia, un filtro de Gabor es una función Gaussiana modulada por una onda sinusoidal plana, como se visualiza en la figura 3.2 [3].

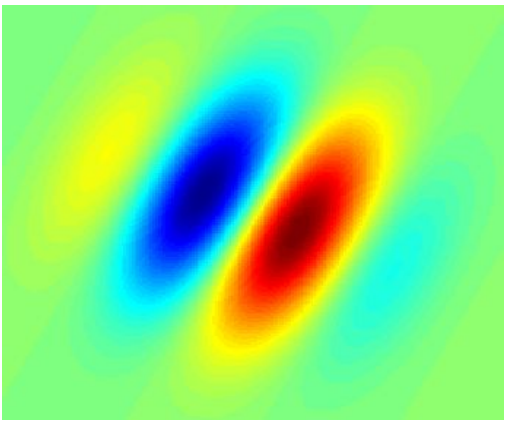
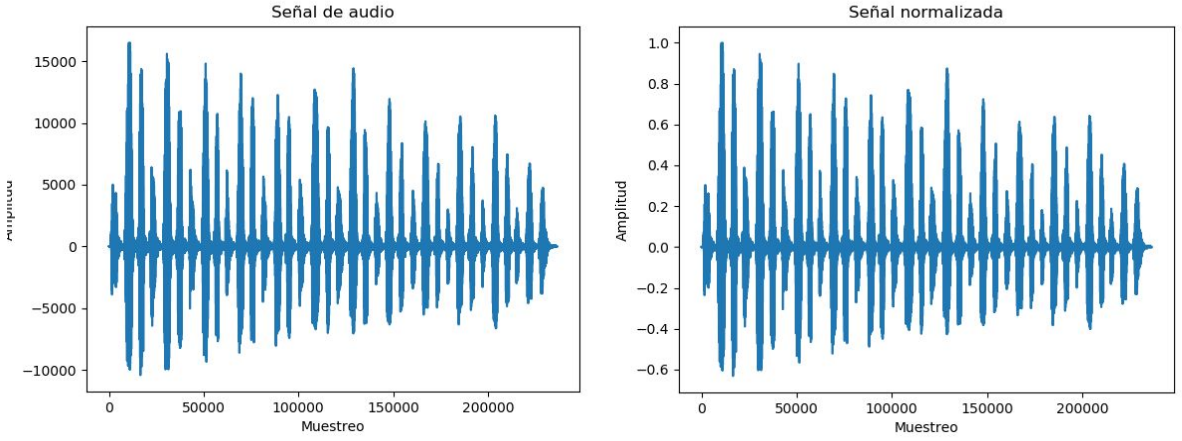


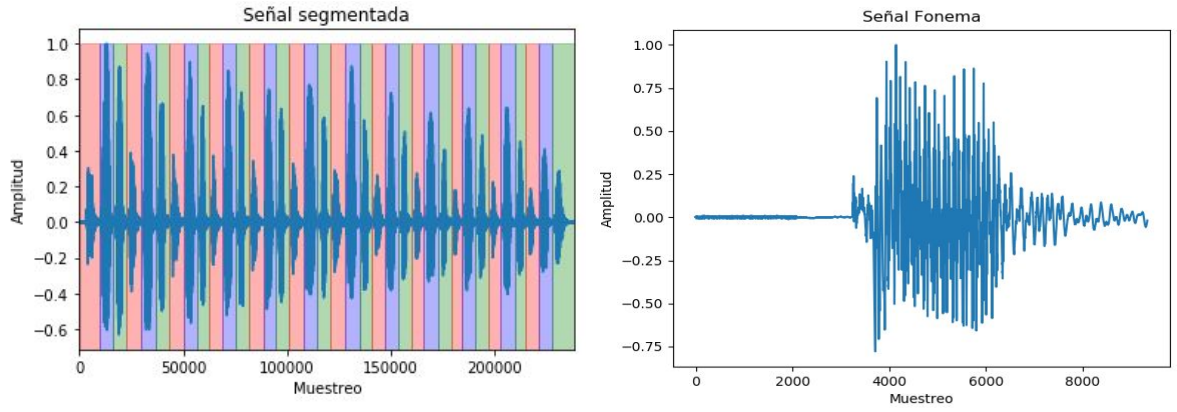
Figura 3.2: Filtro de Gabor bidimensional. Son ampliamente utilizados en procesamiento de imágenes y análisis de texturas. Por medio de dichos filtros pueden ser identificadas frecuencias específicas en determinadas regiones de la imagen.

Dado que en el proceso de producción del habla se generan señales no estacionarias con propiedades que cambian rápidamente en el tiempo, es necesario realizar un proceso de enventanado para analizar pequeños segmentos de 10 ms a 20 ms, dependiendo del tipo de análisis que se quiere realizar. A este proceso se le adiciona también un segmento de solape en la transición entre ventanas, buscando no perder elementos valiosos para el análisis, cuando se pasa de una ventana a otra.



(a) (b)

[3] Esta figura fue tomada de
https://en.wikipedia.org/wiki/Gabor_filter#/media/File:Gabor_filter.png
Última revisión 20/11/2018



(c) (d)

Figura 3.3: Etapa de preprocesado para pronunciación continua de sílabas /pe/-/ta/-/ka/. a) Señal original, b) Normalización de la señal original entre -1 y 1, c) segmentación del audio por medio de filtros de Gabor, d) extracción de uno de los fonemas.

3.2.2 Extracción VOT

Siendo el análisis de características basadas en segmentos VOT, el principal objetivo de este trabajo, lo primero que se debe hacer es localizar los segmentos VOT para cada una de las sílabas producidas por los hablantes. Los principales instantes que se deben hallar son la producción de la oclusiva (ráfaga inicial), el inicio de la vibración del pliegue vocal y la finalización de la vocal (oclusión). El algoritmo para la detección del VOT está basado en el planteamiento realizado en (Montaña, 2018). A continuación se describen los pasos que se siguieron para la estimación del inicio de la vocal:

1. Calcular la envolvente de amplitud de la señal de voz. Este proceso se realiza aplicando la transformada de Hilbert sobre la señal.

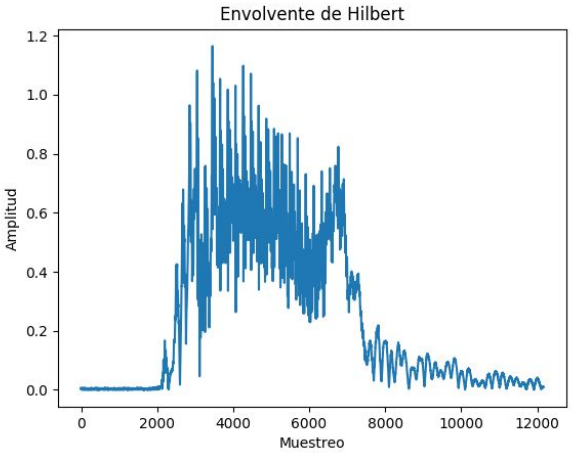


Figura 3.4: Envolvente de amplitud de la señal después de aplicar la transformada de Hilbert para la sílaba /pe/.

2. Suavizar la envolvente para reducir el ruido en amplitud. Para esto se aplica convolución a la envolvente con ventanas Gaussianas de 10 ms.

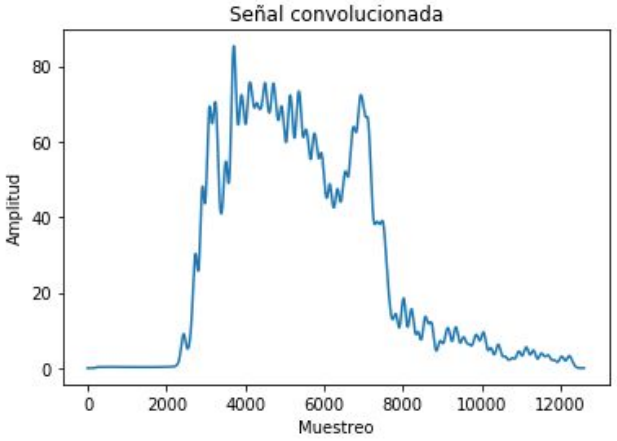


Figura 3.5: Señal suavizada por medio de la convolución de la envolvente con ventanas gaussianas para la sílaba /pe/.

3. Identificar el primer pico prominente por encima del umbral de la envolvente. El valor de dicho umbral es el promedio de la primer mitad de la señal convolucionada.

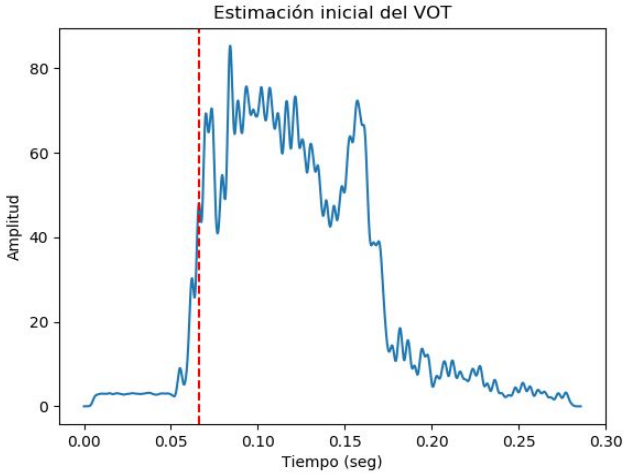


Figura 3.6: Cálculo del inicio de la vocal para la sílaba /pe/. Estimación del inicio de la vocal en la envolvente.

4. Realizar el cálculo de la tasa de cruces por cero (ZCR) de la señal. Esto con la finalidad de corroborar que el pico identificado no se deba a una parada energética del hablante.

5. Mover la estimación al siguiente pico de la amplitud envolvente, si la ZCR en la posición del pico considerado está por encima del promedio de la ZCR.

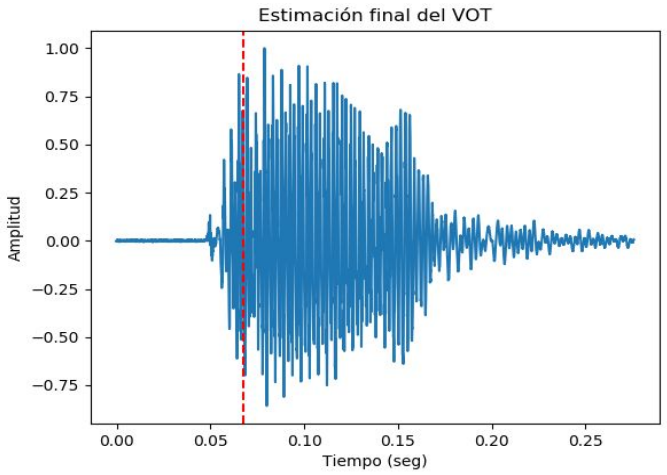


Figura 3.7: Cálculo del inicio de la vocal para la sílaba /pe/. Estimación final de inicio de la vocal aplicando la condición de movimiento de la estimación si se incumple el umbral para la ZCR.

Una vez se tiene estimado el inicio de la vibración, se pasa al cálculo de la ráfaga inicial en la producción de la oclusiva:

1. Calcular el espectrograma para el tramo de tiempo entre el inicio de la señal y el inicio de la vibración ya calculado.
2. Seleccionar la región donde la energía es de al menos el 1% de la energía máxima del intervalo. Esta estimación se ajusta calculando la varianza para pequeños fragmentos de 0.6 ms. Dado que los segmentos explosivos se caracterizan por una alta variabilidad en el tiempo, por medio de la varianza se calcula el momento en que termina el ruido de la señal y se presenta la ráfaga de la consonante, puede ser identificado como una transición abrupta de una región a otra.

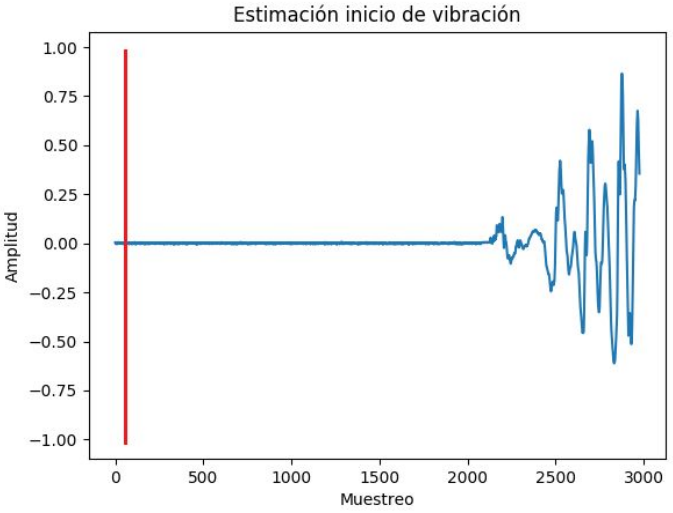


Figura 3.8: Inicio de la vibración de los pliegues vocales para la sílaba /pe/.

Con el cálculo de los dos momentos anteriores se tiene la duración del VOT, para finalizar se pasa a calcular el momento de la oclusión, necesario para la extracción de características temporales y se puede realizar de la siguiente forma:

1. Seleccionar el último tercio de la señal, como segmento de análisis.

2. Identificar el punto de la nueva señal donde la energía es al menos 1% de la energía máxima. Se toma el último instante de tiempo en el que la energía es al menos 1% de la energía media de la nueva señal.

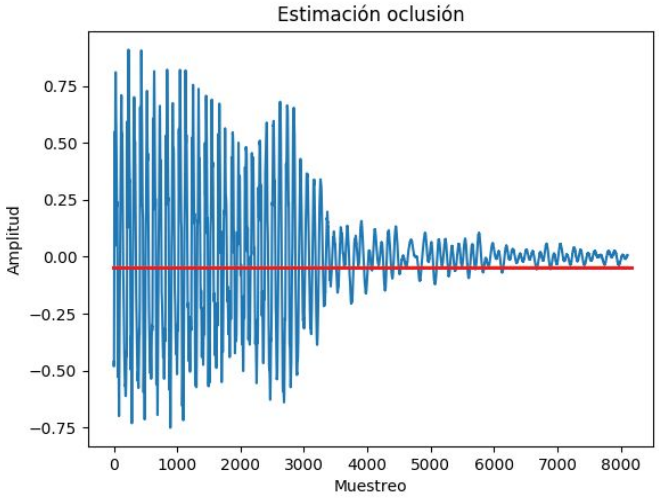


Figura 3.9: Finalización del sonido vocálico para la sílaba /pe/.

De esta forma pueden ser ubicados los tres puntos temporales necesarios para la extracción de características a partir del VOT, en la figura 3.9 se muestra la estimación final.

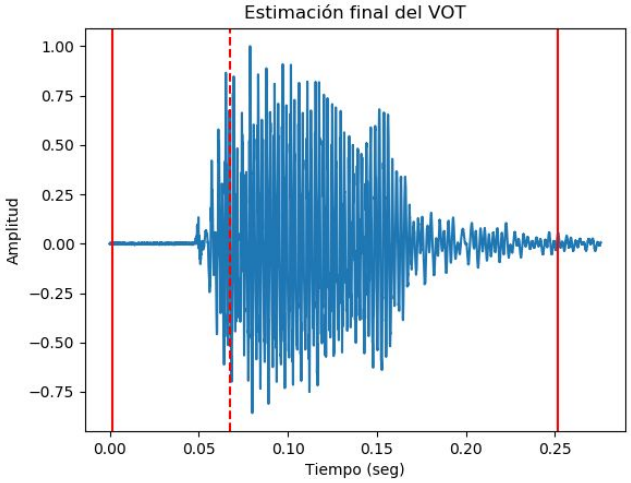


Figura 3.10: Ubicación de los tres momentos necesarios para la extracción de características a partir del VOT: producción de la oclusiva, inicio de la vibración del pliegue vocal y finalización de la vocal, para la sílaba /pe/.

3.2.3 Caracterización

El análisis realizado a los DDK, extrae un conjunto de características orientadas a la clasificación de hablante sano o afectado por la EP. Las características analizadas tienen su base en medidas tanto de tiempo como de frecuencia. Se toman las 36 características propuestas en (Montaña, 2018).

Entre las características a calcular en el dominio del tiempo se tienen 6: VOT, proporción de duración del VOT (VOT-ratio), proporción de duración de las vocales (CV-ratio), cociente de variabilidad de las vocales (VVQ), cociente de variabilidad de la consonante (CVQ) y tasa de articulación (A-rate). Para el caso de VOT, VOT-ratio y CV-ratio, las características finales son los valores medios de los parámetros obtenidos en las distintas sílabas. VOT-ratio es la relación entre el VOT y la longitud de la sílaba. CV-ratio es la relación entre el VOT y la longitud de la vocal. VVQ es la desviación estándar de la duración de las vocales. CVQ es la desviación estándar de la duración de las consonantes. A-rate es el número de sílabas por segundo en un DDK.

Para el análisis en el dominio de frecuencia se tienen 26 características. Dichas características están basadas en 13 filtros MFCC. Se realizó el cálculo de los MFCC, para ventanas de 20 ms con un solape del 50%. Finalmente se calcula la media (13 características) para los MFCC encontrados y la desviación estándar (13 características), para un total de 26 características basadas en MFCC.

Las 4 características siguientes están basadas en el cálculo de los momentos espectrales. Inicialmente se representa el espectro por medio de codificación predictiva lineal (LPC), que codifica las propiedades básicas del espectro en un pequeño número de parámetros. Los momentos espectrales se calculan sobre la suavizada del segmento VOT.

Características basadas en el dominio del tiempo		
Característica	**Unidad**	**Descripción**
VOT	ms	Tiempo de inicio de la vocalización
VOT-ratio	%	Relación VOT y la longitud de la sílaba
CV-ratio	%	Relación entre el VOT y la longitud de la vocal
VVQ	%	Desviación estándar de la duración de las vocales
CVQ	%	Desviación estándar de la duración de las consonantes
A-rate	%	Tasa de articulación
Características basadas en el dominio de la frecuencia		
Característica	**Descripción**	
$\bar{x}MFCC_i, i = 0-12$	Media de $MFCC_i$	
$\sigma MFCC_i, i = 0-12$	Desviación estándar de $MFCC_i$	

Características basadas en los momentos espectrales	
Característica	**Descripción**
Media $\bar{x}$	Tendencia central espectral
Desviación estándar σ	Dispersión espectral
Oblicuidad *skew*	Asimetría espectral
Kurtosis *kurt*	Apuntamiento espectral

Tabla 3.2: Descripción de la caracterización para el VOT

3.3 Clasificación de pacientes con EP y controles

El proceso de clasificación de las características encontradas anteriormente, se basa en la búsqueda de conocer el porcentaje de acierto que se puede alcanzar en la diferenciación de la pertenencia a una clase u otra de las nuevas muestras. Se realiza por medio de métodos de aprendizaje de máquina, tales como: SVM, K-NN y NN.

En la búsqueda de realizar una distribución equilibrada entre los conjuntos de hablantes sanos y pacientes con EP, se utiliza el método de validación cruzada. Consiste en el entrenamiento de las muestras variando el espacio de características, según el tamaño de la matriz final. En este proyecto se realizó con 10 validaciones en un conjunto de 100 muestras (100 hablantes), por lo que el sistema alterna conjuntos de prueba del 10% y conjuntos de entrenamiento del 90% como se visualiza en la figura 3.10. El proceso del método de aprendizaje de máquina se realiza para cada iteración de la validación cruzada y las medidas de desempeño finales serán promediadas entre las resultantes en cada validación.

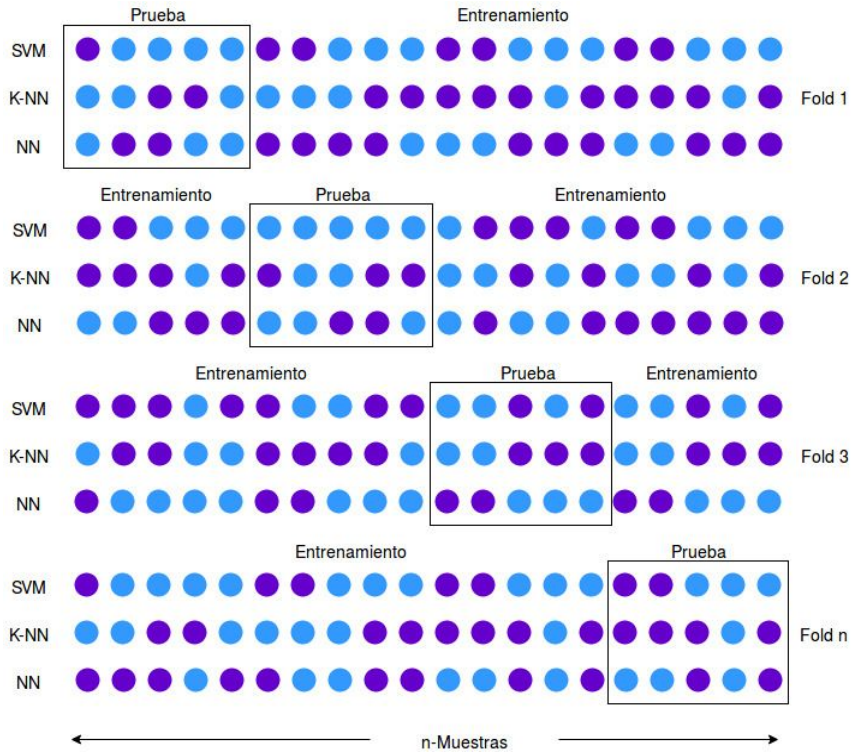


Figura 3.11: Entrenamiento y prueba en un conjunto de n muestras en el que se ejemplifica el método de validación cruzada, dividiendo para cada máquina de aprendizaje entre entrenamiento y prueba.

Después de normalizar las características en el conjunto de entrenamiento, se pasa a entrenar cada una de los métodos. El entrenamiento de la SVM se realiza con un kernel de separación gaussiano. Se utilizan parámetros de margen C con $10^{-3} < C < 10^{4}$ y ancho de banda del Kernel γ con $10^{-6} < \gamma < 10^{4}$. Los parámetros de C y γ se alternan en la ecuación en la búsqueda del mejor hiperplano posible para el conjunto que se está testeando. K-NN se testea con $K = [3, 5, 7, 11, 15]$. Para las NN se utilizan tamaños de capas ocultas de {5, 15, 30}. Dichos valores se combinan en el entrenamiento, optimizando el resultado de la método con la selección de las mejores combinaciones posibles.

Se concluye con la evaluación del desempeño del algoritmo, con una herramienta denominada matriz de confusión, basada en los mismos conceptos vistos para la determinación de las medidas de desempeño del sistema. Se determinan una clase positiva y una negativa, como se visualiza en la tabla 3.3. Las columnas de la matriz representan las predicciones de cada clase, mientras que las filas representan los valores en la clase real.

		Predicción	
		Positivos	**Negativos**
Clase actual	**Positivos**	Verdaderos positivos (VP)	Falsos negativos (FN)

	Negativos	Falsos positivos (FP)	Verdaderos negativos (VN)

Tabla 3.3: Representación de las clases en la matriz de confusión

Capítulo 4

Resultados y análisis

Con la finalidad de tener un punto de comparación para los resultados que se obtengan con el procesamiento de los segmentos VOT. Se realizó un primer procesamiento basado en métodos tradicionales como lo son las transiciones on/off.

4.1 Análisis de voz basado en transiciones on/off

Para este análisis se utiliza la misma base de datos PC-GITA y se utilizan los mismos parámetros básicos en lo que se refiere al preprocesado de la señal, en la búsqueda de generar condiciones óptimas para la extracción de los datos. Se obtuvieron características basadas en las energías de Bark, MFCC y en los momentos espectrales, como se visualiza en la tabla 4.1. Se compilaron de la siguiente forma: 13 MFCCs y la energía logarítmica distribuida en 17 bandas de Bark, luego se multiplicaron por los 4 momentos espectrales, para un total de 120 características.

Características basadas en el dominio de la frecuencia	
Característica	**Descripción**
$MFCC_i, i = 0-12$	Coeficientes $MFCC_i$
$EBARK_i, i = 0-16$	Energías de Bark $EBARK_i$
Características basadas en los momentos espectrales	
Media $\bar{x}$	Tendencia central espectral
Desviación estándar σ	Dispersión espectral
Oblicuidad $skew$	Asimetría espectral
Kurtosis $kurt$	Apuntamiento espectral

Tabla 4.1: Extracción de características para transiciones on/off

La matriz de características obtenida es clasificada por medio los métodos de aprendizaje de máquina descritos anteriormente y con los mismos parámetros

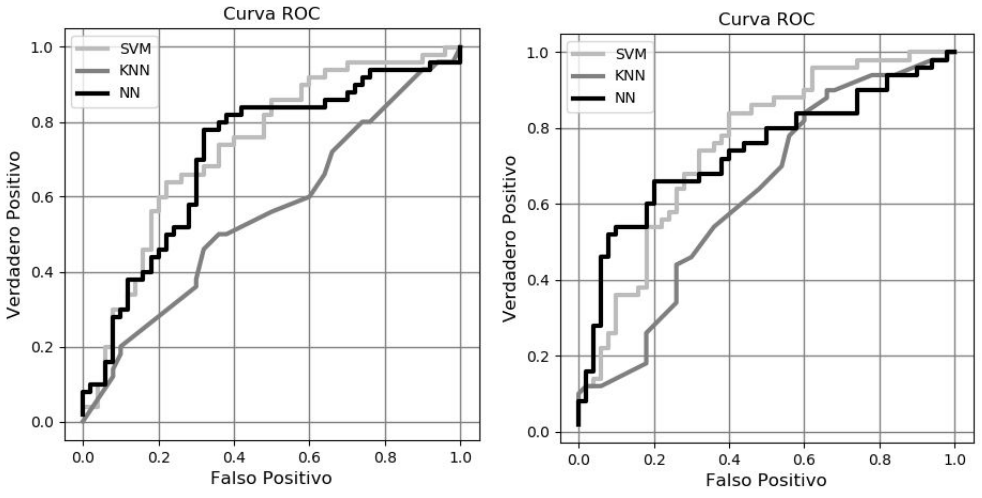
mencionados. Los mejores resultados fueron obtenidos en la SVM, para las repeticiones de /pa-ka-ta/. Se alcanzó una tasa de acierto del 70% y un Auc de 0,74 tabla 4.2 para los segmentos onset y para los segmentos offset se alcanzó una tasa de acierto del 70% y un Auc de 0,75 tabla 4.3.

Transiciones onset												
	/Pa-ta-ka/				/Pa-ka-ta/				/Pe-ta-ka/			
	Acierto	Sensib	Especif	AuC	Acierto	Sensib	Especif	AuC	Acierto	Sensib	Especif	AuC
SVM	64,00%	60,00%	73,30%	0,66	70,00%	67,20%	73,80%	0,74	58,00%	59,10 %	57,10%	0,61
KNN	54,00%	54,80%	53,40%	0,51	53,00%	53,20%	52,80%	0,56	48,00%	48,40 %	47,40%	0,54
NN	69,00%	67,90%	70,20%	0,70	69,00%	67,90%	70,20%	0,70	60,00%	60,90 %	59,30%	0,63

Tabla 4.2: Resultados de Acierto, Especificidad, Sensibilidad y AuC para análisis de transiciones on, con 120 características en cada DDK.

Transiciones offset												
	/Pa-ta-ka/				/Pa-ka-ta/				/Pe-ta-ka/			
	Acierto	Sensib	Especif	AuC	Acierto	Sensib	Especif	AuC	Acierto	Sensib	Especif	AuC
SVM	64,00%	63,50%	64,60%	0,72	70,00%	73,80%	67,20%	0,75	65,00%	69,20%	62,30%	0,63
KNN	59,00%	57,10%	62,20%	0,58	58,00%	59,10%	57,10%	0,63	62,00%	63,00%	61,10%	0,68
NN	69,00%	67,90%	70,20%	0,71	70,00%	69,20%	70,80%	0,75	61,00%	60,40%	61,70%	0,70

Tabla 4.3: Resultados de Acierto, Especificidad, Sensibilidad y AuC para análisis de transiciones off, con 120 características en cada DDK.



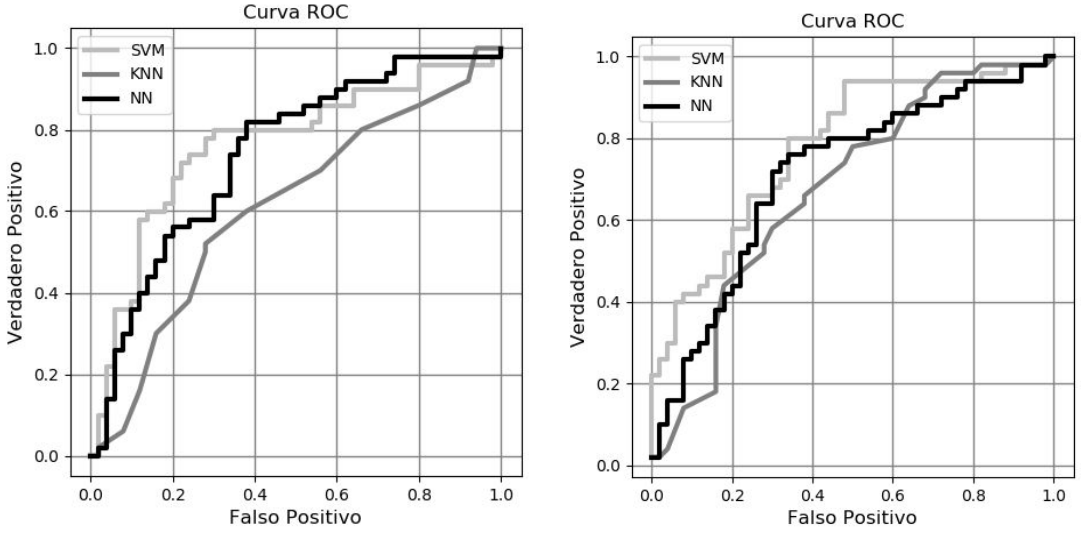
(a) (b)

Figura 4.1: Curvas ROC correspondientes a las mejores tasas de acierto en la tabla 4.2 (/Pa-ka-ta/) y en la tabla 4.3 (/Pa-ka-ta/), respectivamente. a) Transiciones onset (AuC máximo = 0,74) b) Transiciones offset (AuC máximo = 0,75).

En la búsqueda de mejores resultados se agrandó la matriz de características, mezclando las características de las transiciones onset y de las transiciones offset de los tres DDK. De esta forma se obtuvo una matriz de 100 hablantes por 360 características. Los mejores resultados se obtuvieron en la SVM, para las repeticiones de /pa-ka-ta/, como en el experimento anterior. Se alcanzó una tasa de acierto de 74% y un AuC de 0,76 para los segmentos offset, como se visualiza en la tabla 4.4.

/Pa-ta-ka/ - /Pa-ka-ta/ - /Pe-ta-ka/								
	Transiciones offset				Transiciones onset			
	Acierto	Sensib	Especif	AuC	Acierto	Sensib	Especif	AuC
SVM	74,00%	72,20%	76,10%	0,76	72,00%	75,00%	69,60%	0,74
KNN	62,00%	60,00%	65,00%	0,61	63,00%	63,30%	62,70%	0,67
NN	68,00%	66,70%	69,60%	0,73	71,00%	71,40%	70,60%	0,71

Tabla 4.4: Resultados de Acierto, Especificidad, Sensibilidad y AuC para análisis de transiciones on, con 360 características, resultantes de la mezcla de todos los DDK.



(a) (b)

Figura 4.2: Curvas ROC correspondientes a la mezcla de todas las tareas DDK. a) Transiciones offset (AuC máximo = 0,76) b) Transiciones onset (AuC máximo = 0,74), correspondientes a los resultados de la tabla 8.

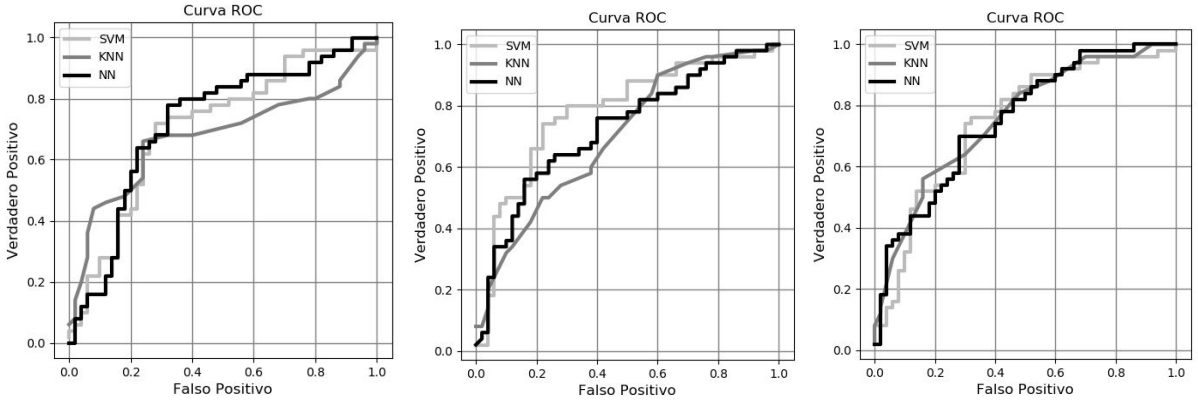
4.2 Aporte del VOT al análisis de voz

Teniendo ya descrito el preprocesado de la señal, extracción de características y clasificación de las mismas, se pasa a exponer los resultados obtenidos para el análisis de los segmentos VOT.

El primer experimento se basó en las 36 características expuestas en la tabla 4.1. Como en los resultados obtenidos en el experimento anterior, la SVM resultó ser el mejor método de clasificación, al arrojar mejores tasas de acierto y de AuC. Para las repeticiones de /pa-ka-ta/, se alcanzó una tasa de acierto máxima de 73% y un AUc de 0,78, como se visualiza en la tabla 4.5:

Segmentos VOT (36 Características)												
	/Pa-ta-ka/				**/Pa-ka-ta/**				**/Pe-ta-ka/**			
	Acierto	**Sensib**	**Especif**	**AuC**	**Acierto**	**Sensib**	**Especif**	**AuC**	**Acierto**	**Sensib**	**Especif**	**AuC**
SVM	70,00%	68,50%	71,70%	0,71	73,00%	70,9%	75,60%	0,78	72,00%	72,90%	71,20%	0,75
KNN	71,00%	69,10%	73,30%	0,68	63,00%	61,00%	65,9%	0,69	67,00%	66,00%	68,10%	0,76
NN	67,00%	68,10%	66,00%	0,68	64,00%	65,2%	63,00%	0,70	70,00%	70,80%	69,20%	0,77

Tabla 4.5: Resultados de Acierto, Especificidad, Sensibilidad y AuC para análisis de segmentos VOT, con 36 características en cada DDK.



(a) (b) (c)

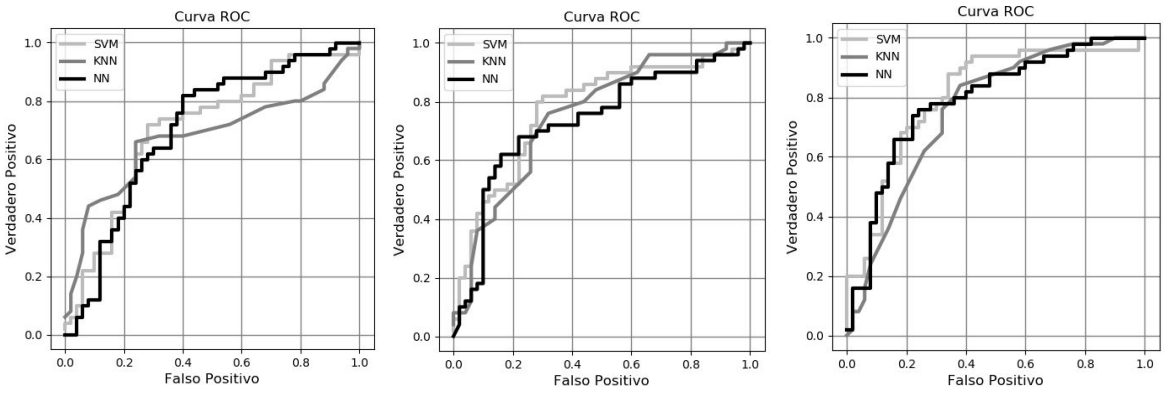
Figura 4.3: Curvas ROC correspondientes a cada tarea DDK en el primer experimento VOT a) /Pa-ta-ka/ (AuC máximo = 0,71) b) /Pa-ka-ta/ (AuC máximo = 0,78) c) /Pe-ta-ka/ (AuC máximo = 0.77), correspondientes a los resultados de la tabla 9.

A pesar de ser una buena tasa de acierto, con este experimento no se logró superar el acierto máximo alcanzado en las transiciones onset/offset. En la búsqueda de mejorar los resultados obtenidos se realizó un segundo experimento, se sumó a la matriz de características, el cálculo de los formantes vocálicos. Se obtuvieron 12 características extraídas para los formantes 1 y 2, calculando su primera y segunda derivada y posteriormente se extrajo la media de todos los segmentos VOT. De esta forma se obtuvo una matriz de 100 hablantes por 48 características.

Segmentos VOT (48 Características)

	/Pa-ta-ka/				/Pa-ka-ta/				/Pe-ta-ka/			
	Acierto	**Sensib**	**Especif**	**AuC**	**Acierto**	**Sensib**	**Especif**	**AuC**	**Acierto**	**Sensib**	**Especif**	**AuC**
SVM	70,00%	68,50%	71,70%	0,71	73,00%	73,50%	72,50%	0,77	74,00%	74,00%	74,00%	0,81
KNN	71,00%	69,10%	73,30%	0,68	70,00%	68,50%	71,70%	0,75	69,00%	69,40%	68,60%	0,76
NN	66,00%	66,80%	66,00%	0,68	67,00%	68,10%	66,00%	0,73	74,00%	76,10%	72,20%	0,78

Tabla 4.6: Resultados de Acierto, Especificidad, Sensibilidad y AuC, para análisis de segmentos VOT, con 48 características en cada DDK.



(a) (b) (c)

Figura 4.4: Curvas ROC correspondientes a cada tarea DDK en el segundo experimento VOT a) /Pa-ta-ka/ (AuC máximo = 0,71) b) /Pa-ka-ta/ (AuC máximo = 0,77) c) /Pe-ta-ka/ (AuC máximo = 0.81), correspondientes a los resultados de la tabla 10.

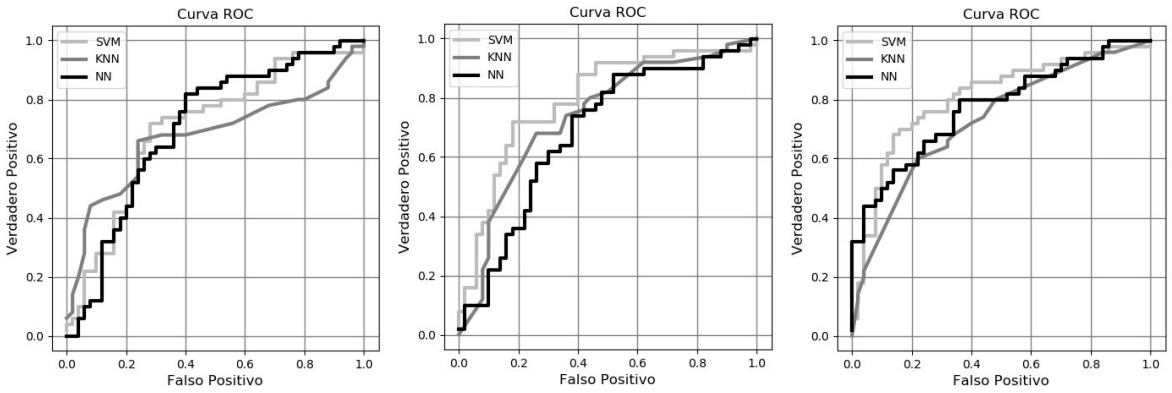
Con el experimento anterior se logró igualar la tasa de acierto máxima alcanzada en los experimentos hechos con las transiciones onset y las transiciones offset, además el AuC subió de forma notable hasta un 0,81 en las repeticiones de /Pe-ta-ka/.

Nuevamente en la búsqueda de mejoras se reforzó el algoritmo en un tercer experimento. Teniendo en cuenta que el cálculo de las 17 energías de Bark generó buenos resultados en anteriores experimentos, se decidió hacer su implementación para los segmentos VOT. Al igual que para los MFCC, al cálculo de las 17 energías de Bark se le extrajeron 17 características de la media y 17 características de la desviación estándar. Finalmente se obtuvo una matriz de 100 pacientes por 82 características. Se alcanzó una tasa de acierto del 77% con AuC de 0,79 para las repeticiones de /Pa-ka-ta/ en la SVM y una tasa de acierto del 76% con AuC de 0,81 para las repeticiones de /Pe-ta-ka/ también para la SVM, como se visualiza en la tabla 4.7.

Segmentos VOT												
	/Pa-ta-ka/				/Pa-ka-ta/				/Pe-ta-ka/			
	Acierto	**Sensib**	**Especif**	**AuC**	**Acierto**	**Sensib**	**Especif**	**AuC**	**Acierto**	**Sensib**	**Especif**	**AuC**

SVM	71,40%	70,60%	72,30%	0,74	77,00%	74,50%	80,00%	0,79	76,00%	74,10%	78,30%	0,81
KNN	68,40%	66,70%	70,50%	0,69	69,00%	71,10%	67,30%	0,74	67,00%	66,70%	67,30%	0,73
NN	65,30%	66,70%	64,20%	0,68	67,00%	69,80%	64,90%	0,69	69,00%	68,60%	69,40%	0,76

Tabla 4.7: Resultados de Acierto, Especificidad, Sensibilidad y AuC, para análisis de segmentos VOT, con 82 características en cada DDK.



(a) (b) (c)

Figura 4.5: Curvas ROC correspondientes a cada tarea DDK en el tercer experimento VOT a) /Pa-ta-ka/ (AuC máximo = 0,74) b) /Pa-ka-ta/ (AuC máximo = 0,80) c) /Pe-ta-ka/ (AuC máximo = 0.81), correspondientes a los resultados de la tabla 11.

Las transiciones on y el VOT pueden estar correlacionados de alguna manera, pero la diferencia está en que las transiciones on consideran cualquier segmento sordo, mientras que los segmentos VOT consideran sólo consonantes plosivas. Además, las transiciones on tienen la misma longitud (40 ms a la izquierda y derecha a partir del límite entre un segmento sordo y sonoro), mientras que el VOT se basa en hallar el tiempo que se demora en pasar de la plosiva al segmento sonoro.

Los resultados demuestran que el mejor clasificador entre los utilizados para los experimentos que se realizaron en este análisis de características articulatorias del habla es la SVM, que junto con las tareas DDK /pa-ka-ta/ y /pe-ta-ka/, aportan las mejores tasas de acierto en la diferenciación entre pacientes con EP y hablantes sanos. La razón por la que se obtienen mejores resultados con la SVM, es que este método es más robusto en comparación a K-NN y RNA.

Aumentar el desempeño de las RNA es posible, para esto es necesario considerar métodos de aprendizaje profundo, los cuales superan los algoritmos tradicionales de aprendizaje automático utilizados por las RNA, con algoritmos de alta complejidad capaces de aprender a partir de la experiencia.

Capítulo 5

Conclusiones

El objetivo principal de esta investigación consistió en el desarrollo de una metodología que permitiera el análisis de la capacidad articulatoria del habla en pacientes con EP por medio del VOT. Para esto se consideraron se consideraron tres tareas de voz que consisten en la repetición rápida de las palabras /pa-ta-ka/, /pe-ta-ka/ y /pa-ka-ta/. El método propuesto, se optimiza con una extracción de características tanto en tiempo como en frecuencia, especialmente con los formantes vocálicos y las energías de Bark, proporcionando cada vez mejores tasas de acierto en la diferenciación de hablantes sanos y pacientes con EP.

La metodología propuesta para el análisis de segmentos VOT en la diferenciación de la voz de pacientes con la EP y hablantes sanos, genera mejoras en comparación con la basada en métodos tradicionales como el análisis de segmentos sonoros y sordos. Para evaluar la articulación de los pacientes se consideraron transiciones onset/offset y segmentos VOT. De acuerdo con los resultados, el desempeño de clasificación fue más alto en el VOT comparado con las transiciones. Por tal motivo el análisis de segmentos VOT se convierte en un concepto importante y que puede ser tenido en cuenta para futuros desarrollos en el campo del análisis de voz de enfermedades que afectan el proceso de generación del habla.

Como lo demuestran los resultados de esta investigación, las características articulatorias siguen teniendo validez para el análisis de problemas en el habla de forma efectiva en la EP. Esto permite pensar a futuro en la contribución que genera el análisis de segmentos VOT a aplicaciones médicas que respaldadas por expertos del área de la salud, generen diagnósticos clínicos de forma certera y ágil.

Trabajos futuros

En la búsqueda de hacer más efectivo el análisis de características en la voz de pacientes con la EP, se debe trabajar en una metodología que permita evaluar el VOT en habla continua. De esta forma los médicos tratantes y los pacientes, pueden llevar un monitoreo permanente del progreso de la enfermedad en cada individuo.

El análisis de los segmentos VOT es posible siempre y cuando la segmentación de sílabas es acertada, por tal motivo es necesario continuar trabajando en mejorar el algoritmo de segmentación automática de sílabas, para tener una mayor precisión en los tiempos de duración de la voz.

Para robustecer la matriz de características y así buscar un mayor acierto en la utilización de métodos de aprendizaje de máquina, se pueden combinar características basadas en articulación con prosodia, inteligibilidad y características fonológicas, aprovechando de manera más efectiva los rasgos que la EP imprime en la voz de los pacientes.

Lista de Figuras

Lista de Tablas

Referencias Bibliográficas

Dorsey, E., Constantinescu, R., Thompson, J. P., et al. (2007). Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology*, *68*(5), 384-386.

Sanchez, J. L., Buritica, O., Pineda, D., Santiago Uribe, C., & Guillermo Palacio, L. (2004). Prevalence of Parkinson's disease and parkinsonism in a Colombian population using the capture-recapture method. *international Journal of Neuroscience*, *114*(2), 175-182.

Pradilla, A., Vesga, A., Boris, E., & León-Sarmiento, F. E. (2003). Estudio neuroepidemiológico nacional (EPINEURO) colombiano. *Revista panamericana de salud pública*, *14*, 104-111.

Hornykiewicz, O. (1998). Biochemical aspects of Parkinson's disease. *Neurology*, *51*(2 Suppl 2), S2-S9.

Logemann, J. A., Fisher, H. B., Boshes, B., & Blonsky, E. R. (1978). Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients. *Journal of Speech and Hearing Disorders*, *43*(1), 47-57.

Skodda, S., Visser, W., & Schlegel, U. (2010). Short-and long-term dopaminergic effects on dysarthria in early Parkinson's disease. *Journal of Neural Transmission*, *117*(2), 197-205.

Martínez, F. 2010. Trastornos del habla y la voz en la enfermedad de Parkinson. Rev Neurol.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., et al. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement disorders*, *23*(15), 2129-2170.

Navarro, T. (2004). Manual de pronunciación española, 28ª edición. Madrid. CSIC

Orozco-Arroyave, J. R. (2016). *Analysis of Speech of People with Parkinson's Disease* (Vol. 41). Logos Verlag Berlin GmbH.

Gutiérrez-Calderón, J. A., Gama-Melo, E. N., Amaya-Hurtado, D. & Avilés-Sánchez, O. F. (2013) Desarrollo de interfaces para la detección del habla sub-vocal. Bogotá, Colombia, 21 de 05 de 2013.

Muzio, G. (2013). Neurobiología de la motivación. Figura 1. Recuperado de https://bluesmarteurope.wordpress.com/2013/10/05/neurobiologia-de-la-motivacion.

Correa, J. (2016). Análisis del efecto del Parkinson en el temblor de la voz: envolvente espectral. Figura 2, pp. 22-23.

Prieto-Bayón, R. (2015). Repercusión de la estimulación cerebral profunda en el habla y la voz de un paciente con Parkinson. Efectos de la intervención logopédica.

Benesty, J., Sondhi, M. M., & Huang, Y. (Eds.). (2007). Springer handbook of speech processing. Springer.

Herrero, M., T. 2011. Enfermedad del Parkinson.

Rabiner, L. & Juang, B. (1993). Fundamentals of speech recognition. Prentice-Hall.

Orozco-Arroyave, J. R., Vásquez, J. C. & Vargas-Bonilla, J. F. (2017). NeuroSpeech: An open-source software for Parkinson's speech analysis.

Hanson, D. G., Gerratt, B. R., & Ward, P. H. (1984). Cinegraphic observations of laryngeal function in Parkinson's disease. *The Laryngoscope*, *94*(3), 348-353.

Tsanas, A., Little, M. A., Fox, C., & Ramig, L. O. (2014). Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *22*(1), 181-190.

Ackermann, H., & Ziegler, W. (1991). Articulatory deficits in parkinsonian dysarthria: an acoustic analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, *54*(12), 1093-1098.

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

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 (TASLP)*, *22*(9), 1366-1378.

Orozco-Arroyave, J.R., Hönig, F., Arias-Londoño, J.D. & Vargas-Bonilla, J.F. (2016). Automatic detection of Parkinson's disease in running speech spoken in three different languages, J. Acoust. Soc. Am. 139 (1) 481–500.

Ho, A. K., Iansek, R., Marigliani, C., Bradshaw, J. L., & Gates, S. (1999). Speech impairment in a large sample of patients with Parkinson's disease. *Behavioural neurology*, *11*(3), 131-137.

Bocklet, T., Steidl, S., Nöth, E., & Skodda, S. (2013). Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues. In *Interspeech* (pp. 1149-1153).

Rodríguez-Violante, M. & Cervantes-Arriaga, A. (2014). La escala unificada de la Enfermedad del Parkinson modificada por la sociedad de Trastornos del Movimiento (MDS-UPDRS): Aplicación clínica e investigación.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., et al. (2008). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Movement disorders, 23(15), 2129-2170.

Postuma, R. B., Lang, A. E., Gagnon, J. F., Pelletier, A. & Montplaisir, J. (2012). How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder. Brain , vol. 135, pp. 1860–1870.

Baumgartner, C. A., Sapir, S., & Ramig, L. O. (2001). Voice quality changes following phonatory-respiratory effort treatment (LSTVT) versus respiratory effort treatment for individuals with Parkinson disease. J. Voice, vol. 15, pp. 105–114.

Ackerman, H., Koznack, J., & Hertrich, I. (1997). The temporal control of repetitive articulatory movements in Parkinson's disease. Brain Lang., vol. 57, pp. 312–319.

Fletcher, S. (1972). Time-by-count measurement of diadochokinetic syllable rate," J. Speech Hear. R., vol. 15, pp. 757–762.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. Language and speech, 1(3), 153-167.

Martínez-Sánchez, F. (2010). Trastornos del habla y la voz en la enfermedad del Parkinson. Rev Neurol, 51(9), 542-550.

Adjarian, H. (1899). Les explosives de l'ancien arménien étudiées dans les dialectes modernes, La Parole. Revue internationale de Rhinologie, Otologie, Laryngologie et Phonétique expérimentale, pp. 119-127.

Lin, C. & Wang. (2011). H. Automatic estimation of voice onset time for word-initial stops by applying random forest to onset detection. J. Acoust. Soc. Am. 130 (1) 514–525 .

Lisker, L. & Abramson, A. (1964). A cross-language study of voicing in initial stops: acoustical measurements, Word Vol. 20, 384-422 .

Li Zuo, W., Wang, Z. & Chen, H. (2013). Effective detection of Parkinson's disease using an adaptive fuzzy k -nearest neighbor approach. Biomedical Signal Processing and Control 8 (2013) 364–373.

Carmona, E. (2013). Tutorial sobre Máquinas de Vectores Soporte (SVM).

Matich, D., Ruiz, C. & Basualdo, M. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones.

Swets, J., Dawes R., & Monahan J. (2000). Psychological science and improve diagnostic decisions.

Villa, T., Belalcazar, E., Bedoya, S., Garcés, J & Orozco, J. (2012). Automatic Detection of Laryngeal Pathologies using Cepstral analysis in Mel and Bark scales.

Villamil, I. & Vizcaya, P. (2005). Aplicaciones en reconocimiento de voz utilizando HTK.

Villa, T. & Arias, J. (2015). Metodología de análisis tiempo-frecuencia para la evaluación automática de la voz de pacientes con enfermedad de Parkinson.

San Juan, E., Watkins, F. & kaschel, H. (2013). Sistema de ayuda visual para apoyar aprendizaje de fonemas españoles. RIELAC, Vol.XXXIV, pp.87 - 99.

Bradlow, A.R. (1995). A comparative acoustic study of English and Spanish vowels. Journal of the Acoustical Society of America, 97, pp. 1916-1924.

Martínez, C. (2003). En torno a las vocales del español: análisis y reconocimiento.

Paredes, O., Romo, R., Vélez, H. & Morales, J. (2017). Análisis estadístico de los espectros de frecuencia de las regiones reguladoras del ENCODE. Revista Mexicana de Ingeniería Biomédica, pp. 641-642.

Arriagada, M. (2015). Comparación de métricas de distancia en el algoritmo K-Vecinos Más Cercanos para el problema de Reconocimiento Automático de Dígitos Manuscritos.

Montaña, D., Campos, Y. & Pérez, C. (2018). A Diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of Parkinson's disease.